A protocol for the use of case reports/studies and case series in systematic reviews for clinical toxicology

Supplemental Materials

Appendix 1. Instructions for making risk of bias determinations for case-series/control studies

Most of the text from these instructions and criteria for judging risk of bias has been adopted verbatim or adapted from one of the latest Navigation Guide systematic reviews.

INSTRUCTIONS:

o *Please evaluate each individual study for the following nine risk of bias domains. Please answer "**low risk**," "**probably low risk**," "**probably high risk**," "**high risk**," or "**not applicable**" and provide details/justification for your rating. If there is empirical evidence or other knowledge that informs the direction of bias, please include this in your answer as well; however, if there is not enough information to do so please do not guess at the direction of bias.*

o *Additionally, please note that some internal validity issues could potentially be appropriately captured in several different risk of bias considerations. In this situation, please select the single most appropriate domain to evaluate this potential bias, to avoid double-counting the same internal validity concern.*

1. *Are the study groups at risk of not representing their source populations in a manner that might introduce selection bias?*

The source population is viewed as the population for which study investigators are targeting their study question of interest.

Examples of considerations for this risk of bias domain include: 1) the context of the case report; 2) level of detail reported for participant inclusion/exclusion (including details from previously published papers referenced in the article), with inclusion of all relevant consecutive patients in the considered period; 3) exclusion rates, attrition rates and reasons.

a. **Criteria for a judgment of LOW risk of bias (i.e., answer: No"): EITHER:**

- The descriptions of the source population, inclusion/exclusion criteria, recruitment and enrolment procedures, participation and follow-up rates were sufficiently detailed, and adequate data were supplied on the distribution of relevant study sample and population characteristics to support the assertion that risk of selection effects was minimal.

OR

- Although the descriptions and/or data as indicated in "a" above suggested the potential for selection effects, adequate support was given indicating that potential selection effects were not differential across both exposure and outcome.

OR

- Although the descriptions and/or data as indicated in "a" above suggested the potential for selection effects and there was no support indicating that potential selection effects were not differential across both exposure and outcome, selection factors appeared to be well-understood, were measured in the data set, and appropriate adjustment post hoc techniques were used to control for selection bias.

b. Criteria for the judgment of PROBABLY LOW risk of bias (i.e., answer: "Probably No"):

There is insufficient information about participant selection to permit a judgment of low risk of bias, but there is indirect evidence which suggests that inclusion/exclusion criteria, recruitment and enrollment procedures, and participation and follow-up rates were consistent across groups as described by the criteria for a judgment of low risk of bias.

c. Criteria for the judgment of HIGH risk of bias (i.e., answer: "Yes"):

- There were indications from descriptions of the source population, inclusion/exclusion criteria, recruitment and enrollment procedures, participation and follow-up rates, OR
- Data on the distribution of relevant study sample and population characteristics that risk of selection effects were substantial; AND
- There was no support to indicate that potential selection effects were not differential across both exposure and outcome; AND
- Adjustment post hoc techniques were not used to control for selection bias.

d. Criteria for the judgment of PROBABLY HIGH risk of bias (i.e., answer: "Probably Yes"):

There is insufficient information about participant selection to permit a judgment of

high risk of bias, but there is indirect evidence which suggests that inclusion/exclusion criteria, recruitment and enrollment procedures, and participation and follow-up rates were inconsistent across groups, as described below by the criteria for a judgment of high risk of bias.

e. **Criteria for the judgment of NOT APPLICABLE (risk of bias domain is not applicable to study):**

There is evidence that participant selection is not an element of study design capable of introducing risk of bias in the study.

2. Were exposure/ Intervention (toxic, treatment) assessment methods lacking accuracy?

The following list of considerations represents a collection of factors proposed by experts in various fields that may potentially influence the internal validity of the exposure assessment in a systematic manner (not those that may randomly affect overall study results). These should be interpreted only as suggested considerations, and should not be viewed as scoring or a checklist. Considering there are no control in such designs, this should be evaluated carefully to be sure the report really bring something to the actual knowledge.

