*Last updated: 24 October 2021*

# Collaboration for Environmental Evidence Critical Appraisal Tool Version 0.3 (Prototype)

**Contents**

*Please cite as: Konno K, Livoreil B, Pullin AS. 2021. Collaboration for Environmental Evidence Critical Appraisal Tool Version 0.3 (Prototype).*

## Part A: General Description of the Tool

We are currently developing a critical appraisal tool for evaluating 'risk of bias' (or threats to internal validity) of primary studies assessing effectiveness of interventions or impacts of exposures in environmental management, and here we provide a third draft of the tool. The tool is still under development and requires initial testing, but it may help environmental evidence synthesists conduct critical appraisal. Application of the tool involves: (a) answering the checklist questions within individual risk-of-bias criteria (see below); (b) judging the risk of bias within individual risk-of-bias criteria; and (c) making an overall judgement about risk of bias for the study findings (estimates of effectiveness of intervention or impact of exposure).

### *Understanding Risk of Bias*

Bias or inaccuracy is referred to as deviation from the truth [1] or a systematic error in study's findings (including inference that is made in the study) [2]. Although the distinction between bias and imprecision (i.e., random error) should be kept clearly in mind when applying this tool, biases in real-world data cannot often be quantified, and thus perfectly distinguishing between them is not possible unless true values are known [1]. This is why the concept of risk of bias (measure of threats to internal validity) was developed [2,3], and it is now widely applied to evaluate how susceptible studies are to biases in the healthcare sector rather than trying to quantify biases. By applying this tool, study findings will always have risks of bias since it is not often possible to prove that there are no biases in findings. Having risks of bias does not mean that the findings are biased but there are always possibilities for findings being biased to some extent. The spectrum of risk of bias we provide is categorical (low, medium, and high; see below).

### *Scope of the Tool*

The tool is designed for environmental management research such as applied ecology, biodiversity conservation and conservation genetics, soil, water and air pollution, agriculture, park and protected area management, environmental epidemiology and pathogen control, species invasions, river and wetland management, exploitation of natural resources and fisheries, waste management, sustainable energy and consumption, and broader contexts of environmental sustainability may also be relevant if outcomes of interest are measured quantitatively.

In general, 'medical research involving human subjects' is beyond the scope of this tool. Such research often accords with the Declaration of Helsinki (https://www.wma.net/policies-post/wma-declaration-of-helsinki-ethical-principles-for-medical-research-involving-human-subjects/) and provides ethics statement (name of ethics committee and approval or project ID). Although implications for environmental management may be provided in such research, they are beyond the scope of this tool. For example, the impact of exposure to pesticides on human urinary biomarkers may be of interest of U.S. Environmental Protection Agency. In such cases, we refer risk-of-bias assessors (who assess risk of bias, 'assessors' hereon) to the risk of bias assessment tools developed in the healthcare sector (e.g., 'RoB 2' [4] or 'ROBINS-I' [5]).

Purely laboratory-based biological research (*in vivo* and *in vitro* experiments) is also beyond the scope of this tool. The checklist below may be used to check if the use of this tool is appropriate for a planned environmental evidence review (**Figure A1**). If any of the items in 'beyond the scope' applies, the planned review is not within the scope even when all of the items in 'within the scope' apply.

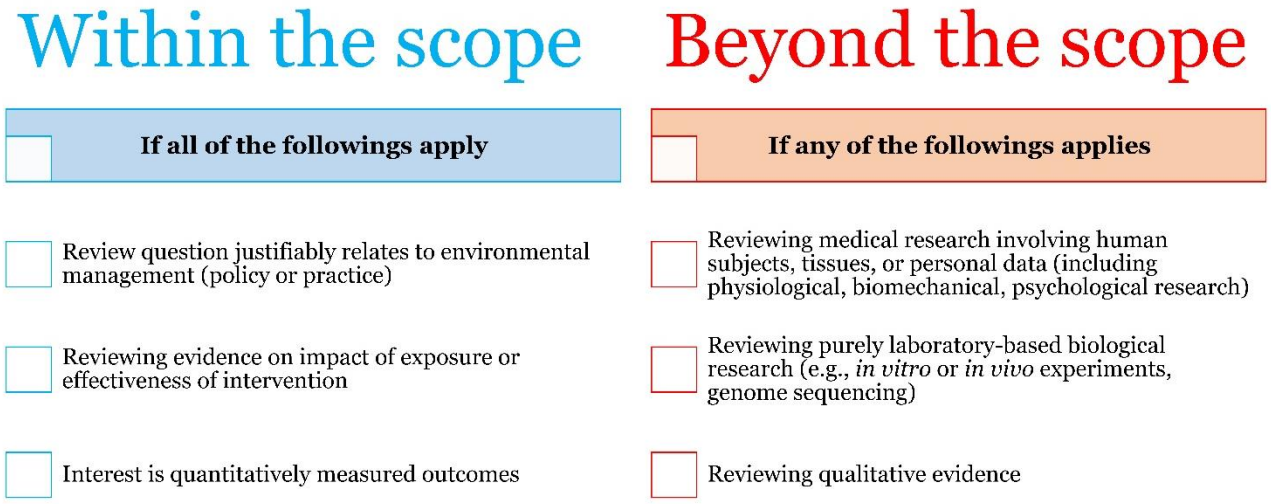


***Figure A1.*** *Checklist for the scope of the tool. Note: If any of the conditions in 'beyond the scope' applies, a planned review is not within the scope even when all of the conditions in 'within the scope' apply.*

### *Assumptions of the Tool*

#### Subjectivity

In any assessment of risk, it inevitably requires subjective judgements to be made [6], and so application of this tool does not make conduct of critical appraisal 100% objective. For example, this tool asks assessors for some subjective judgements using terms such as 'acceptable', 'tolerable' and 'accurate enough', and so there is some degree of freedom for assessors how certain aspects of studies, that might be sources of bias, are judged for risk of bias. This subjectivity allows some flexibility to reflect the extent of acceptability of risk of bias based on the nature of primary research and evidence synthesis. We thus strongly encourage assessors to check consistency between them before and when formally applying the tool. Please also note that reading this document alone should not be considered as formal training of critical appraisal. Assessors may need to consult evidence synthesis and subject experts before formally applying this tool.

#### Study Designs

For a primary study assessing effectiveness or impact, its design (i.e., type of comparison made) may be referred to as one of the followings: 'control-impact' (CI), 'before-after' (BA), 'before-after control-impact' (BACI) and 'randomised controlled trial' (RCT) (**Figure A2**). While relatively more rigorous study designs (e.g. experimental BACI and RCT compared to CI or BA) may be employed to increase precision (i.e. to control variances [7]) and accuracy, it is theoretically possible that study findings are not biased if sources of bias are sufficiently eliminated in any study designs, or approaches are justified on a case by case basis based on addressed causal structure and available data [8,9]. Thus, this tool does not assume that differences in study designs alone affect accuracy of study findings. In other words, the tool does not assume that study designs themselves are biased, but it assumes that effect estimates may be biased if sources of bias are not controlled for or not taken into account. Hence, the tool does not require any coded data on study designs for assessing risk of bias although such coded data may be useful for sensitivity or subgroup analysis. Note that, in this document, we refer the term 'comparator' as an overarching comparison group which may include control (no intervention/exposure), baseline (before intervention/exposure) or alternative intervention or exposure that is compared with to estimate effectiveness or impact (**Figure A2**).

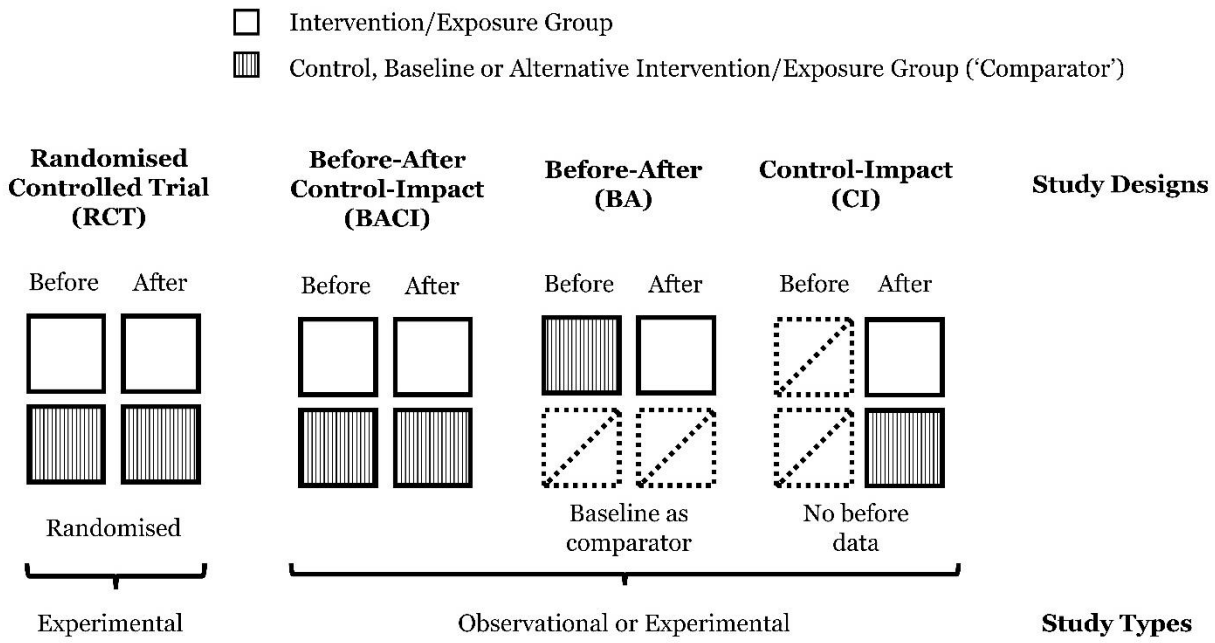


***Figure A2.*** *Different study designs (type of comparisons). Dotted lines indicate the lack of data. Striped boxes may be referred to as comparators.*

**Experimental Vs. Observational**

Primary studies may be experimental (i.e., interventions or exposures are formally applied in studies to evaluate effects) or observational (i.e., comparisons of groups of interest without formal application of interventions or exposures in studies). The tool does not allow assessors to assess risk of bias based solely on type of studies (experimental or observational), and therefore the same logic above applies to type of studies. In other words, the tool does not assume that study types themselves are biased, but it assumes that effect estimates may be biased if sources of bias are not controlled for or not taken into account. Note that there are two risk-of-bias criteria only applicable to one type of study (one for observational studies only and one for experimental studies only), and so coding study types before applying the tool may make critical appraisal relatively efficient.

**Intervention or Exposure**

This tool can be used for both intervention and exposure studies, and there is no difference in the process of risk-of-bias assessment between exposure and intervention studies. The terms intervention and exposure are used at the same time in the tool as both effectiveness and impacts are measures of effects of causes. Interventions may be characterised based on desired or intended outcomes that are indeed measured in studies (e.g., intervention to remove pollutants and removal of pollutants is measured) while exposures may be characterised based on undesired or unintended outcomes (e.g., exposure of soil fauna to fertilisers when fertilisers are to increase crop yield). In most cases, the distinction between interventions and exposures may be made based on specific outcomes (i.e., endpoints) that are measured in relation to causes of interest. In environmental management, exposure studies may be conducted experimentally, and this capability highlights a major distinction from 'medical research involving human subjects' in which experimental exposure studies are not often feasible due to ethical concerns. The distinction between interventions and exposures is quite obvious in the health sector while it may not be so obvious in the environmental sector.

***Individual Risk-of-Bias Criteria***

There are diverse definitions of forms of bias, and they may be used differently by different people in different disciplines, and there are some overlaps in meaning among some forms of bias. Thus, it is necessary to determine what sources of bias are investigated and when the sources of bias may

occur. In this tool we chose to use seven risk-of-bias criteria (equivalent of 'domains', the term used in the health sector [4,5]) to draw borderlines among them, which are defined as follow:

1. Risk of confounding biases (risk of biases due to uncontrolled (or inappropriately controlled) variable (confounder) that influences both the intervention/exposure and the outcome)
2. Risk of post-intervention/exposure selection biases (risk of biases arising from systematic differences in the selection of subjects or areas into the study or analysis after intervention or exposure)
3. Risk of misclassified comparison biases (risk of biases arising from misclassification or measurement of intervention, exposure and/or comparator; applicable to observational studies only)
4. Risk of performance biases (risk of biases due to altered treatment procedure of interest; applicable to experimental studies only)
5. Risk of detection biases (risk of biases arising from systematic differences in measurement of outcomes)
6. Risk of outcome reporting biases (risk of biases in reporting of study findings)
7. Risk of outcome assessment biases (risk of biases due to error in applied statistical methods)

Note that Criterion 3 is only applicable to observational studies and Criterion 4 is only applicable to experimental studies, and so assessors will need to answer checklist questions of six risk-of-bias criteria for a study's findings. Please also note that, to avoid potential confusion, we do not mean to define forms of bias here, rather we mean to define the criteria used in the tool to fit for the purpose of risk-of-bias assessment. Thus, we intentionally use plurals rather than singular terms to highlight our intention (e.g., 'confounding biases' is used rather than 'confounding bias') as singular terms may imply definitions of forms of bias.

**Criterion 1: Risk of Confounding Biases**

This criterion is concerned with biases due to uncontrolled (or inappropriately controlled) variable (confounder) that influences both the intervention/exposure and the outcome. Causal directed acyclic graph (DAG; also known as causal model or causal diagram) can be a useful tool for investigating the potential of confounding [10,11]. In the language of DAG, the presence of confounding is illustrated as unblocked 'backdoor path' that affects both the intervention/exposure and the outcome [11]. If there is confounding, confounder (also known as concomitant or covariate [8]) will distort the association between an intervention/exposure and an outcome. More technically, if the assumption of exchangeability between groups does not hold due to presence of shared causes of the intervention/exposure and the outcome, effect estimates will be biased due to confounding [12]. An example of confounding is that if the effect of population density of salmonids on their survival to be studied, presence of predators may be a confounder as it may influence both survival and population density [13] (**Figure A3**). Controlling for the presence of predators thus may reduce the risk of confounding bias in this instance. When a confounder is controlled for, such as through stratification, conditional exchangeability between groups may hold (i.e., the assumption of randomness may not be violated within a stratum or strata of the confounder variable), and thus a valid effect estimate may be provided within a stratum or strata [9].

