Navigation Guide Protocol for Rating the Quality and Strength of Human and Non-Human Evidence

University of California, San Francisco

Program on Reproductive Health and the Environment

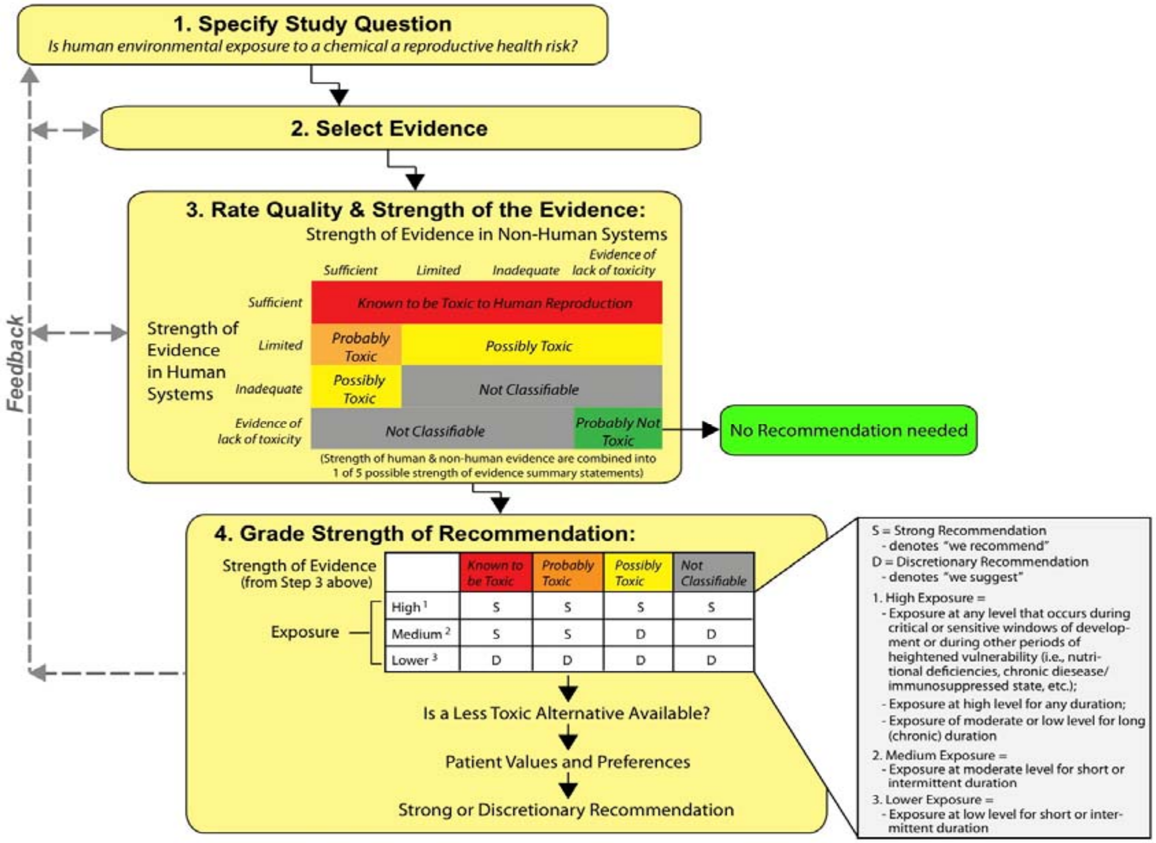
December 2012

Contents

I. Overview

This protocol addresses Step 3 in Navigation Guide, rating the quality and strength of the human and non-human evidence (Figure 1). Except as noted, the protocol is directly excerpted from the Cochrane Handbook and/or GRADE [1].