List of Considerations:

Possible sources of exposure assessment metrics:

1) *Identification of the exposure*
2) *Dose evaluation*
3) *Toxicological values*
4) *Clincial effects**
5) *Biological effects**
6) *Treatments given (dose, timing, route)*

* Some clinical and biological effects might be related to exposure

For each, overall considerations include:

1) *What is the quality of the source of the metric being used?*
2) *Is the exposure measured in the study a surrogate for the exposure?*
3) *What was the temporal coverage (i.e., short or long-term exposure)?*
4) *Did the analysis account for prediction uncertainty?*
5) *How was missing data accounted for, and any data imputations incorporated?*

6) Were sensitivity analyses performed?

a. Criteria for a judgment of LOW risk of bias (i.e., answer: "No"):

The reviewers judge that there is low risk of exposure misclassification, i.e.:

- There is high confidence in the accuracy of the exposure assessment methods, such as methods that have been tested for validity and reliability in measuring the targeted exposure.

b. Criteria for the judgment of PROBABLY LOW risk of bias (i.e., answer: "Probably No"):

There is insufficient information about the exposure assessment methods to permit a judgment of low risk of bias, but there is indirect evidence that suggests that methods were robust, as described by the criteria for a judgment of low risk of bias.

c. Criteria for the judgment of HIGH risk of bias (i.e., answer: "Yes"):

The reviewers judge that there is high risk of exposure misclassification and any one of the following:

- There is low confidence in the accuracy of the exposure assessment methods; or
- Less-established or less direct exposure measurements are not validated and are suspected to introduce bias that impacts the outcome assessment; or
- Uncertain how exposure information was obtained; or:

d. Criteria for the judgment of PROBABLY HIGH risk of bias (i.e., answer: "Probably Yes"):

There is insufficient information about the exposure assessment methods to permit a judgment of high risk of bias, but there is indirect evidence that suggests that methods were not robust, as described by the criteria for a judgment of high risk of bias.

e. Criteria for the judgment of NOT APPLICABLE (risk of bias domain is not applicable to study):

There is evidence that exposure assessment methods are not capable of introducing risk of bias in the study.

3. *Were outcome assessment methods lacking accuracy?*

This item is similar to actual Navigation guideline.

a. Criteria for a judgment of LOW risk of bias (i.e., answer: "No"):

The reviewers judge that there is low risk of outcome misclassification, i.e.:

- Outcomes were assessed and defined consistently across all study participants, using valid and reliable measures and with a sufficient period of time. Note that all outcome assessment measures captured in the PECO statement are considered beforehand to be valid and reliable, unless other information provided within the study warrants a consideration otherwise; OR
- Less-established or less direct outcome measurements are validated against well-established or direct methods (e.g. self-reported alcohol consumption); OR
- Appropriate sensitivity analyses were conducted that suggest the influence of outcome misclassification would be minimal
- AND, if applicable, appropriate Quality Assurance/Quality Control (QA/QC) for methods is described and is satisfactory.

b. Criteria for the judgment of PROBABLY LOW risk of bias (i.e., answer: "Probably No"):

There is insufficient information about the outcome assessment methods to permit a judgment of low risk of bias, but there is indirect evidence which suggests that methods were robust, as described by the criteria for a judgment of low risk of bias. Appropriate QA/QC for methods are not described but the review authors judge that the outcome and the outcome assessment are objective and uniform across study groups.

c. Criteria for the judgment of HIGH risk of bias (i.e., answer: "Yes"):

The reviewers judge that there is high risk of outcome misclassification and any one of the following:

- There is low confidence in the accuracy of the outcome assessment methods; or
- Less-established or less direct outcome measurements are not validated and are suspected to introduce bias that impacts the outcome assessment
- Uncertain how outcome information was obtained

d. Criteria for the judgment of PROBABLY HIGH risk of bias (i.e., answer: "Probably Yes"):

There is insufficient information about the outcome assessment methods to permit a judgment of high risk of bias, but there is indirect evidence which suggests that methods were not robust, as described by the criteria for a judgment of high risk of bias.

e. Criteria for the judgment of NOT APPLICABLE (risk of bias domain is not

applicable to study):

There is evidence that outcome assessment methods are not capable of introducing risk of bias in the study.