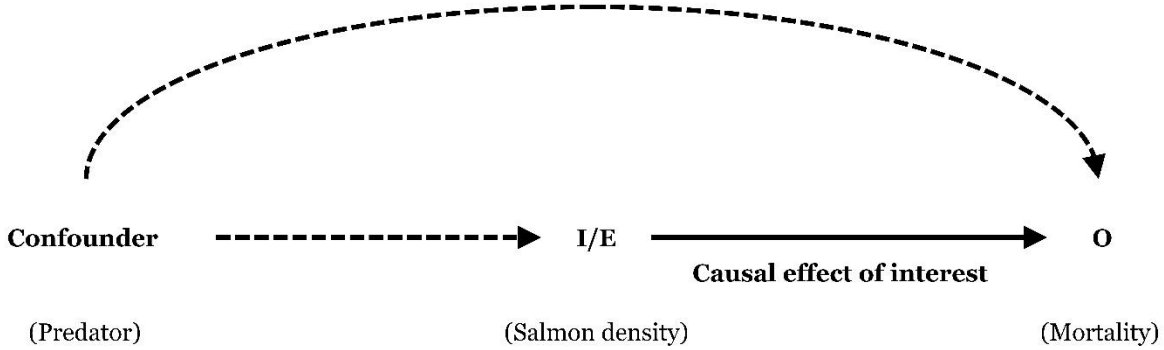

***Figure A3.*** *Causal directed acyclic graph (DAG) showing a simple example of confounding. A confounder (predator) distorts the association between the intervention or exposure (I/E; salmon density) and the outcome (O; mortality) even when the true effect is null. In causal DAG, such distorted association due to confounding may be described using a unblocked 'backdoor path' (dotted lines)* [9].

Awareness of potential confounders (i.e., hypothetical candidates) may be gained by trying to understand variability in effect. For example, morphological differences in species (e.g., colour) may be attributable to variability in effect. In other words, the difference in morphology may be a covariate. When such factor is not controlled (or inappropriately controlled) for before intervention/exposure, resulting effect estimates may be biased if confounding is not adjusted for (**Figure A4**). Note such baseline confounding is sometimes referred to as 'selection bias' [5,9]. Baseline confounding due to selection of subjects or areas into studies can be dealt before intervention or exposure (e.g., stratified sampling in which the population of inference is divided into subpopulations [14]), and thus it needs to be assessed as baseline confounding in this criterion in our tool.

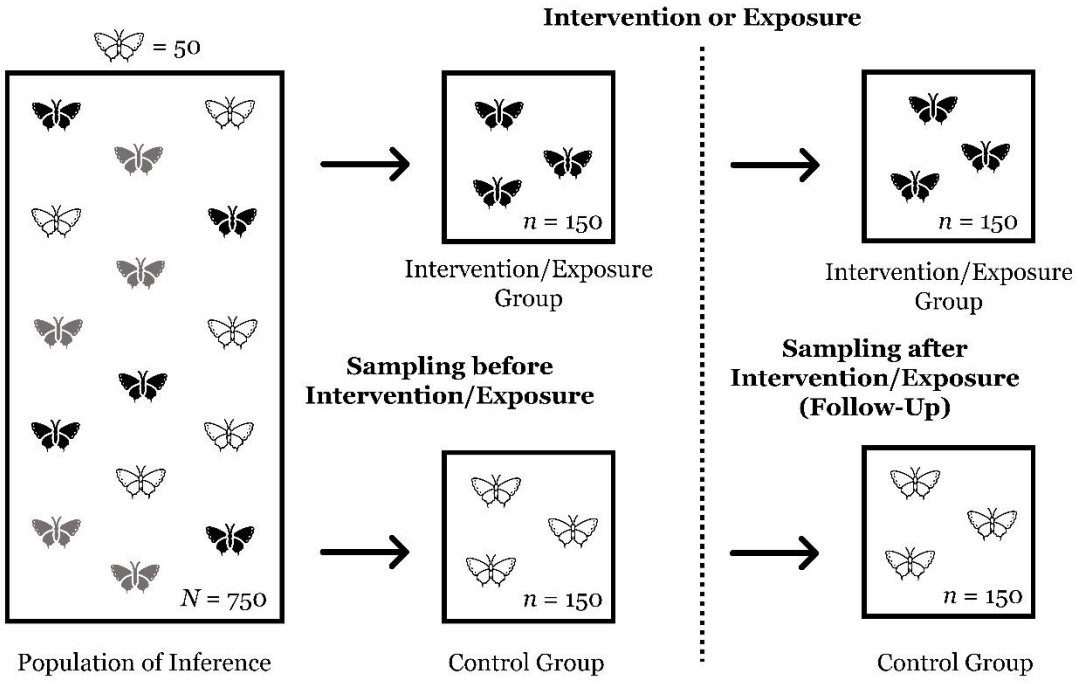


***Figure A4.*** *Theoretical example of baseline confounding that is sometimes referred to as 'selection bias'. Population of inference (target population) consists of black, grey and white butterflies (N = 750). However, different colours are sampled between groups (150 black butterflies in the intervention or exposure group and 150 white butterflies in the control group). In this case, the effect may be over- or underestimated when the colour is a covariate. This type of 'selection bias' needs to be assessed as baseline confounding in* ***Criterion 1*** *in our tool.*

Assessment of risk of confounding biases requires subject knowledge for determining potential confounders of the addressed causal structure [9,11]. We suggest assessors to list all potential confounders of an addressed causal structure using **Appendix A** and to develop a general causal model before formally assessing risk of confounding biases. Note that developed general causal model may need to be remodelled for each study to detect risk of confounding biases (e.g., inappropriate adjustment of relevant variables). Assessors who are not familiar with causal DAGs may wish to consult Pearl 2009 [8], Hernán & Robins 2020 [9], or relevant literature cited in them for guidance.

**Criterion 2: Risk of Post-Intervention/Exposure Selection Biases**

This criterion is concerned with biases arising from systematic differences in the selection of subjects or areas into the study or analysis after intervention or exposure. This is a distinct phenomenon from that of baseline confounding ('selection bias') described above [5], and thus we explicitly use the term 'post-intervention/exposure selection bias' to distinguish from baseline

confounding. A statistical population of inference often consists of subjects or areas with different characteristics. These characteristics may be effect modifiers (i.e., effect may vary across levels of variables). If subjects or areas with different characteristics are followed-up or selected after intervention or exposure when the characteristics are indeed an effect modifier, an effect may be over- or underestimated, depending on the addressed causal structure and subjects' or areas' characteristics (**Figure A5**). An example of post-intervention/exposure selection bias is that if the effect of insect herbivory on losses of woody plant foliage to be studied, haphazard leaf selection may be more likely to result in systematic differences in the selection of leaves compared to random or systematic selection, and thus may affect the estimate of effect [15]. Assessors who are not familiar with random or systematic sampling techniques may wish to consult Cochran 1977 [14] for details.

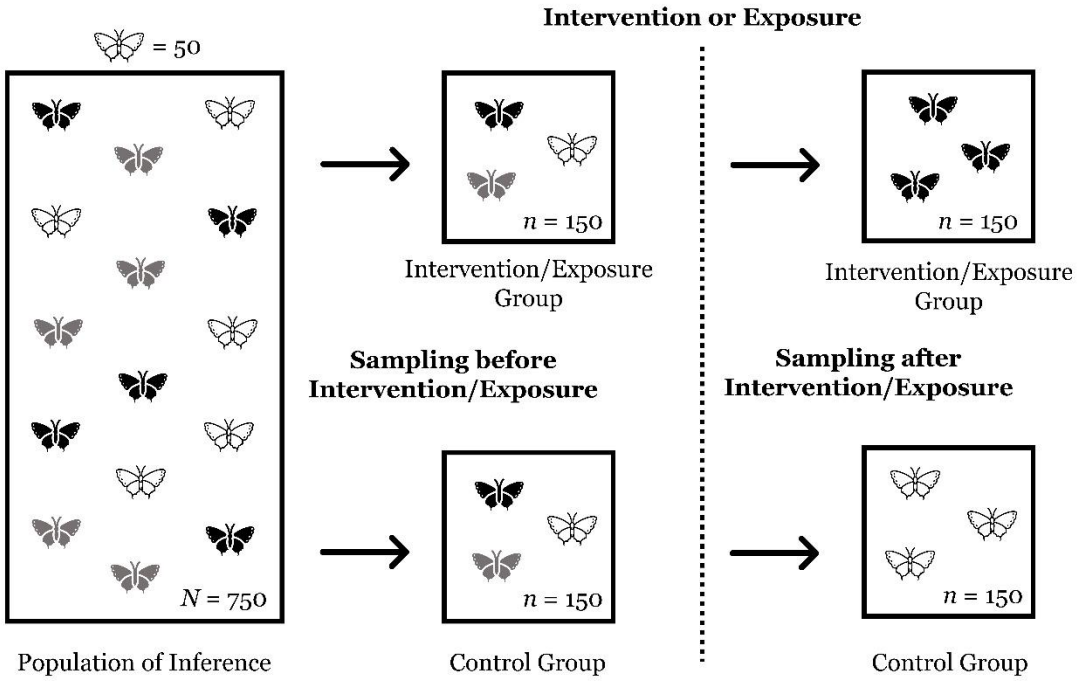


***Figure A5.*** *Theoretical example of post-intervention/exposure selection bias. Population of inference consists of black, grey and white butterflies (N = 750). Before an intervention or exposure occurs, stratified baseline sampling of butterflies is conducted (n = 150 for each group and 50 butterflies for each colour). However, if only certain colours of butterflies are sampled after intervention or exposure and the colours are different between groups (i.e., exchangeability does not hold), the effect may be over- or underestimated when the colour is an effect modifier.*

Another source of post-intervention/exposure selection bias is systematic differences in missing data between intervention/exposure and comparator groups (also known as attrition bias [5,16]). Attrition bias may occur if subjects or areas are missed unequally between groups, and effects of interventions or exposures vary among levels of certain variables [5,9]. For example, if the effect of different light intensities on leaf lifespan of evergreen woody plants to be studied, missing data on individuals with certain characteristics, for example certain maturity or age of plants (i.e., with certain qualitative or quantitative values that describe how mature they are or how old they are), for one group may result in a systematic difference in characteristics of groups which may in turn result in over- or underestimation of the effect as life stage or age may be an effect modifier [17]. However, if percentages of missing data are the same (or nearly the same) and the collected data still allow a valid comparison within a stratum or strata (e.g., immature plants are missed for both groups but data on all mature plants are collected successfully for both groups), the missing data will not affect the estimate of effect from the view of bias due to systematic differences in missing data (**Figure A6**). Note that unequal loss of data may occur even after data are successfully collected from all subjects or areas included in a study (e.g., during copying data from the original record to a form for data analysis) [14,18].

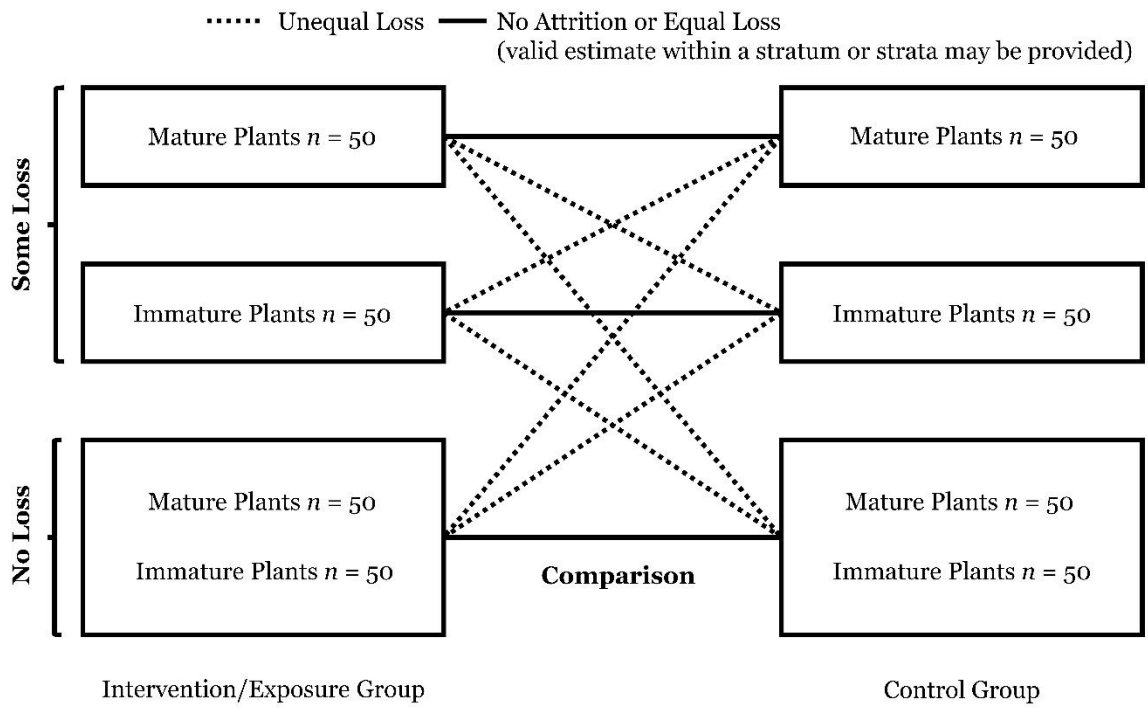


***Figure A6.*** *Simple examples of no, equal and unequal loss of subjects. If unequal loss occurs (dotted lines), an estimate of effectiveness or impact may be biased when maturity is an effect modifier. If there is no attrition or equal loss occurs, the estimate (for strata or a stratum) will not be biased from the view of bias due to systematic differences in missing data between intervention/exposure and comparator groups.*

**Criterion 3: Risk of Misclassified Comparison Biases (Observational Studies Only)**
This criterion is concerned with biases arising from misclassification or measurement of intervention, exposure and/or comparator (also known as measurement bias or information bias [9]). This criterion is only applicable to observational studies. If experimental treatments are applied, this criterion will be 'not applicable', and experimental studies will need to be assessed in **Criterion 4** below instead. Accurate and precise definitions of intervention or exposure and comparator groups are necessary for avoiding misclassification. Intervention or exposure and comparator may be misclassified by researchers who usually classify them based on available information and data which may not always be accurate. For example, if the effect of pesticides on the mortality of honey bees (*Apis mellifera*) to be assessed in an observational study, defining what pesticides (compounds) are relevant and what count as exposure (e.g. compounds found in pollen, beeswax or honey bees) may be necessary so that no meaningful vagueness will remain. When certain individuals are classified in 'no exposure' group as a comparator, evidence of 'no exposure' and descriptions of what compounds are screened for in what objects (pollen, beeswax or honey bees) may be necessary because, for example, as the number of screened compounds increases, the number of detected compounds may increase [19]. Thus, if screened compounds are not extensive in 'no exposure' comparator group in a such observational study and researchers failed to detect pesticides residue in the 'no exposure' group (when there is residue), it will lead to a misclassified comparison and the effect may be biased towards null (no effect) in this instance as this comparison is actually 'exposure to pesticides' vs. 'exposure to pesticides' (**Figure A7**). However, note that if researchers are interested in specific compounds only, and there is supporting information of a correctly classified comparison, then the existence of other pesticides in samples, that are not of the focus, may be of an issue of confounding (i.e., the influence of the existence of other pesticides may be a confounder) and thus such aspects should be dealt in **Criterion 1** above.