Figure 1. Navigation Guide Methodology



1. Specify Study Question
Is human environmental exposure to a chemical a reproductive health risk?
2. Select Evidence
3. Rate Quality & Strength of the Evidence:
Feedback
No Recommendation needed
4. Grade Strength of Recommendation:
Is a Less Toxic Alternative Available?
Patient Values and Preferences
Strong or Discretionary Recommendation

Strength of Evidence in Non-Human Systems

Strength of Evidence in Human Systems	Sufficient	Limited	Inadequate	Evidence of lack of toxicity
Sufficient	Known to be Toxic to Human Reproduction			
Limited	Probably Toxic	Possibly Toxic		
Inadequate	Possibly Toxic	Not Classifiable		
Evidence of lack of toxicity	Not Classifiable			Probably Not Toxic

(Strength of human & non-human evidence are combined into 1 of 5 possible strength of evidence summary statements)

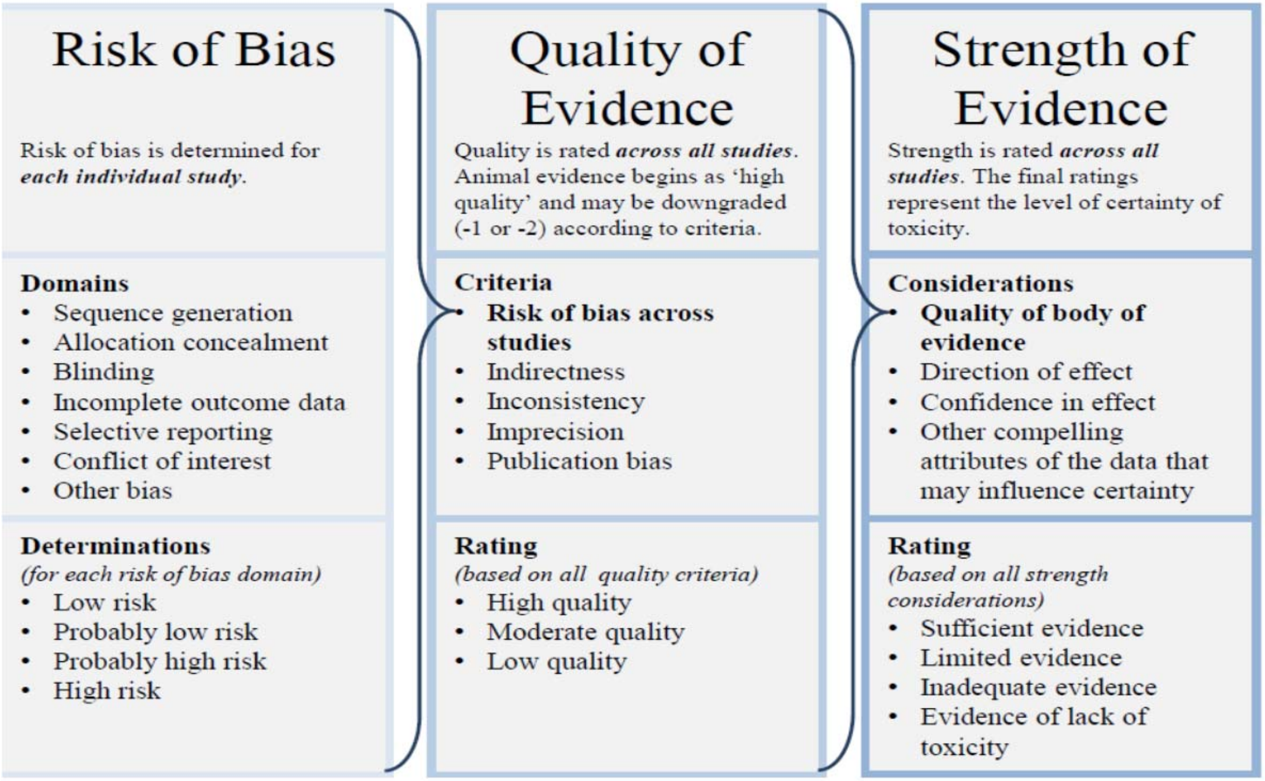
Strength of Evidence (from Step 3 above)		Known to be Toxic	Probably Toxic	Possibly Toxic	Not Classifiable
Exposure	High[1]	S	S	S	S
	Medium[2]	S	S	D	D
	Lower[3]	D	D	D	D

S = Strong Recommendation
- denotes "we recommend"

D = Discretionary Recommendation
- denotes "we suggest"

1. High Exposure =
   - Exposure at any level that occurs during critical or sensitive windows of development or during other periods of heightened vulnerability (i.e., nutritional deficiencies, chronic diesease/ immunosuppressed state, etc.);
   - Exposure at high level for any duration;
   - Exposure of moderate or low level for long (chronic) duration
2. Medium Exposure =
   - Exposure at moderate level for short or intermittent duration
3. Lower Exposure =
   - Exposure at low level for short or intermittent duration

Figure 2. Overview of Navigation Guide Method for Grading Non-Human Evidence



Risk of Bias

Risk of bias is determined for ***each individual study***.