4. *Was potential confounding inadequately incorporated?*

This is a very important issue for case reports and series. Indeed, because there are no control, cases reports and series to be considered will need to be well-documented about other context or treatment that might explain the outcome. List of important potential confounders, should be collectively generated by review authors prior to the initiation of screening for studies based on expert opinion and knowledge gathered from the literature:

Tier I: Important confounders

- *Other associated treatment (i.e. / intoxication, insufficient dose, history or context, ...)*
- *Medical history*

Tier II: Other potentially important confounders and effect modifiers:

- *Age, sex, country*

   a. **Criteria for a judgment of LOW risk of bias (i.e., answer: "No"):**

   - The study appropriately assessed all important confounders. The determination of specific confounders may also be informed by, but not limited to, the studies included in the overall review,

   - AND the important potential confounders were measured consistently across study groups using valid and reliable methods, or the influence of covariate measurement error was determined, through sensitivity analysis, to be minimal.

   b. **Criteria for the judgment of PROBABLY LOW risk of bias (i.e., answer: "Probably No"):**

   - The study appropriately accounted for most but not all of the important confounders or used appropriate statistical techniques;

   - AND some of the other potentially important confounders relevant using appropriate statistical techniques,

   - OR reported that these confounders were evaluated and omitted because inclusion did not substantially affect the results;

   - AND this is not expected to introduce substantial bias.

   c. **Criteria for the judgment of HIGH risk of bias (i.e., answer: "Yes"):**

- The study did not account for or evaluate multiple important confounders (gender, age, socioeconomic status),

- AND did not account for or evaluate multiple other potentially important confounders relevant (type of industry type, type of occupation, country),

- OR the important potential confounders were inappropriately measured and/or inappropriately analyzed across study groups.

d. Criteria for the judgment of PROBABLY HIGH risk of bias (i.e., answer: "Probably Yes"):

- The study evaluated some but not all of the important confounders (gender, age, socioeconomic status),

- AND some but not all of the other potentially important confounders relevant (type of industry type, type of occupation, country), OR used questionable statistical techniques for confounder adjustment;

- AND this is expected to introduce substantial bias.

e. Criteria for the judgment of NOT APPLICABLE (risk of bias domain is not applicable to study):

There is evidence that outcome assessment methods are not capable of introducing risk of bias in the study.

5. Were incomplete outcome data inadequately addressed?

This item is similar to actual Navigation guideline.

a. Criteria for a judgment of LOW risk of bias (i.e., answer: "No"):

Participants were followed long enough to obtain outcome measurements

OR any one of the following:

- No missing outcome data; or
- Reasons for missing outcome data unlikely to be related to true outcome (for survival data, censoring unlikely to introduce bias); or
- Attrition or missing outcome data balanced in numbers; or
- For dichotomous outcome data, the proportion of missing outcomes compared with observed event risk not enough to have a relevant impact on the exposure effect estimate; or

- Missing data have been imputed using appropriate methods

b. Criteria for the judgment of PROBABLY LOW risk of bias (i.e., answer: "Probably No"):

There is insufficient information about incomplete outcome data to permit a judgment of low risk of bias, but there is indirect evidence which suggests incomplete outcome data was adequately addressed, as described by the criteria for a judgment of low risk of bias.

c. Criteria for the judgment of HIGH risk of bias (i.e., answer: "Yes"):

Participants were not followed long enough to obtain outcome measurements OR 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 exposure groups; or
- For dichotomous outcome data, the proportion of missing outcomes compared with observed event risk enough to induce biologically relevant bias in intervention effect estimate; or
- For continuous outcome data, plausible effect size (difference in means or standardized difference in means) among missing outcomes enough to induce biologically relevant bias in observed effect size; or
- Potentially inappropriate application of imputation.

d. Criteria for the judgment of PROBABLY HIGH risk of bias (i.e., answer: "Probably Yes"):

There is insufficient information about incomplete outcome data to permit a judgment of high risk of bias, but there is indirect evidence which suggests incomplete outcome data was not adequately addressed, as described by the criteria for a judgment of high risk of bias.

e. Criteria for the judgment of NOT APPLICABLE (risk of bias domain is not applicable to study):

There is evidence that incomplete outcome data is not capable of introducing risk of bias in the study.