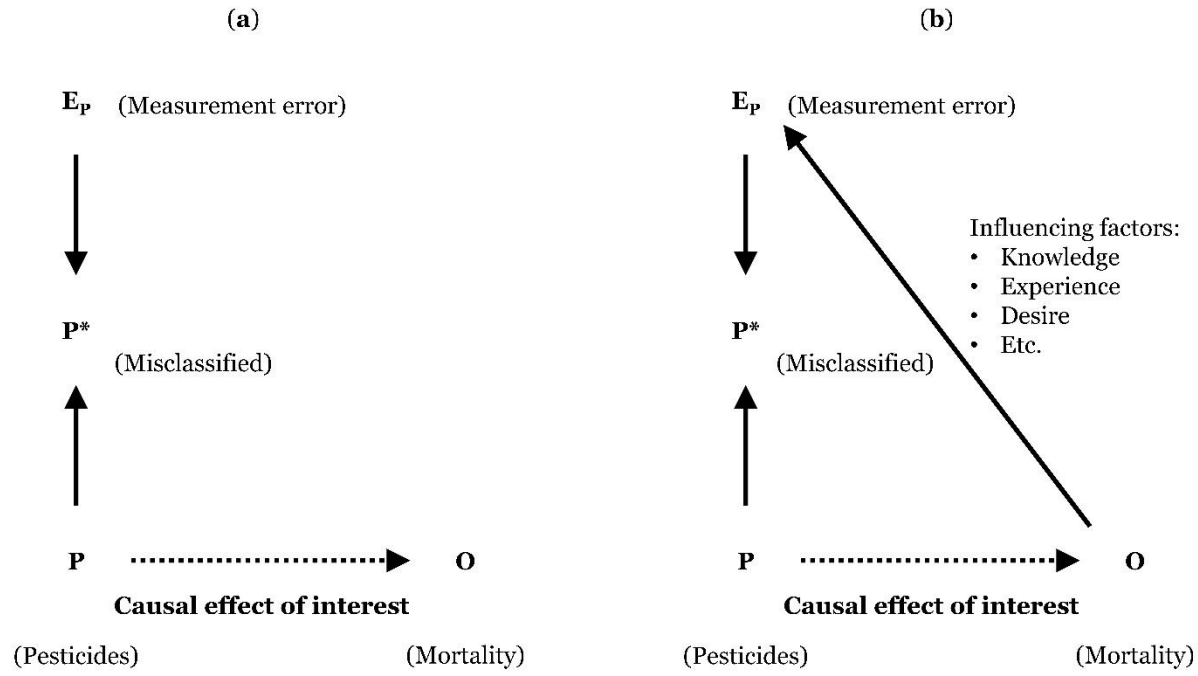


***Figure A7.*** *A theoretical example of misclassified comparison bias (also known as misclassification bias, measurement bias or information bias* [9]*). If an intended comparison (P) is not made due to measurement error ($E_P$), it will result in an inappropriate comparison (P*), and therefore it will bias the effect estimate even when there is no bias in the measurement of the outcome (O). (**a**) Researchers may just misclassify available information, or (**b**) measurements may be influenced by knowledge, experience, desire, etc.*

**Criterion 4: Risk of Performance Biases (Experimental Studies Only)**

This criterion is concerned with biases due to altered treatment procedure of interest or deviated initiation, implementation or discontinuation either by persons who apply experimental treatments or subjects who receive the treatments (also known as measurement bias [9]). This criterion is only applicable to studies that apply experimental treatments. If no experimental treatments are applied (e.g., observational site comparison studies), this criterion will be 'not applicable'. In this tool, 'treatments' are defined as different procedures to be compared for measuring effects [18]. The term 'control' is often used loosely to denote 'no treatment' in scientific literature, however, it is technically a treatment (a procedure to be compared against) [18], and thus both intervention/exposure and control (or alternative intervention/exposure) need to be assessed in this criterion. Performance bias may arise if treatments are altered either intentionally or unintentionally after the procedures have taken place (this is not necessarily initiation of intervention/exposure as the time lag between the start of procedure and the start of intervention may differ and the start of intervention may be recorded incorrectly) and if the alterations are not taken into account (e.g., altered procedures are not reflected on data collection sheet, or areas that have been applied altered procedures are not excluded). For example, if nitrogen fertiliser is applied more than initially planned and this deviated implementation is not recorded, a difference in crop yield (measure of effect) may be overestimated (**Figure A8**). In medical randomised controlled trials, it is recommended to apply double-blinding (both participants and healthcare provider) to avoid performance biases because knowledge of healthcare intervention or allocation may affect procedures [5]. However, if there are no deviations from procedures of interest (or there is supporting information of 'no alterations') and procedures are successfully implemented for all subjects or areas without any influence of awareness, then awareness of the procedures alone will not affect an estimate of an effect from the view of biases due to altered treatment procedure (note outcome measurers' awareness of details of study is addressed below in **Criterion 5**).

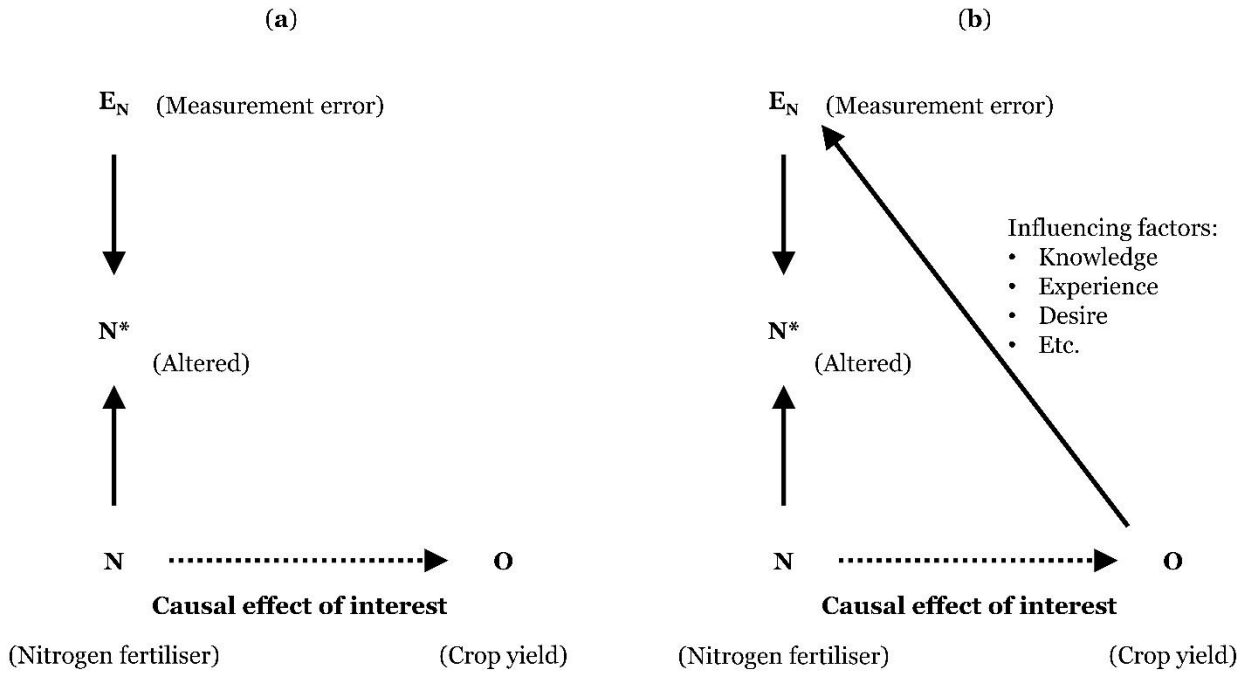


***Figure A8.*** *A theoretical example of performance bias (also known as measurement bias* [9]*). If an intended intervention (N) is influenced by measurement error ($E_N$), it will result in altered intervention (N*), and therefore it will bias the effect estimate even when there is no bias in the measurement of the outcome (O). (**a**) Researchers may fail to apply intended intervention or mismeasure intervention, or (**b**) measurements may be influenced by knowledge, experience, desire, etc.*

**Criterion 5: Risk of Detection Biases**

This criterion is concerned with biases arising from systematic differences in measurements of outcomes (also known as measurement bias [9]). Systematic errors in measurement of outcomes may occur if outcome data are determined differently between groups, either intentionally (e.g. influence of desire to obtain a certain direction of effect) or unintentionally (e.g. due to cognitive bias or human errors) (**Figure A9**). For example, if different extractants (e.g. water extractant and calcium chloride extractant) are used to obtain soil organic carbon concentration data (as outcome of interest) in different groups, the effect may be over- or underestimated due to the systematic difference in the process of measuring the outcomes [20]. When studying complex systems, and especially when many steps are involved in measuring outcomes, each calibration method or applied instrument may need to be the same between groups because, for example, devices used in a specific step may be biased and thus differently affect outcome data between groups [14]. Note if the same biased device is used to measure outcomes in both groups, the bias will cancel out, and it will not affect estimate of effectiveness or impact. For example, if true outcome measurements of dissolved organic carbon concentrations are 10 mg/L for both groups, and the biased outcome measurements are 20 mg/L for both groups, the differences between the groups are the same (0 mg/L), and hence there is no bias in the estimate of the effect in this case [5].

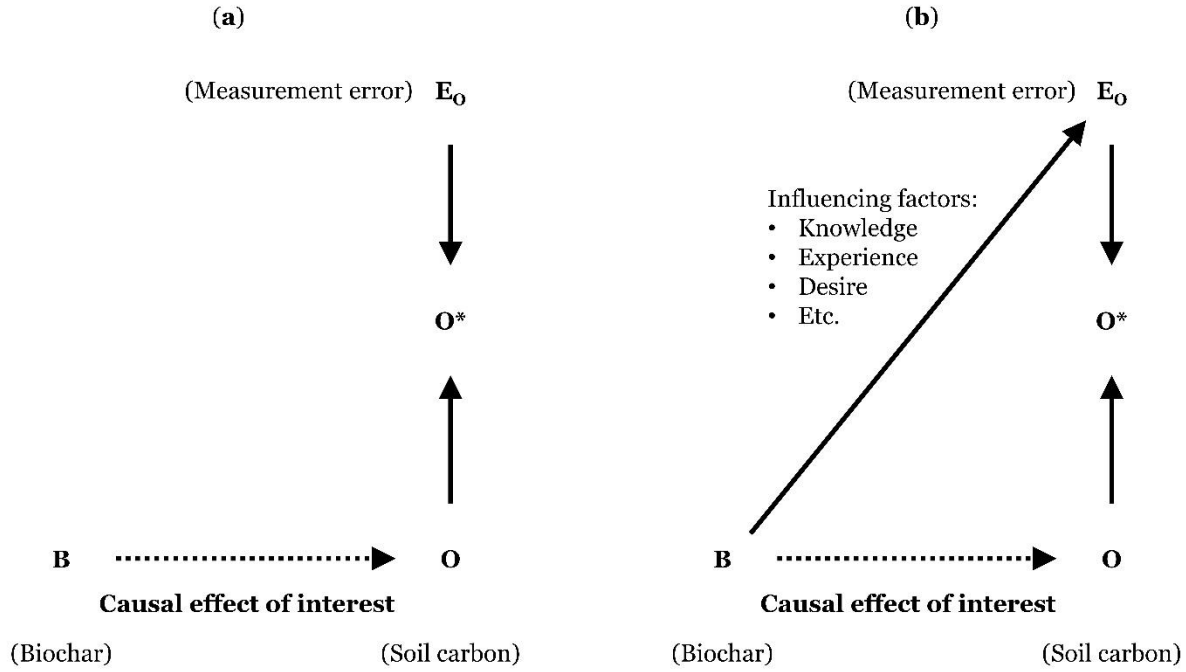


***Figure A9.*** *A theoretical example of detection bias (also known as measurement bias* [9]*). If the outcome (O) is determined differently between groups due to measurement error ($E_O$), it will result in a biased outcome (O*).* ***(a)*** *Researchers may fail to measure the outcomes, or* ***(b)*** *measurements may be influenced by knowledge, experience, desire, etc.*

**Criterion 6: Risk of Outcome Reporting Biases**

This criterion is concerned with biases in reporting of study findings. Outcome reporting biases may arise if findings are selectively disclosed in reporting that does not reflect actual findings. Selective disclosure may appear at three different levels [4]:

- Selective disclosure of findings from multiple measurements
- Selective disclosure of findings from multiple subgroups or subpopulations
- Selective disclosure of findings from multiple analyses

In every way, it is theoretically possible that findings are selectively disclosed to only report statistically significant, interesting and/or desired results of the parties or individuals who involved in research. Outcome reporting biases may especially be suspected when authors do not report any non-significant results or no/small effects, or there are discrepancies between predefined hypotheses and methods and ones reported after studies are conducted.

**Criterion 7: Risk of Outcome Assessment Biases**

This criterion is concerned with biases due to error in applied statistical methods. There is currently no such criterion (domain) in widely applied risk-of-bias assessment tools in the healthcare sector (RoB 2 [4] and ROBINS-I [5]). Steenland et al. argued that there is no dedicated criterion for assessing the ability of obtaining unbiased results of inferential statistics, including appropriateness of chosen statistical methods, in four risk-of-bias assessment tools in health sciences. They thus suggested that this kind of criterion should preferably be added in the risk-of-bias assessment tools [21]. We agree with the suggestion and so established a dedicated criterion in our tool and call this criterion 'risk of outcome assessment biases' defined as biases due to errors in applied analyses, and applied analyses specifically mean applied statistical methods in our tool. The checklist questions ask about four specific issues:

- Data analysts' awareness of exposure or intervention received by subjects or areas
- Errors in applied descriptive statistics (e.g., miscalculations of sample sizes, means, variance)

- Errors in applied inferential statistics (including null hypothesis testing, estimation, coding)
- Violation of assumption for the applied inferential statistics and appropriateness of the applied statistical methods (e.g., criteria for normality and equal variances are not satisfied, inappropriate choice of statistical tests)

It is possible that all the other criteria (1–6) are considered to be low risks of bias but there are errors in applied descriptive or inferential statistics. If there is any error at this stage, it is very likely that the outcome of a study (measure of effectiveness or impact, or inference) will be changed (no matter how trivial or substantial it is), and thus the algorithm suggests high risk of bias when there is any error to reflect that concern. There may also be a case where no inferential statistics (e.g., hypothesis testing) is conducted in a study for specific findings of assessors' interest. For example, a meta-analysis review is being carried out and the assessors are only interested in specific comparable raw data that are only a part of the study (and thus there is no dedicated hypothesis testing for the findings of interest). In such case, assessors may select 'not applicable' for the third and fourth bullet points above (see below for the instructions).