Domains

- Sequence generation
- Allocation concealment
- Blinding
- Incomplete outcome data
- Selective reporting
- Conflict of interest
- Other bias

Determinations
(for each risk of bias domain)

- Low risk
- Probably low risk
- Probably high risk
- High risk

Quality of Evidence

Quality is rated ***across all studies***. Animal evidence begins as 'high quality' and may be downgraded (-1 or -2) according to criteria.

Criteria

- **Risk of bias across studies**
- Indirectness
- Inconsistency
- Imprecision
- Publication bias

Rating
(based on all quality criteria)

- High quality
- Moderate quality
- Low quality

Strength of Evidence

Strength is rated ***across all studies***. The final ratings represent the level of certainty of toxicity.

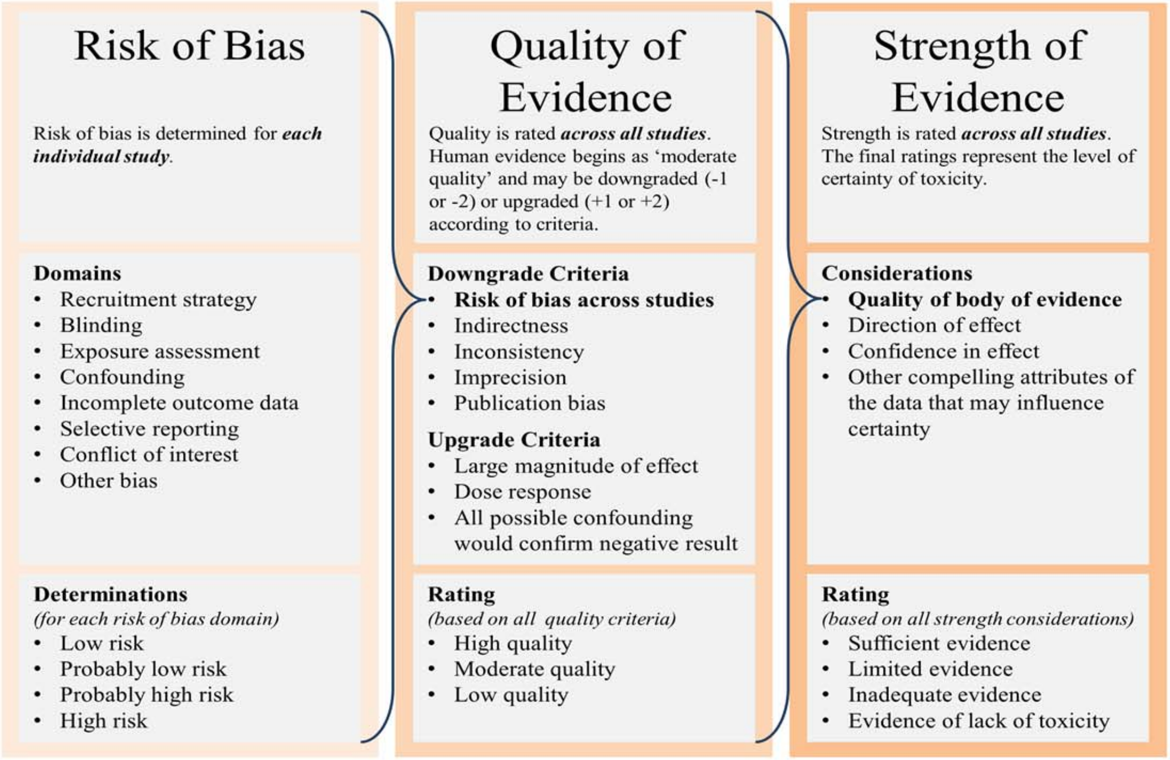
Considerations

- **Quality of body of evidence**
- Direction of effect
- Confidence in effect
- Other compelling attributes of the data that may influence certainty

Rating
(based on all strength considerations)

- Sufficient evidence
- Limited evidence
- Inadequate evidence
- Evidence of lack of toxicity

Figure 3. Overview of Navigation Guide Method for Grading Human Evidence



Risk of Bias

Risk of bias is determined for ***each individual study***.