6. Does the study report appear to have selective outcome reporting?

This item is similar to actual Navigation guideline, though very unlikely in published case reports/studies and series.

a. Criteria for a judgment of LOW risk of bias (i.e., answer: "No"):

All of the study's pre-specified (primary and secondary) outcomes outlined in the pre-published protocol or the published manuscript's methods, abstract, and/or introduction section that are of interest in the review have been reported in the pre-specified way.

b. Criteria for the judgment of PROBABLY LOW risk of bias (i.e., answer: "Probably No"):

There is insufficient information about selective outcome reporting to permit a judgment of low risk of bias, but there is indirect evidence which suggests the study was free of selective reporting, as described by the criteria for a judgment of low risk of bias. This includes if a pre-published protocol is not available but the study's pre-specified (primary and secondary) outcomes outlined in the published manuscript's methods, abstract, and/or introduction section that are of interest in the review have been reported in the pre-specified way.

c. Criteria for the judgment of HIGH risk of bias (i.e., answer: "Yes"): Any one of the following:

- Not all of the study's pre-specified primary outcomes (as outlined in the pre-published protocol or published manuscript's methods, abstract, and/or introduction) have been reported; OR
- One or more primary outcomes is reported using measurements, analysis methods or subsets of the data (e.g. subscales) that were not pre-specified; OR
- One or more reported primary outcomes were not pre-specified (unless clear justification for their reporting is provided, such as an unexpected effect); OR
- One or more outcomes of interest are reported incompletely

d. Criteria for the judgment of PROBABLY HIGH risk of bias (i.e., answer: "Probably Yes"):

There is insufficient information about selective outcome reporting to permit a judgment of high risk of bias, but there is indirect evidence which suggests the study was not free of selective reporting, as described by the criteria for a judgment of high risk of bias. This includes if a pre-published protocol is not available and the study's pre-specified (primary and secondary) outcomes outlined in the published manuscript's methods, abstract, and/or introduction section that are of interest in the review have not been reported in the pre-specified way.

e. Criteria for the judgment of NOT APPLICABLE (risk of bias domain is not applicable to study):

There is evidence that selective outcome reporting is not capable of introducing risk of bias in the study.

7. Did the study receive any support from a company, study author, or other entity having a financial interest?

This item is similar to actual Navigation guideline.

a. Criteria for a judgment of LOW risk of bias (i.e., answer: "No"):

The study did not receive support from a company, study author, or other entity having a financial interest in the outcome of the study. Examples include the following:

- Funding source is limited to government, non-profit organizations, or academic grants funded by government, foundations and/or non-profit organizations;
- Chemicals or other treatment used in study were purchased from a supplier;
- Company affiliated staff are not mentioned in the acknowledgements section;
- Authors were not employees of a company with a financial interest in the outcome of the study;
- Company with a financial interest in the outcome of the study was not involved in the design, conduct, analysis, or reporting of the study and authors had complete access to the data;
- Study authors make a claim denying conflicts of interest;
- Study authors are unaffiliated with companies with financial interest, and there is no reason to believe a conflict of interest exists;
- All study authors are affiliated with a government agency (are prohibited from involvement in projects for which there is a conflict of interest or an appearance of conflict of interest).

b. Criteria for the judgment of PROBABLY LOW risk of bias (i.e., answer: "Probably No"):

There is insufficient information to permit a judgment of low risk of bias, but there is indirect evidence which suggests the study was free of support from a company, study author, or other entity having a financial interest in the outcome of the study, as described by the criteria for a judgment of low risk of bias.

c. Criteria for the judgment of HIGH risk of bias (i.e., answer: "Yes"):

The study received support from a company, study author, or other entity having a financial interest in the outcome of the study. Examples of support include:

- Research funds;
- Chemicals, equipment or testing provided at no cost;
- Writing services;
- Author/staff from study was employee or otherwise affiliated with company with financial interest;
- Company limited author access to the data;
- Company was involved in the design, conduct, analysis, or reporting of the study;
- Study authors claim a conflict of interest

d. Criteria for the judgment of PROBABLY HIGH risk of bias (i.e., answer: "Probably Yes"):

There is insufficient information to permit a judgment of high risk of bias, but there is indirect evidence which suggests the study was not free of support from a company, study author, or other entity having a financial interest in the outcome of the study, as described by the criteria for a judgment of high risk of bias.

e. Criteria for the judgment of NOT APPLICABLE (risk of bias domain is not applicable to study):

There is evidence that conflicts of interest are not capable of introducing risk of bias in the study.

8. *Did the study appear to have other problems that could put it at a risk of bias?*

This item is similar to actual Navigation guideline.

a. Criteria for a judgment of LOW risk of bias (i.e. answer: "No"):

The study appears to be free of other sources of bias.

b. Criteria for the judgment of PROBABLY LOW risk of bias (i.e., answer: "Probably No"):

There is insufficient information to permit a judgment of low risk of bias, but there is indirect evidence which suggests the study was free of other threats to validity.

c. Criteria for the judgment of HIGH risk of bias (i.e., answer: "Yes"):

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

- Stopped early due to some data-dependent process (including a formal-stopping rule); OR

- The conduct of the study is affected by interim results (e.g. recruiting additional participants from a subgroup showing greater or lesser effect); OR

- Has been claimed to have been fraudulent; OR

- Had some other problem.

d. Criteria for the judgment of PROBABLY HIGH risk of bias (i.e., answer: "Probably Yes"):

There is insufficient information to permit a judgment of high risk of bias, but there is indirect evidence which suggests the study was not free of other threats to validity, as described by the criteria for a judgment of high risk of bias.

Appendix 2. Instructions for making quality of evidence

We will assess the quality of evidence for the entire body of evidence by outcome/exposure. We will adopt or adapt from the latest Navigation Guide instructions for grading, including criteria (and keep GRADE system for randomized control studies). We will downgrade for the following five GRADE reasons: (i) risk of bias; (ii) inconsistency; (iii) indirectness; (iv) imprecision; and (v) publication. We will grade the evidence, using the three Navigation Guide quality of evidence ratings: "high" "moderate" and "low". Within each of the relevant reasons for downgrading, we will rate any concern per reason as "none", "serious" or "very serious".

We will start at "high" for randomized control studies, "moderate" for observational studies, and "low" for case reports/ series. Indeed, if case reports/studies and series might be considered, it is only on low evidence.

We will downgrade for no concern by nul (0), for a serious concern by one grade (-1), and for a very serious concern by two grades (-2).

Grading Quality

Each of the categories to consider in downgrading or upgrading the evidence is described in detail below. Please record your results on the chart at the end of each category, including a brief explanation for your ratings.

1. Down grade Categories

Category 1. Quality of Study Limitations (Risk of Bias)

Possible ratings: 0=no change; -1 or -2=downgrade 1 or 2 levels respectively

The evidence from studies can be rated down if most of the relevant evidence comes from studies that suffer from a high risk of bias. Risk of bias is rated by outcome across studies. Study limitations for each outcome for individual studies and across studies are summarized in the heat maps. GRADE outlines the following principles for moving from risk of bias in individual studies to rating certainty of evidence across studies.

1. In deciding on the overall certainty of evidence, one does not average across studies (for instance if some studies have no serious limitations, some serious limitations, and some very serious limitations, one does not automatically rate quality down by one level because of an average rating of serious limitations).

Rather, judicious consideration of the contribution of each study, with a general guide to focus on the high-quality studies is warranted.

2. This judicious consideration requires evaluating the extent to which each study contributes toward the estimate of magnitude of effect. The contribution that each study makes will usually reflect study sample size and number of outcome events. Larger studies with many events will contribute more, much larger studies with many more events will contribute much more.

3. One should be conservative in the judgment of rating down. That is, one should be confident that there is substantial risk of bias across most of the body of available evidence before one rates down for risk of bias.