***Checklist Questions in Risk-of-Bias Criteria***

The tool provides multiple checklist questions which are designed to help judgement about risk of bias within each risk-of-bias criterion. Assessors are required to answer all checklist questions in 'general' category, as well as in 'conditional' category if conditions are met. There is also an 'optional' category for each risk-of-bias criterion to allow assessors to communicate the magnitude and direction of potential bias and/or to consider results of quantitative assessment of risk of bias (e.g., through simulations) if conducted. If assessors feel that optional checklist questions are worth answering, they can answer them or otherwise 'skip' them. We suggest that detailed rationale or empirical evidence be provided when predicting magnitude and direction of bias. The response options for the majority of checklist questions are fixed. These are:

- Yes (Y)
- Seemingly yes (SY)
- Seemingly no (SN)
- No (N)
- Unclear
- Not applicable (NA)

Assessors may select a 'yes' or 'no' when there is an explicit claim or information in available materials of a study. Assessors may select 'seemingly yes' or 'seemingly no' when they infer or suspect that it seems to be yes or no based on available materials of a study. Otherwise, please select an 'unclear' when a checklist question is applicable. When 'unclear' is selected, the default algorithm suggests assessors a certain direction (Y/SY or N/SN). In this default algorithm, selecting an 'unclear' response equates to selecting a response that suggests higher risk of bias (i.e., higher risk of bias is favoured when 'unclear' is selected). When conditions are not met, assessors can select 'not applicable'. Responses to all checklist questions will be needed for judging risk of bias in the seven individual risk-of-bias criteria, as well as for judging overall risk of bias for the study findings, so please make sure to record your responses as you go.

***Risk-of-Bias Judgement within Risk-of-Bias Criteria***

Once assessors have responded to all checklist questions within a risk-of-bias criterion, they will have to judge a risk of bias for the criterion. The levels of risk of bias can be selected from the following:

- Low risk of bias (Low)
- Medium risk of bias (Med)
- High risk of bias (High)

The process of making a judgement about risk of bias is rather straightforward with the default setting. We will provide a roadmap-like diagram (algorithm) for suggested judgement in each risk-of-bias criterion. If an optional question has been answered through quantitative assessment (e.g. through simulation), assessor's judgement about risk of bias may be upgraded (e.g. from high to medium) or downgraded (e.g. medium to high) from the suggested judgement, depending on result of quantitative assessment. Again, we suggest that detailed rationale or empirical evidence be provided when predicting magnitude and direction of bias.

***Overall Risk of Bias Judgement***

Once assessors have judged six risk-of-bias criteria (it is not seven criteria because one criterion will be 'not applicable') for a study's findings, they can make an overall judgement about risk of bias. The same levels of risk of bias as above will be used as follows:

- Overall low risk of bias: a study is considered to have low risk of bias for all assessed risk-of-bias criteria for the findings
- Overall medium risk of bias: a study is considered to have medium risk of bias in at least one risk-of-bias criterion, but not to have high risk of bias for any risk-of-bias-criteria for the findings
- Overall high risk of bias: a study is considered to have high risk of bias in at least one risk-of-bias criterion for the findings

Please note this three-level classification of overall risk of bias may be too simple for some evidence syntheses because it does not take into account the frequencies of low and medium risk of bias within an overall medium risk of bias. For example, one study's findings are judged to have low risk of bias for five criteria and medium risk of bias for one criterion (this will result in an overall medium risk of bias) will be the same as another study's findings that are judged to have medium risk of bias for the six risk-of-bias criteria (this will also result in an overall medium risk of bias). Thus, recording and communicating the extent of risk of bias within an overall medium risk of bias may be useful.

If some or all of the optional checklist questions are answered through quantitative assessment of risk of bias (e.g., through simulations), some adjustment of overall judgement of risk of bias may be applied. In such case, we recommend assessors to note detailed rationale and justification for drawing an overall conclusion of risk of bias for the study findings so that transparent communication of risk of bias for the study findings will be ensured.

***Customisation (Optional)***

Users may wish to customise the tool based on detailed rationale and justification, so we provide some potential customisation.

Regarding the process of making a judgement about risk of bias within each risk-of-bias criterion, the default algorithm is the recommended setting. However, assessors can change the algorithm if they so wish. For example, in the default setting, selecting a 'unclear' response equates to selecting a response that directs to higher risk of bias (i.e., higher risk of bias is favoured when 'unclear' is selected). If assessors do not feel this setting is right, they can customise. We suggest assessors to record any deviations from the default setting and report reasons for the deviations.

Regarding making an overall judgement about risk of bias, some levels of risk of bias can be modified or added. For example, it is possible for assessors to divide overall medium risk of bias into multiple levels using the frequencies of low and medium risk of bias for better communicating the risk of bias (e.g., splitting into medium-low and medium-high). The same logic may apply to an overall high risk of bias if assessors would like to investigate the high risk of bias further (e.g. dividing an overall high risk of bias into multiple levels). We suggest assessors to record such customisation and report their rationale behind it.

***Tutorial***

In the current prototype version, assessors will normally need one copy of the Excel or portable document format (PDF) file (**Part B** of this document) for each study's estimate of effectiveness of intervention or impact of exposure. For example, if a review team has 10 separate estimates of effectiveness of intervention or impact of exposure, they will need 10 copies of the Excel or **Part B** of this PDF file to record their responses. Please note that **Part B** of this PDF file provides some elaboration for checklist questions which is not included in the Excel file.

**The PDF File**

It should be straightforward to apply the tool so please follow the instructions provided below in **Part B** of this PDF file. How to record assessors' decisions is entirely up to them. They can work with digital or hard copies. They may use an Excel sheet called 'All_decisions' described below to record or merge all decisions.

**The Excel File**

The checklist questions for individual risk-of-bias criteria are provided in each of the 'Criterion' sheets. In the individual Criterion sheets (i.e., from 'Criterion1' to 'Criterion7'), there is a drop-down list (shaded in grey) for each checklist question so assessors can just select a response (**Figure A10**).

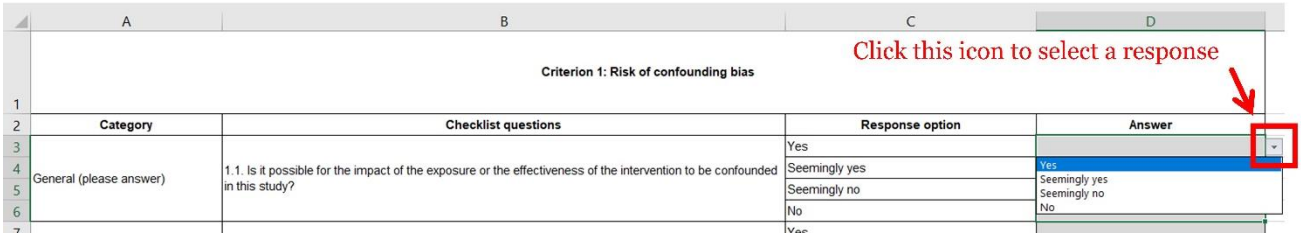


***Figure A10.*** *How to answer a checklist question in the Excel version.*

Once assessors have answered the checklist questions, they will need to use the algorithm provided in the sheet to make a judgement about risk of bias for the risk-of-bias criterion. There is also a drop-down list for this, so assessors can just select the appropriate risk of bias (**Figure A11**).

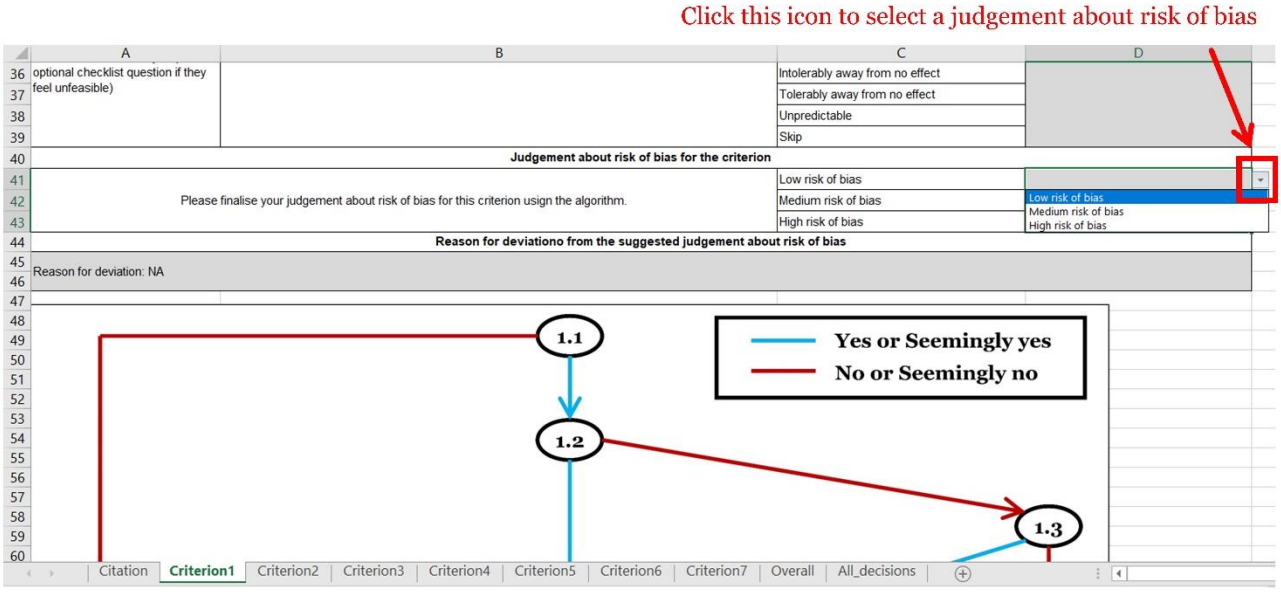


***Figure A11.*** *How to select a judgement about risk of bias within a risk-of-bias criterion in the Excel version.*

Once assessors have judged risk of bias for all risk-of-bias criteria (from 'Criterion1' to 'Criterion7'), they can make an overall judgement about risk of bias for the study findings in the 'Overall' sheet.

Judgement about risk of bias for individual risk-of-bias criteria should automatically appear in the corresponding cells shaded in grey (**Figure A12**). The criteria are also provided in the sheet so making an overall judgement about risk of bias should be straightforward. In the same way as selecting a judgement about risk of bias within a risk-of-bias criterion, assessors can select overall risk of bias using a drop-down list shaded in grey.

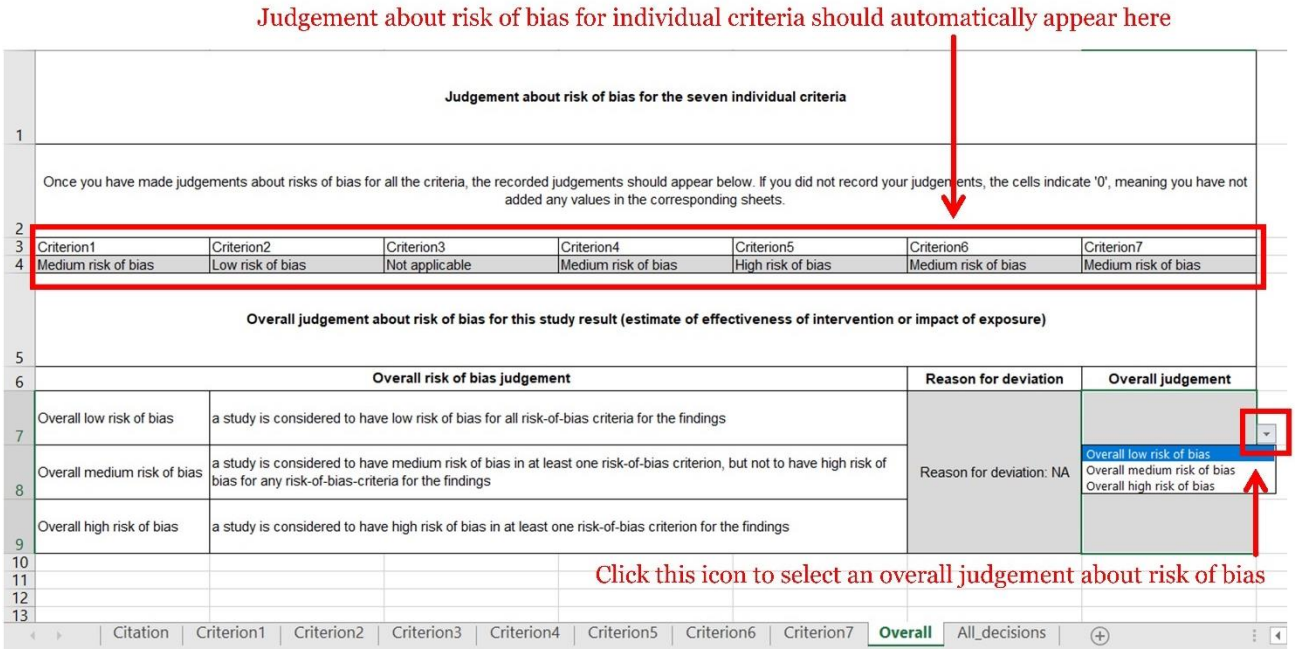


***Figure A12.*** *How to make an overall judgement about risk of bias in the Excel version.*

Once assessors have answered all checklist questions and have made judgements about risk of bias within individual risk-of-bias criteria and overall risk of bias, all decisions should automatically appear in one row (shaded in grey) in the 'All_decisions' sheet. Assessors should thus be able to merge their decisions on all included studies with ease (e.g., pasting the row to your data extraction sheet).

## Part B: Applying the Tool

### *Criterion 1: Risk of Confounding Biases*

This criterion is concerned with biases due to uncontrolled (or inappropriately controlled) variable (confounder) that influences both the intervention/exposure and the outcome. We suggest assessors (and review teams) to complete **Appendix A** and develop a causal model before answering the checklist questions below to make assessment more objective.

### Answering the Checklist Questions

Please answer the checklist questions in **Table B1** and record your responses.