Domains

- Recruitment strategy
- Blinding
- Exposure assessment
- Confounding
- Incomplete outcome data
- Selective reporting
- Conflict of interest
- Other bias

Determinations

(for each risk of bias domain)

- Low risk
- Probably low risk
- Probably high risk
- High risk

Quality of Evidence

Quality is rated ***across all studies***. Human evidence begins as 'moderate quality' and may be downgraded (-1 or -2) or upgraded (+1 or +2) according to criteria.

Downgrade Criteria

- **Risk of bias across studies**
- Indirectness
- Inconsistency
- Imprecision
- Publication bias

Upgrade Criteria

- Large magnitude of effect
- Dose response
- All possible confounding would confirm negative result

Rating

(based on all quality criteria)

- High quality
- Moderate quality
- Low quality

Strength of Evidence

Strength is rated ***across all studies***. The final ratings represent the level of certainty of toxicity.

Considerations

- **Quality of body of evidence**
- Direction of effect
- Confidence in effect
- Other compelling attributes of the data that may influence certainty

Rating

(based on all strength considerations)

- Sufficient evidence
- Limited evidence
- Inadequate evidence
- Evidence of lack of toxicity

1. Co-authors will independently review the final data and independently rate the quality of evidence according to *a priori* criteria set forth in the protocol.
2. Co-authors will compare their results. Any discrepancies between the co-authors' decisions will be resolved through discussion. The senior author (TW) will be the ultimate arbiter of the discrepancies that cannot be resolved through consensus among the co-authors. The final judgments of all reviewers will be documented.
3. The initial quality level of non-human experimental data is considered "high" consistent with GRADE guidelines for rating experimental studies (i.e., randomized controlled trials). The initial quality level of human observational data is considered "moderate". This is in contrast to GRADE guidelines, developed for clinical interventions, which assign observational studies an initial rating of "low" quality [2]. There is variability in the quality of studies, however, and not all observational studies may be low quality [3]. In environmental health, human observational data are the "best" data available for decision-making, and in this regard they are comparable to human randomized controlled trials (RCTs) in the clinical sciences. Because ethics virtually precludes human RCTs in environmental health, beginning human observational studies at "moderate" quality captures the value of these data relative to what data are available. In addition, human observational studies are recognized as being a reliable source of evidence in the clinical sphere, as not all healthcare decisions are, or can be, based on RCTs; [4] recognition of the absolute value of human observational data evidence-based to clinical decision-making is also increasing [5, 6].
4. "Fetal growth" is the outcome being assessed in this review.
    - In humans, the outcome fetal growth includes all the following measures: birth weight, birth length, head circumference, and ponderal index; all of these measures are sufficiently similar to rate together as a measure of the same outcome.
    - In non-human mammalians, the outcome "fetal" growth includes all the following measures: "Fetal" data, which refers to when outcome measurements are taken from progeny near-term (i.e., E18 for mice, E21 for rats). "Pup" data, which refers to when outcome measurements are taken from progeny at or soon after birth.
    - In non-human non-mammalians, the outcome fetal growth is equivalent to "embryonic" growth and includes measures of weight, length or volume, depending on the model system.
5. For the purpose of the PFOA case study, there are 3 populations for which we are rating the quality of evidence for PFOA's effect on fetal growth: (1) the quality of human evidence for fetal growth; (2) the quality of mammalian animal evidence for fetal growth; and (3) the quality of non-human, non-mammalian evidence for fetal growth.
6. There are 5 categories that can lead to **downgrading** quality of evidence for an outcome: risk of bias (study limitations); indirectness; inconsistency; imprecision; and publication bias. According to GRADE, these 5 categories address nearly all issues that bear on the quality of evidence [2]. GRADE states that these categories were arrived at through a case- based process by members of GRADE, who identified a broad range of issues and factors related to the assessment of the quality of studies. All potential factors were considered, and through an iterative process of discussion and review, concerns were scrutinized and solutions narrowed by consensus to these five categories. GRADE also

defines 3 categories that can lead to upgrading quality of evidence for an outcome: large effect; confounding would minimize effect; and dose response.