4. The risk of bias should be considered in the context of other limitations. If, for instance, reviewers find themselves in a close-call situation with respect to two quality issues (risk of bias and, say, precision), GRADE suggests rating down for at least one of the two.

5. Notwithstanding the first four principles, reviewers will face close-call situations. You should acknowledge that you are in such a situation, make it explicit why you think this is the case, and make the reasons for your ultimate judgment apparent.

Rating for Risk of Bias (Study Limitations) 0 no change -1 decrease quality 1 level -2 decrease quality 2 levels		Rationale for your judgment
Randomized control studies		
Observational studies (with control)		
Case reports / series		

Category 2. Indirectness of Evidence

Possible ratings: 0=no change; -1 or -2=downgrade 1 or 2 levels

Certainty of evidence (your confidence in estimates of effect) may decrease when substantial differences exist between the population, exposure, or outcomes measured in the research studies under consideration in the review.

Evidence is direct when it directly compares the exposures in which we are interested in the populations in which we are interested and measures outcomes important to the study question (in GRADE the outcomes must be important to patients).

Based on GRADE, evidence can be indirect in one of three ways.

1. The population studied differs from the population of interest (the term applicability is often used for this form of indirectness). GRADE states that in general, one should not rate down for population differences unless one has compelling reason to think that the biology in the population of interest is so different than the population tested that the magnitude of effect will differ substantially. According to GRADE, most often, this will not be the case.

2. The intervention (exposure) tested may differ from the exposure of interest, i.e., a difference in the chemical, route and/or dose. Decisions regarding indirectness of populations and exposure depend on an understanding of whether biological or social factors are sufficiently different that one might expect substantial differences in the magnitude of effect. GRADE also states, "As with all other aspects of rating certainty of evidence, there is a continuum of similarity of the intervention that will require judgment. It is rare, and usually unnecessary, for the intended populations and interventions to be identical to those in the studies, and we should only rate down if the differences are considered sufficient to make a difference in outcome likely."

3. Outcomes may differ from those of primary interest; for instance, surrogate outcomes that are not themselves important, but measured in the presumption that changes in the surrogate reflect changes in an important outcome. The difference between desired and measured outcomes may relate to time frame. When there is a discrepancy between the time frame of measurement and that of interest, whether to rate down by one or two levels will depend on the magnitude of the discrepancy. Another source of indirectness related to measurement of outcomes is the use of substitute or surrogate endpoints in place of the exposed population's important outcome of interest. In general, the use of a surrogate outcome requires rating down the certainty of evidence by

one, or even two, levels. Consideration of the biology, mechanism, and natural history of the disease can be helpful in making a decision about indirectness. Surrogates that are closer in the putative causal pathway to the adverse outcomes warrant rating down by only one level for indirectness. GRADE states that rarely, surrogates are sufficiently well established that one should choose not to rate down certainty of evidence for indirectness. In general, evidence based on surrogate outcomes should usually trigger rating down, whereas the other types of indirectness will require a more considered judgment.

Rating for Indirectness		Rationale for your judgment
0 no change -1 decrease quality 1 level -2 decrease quality 2 levels		
Randomized control studies		
Observational studies (with control)		
Case reports / series		

Category 3. Inconsistency of Evidence

Possible ratings: 0 = no change; -1 or -2=downgrade 1 or 2 levels

According to Cochrane, "when studies yield widely differing estimates of effect (heterogeneity or variability in results) investigators should look for robust explanations for that heterogeneity. When heterogeneity exists and affects the interpretation of results, but authors fail to identify a plausible explanation, the quality of the evidence decreases."

Based on GRADE, **a body of evidence is not rated up in quality if studies yield consistent results**, **but may be rated down in quality if inconsistent**. Their stated reason is that a consistent bias will lead to consistent, spurious findings.

GRADE suggests rating down the certainty of evidence if large inconsistency (heterogeneity) in study results remains after exploration of a priori hypotheses that might explain heterogeneity. Judgment of the extent of heterogeneity is based on similarity of point estimates, extent of overlap of confidence intervals, and statistical

criteria. GRADE's recommendations refer to inconsistencies in effect size, specifically to relative measures (risk ratios and hazard ratios or odds ratios), not absolute measures.