***Table B1.*** *Checklist questions for risk of confounding biases.*

| Category | Checklist Questions | Answer (Tick One Applies) |
|---|---|---|
| General (please answer) | **1.1.** Is it possible for the impact of the exposure or the effectiveness of the intervention to be confounded in this study?<br><br>(E.g., Select Y/SY when randomisation is considered to deconfound.) | ☐ Yes<br>☐ Seemingly yes<br>☐ Seemingly no<br>☐ No |
| Conditional (answer if Y/SY to 1.1, otherwise select 'Not applicable') | **1.2.** Did the author(s) control for all the potential confounders?<br><br>(All potential confounders should be listed in **Appendix A**.) | ☐ Yes<br>☐ Seemingly yes<br>☐ Seemingly no<br>☐ No<br>☐ Unclear (No)<br>☐ Not applicable |
| Conditional (answer if N/SN/Unclear to 1.2, otherwise select 'Not applicable') | **1.3.** Is there any justifiable reason for not controlling for all the potential confounders (so that omission of some of the potential confounders is unlikely to influence the assessment of the effectiveness or impact)?<br><br>(E.g., select Y/SY when there is evidence that omission of some of the potential confounders does not affect the assessment of effectiveness or impact. This may be the case if adjusting all potential confounders will lead to overadjustment, or an 'instrumental variable' is used for estimating the effectiveness or impact, etc. Instrumental variable is a variable that (1) is not associated with the confounder(s), (2) is associated with the intervention/exposure but (3) does not directly influence the outcome. If used appropriately, it enables valid estimation. See Hernán & Robins 2020 for guidance.) | ☐ Yes<br>☐ Seemingly yes<br>☐ Seemingly no<br>☐ No<br>☐ Not applicable |
| Conditional (answer if Y/SY to 1.2 or 1.3, otherwise select 'Not applicable') | **1.4.** Were the potential confounders, that were controlled for, (and/or the instrumental variable used if applicable) likely to be measured accurately and precisely enough?<br><br>(Measurements of factors may be nominal (categorical), ordinal (ranks) or scale.) | ☐ Yes<br>☐ Seemingly yes<br>☐ Seemingly no<br>☐ No<br>☐ Unclear (No)<br>☐ Not applicable |
| Conditional (answer if you have answered 1.4, otherwise select 'Not applicable') | **1.5.** Did the author(s) analyse the effect appropriately by taking into account the potential confounders, as well as the issue of accuracy and precision of the measurements of the potential confounders (and the instrumental variable if applicable)?<br><br>(Examples of appropriate adjustment techniques for confounding may include stratification, matching, inverse probability weighting, standardisation, G-estimation, and instrumental variable estimation.) | ☐ Yes<br>☐ Seemingly yes<br>☐ Seemingly no<br>☐ No<br>☐ Unclear (No)<br>☐ Not applicable |

| | | |
|---|---|---|
| Optional (It is suggested that detailed rationale or empirical evidence be provided when predicting magnitude and direction of bias. Assessors may skip this optional checklist question if they feel unfeasible) | **1.6.** What are the predicted magnitude and the direction of biases due to confounding?<br><br>(Note quantitative assessment (e.g., through simulation) may be conducted by risk-of-bias assessor(s) to predict the magnitude and direction of bias for this study result.) | ☐ Intervention or exposure intolerably favoured *<br>☐ Intervention or exposure tolerably favoured **<br>☐ Comparator intolerably favoured *<br>☐ Comparator tolerably favoured **<br>☐ Intolerably towards no effect *<br>☐ Tolerably towards no effect **<br>☐ Intolerably away from no effect *<br>☐ Tolerably away from no effect **<br>☐ Unpredictable<br>☐ Skip |

** Intolerable means that the study result should not be considered as valid enough in relation to the predicted magnitude of bias. ** Tolerable means that the study result could be considered as valid enough in relation to the predicted magnitude of bias.*

Once you have answered the checklist questions, please use the diagram below (**Figure B1**) to finalise your judgement about risk of bias for this criterion.

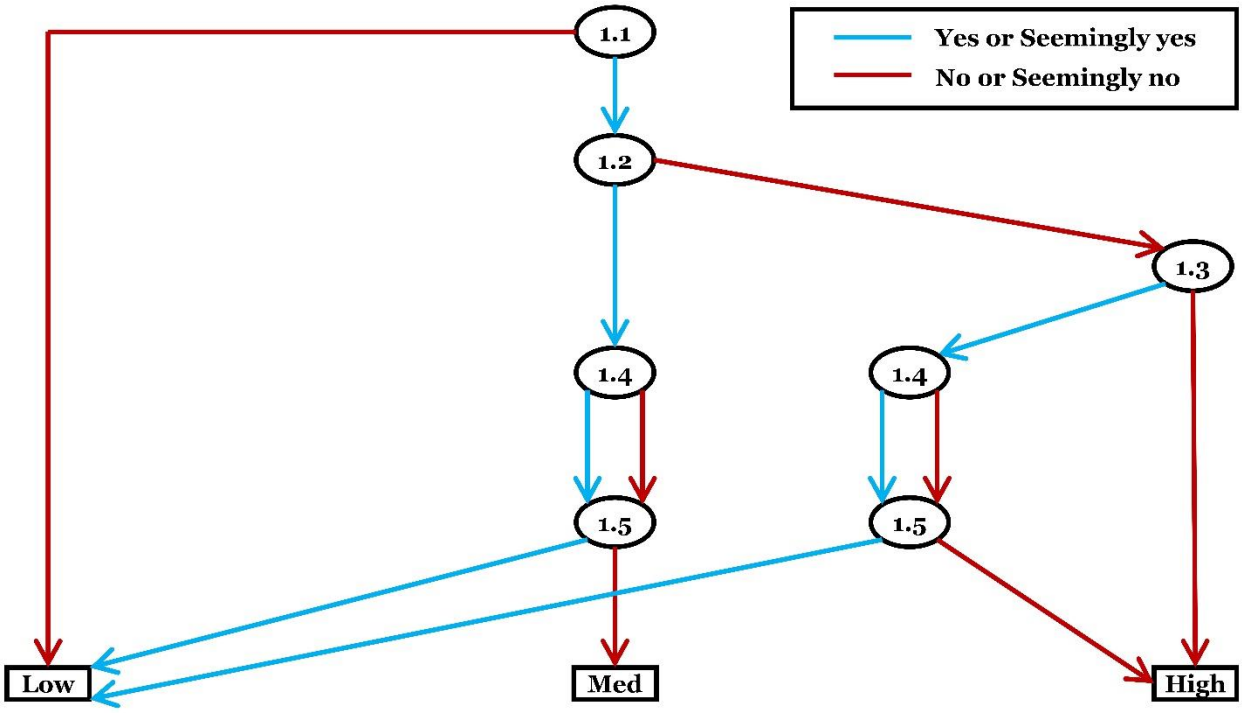


***Figure B1.*** *Roadmap diagram for making judgement about risk of confounding biases. Note: if the optional question has been answered through quantitative assessment (e.g., through simulation), assessor's judgement about risk of bias for this criterion may be upgraded or downgraded from the suggested judgement, depending on result of quantitative assessment.*

Please record your judgement about risk of bias for this criterion using **Box B1** below.

***Box B1.*** *Judgement about risk of confounding biases.*

| |
|---|
| ☐ Low risk of bias (reason for deviation from the suggested judgement: _____) |
| ☐ Medium risk of bias (reason for deviation from the suggested judgement: _____) |
| ☐ High risk of bias (reason for deviation from the suggested judgement: _____) |

| Quantitative prediction of magnitude of bias (if available): _____ |
|---|

### *Criterion 2: Risk of Post-Intervention/Exposure Selection Biases*

This criterion is concerned with biases arising from systematic differences in the selection of subjects or areas into the study or analysis after intervention or exposure.

### Answering the Checklist Questions

Please answer the checklist questions in **Table B2** and record your responses.

***Table B2.*** *Checklist questions for risk of post-intervention/exposure selection biases.*

| Category | Checklist Questions | Answer (Tick One Applies) |
|---|---|---|
| General (please answer whichever suitable) | **2.1.a.** Was the selection of subjects or areas after intervention or exposure random or systematic (i.e., based on random or systematic sampling), and exchangeability between groups could be assumed based on the selection approach?<br><br>(This applies when an attempt was not made to collect data of the entire, or nearly entire, population of inference. Exchangeability might be assumed if the two groups were exchangeable (i.e., comparable, the hypothetical effect estimate under the exchanged condition (i.e., swap) would not also be affected by post-intervention/exposure selection bias.)<br><br>OR<br><br>**2.1.b.** Was the entire (or nearly entire) population of inference followed-up after intervention or exposure, and exchangeability between before and after groups could be assumed?<br><br>(This applies when an attempt was made to collect data of the entire, or nearly entire, population of inference. Exchangeability might be assumed if the two groups were exchangeable (i.e., comparable, the hypothetical effect estimate under the exchanged condition (i.e., swap) would not also be affected by post-intervention/exposure selection bias.) | ☐ Answer 2.1.a<br>☐ Answer 2.1.b<br><br>☐ Yes<br>☐ Seemingly yes<br>☐ Seemingly no<br>☐ No<br>☐ Unclear (No) |
| General (please answer) | **2.2.** Was/were the researcher(s) unaware (or blinded) of the selection of subjects or areas?<br><br>(Non-blinding of selection might be a factor that influences the selection after intervention/exposure.) | ☐ Yes<br>☐ Seemingly yes<br>☐ Seemingly no<br>☐ No<br>☐ Unclear (No) |
| General (please answer) | **2.3.** After the start of the intervention/exposure or during the analysis, were any subjects or areas excluded or lost from the study or analysis?<br><br>(When some subjects or areas, or collected data are excluded, it might increase the risk of post-intervention/exposure selection bias.) | ☐ Yes<br>☐ Seemingly yes<br>☐ Seemingly no<br>☐ No<br>☐ Unclear (Yes) |

| | | |
|---|---|---|
| Conditional (answer if N/SN/Unclear to 2.1, or Y/SY/Unclear to 2.3, otherwise select 'Not applicable') | **2.4.** Were the subjects or areas included in the study (or analysis) comparable between groups and so they allowed a valid comparison to be made (i.e., exchangeability or conditional exchangeability between groups could be assumed)?<br><br>(Select N/SN when groups are not comparable. For example, if the effect of air pollutants on plant growth to be studied, 'resistance to pollutants' may differ. 'Less resistant' individual plants may die, and 'pollutant-resistant' individuals may survive, and when only 'pollutant-resistant' individuals are included in the exposure group and both 'less resistant' and 'pollutant-resistant' individuals are included in the control group, the effect estimate may be biased. If only 'pollutant-resistant' individuals are selected in both groups, conditional exchangeability may hold.) | ☐ Yes<br>☐ Seemingly yes<br>☐ Seemingly no<br>☐ No<br>☐ Unclear (No)<br>☐ Not applicable |
| Conditional (answer if N/SN/Unclear to 2.4, otherwise select 'Not applicable') | **2.5.** Were the difference(s) between groups likely to be explained by the intervention/exposure or a variable influenced by the intervention/exposure (including the outcome)?<br><br>(Note some variables influenced by the intervention or exposure might be unmeasured in the study. Such (unmeasured) variables are sometimes called latent variables.) | ☐ Yes<br>☐ Seemingly yes<br>☐ Seemingly no<br>☐ No<br>☐ Unclear (Yes)<br>☐ Not applicable |
| Conditional (answer if Y/SY/Unclear to 2.5, otherwise select 'Not applicable') | **2.6.** Did the author(s) adjust for the potential post-intervention/exposure selection bias in an appropriate way?<br><br>(E.g., stratification to pool stratum-specific outcomes) | ☐ Yes<br>☐ Seemingly yes<br>☐ Seemingly no<br>☐ No<br>☐ Unclear (No)<br>☐ Not applicable |
| Optional (It is suggested that detailed rationale or empirical evidence be provided when predicting magnitude and direction of bias. Assessors may skip this optional checklist question if they feel unfeasible) | **2.7.** What are the predicted magnitude and the direction of biases arising from systematic differences in the selection of subjects or areas into the study or analysis after intervention or exposure?<br><br>(Note quantitative assessment (e.g., through simulation) may be conducted by risk-of-bias assessor(s) to predict the magnitude and direction of bias for this study result.) | ☐ Intervention or exposure intolerably favoured *<br>☐ Intervention or exposure tolerably favoured **<br>☐ Comparator intolerably favoured *<br>☐ Comparator tolerably favoured **<br>☐ Intolerably towards no effect *<br>☐ Tolerably towards no effect **<br>☐ Intolerably away from no effect *<br>☐ Tolerably away from no effect **<br>☐ Unpredictable<br>☐ Skip |

** Intolerable means that the study result should not be considered as valid enough in relation to the predicted magnitude of bias. ** Tolerable means that the study result could be considered as valid enough in relation to the predicted magnitude of bias.*

Once you have answered the checklist questions, please use the diagram below (**Figure B2**) to finalise your judgement about risk of bias for this criterion.

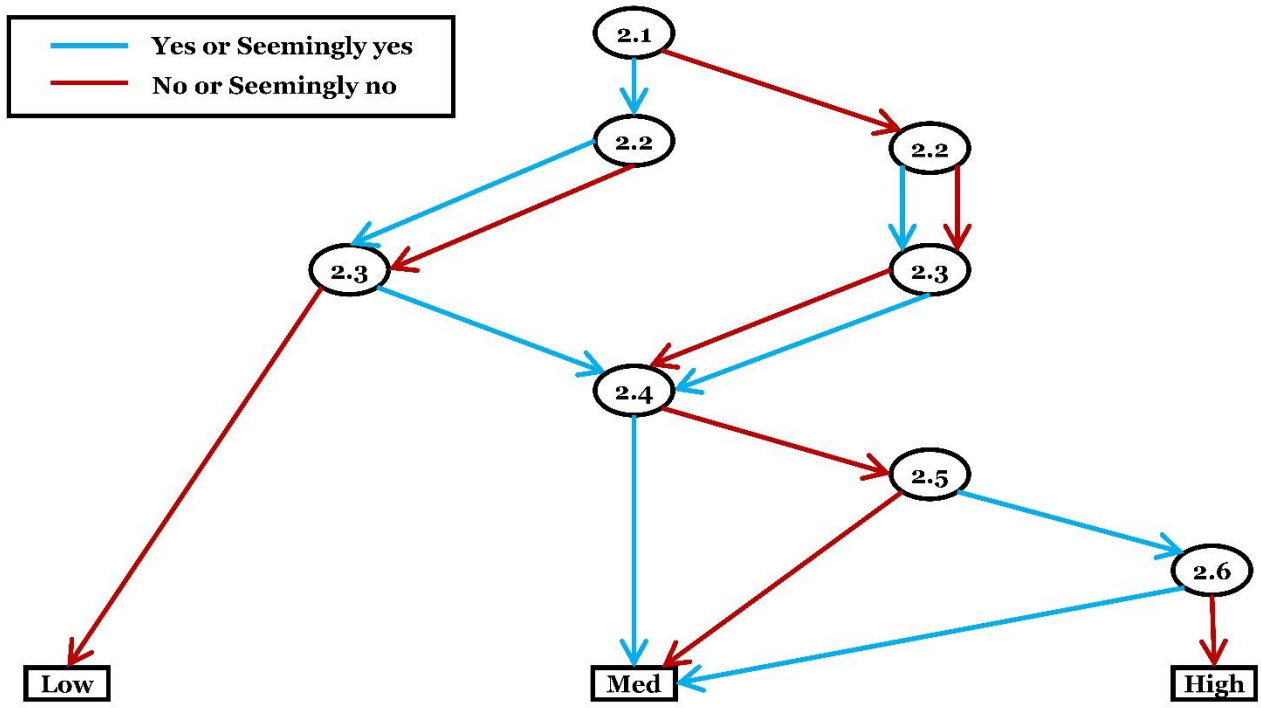


***Figure B2.*** *Roadmap diagram for making judgement about risk of post-intervention/exposure selection biases. Note: if the optional question has been answered through quantitative assessment (e.g., through simulation), assessor's judgement about risk of bias for this criterion may be upgraded or downgraded from the suggested judgement, depending on result of quantitative assessment.*

Please record your judgement about risk of bias for this criterion using **Box B2** below.