7. While GRADE specifies systematic review authors consider quality of evidence under a number of discrete categories and to either rate down or not on the basis of each category, they also state that rigid adherence to this approach ignores the fact that quality is actually a continuum and that an accumulation of limitations across categories can ultimately provide the impetus for rating down in quality [7]. Thus authors who decide to rate down quality by a single level will specify the one category most responsible for their decision while documenting all factors that contributed to the final decision to rate down quality.
8. The quality of evidence rating for human and non-human data will be translated into strength of evidence ratings for each stream of evidence.
9. The strength of evidence for human and non-human data will be combined into an overall statement of toxicity, i.e., known to be toxic to fetal growth; probably toxic to fetal growth; possibly toxic to fetal growth; known to be not-toxic to fetal growth.

II. Rate the Quality of Evidence

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.

Category 1. Rate the Quality of Study Limitations (Risk of Bias)[8]

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

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 quality of evidence across studies.

1. In deciding on the overall quality 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.[a]

[a] Note: Limitations to GRADE's risk of bias assessments as stated by GRADE: “First, empirical evidence supporting the criteria is limited. Attempts to show systematic difference between studies that meet and do not meet specific criteria have shown inconsistent results. Second, the relative weight one should put on the criteria remains uncertain. The GRADE approach is less comprehensive than many systems, emphasizing simplicity and parsimony over completeness. GRADE's approach does not provide a quantitative rating of risk of bias. Although such a rating has advantages, we share with the Cochrane Collaboration methodologists a reluctance to provide a risk of bias score that, by its nature, must make questionable assumptions about the relative extent of bias associated with individual items and fails to consider the context of the individual items.”

2. This judicious consideration requires evaluating the extent to which each study contributes toward the estimate of magnitude of effect. This contribution 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.

Risk of Bias (Study Limitations) Rating 0 no change -1 decrease quality 1 level -2 decrease quality 2 levels		**Rationale for your judgment**
Human		
Non-Human Mammalian		
Non-Human Non-Mammalian		

Category 2. Rate Indirectness of Evidence

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

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

Evidence is direct when it directly compares the exposures in which we are interested when applied to 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 [9] (as modified to reflect our "PECO" instead of "PICO" question), evidence can be indirect in one of three ways.[b]

[b] GRADE includes a fourth type of indirectness that occurs when there are no direct (i.e., head-to-head) comparisons between two or more interventions of interest. This criterion is not relevant to our study

1. The population studied differs from the population of interest (the term applicability is often used for this form of indirectness). **Please note the Navigation Guide's *a priori* assumption is that mammalian evidence of a health effect/lack of health effect is deemed to be direct evidence of human health with regards to directness of the population**. This is a marked departure from GRADE,[c] based on empirical evidence in environmental health science that the reliability of experimental animal (mammalian) data for reproductive and developmental health has been well established though multiple studies of concordance between mammalian animals and humans after exposure to a variety of chemical agents [10-14]. Presently, there is no example of a chemical agent that has adversely affected human reproduction or development but has not caused the same or similar adverse effects in mammalian animal models [12]. The National Academy of Sciences (NAS) has recognized the importance of animal data in identifying potential developmental risks. According to the NAS, studies of comparison between developmental effects in animals and humans find that "there is concordance of developmental effects between animals and humans and that humans are as sensitive or more sensitive than the most sensitive animal species [15]." 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. In applying this GRADE principle to the Navigation Guide, non-human evidence would be rated down as indirect when it is a biologically inappropriate non-human model system for the health outcome under study.