Based on GRADE, reviewers should consider rating down for inconsistency when:

1. Point estimates vary widely across studies;

2. Confidence intervals (CIs) show minimal or no overlap;

3. The statistical test for heterogeneity-which tests the null hypothesis that all studies in a meta- analysis have the same underlying magnitude of effect-shows a low p-value;

4. The I2 -which quantifies the proportion of the variation in point estimates due to among-study differences-is large. (i.e., the I2 index quantifies the degree of heterogeneity in a meta-analysis).

GRADE states that inconsistency is important only when it reduces confidence in results in relation to a particular decision. Even when inconsistency is large, it may not reduce confidence in results regarding a particular decision. For example, studies that are inconsistent related to the magnitude of a beneficial or harmful effect (but are in the same direction) would not be rated down; in instances when results are inconsistent as to whether there is a benefit or harm of treatment, GRADE would rate down the certainty of evidence as a result of variability in results, because the meaning of the inconsistency is so relevant to the decision to treat or not to treat.

Rating for Inconsistency 0 no change -1 decrease quality 1 level -2 decrease quality 2 levels		Rationale for your judgment
Randomized control studies		
Observational studies (with control)		
Case reports / series		

Category 4. Imprecision of Evidence

Possible ratings: 0=no change; -1 or -2=downgrade 1 or 2 levels

Cochrane states that when studies have few participants and few events, and thus have wide confidence intervals (CIs), authors can lower their rating of the certainty of evidence. These ratings of precision are made as judgments by review authors. The ratings are made by looking across studies, or, if available, on the results of a meta-analysis.

GRADE defines evidence quality differently for systematic reviews and guidelines. For systematic reviews, quality refers to confidence in the estimates of effect. For guidelines, quality refers to the extent to which confidence in the effect estimate is adequate to support a particular decision. For the purpose of step 3 of Navigation Guide, we will use the systematic review definition, because the decision phase does not occur until step 4 when recommendations for prevention are made. Thus, when reviewing the data for imprecision, evaluate your confidence in the estimate of the effect.

According to GRADE, to a large extent, CIs inform the impact of random error on evidence quality. Thus, when considering imprecision, the issue is whether the CI around the estimate of exposure effect is sufficiently narrow. If it is not, GRADE rates down the evidence quality by one level (for instance, from high to moderate). If the CI is very wide, GRADE might rate down by two levels.

Rating for Imprecision 0 no change -1 decrease quality 1 level -2 decrease quality 2 levels		Rationale for your judgment
Randomized control studies		
Observational studies (with control)		
Case reports / series		

Category 5. Publication Bias

Possible ratings: 0 = no change; -1 or -2=downgrade 1 or 2 levels

GRADE assess publication bias in a similar manner. Whereas "selective outcome reporting" is assessed for each study included in the review as part of the risk of bias assessment, "publication bias" is assessed on the body of evidence. GRADE states

that "when an entire study remains unreported and the results relate to the size of the effect- publication bias- one can assess the likelihood of publication bias only by looking at a group of studies."

Cochrane's definition of publication bias is "the *publication* or *non-publication* of research findings depending on the nature and direction of the results." Cochrane and GRADE are primarily concerned with *overestimates* of true effects of treatments or pharmaceuticals, especially related to "small studies effects", i.e., the tendency for estimates of an intervention to be more beneficial in smaller studies. There is empirical evidence in the clinical sciences that publication and other reporting biases result in over estimating the effects of interventions.

In contrast, in environmental health, we are primarily concerned with *underestimating* the true effects of a chemical exposure, since in many cases population wide exposure has already occurred. We are also concerned that studies finding no association are less likely to be published because journals are less likely to publish "negative" findings.