***Box B2.*** *Judgement about risk of post-intervention/exposure selection biases.*

- ☐ Low risk of bias (reason for deviation from the suggested judgement: _____)
- ☐ Medium risk of bias (reason for deviation from the suggested judgement: _____)
- ☐ High risk of bias (reason for deviation from the suggested judgement: _____)

Quantitative prediction of magnitude of bias (if available): _____

### *Criterion 3: Risk of Misclassified Comparison Biases (Observational Studies Only)*

This criterion is concerned with biases arising from misclassification or measurement of intervention, exposure and/or comparator.

### Answering the Checklist Questions

Please answer the checklist questions in **Table B3** and record your responses.

***Table B3.*** *Checklist questions for risk of misclassified comparison biases.*

| Category | Checklist Questions | Answer (Tick One Applies) |
|---|---|---|
| Conditional (answer if type of the study is observational, otherwise select 'not applicable') | **3.1.** Were the intervention or exposure (group) and the comparator (group) of interest sufficiently well-defined so that no meaningful vagueness remains for the intended assessment of causal effect of interest?<br><br>(Select N/SN if the intervention/exposure and comparator of interest are vaguely or poorly defined. Such definitions may be provided in a study protocol.) | ☐ Yes<br>☐ Seemingly yes<br>☐ Seemingly no<br>☐ No<br>☐ Unclear (No)<br>☐ Not applicable |
| Conditional (answer if type of the study is | **3.2.** Were the observed intervention or exposure (group) and the comparator (group) appropriate for the intended assessment of causal effect (i.e., causal effect of interest)? | ☐ Yes<br>☐ Seemingly yes<br>☐ Seemingly no |

| | | |
|---|---|---|
| observational, otherwise select 'not applicable') | (Select N/SN when measure or classification of exposure or intervention is unlikely to be accurate or precise enough, for example, when the use of an imprecise or inaccurate biomarker is used as a measure of exposure.) | ☐ No<br>☐ Unclear (No)<br>☐ Not applicable |
| Conditional (answer if type of the study is observational, otherwise select 'not applicable') | **3.3.** Might measure or classification of the observed exposure, intervention or comparator (group) have been incorrect due to error or influence of some knowledge, experience or desire?<br><br>(Examples may include intentional misclassification of exposure to yield a desired outcome, unintentional misclassification due to prior knowledge or cognitive bias.) | ☐ Yes<br>☐ Seemingly yes<br>☐ Seemingly no<br>☐ No<br>☐ Unclear (Yes)<br>☐ Not applicable |
| Optional (It is suggested that detailed rationale or empirical evidence be provided when predicting magnitude and direction of bias. Assessors may skip this optional checklist question if they feel unfeasible. Select 'not applicable' if experimental treatments are applied.) | **3.4.** What are the predicted magnitude and the direction of biases arising from misclassification or measurement of intervention, exposure and/or comparator?<br><br>(Note quantitative assessment (e.g., through simulation) may be conducted by risk-of-bias assessor(s) to predict the magnitude and direction of bias for this study result.) | ☐ Intervention or exposure intolerably favoured *<br>☐ Intervention or exposure tolerably favoured **<br>☐ Comparator intolerably favoured *<br>☐ Comparator tolerably favoured **<br>☐ Intolerably towards no effect *<br>☐ Tolerably towards no effect **<br>☐ Intolerably away from no effect *<br>☐ Tolerably away from no effect **<br>☐ Unpredictable<br>☐ Not applicable<br>☐ Skip |

** Intolerable means that the study result should not be considered as valid enough in relation to the predicted magnitude of bias. ** Tolerable means that the study result could be considered as valid enough in relation to the predicted magnitude of bias.*

Once you have answered the checklist questions, please use the diagram below (**Figure B3**) to finalise your judgement about risk of bias for this criterion.

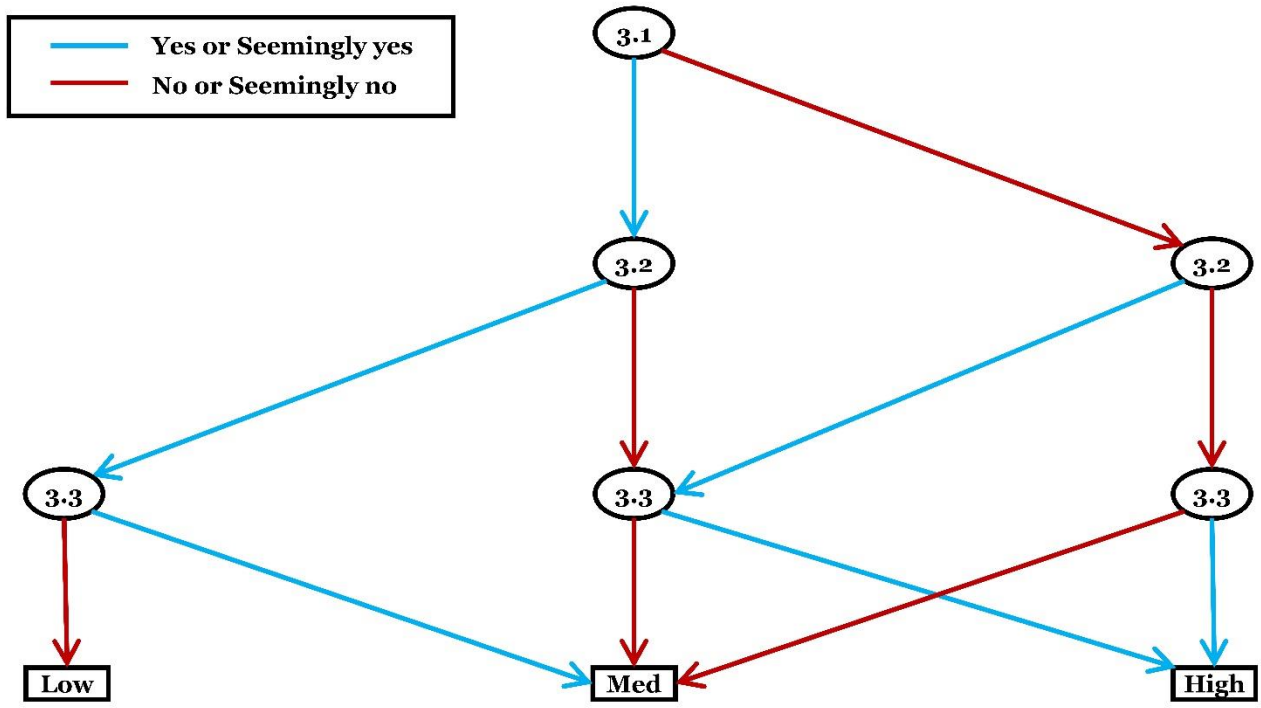


***Figure B3.*** *Roadmap diagram for making judgement about risk of misclassified comparison biases. Note: if the optional question has been answered through quantitative assessment (e.g., through simulation), assessor's judgement about risk of bias for this criterion may be upgraded or downgraded from the suggested judgement, depending on result of quantitative assessment.*

Please record your judgement about risk of bias for this criterion using **Box B3** below.

***Box B3.*** *Judgement about risk of misclassified comparison biases.*

- ☐ Low risk of bias (reason for deviation from the suggested judgement: _____)
- ☐ Medium risk of bias (reason for deviation from the suggested judgement: _____)
- ☐ High risk of bias (reason for deviation from the suggested judgement: _____)
- ☐ Not applicable

Quantitative prediction of magnitude of bias (if available): _____

### *Criterion 4: Risk of Performance Biases (Experimental Studies Only)*

This criterion is concerned with biases due to altered treatment procedure of interest.

#### Answering the Checklist Questions

Please answer the checklist questions in **Table B4** and record your responses.

***Table B4.*** *Checklist questions for risk of performance biases.*

| Category | Checklist Questions | Answer (Tick One Applies) |
|---|---|---|
| Conditional (answer if experimental treatments are applied in the study, otherwise select 'not applicable') | **4.1.** Were any of the persons, who applied or received treatments (intervention, exposure, alternative intervention, alternative exposure, or control), aware of the hypothesis that was being tested or the comparison that was being made?<br><br>(Awareness of intervention, exposure, alternative intervention, alternative exposure or control treatment and the outcome of interest might be a factor that influences the treatment procedure.) | ☐ Yes<br>☐ Seemingly yes<br>☐ Seemingly no<br>☐ No<br>☐ Unclear (Yes)<br>☐ Not applicable |
| Conditional (answer if | **4.2.** Were there any alterations of intervention/exposure or control treatments of interest that might have an impact on | ☐ Yes<br>☐ Seemingly yes |

| | | |
|---|---|---|
| experimental treatments are applied in the study, otherwise select 'not applicable') | the effectiveness of the intervention or the impact of the exposure?<br><br>(Examples of alterations may include deviated initiation, implementation and/or discontinuation. E.g., select Y/SY if starting time for the intervention/exposure is deviated from a specified time window) | □ Seemingly no<br>□ No<br>□ Unclear (Yes)<br>□ Not applicable |
| Conditional (answer if Y/SY/Unclear to 4.2, otherwise select 'Not applicable') | **4.3.** Were these deviated treatments unbalanced between intervention or exposure groups (when comparing two interventions or exposures), or inaccurately taken into account (when comparing intervention or exposure vs. control, and thus it might have influenced the estimate of impact or effectiveness?<br><br>(E.g., select Y/SY when nitrogen fertilizer was mistakenly applied more than initially planned for one group, but this deviation is not reflected on the data collection sheet, i.e., not occurred as recorded.) | □ Yes<br>□ Seemingly yes<br>□ Seemingly no<br>□ No<br>□ Unclear (Yes)<br>□ Not applicable |
| Conditional (answer if experimental treatments are applied in the study, otherwise select 'not applicable') | **4.4.** Were both exposure/intervention and comparator treatments initiated and implemented as intended (or occurred as recorded) for all, or nearly all, subjects or areas?<br><br>(When intervention, exposure, alternative intervention, alternative exposure or control treatment is not successful, it might be a source of performance bias.) | □ Yes<br>□ Seemingly yes<br>□ Seemingly no<br>□ No<br>□ Unclear (No)<br>□ Not applicable |
| Conditional (answer if Y/SY/Unclear to 4.2, or N/SN/Unclear to 4.4, otherwise select 'Not applicable') | **4.5.** Are the used analysis methods of the impact of the exposure or the effectiveness of the intervention appropriate in relation to bias due to altered treatment procedure of interest?<br><br>(E.g., select N/SN if the altered treatment procedure is not taken into account in the analysis.) | □ Yes<br>□ Seemingly yes<br>□ Seemingly no<br>□ No<br>□ Unclear (No)<br>□ Not applicable |
| Optional (It is suggested that detailed rationale or empirical evidence be provided when predicting magnitude and direction of bias. Assessors may skip this optional checklist question if they feel unfeasible. Select 'not applicable' if no experimental treatments are applied) | **4.6.** What are the predicted magnitude and the direction of biases due to altered treatment procedure of interest?<br><br>(Note quantitative assessment (e.g., through simulation) may be conducted by risk-of-bias assessor(s) to predict the magnitude and direction of bias for this study result.) | □ Intervention or exposure intolerably favoured *<br>□ Intervention or exposure tolerably favoured **<br>□ Comparator intolerably favoured *<br>□ Comparator tolerably favoured **<br>□ Intolerably towards no effect *<br>□ Tolerably towards no effect **<br>□ Intolerably away from no effect *<br>□ Tolerably away from no effect **<br>□ Unpredictable<br>□ Not applicable<br>□ Skip |

** Intolerable means that the study result should not be considered as valid enough in relation to the predicted magnitude of bias. ** Tolerable means that the study result could be considered as valid enough in relation to the predicted magnitude of bias.*

Once you have answered the checklist questions, please use the diagram below (**Figure B4**) to finalise your judgement about risk of bias for this criterion.