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 quality 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

question related to toxicity of PFOA; it could be relevant to future case studies.

[c] According to GRADE, in general, GRADE rates animal evidence down two levels for indirectness. They note that animal studies may, however, provide an important indication of drug toxicity. GRADE states, "Although toxicity data from animals does not reliably predict toxicity in humans, evidence of animal toxicity should engender caution in recommendations." However, GRADE does not preclude rating non-human evidence as high quality. They state, "Another type of nonhuman study may generate high- quality evidence. Consider laboratory evidence of change in resistance patterns of bacteria to antimicrobial agents (e.g., the emergence of methicillin-resistant staphylococcus aureus-MRSA). These laboratory findings may constitute high-quality evidence for the superiority of antibiotics to which MRSA is sensitive vs. methicillin as the initial treatment of suspected staphylococcus sepsis in settings in which MRSA is highly prevalent."

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 quality 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 quality 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.

Indirectness Rating 0 no change -1 decrease quality 1 level -2 decrease quality 2 levels		**Rationale for your judgment**
Human		
Non-Human Mammalian		
Non-Human Non-Mammalian		

Category 3. Rate 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 effects the interpretation of results, but authors fail to identify a plausible explanation, the quality of the evidence decreases."

Based on GRADE [16], **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 quality 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 I^2 -which quantifies the proportion of the variation in point estimates due to among-study differences-is large. (I.e., the I^2 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 quality 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.

Inconsistency Rating 0 no change -1 decrease quality 1 level -2 decrease quality 2 levels		**Rationale for your judgment**
Human		
Non-Human Mammalian		
Non-Human Non-Mammalian		

Category 4. Rate 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 quality of evidence. These ratings of precision are made as judgments by review authors.

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 [17]. 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. When considering the quality of evidence, 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.

Imprecision Rating 0 no change -1 decrease quality 1 level -2 decrease quality 2 levels		**Rationale for your judgment**
Human		
Non-Human Mammalian		
Non-Human Non-Mammalian		

Category 5. Rate Publication Bias

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

GRADE [8] and Cochrane [1] 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 [1].

In contrast, with the Navigation Guide, we are primarily concerned with *underestimating* the true effects of a chemical exposure, since in many cases population wide exposure has already occurred. 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."

Publication Bias Rating 0 no change -1 decrease quality 1 level -2 decrease quality 2 levels		**Rationale for your judgment**
Human		
Non-Human Mammalian		
Non-Human Non-Mammalian		

Category 6. Rate Factors that Can Increase Quality of Evidence

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

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 [18].

GRADE specifies 3 categories for increasing the quality of evidence [18]:

1. Large magnitude of effect. Modeling studies suggests that confounding (from non-random allocation) alone is unlikely to explain associations with a relative risk (RR) greater than 2 (or less than 0.5), and very unlikely to explain associations with an

RR greater than 5 (or less than 0.2). Thus, these are the definitions of “large magnitude of effect” used 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 modeling 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 quality 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.”

2. Dose-response gradient. 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.

3. All plausible residual confounders or biases would reduce a demonstrated effect, or suggest a spurious effect when results show no effect. GRADE provides the following example of grading up evidence when observational studies have failed to demonstrate an association. Observational studies failed to confirm an association between vaccination and autism. This lack of association occurred despite the empirically confirmed bias that parents of autistic children diagnosed after the publicity associated with the article that originally suggested this relationship would be more likely to remember their vaccine experience than parents of children diagnosed before the publicity and presumably, than parents of non-autistic children. The negative findings despite this form of recall bias suggest rating up the quality of evidence.

Large Magnitude of Effect Rating 0 no change +1 increase quality 1 level +2 increase quality 2 levels		**Rationale for your judgment**
Human		

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.