Applying this inverted concern to GRADE's assessment for publication bias, leads to these considerations when rating publication bias:

- Early *negative* studies, particularly if small in size, are suspect. (GRADE is concerned with early *positive* studies).
- Authors of systematic reviews should suspect publication bias when studies are uniformly small, particularly when sponsored by the industry. (Same as GRADE)
- Empirical examination of patterns of results (e.g., funnel plots) may suggest publication bias but should be interpreted with caution. (Same as GRADE)
- More compelling than any of these theoretical exercises is authors' success in obtaining the results of some unpublished studies and demonstrating that the published and unpublished data show different results. (Same as GRADE)
- Comprehensive searches of the literature including unpublished studies, i.e., the grey literature, and a search for research in other languages are important to addressing publication bias. Note that Cochrane also states "comprehensive searching is not sufficient to prevent some substantial potential biases."

Rating for Publication Bias 0 no change -1 decrease quality 1 level -2 decrease quality 2 levels	Rationale for your judgment
Randomized control studies	
Observational studies (with control)	
Case reports / series	

2. *Upgrade Categories*

GRADE states that the circumstances for upgrading likely occur infrequently and are primarily relevant to observational and other non-randomized studies. Although it is possible to rate up results from randomized controlled trials, GRADE has yet to find a compelling circumstance for doing so. GRADE specifies 3 categories for increasing the certainty of evidence: Large Magnitude of Effect, Dose Response, and Residual Confounding Increases Confidence

Category 6. Large Magnitude of Effect

Possible ratings: 0 = no change; +1 or +2=upgrade 1 or 2 levels

Modelling studies suggests that confounding (from non-random allocation) alone is unlikely to explain associations with a relative risk (RR) or odd ratio (OR) greater than 2 (or less than 0.5), and very unlikely to explain associations with an RR (or OR) greater than 5 (or less than 0.2). Thus, these are the definitions of "large magnitude of effect" used by GRADE to upgrade 1 or 2 levels, respectively. Also, GRADE is more likely to rate up if the effect is rapid and out of keeping with prior trajectory; usually supported by indirect evidence. GRADE presents empirical evidence to support these conclusions, and states that "although further research is warranted, both modelling and empirical work suggest the size of bias from confounding is unpredictable in direction but bounded in size.

Hence, the GRADE group has previously suggested guidelines for rating certainty of evidence up by one category (typically from low to moderate) for associations greater than 2, and up by two categories for associations greater than 5."

At present, we do not have an empirically defined "large magnitude of effect." Therefore, for the purpose of this case study, review authors should assess whether the results indicate a large magnitude of effect using their expert judgment of "large effects" in environmental health and state their definition for discussion by the group.

Rating for Large Magnitude of Effect 0 no change +1 increase quality 1 level +2 increase quality 2 levels	Rationale for your judgment
Randomized control studies	
Observational studies (with control)	
Case reports / series	

Category 7. Dose-response

Possible ratings: 0 = no change; +1 or +2=upgrade 1 or 2 levels

Possible considerations include consistent dose response gradients in one or multiple studies, and/or dose response across studies, depending on the overall relevance to the body of evidence.

Rating for Residual Confounding Increases Confidence 0 no change +1 increase quality 1 level +2 increase quality 2 levels	Rationale for your judgment
Randomized control studies	
Observational studies (with control)	
Case reports / series	

Category 8. Residual Confounding Increases Confidence

Possible ratings: 0=no change; +1 or +2=upgrade 1 or 2 levels

Upgrade if consideration of all plausible residual confounders, biases, or effect modification would underestimate the effect or suggest a spurious effect when results show no effect. If a study reports an association despite the presence of residual

confounding, biases or effect modification that would diminish the association, confidence in the association is increased. GRADE provides an illustrative example related to bias: rating up observational evidence finding lack of association between vaccination and autism, which occurred despite empirically confirmed bias that parents of autistic children may be more likely to remember their vaccine experience. The negative findings despite this form of recall bias suggest rating up the certainty of evidence.

Rating for Residual Confounding Increases Confidence 0 no change +1 increase quality 1 level +2 increase quality 2 levels		Rationale for your judgment
Randomized control studies		
Observational studies (with control)		
Case reports / series		

The results of the reviewers' ratings by population will be compiled and discussed leading to a final decision on overall quality of human evidence. The rationale for the decision will be fully documented.

Final decision on overall quality of evidence:
---- High
---- Moderate
---- Low