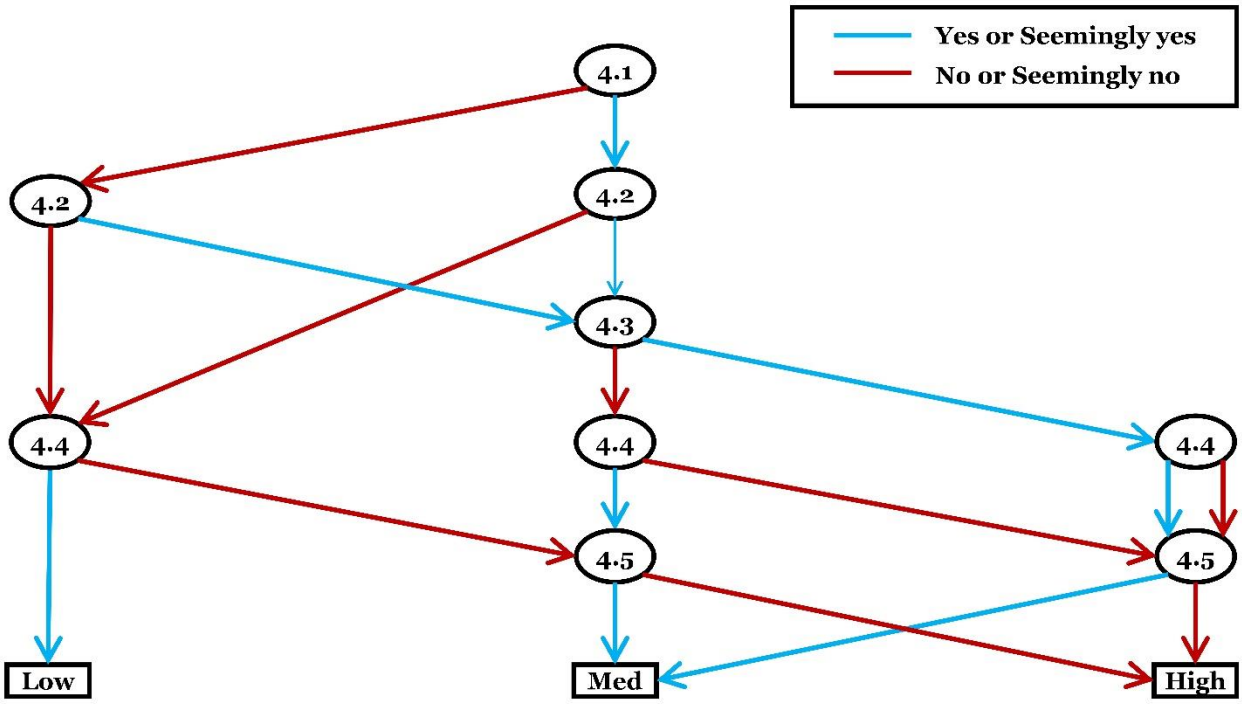


***Figure B4.*** *Roadmap diagram for making judgement about risk of performance biases. Note: if the optional question has been answered through quantitative assessment (e.g., through simulation), assessor's judgement about risk of bias for this criterion may be upgraded or downgraded from the suggested judgement, depending on result of quantitative assessment.*

Please record your judgement about risk of bias for this criterion using **Box B4** below.

***Box B4.*** *Judgement about risk of performance biases.*

- ☐ Low risk of bias (reason for deviation from the suggested judgement: _____)
- ☐ Medium risk of bias (reason for deviation from the suggested judgement: _____)
- ☐ High risk of bias (reason for deviation from the suggested judgement: _____)
- ☐ Not applicable

Quantitative prediction of magnitude of bias (if available): _____

### *Criterion 5: Risk of Detection Biases*

This criterion is concerned with biases arising from systematic differences in measurement of outcomes.

#### Answering the Checklist Questions

Please answer the checklist questions in **Table B5** and record your responses.

***Table B5.*** *Checklist questions for risk of detection biases.*

| Category | Checklist Questions | Answer (Tick One Applies) |
|---|---|---|
| General (please answer) | **5.1.** Was there any way for the outcome measure to be affected by knowledge of the exposure, intervention, subjects or areas, or desire for certain outcome?<br><br>(E.g., select Y/SY if data collectors who measured the outcome or human subjects who report their outcomes were aware of the details of the study.) | ☐ Yes<br>☐ Seemingly yes<br>☐ Seemingly no<br>☐ No<br>☐ Unclear (Yes) |

| | | |
|---|---|---|
| General (please answer) | **5.2.** Was the measured outcome appropriate for the intended assessment of causal effect (i.e., causal effect of interest)?<br><br>(E.g., select Y/SY if the measured outcome is consistent with the intended outcome; select N/SN if measured outcome is different from the pre-specified outcome or inappropriate for the intended assessment of causal effect.) | ☐ Yes<br>☐ Seemingly yes<br>☐ Seemingly no<br>☐ No<br>☐ Unclear (No) |
| General (please answer) | **5.3.** Were the methods for measuring the outcome data the same across the groups?<br><br>(E.g., select N/SN if the outcome was measured at different time windows between groups (e.g., 2 weeks after intervention vs. 4 weeks after control treatment). When exactly the same methods cannot be used among the groups due to the nature of the study, assessors may select Y/SY if the methods are sufficiently comparable. E.g., if a study measure bird species diversity in fields, and detectability of species is slightly different between groups but the slight difference in detectability can be considered comparable.) | ☐ Yes<br>☐ Seemingly yes<br>☐ Seemingly no<br>☐ No<br>☐ Unclear (No) |
| Conditional (answer if N/SN/Unclear to 5.3) | **5.4.** Were the potential differences in measured outcomes between groups investigated and adjusted/corrected if necessary?<br><br>(Differences in measured outcomes between groups can be a source of detection bias.) | ☐ Yes<br>☐ Seemingly yes<br>☐ Seemingly no<br>☐ No<br>☐ Unclear (No)<br>☐ Not applicable |
| Optional (It is suggested that detailed rationale or empirical evidence be provided when predicting magnitude and direction of bias. Assessors may skip this optional checklist question if they feel unfeasible) | **5.4.** What are the predicted magnitude and the direction of biases arising from systematic differences in measurement of outcomes?<br><br>(Note quantitative assessment (e.g., through simulation) may be conducted by risk-of-bias assessor(s) to predict the magnitude and direction of bias for this study result.) | ☐ Intervention or exposure intolerably favoured *<br>☐ Intervention or exposure tolerably favoured **<br>☐ Comparator intolerably favoured *<br>☐ Comparator tolerably favoured **<br>☐ Intolerably towards no effect *<br>☐ Tolerably towards no effect **<br>☐ Intolerably away from no effect *<br>☐ Tolerably away from no effect **<br>☐ Unpredictable<br>☐ Skip |

** Intolerable means that the study result should not be considered as valid enough in relation to the predicted magnitude of bias. ** Tolerable means that the study result could be considered as valid enough in relation to the predicted magnitude of bias.*

Once you have answered the checklist questions, please use the diagram below (**Figure B5**) to finalise your judgement about risk of bias for this criterion.

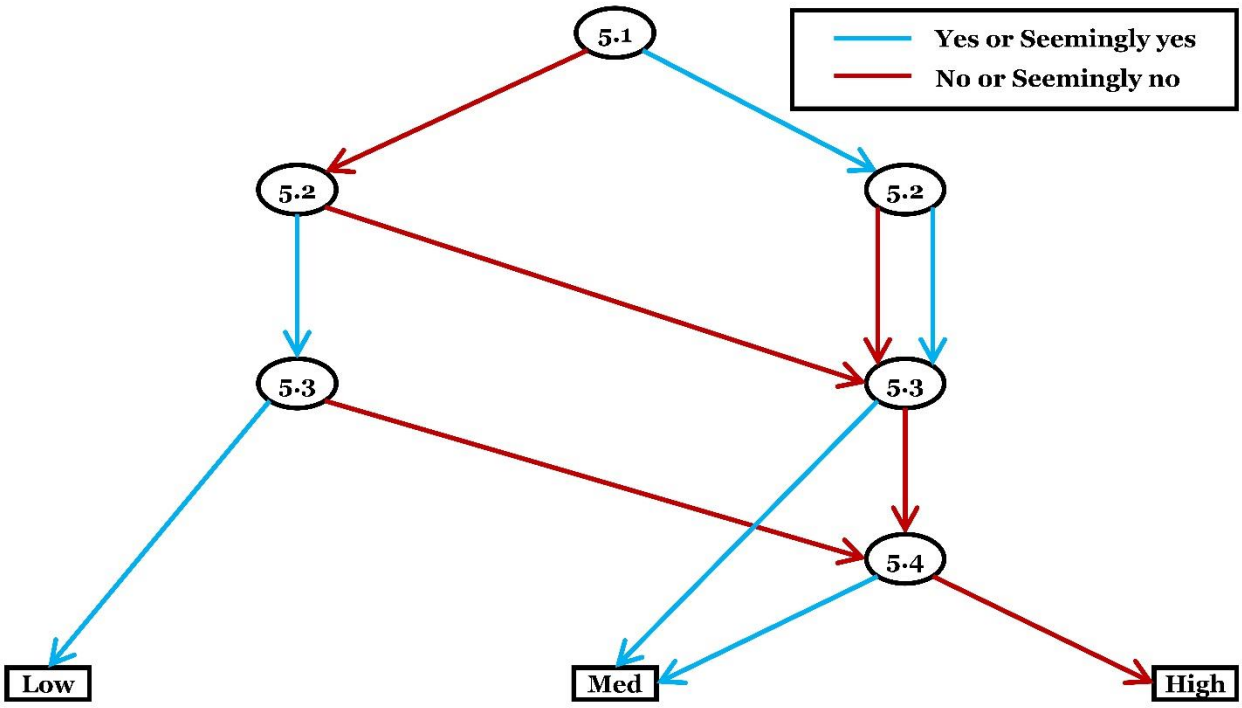


***Figure B5.*** *Roadmap diagram for making judgement about risk of detection biases. Note: if the optional question has been answered through quantitative assessment (e.g., through simulation), assessor's judgement about risk of bias for this criterion may be upgraded or downgraded from the suggested judgement, depending on result of quantitative assessment.*

Please record your judgement about risk of bias for this criterion using **Box B5** below.

***Box B5.*** *Judgement about risk of detection biases.*

- ☐ Low risk of bias (reason for deviation from the suggested judgement: _____)
- ☐ Medium risk of bias (reason for deviation from the suggested judgement: _____)
- ☐ High risk of bias (reason for deviation from the suggested judgement: _____)

Quantitative prediction of magnitude of bias (if available): _____

### *Criterion 6: Risk of Outcome Reporting Biases*

This criterion is concerned with biases in reporting of study findings.

#### Answering the Checklist Questions

Please answer the checklist questions in **Table B6** and record your responses.

***Table B6.*** *Checklist questions for risk of outcome reporting biases.*

| Category | Checklist Questions | Answer (Tick One Applies) |
|---|---|---|
| General (please answer) | **6.1.** Are the reported relevant outcome data (or effect estimate) likely to be of (or based on) selected measurements of the outcome?<br><br>(I.e., only a part of measured outcomes is reported. E.g., only 80 measured outcomes are reported when there are 100, or the effect estimate is based on 80 measured outcomes when there are 100.) | ☐ Yes<br>☐ Seemingly yes<br>☐ Seemingly no<br>☐ No<br>☐ Unclear (Yes) |
| General (please answer) | **6.2.** Are relevant outcome data likely to be unreported for some subgroup(s)? | ☐ Yes<br>☐ Seemingly yes<br>☐ Seemingly no<br>☐ No<br>☐ Unclear (Yes) |

| | | |
|---|---|---|
| | (I.e., only outcome data on certain subjects or areas with certain characteristic(s) (e.g., taxonomic group) or in certain conditions (e.g., intervention intensity) are available.) | |
| General (please answer) | **6.3.** Is/are the analysis/analyses of the causal relationship of interest (intervention-outcome or exposure-outcome) likely to be partially reported?<br><br>(I.e., there is/are other relevant analysis/analyses of the causal relationship that is/are not reported.) | ☐ Yes<br>☐ Seemingly yes<br>☐ Seemingly no<br>☐ No<br>☐ Unclear (Yes) |
| Optional (It is suggested that detailed rationale or empirical evidence be provided when predicting magnitude and direction of bias. Assessors may skip this optional checklist question if they feel unfeasible) | **6.4.** What are the predicted magnitude and the direction of biases in reporting of study findings?<br><br>(Note quantitative assessment (e.g., through simulation) may be conducted by risk-of-bias assessor(s) to predict the magnitude and direction of bias for this study result.) | ☐ Intervention or exposure intolerably favoured *<br>☐ Intervention or exposure tolerably favoured **<br>☐ Comparator intolerably favoured *<br>☐ Comparator tolerably favoured **<br>☐ Intolerably towards no effect *<br>☐ Tolerably towards no effect **<br>☐ Intolerably away from no effect *<br>☐ Tolerably away from no effect **<br>☐ Unpredictable<br>☐ Skip |

** Intolerable means that the study result should not be considered as valid enough in relation to the predicted magnitude of bias. ** Tolerable means that the study result could be considered as valid enough in relation to the predicted magnitude of bias.*

Once you have answered the checklist questions, please use the diagram below (**Figure B6**) to finalise your judgement about risk of bias for this criterion.

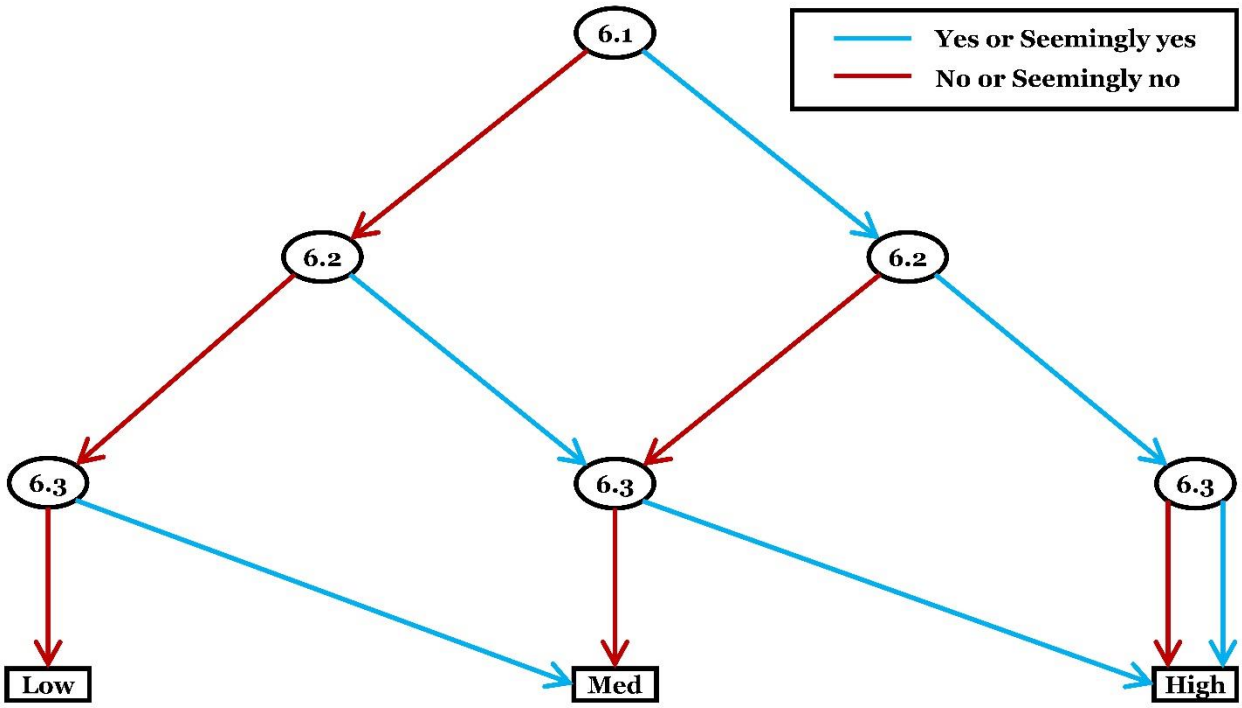

***Figure B6.*** *Roadmap diagram for making judgement about risk of outcome reporting biases. Note: if the optional question has been answered through quantitative assessment (e.g., through simulation), assessor's judgement about risk of bias for this criterion may be upgraded or downgraded from the suggested judgement, depending on result of quantitative assessment.*

Please record your judgement about risk of bias for this criterion using **Box B6** below.