1. **Final decision on overall quality of human evidence:**

(Example: Moderate quality is upgraded 1 step to high for Xyz reason(s))
---- High
---- Moderate
---- Low
---- Very

2. **Final decision on overall quality of non-human mammalian evidence:**

(Example: High quality is downgraded 1 step to moderate for Xyz reason(s))
---- High
---- Moderate
---- Low
---- Very

3. **Final decision on overall quality of non-human non-mammalian evidence:**

(Example: High quality is downgraded 1 step to moderate for Xyz reason(s))
---- High
---- Moderate
---- Low
---- Very

III. Rate the Strength of Evidence

The evidence quality ratings will be translated into strength of evidence for each population based on a combination of four criteria: (1) Quality of body of evidence; (2) Direction of effect; (3) Confidence in effect; and (4) Other compelling attributes of the data that may influence certainty (Figures 2 and 3). These strength of evidence ratings are linked to Tables 1 and 2, below, where their meaning is defined.

IV. Combine Strength of Evidence For Human and Non-human Evidence

The final step in the process is to combine the strength of the evidence according to the chart in Figure 1. Combining the strength of evidence for human and non-human data will be produce an overall statement of toxicity, i.e., known to be toxic to fetal growth; probably toxic to fetal growth; possibly toxic to fetal growth; known to be not-toxic to fetal growth.

Table 1. Navigation Guide Criteria for Evaluating Strength of Human Evidence

Sufficient evidence of toxicity	The available evidence usually includes consistent results from well-designed, well-conducted studies, and the conclusion is unlikely to be strongly affected by the results of future studies[b]. For human evidence[a] a positive relationship is observed between exposure and outcome where chance, bias, and confounding, can be ruled out with reasonable confidence.
Limited Evidence of Toxicity	The available evidence is sufficient to determine the effects of the exposure, but confidence in the estimate is constrained by such factors as: the number, size, or quality of individual studies, the confidence in the effect, or inconsistency of findings across individual studies[b]. As more information becomes available, the observed effect could change, and this change may be large enough to alter the conclusion. For human evidence a positive relationship is observed between exposure and outcome where chance, bias, and confounding cannot be ruled out with reasonable confidence.
Inadequate Evidence of Toxicity	Studies permit no conclusion about a toxic effect. The available evidence is insufficient to assess effects of the exposure. Evidence is insufficient because of: the limited number or size of studies, low quality of individual studies, or inconsistency of findings across individual studies. More information may allow an estimation of effects.
Evidence of Lack of Toxicity	The available evidence includes consistent results from well-designed, well-conducted studies, and the conclusion is unlikely to be strongly affected by the results of future studies[b]. For human evidence more than one study showed no effect on the outcome of interest at the full range of exposure levels that humans are known to encounter, where bias and confounding can be ruled out with reasonable confidence. The conclusion is limited to the age at exposure and/or other conditions and levels of exposure studied.

[a] The Navigation Guide rates the quality and strength of evidence of human and non-human evidence streams separately as “sufficient”, “limited”, “inadequate” or “evidence of lack of toxicity” and then these two ratings are combined to produce one of five possible statements about the overall strength of the evidence of a chemical’s reproductive/developmental toxicity. The methodology is adapted from the criteria used by the International Agency for Research on Cancer (IARC) to categorize the carcinogenicity of substances [19] except as noted.

Table 2. Navigation Guide Criteria for Evaluating Strength of Non-Human Evidence

Sufficient evidence of toxicity	The available evidence usually includes consistent results from well-designed, well-conducted studies, and the conclusion is unlikely to be strongly affected by the results of future studies. For non-human[a] evidence positive association has been established through either multiple positive results or a single appropriate study[b] in a single species.
Limited evidence of toxicity	The available evidence is sufficient to determine the effects of the exposure, but confidence in the estimate is constrained by such factors as: the number, size, or quality of individual studies, the confidence in the effect, or inconsistency of findings across individual studies. As more information becomes available, the observed effect could change, and this change may be large enough to alter the conclusion. For non-human evidence the data suggest an effect, but only in a single study; or there are other important limitations in the quality of the body of evidence as specified.
Inadequate evidence of toxicity	Studies permit no conclusion about a toxic effect. The available evidence is insufficient to assess effects of the exposure. Evidence is insufficient because of: the limited number or size of studies, low quality of individual studies, or inconsistency of findings across individual studies. More information may allow an estimation of effects.
Evidence of lack of toxicity	The available evidence includes consistent results from well-designed, well-conducted studies, and the conclusion is unlikely to be strongly affected by the results of future studies. For non-human evidence data on an adequate array of endpoints from more than one study with two species showed no adverse effects at doses that were minimally toxic in terms of inducing an adverse effect. Information on pharmacokinetics, mechanisms, or known properties of the chemical class may also strengthen the evidence.[c] Adequate studies in at least two species show that the exposure is not toxic. Conclusion is limited to the species, age at exposure, and/or other conditions and levels of exposure studied.