***Box B6.*** *Judgement about risk of outcome reporting biases.*

- ☐ Low risk of bias (reason for deviation from the suggested judgement: _____)
- ☐ Medium risk of bias (reason for deviation from the suggested judgement: _____)
- ☐ High risk of bias (reason for deviation from the suggested judgement: _____)

Quantitative prediction of magnitude of bias (if available): _____

### *Criterion 7: Risk of Outcome Assessment Biases*

This criterion is concerned with biases due to error in applied statistical methods.

### Answering the Checklist Questions

Please answer the checklist questions in **Table B7** and record your responses.

***Table B7.*** *Checklist questions for risk of outcome assessment biases.*

| Category | Checklist Questions | Answer (Tick One Applies) |
|---|---|---|
| General (please answer) | **7.1.** Was/were the person(s), who estimated the effectiveness of the intervention or the impact of the exposure, aware of the exposure or intervention received by subjects or areas?<br><br>(E.g., select Y/SY if analysts were aware of the details of the study.) | ☐ Yes<br>☐ Seemingly yes<br>☐ Seemingly no<br>☐ No<br>☐ Unclear (Yes) |
| General (please answer) | **7.2.** Is it likely that there is/are error(s) or inappropriate methods in the applied descriptive statistical analyses?<br><br>(E.g., miscalculations of sample sizes, means, medians, variances, ranges for intervention/exposure and comparator groups, error in converting analogue to digital data.) | ☐ Yes<br>☐ Seemingly yes<br>☐ Seemingly no<br>☐ No<br>☐ Unclear (Yes) |
| Conditional (answer if inferential statistics are applied, otherwise select 'not applicable') | **7.3.** Is it likely that there is/are error(s) in the applied inferential statistics (including null hypothesis testing, estimation, coding)?<br><br>(E.g., miscalculations of differences between intervention/exposure and comparator, errors in coding, etc.) | ☐ Yes<br>☐ Seemingly yes<br>☐ Seemingly no<br>☐ No<br>☐ Unclear (Yes)<br>☐ Not applicable |
| Conditional (answer if inferential statistics are applied, otherwise select 'not applicable') | **7.4.** Were assumptions for the applied inferential statistics violated or the applied inferential statistical methods inappropriate for the inferential goal(s)?<br><br>(E.g., use of inappropriate sample sizes to test the hypothesis, normality not assumed when conducting a parametric test, equal or unequal variances not tested when testing for a difference, no justification for the choice of dependent and independent variables, a Pearson's correlation test was used when analysing a causal relationship, inappropriate comparison of multiple models to support the provided statement when some of the models do not relate to impact or effectiveness, inappropriate modelling which may affect an estimate of effectiveness or impact.) | ☐ Yes<br>☐ Seemingly yes<br>☐ Seemingly no<br>☐ No<br>☐ Unclear (Yes)<br>☐ Not applicable |
| Optional (It is suggested that | **7.5.** What are the predicted magnitude and the direction of biases due to error in applied statistical methods? | ☐ Intervention or exposure intolerably favoured * |

| | | |
|---|---|---|
| detailed rationale or empirical evidence be provided when predicting magnitude and direction of bias. Assessors may skip this optional checklist question if they feel unfeasible) | (Note quantitative assessment (e.g., through simulation) may be conducted by risk-of-bias assessor(s) to predict the magnitude and direction of bias for this study result.) | ☐ Intervention or exposure tolerably favoured **<br>☐ Comparator intolerably favoured *<br>☐ Comparator tolerably favoured **<br>☐ Intolerably towards no effect *<br>☐ Tolerably towards no effect **<br>☐ Intolerably away from no effect *<br>☐ Tolerably away from no effect **<br>☐ Unpredictable<br>☐ Skip |

** Intolerable means that the study result should not be considered as valid enough in relation to the predicted magnitude of bias. ** Tolerable means that the study result could be considered as valid enough in relation to the predicted magnitude of bias.*

Once you have answered the checklist questions, please use the diagram below (**Figure B7**) to finalise your judgement about risk of bias for this criterion.

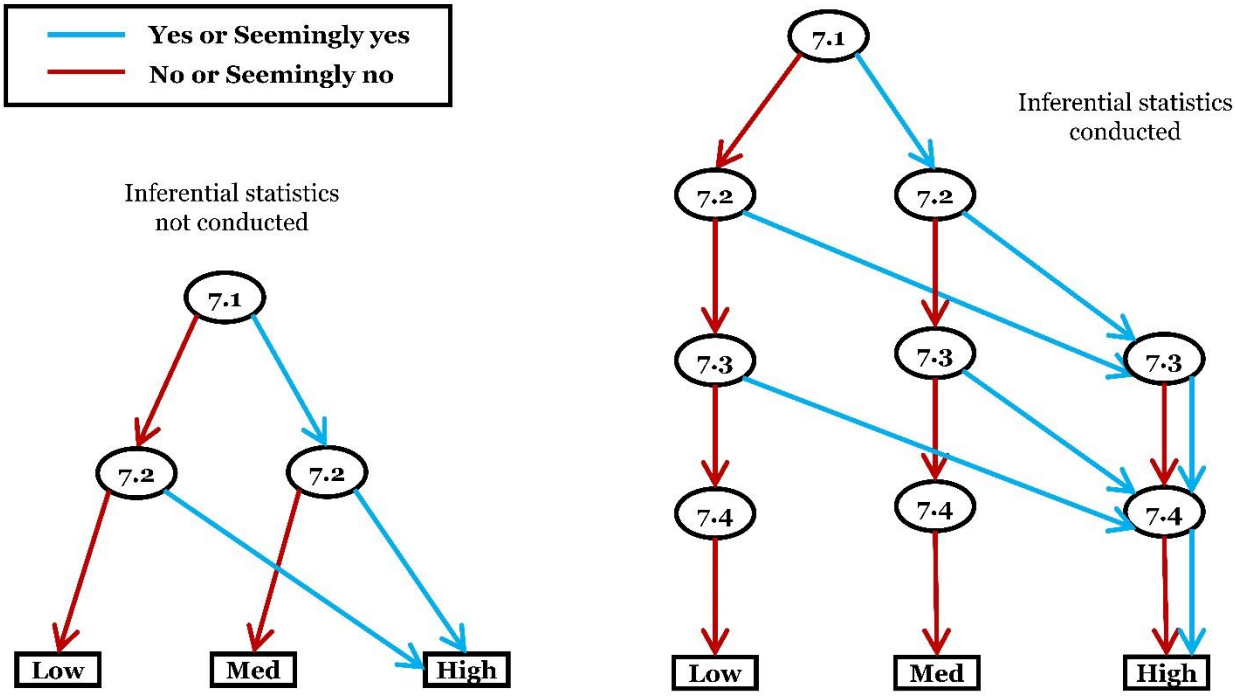


***Figure B7.*** *Roadmap diagram for making judgement about risk of outcome assessment biases. Note: if the optional question has been answered through quantitative assessment (e.g., through simulation), assessor's judgement about risk of bias for this criterion may be upgraded or downgraded from the suggested judgement, depending on result of quantitative assessment.*

Please record your judgement about risk of biases for this criterion using **Box B7** below.

***Box B7.*** *Judgement about risk of outcome assessment biases.*

☐ Low risk of bias (reason for deviation from the suggested judgement: _____)
☐ Medium risk of bias (reason for deviation from the suggested judgement: _____)
☐ High risk of bias (reason for deviation from the suggested judgement: _____)
Quantitative prediction of magnitude of bias (if available): _____

### *Making an Overall Judgement about Risk of Bias for the Study Result*

Once you have judged risk of bias for all criteria, please make an overall judgement about risk of bias for this study result using **Box B8** below. If some or all of the optional questions have been answered through quantitative assessment (e.g., through simulation), assessor's overall judgement about risk of bias may be upgraded or downgraded from the suggested judgement, depending on result of quantitative assessment. Also, if there is a justifiable reason for upgrading or downgrading overall judgement even when no quantitative assessments are conducted, assessors may deviate from the suggested judgement and record the reason.

***Box B8.*** *Overall judgement about risk of bias for the study result.*

- ☐ Overall low risk of bias: a study is considered to have low risk of bias for all risk-of-bias criteria for the findings (reason for deviation from the suggested judgement: _____)
- ☐ Overall medium risk of bias: a study is considered to have medium risk of bias in at least one risk-of-bias criterion, but not to have high risk of bias for any risk-of-bias-criteria for the findings (reason for deviation from the suggested judgement: _____)
- ☐ Overall high risk of bias: a study is considered to have high risk of bias in at least one risk-of-bias criterion for the findings (reason for deviation from the suggested judgement: _____)

Quantitative prediction of magnitude of bias (if available): _____

## Appendix A: Listing All Potential Confounders

| **Addressed Review Question and Its Elements** | |
|---|---|
| Review Question | |
| Population | |
| Intervention/Exposure | |
| Comparator | |
| Outcome | |
| **Determined All Potential Confounders (Reason for the choice of potential confounders should be provided)** | |
| Potential Confounder 1 | (Reason: ) |
| Potential Confounder 2 | (Reason: ) |
| Potential Confounder 3 | (Reason: ) |
| Potential Confounder 4 | (Reason: ) |
| Potential Confounder 5 | (Reason: ) |
| Potential Confounder 6 | (Reason: ) |
| Potential Confounder 7 | (Reason: ) |
| Potential Confounder 8 | (Reason: ) |
| Potential Confounder 9 | (Reason: ) |
| Potential Confounder 10 | (Reason: ) |
| Potential Confounder 11 | (Reason: ) |
| Potential Confounder 12 | (Reason: ) |

## References


1. Eisenhart, C. Expression of the uncertainties of final results. *Science* **1968**, *160*, 1201–1204, doi:10.1126/science.160.3833.1201.

2. Higgins, J.; Thomas, J.; Chandler, J.; Cumpston, M.; Li, T.; Page, M.; Welch, V. *Cochrane Handbook for Systematic Reviews of Interventions version 6.0*; Cochrane, 2019.

3. *CEE Guidelines and Standards for Evidence Synthesis in Environmental Management Version 5.0*; Collaboration for Environmental Evidence, 2018, doi:http://www.environmentalevidence.org/information-for-authors.

4. Sterne, J.A.C.; Savović, J.; Page, M.J.; Elbers, R.G.; Blencowe, N.S.; Boutron, I.; Cates, C.J.; Cheng, H.Y.; Corbett, M.S.; Eldridge, S.M.; et al. RoB 2: A revised tool for assessing risk of bias in randomised trials. *BMJ* **2019**, *366*, l4898, doi:10.1136/bmj.l4898.

5. Sterne, J.A.; Hernán, M.A.; Reeves, B.C.; Savović, J.; Berkman, N.D.; Viswanathan, M.; Henry, D.; Altman, D.G.; Ansari, M.T.; Boutron, I.; et al. ROBINS-I: A tool for assessing risk of bias in non-randomised studies of interventions. *BMJ* **2016**, *355*, i4919, doi:10.1136/bmj.i4919.

6. The Royal Society *Risk Analysis, Perception and Management*; The Royal Society: London, 1992.

7. Kerlinger, F. *Foundations of Behavioral Research*; 2nd ed.; Holt: New York, 1973.

8. Pearl, J. *Causality*; 2nd Editio.; Cambridge University Press: New York, 2009.

9. Hernán, M.A.; Robins, J.M. *Causal Inference: What If*; Chapman & Hall/CRC: Boca Raton, 2020.

10. Pearl, J. Causal diagrams for empirical research. *Biometrika* **1995**, *82*, 669–688, doi:10.1093/biomet/82.4.669.

11. Greenland, S.; Pearl, J.; Robins, J.M. Causal diagrams for epidemiologic research. *Epidemiology* **1999**, *10*, 37–48, doi:10.1097/00001648-199901000-00008.

12. Greenland, S.; Robins, J.M. Identifiability, exchangeability, and epidemiological confounding. *Int. J. Epidemiol.* **1986**, *15*, 413–419, doi:10.1093/ije/15.3.413.

13. Matte, J.M.; Fraser, D.J.; Grant, J.W.A. Density-dependent growth and survival in salmonids: Quantifying biological mechanisms and methodological biases. *Fish Fish.* **2020**, *21*, 588–600, doi:10.1111/faf.12448.

14. Cochran, W.G. *Sampling techniques*; Third Edit.; Wiley, 1977; ISBN 9780471162407.

15. Zvereva, E.L.; Kozlov, M. V. Biases in studies of spatial patterns in insect herbivory. *Ecol. Monogr.* **2019**, *89*, e01361, doi:10.1002/ecm.1361.

16. Nunan, D.; Aronson, J.; Bankhead, C. Catalogue of bias: attrition bias. *BMJ evidence-based Med.* **2018**, *23*, 21–22, doi:10.1136/ebmed-2017-110883.

17. Konno, K.; Akasaka, M.; Koshida, C.; Katayama, N.; Osada, N.; Spake, R.; Amano, T. Ignoring non-English-language studies may bias ecological meta-analyses. *Ecol. Evol.* **2020**, *10*, 6373–6384, doi:10.1002/ece3.6368.

18. Cochran, W.; Cox, G. *Experimental designs*; 2nd Editio.; John Wiley & Sons: New York, 1957.

19. Toselli, G.; Sgolastra, F. Seek and you shall find: An assessment of the influence of the analytical methodologies on pesticide occurrences in honey bee-collected pollen with a systematic review. *Chemosphere* **2020**, *258*, 127358, doi:10.1016/j.chemosphere.2020.127358.

20. Li, S.; Zheng, X.; Liu, C.; Yao, Z.; Zhang, W.; Han, S. Influences of observation method, season, soil depth, land use and management practice on soil dissolvable organic carbon concentrations: A meta-analysis. *Sci. Total Environ.* **2018**, *631–632*, 105–114, doi:10.1016/J.SCITOTENV.2018.02.238.

21. Steenland, K.; Schubauer-Berigan, M.K.; Vermeulen, R.; Lunn, R.M.; Straif, K.; Zahm, S.; Stewart, P.; Arroyave, W.D.; Mehta, S.S.; Pearce, N. Risk of bias assessments and evidence syntheses for observational epidemiologic studies of environmental and occupational exposures: Strengths and limitations. *Environ. Health Perspect.* **2020**, *128*, 095002, doi:10.1289/EHP6980.