[a] The Navigation Guide rates the quality and strength of evidence of human and non-human evidence streams separately as ‘sufficient’, ‘limited’, ‘inadequate’ or ‘evidence of lack of toxicity’ and then these 2 ratings are combined to produce one of five possible statements about the overall strength of the evidence of a chemical’s reproductive/developmental toxicity. These definitions are adapted from the criteria used by the International Agency for Research on Cancer (IARC) to categorize the carcinogenicity of substances [19] except as noted.

[b] IARC’s criteria for sufficient evidence of carcinogenicity in animals requires multiple positive results (species, studies, sexes). The Navigation Guide integrates USEPA’s minimum criteria for animal data for a reproductive or developmental hazard, i.e., data demonstrating an adverse reproductive effect in a single appropriate, well-executed study in a single test species [20]. The Navigation Guide also incorporates USEPA’s ”sufficient evidence category” which includes data that “collectively provide enough information to judge whether or not a reproductive hazard exists within the context of effect as well as dose, duration, timing, and route of exposure. This category may include both human and experimental animal evidence” [20]. The USEPA statement for developmental hazards is slightly different but includes the same relevant information regarding dose, duration, timing, etc [21].

[c] Based on minimum data requirements according to USEPA Guidelines for Reproductive Toxicity [20]

V. References

1. Higgins, J.P.T. and S. Green, *Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0,* 2011: The Cochrane Collaboration.
2. Balshem, H., et al., *GRADE guidelines: 3. Rating the quality of evidence.* J Clin Epidemiol, 2011. **64**(4): p. 401-6.
3. Viswanathan, M., et al., *Assessing the Risk of Bias of Individual Studies in Systematic Reviews of Health Care Interventions.*, in *Agency for Healthcare Research and Quality Methods Guide for Comparative Effectiveness Reviews,* 2012.
4. Eden, J., et al., *Knowing what works in health care: a roadmap for the nation*, ed. Institute of Medicine, 2008: National Academies Press.
5. Peterson, E.D., *Research methods to speed the development of better evidence-the registries example.*, in *Evidence-based medicine and the changing nature of health care: 2007 IOM annual meeting summary,* 2008, The National Academies Press: Washington, DC.
6. Halvorson, G.C., *Electronic medical records and the prospect of real time evidence development.*, in *Evidence-based medicine and the changing nature of health care: 2007 IOM annual meeting summary,* 2008, The National Academies Press: Washington, DC.
7. Guyatt, G.H., et al., *GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables.* J Clin Epidemiol, 2011. **64**(4): p. 383-94.
8. Guyatt, G.H., et al., *GRADE guidelines: 5. Rating the quality of evidence-publication bias.* J Clin Epidemiol, 2011.
9. Guyatt, G.H., et al., *GRADE guidelines: 8. Rating the quality of evidence-indirectness.* J Clin Epidemiol, 2011.
10. Hemminki, K. and P. Vineis, *Extrapolation of the evidence on teratogenicity of chemicals between humans and experimental animals: chemicals other than drugs.* Teratog Carcinog Mutagen, 1985. **5**(4): p. 251-318.
11. Nisbet, I.C.T. and N.J. Karch, *Chemical Hazards to Human Reproduction,* 1983, Park Ridge, NJ: Noyes Data Corp.
12. Kimmel, C.A., et al., *Reliability of Experimental Studies for Predicting Hazards to Human Development*, in *NCTR Technical Report for Experiment No. 6015,* 1984: Jefferson, AR.
13. Nemec, M.D., et al., *Significance, Reliability, and Interpretation of Developmental and Reproductive Toxicity Study Findings*, in *Developmental Reproductive Toxicology: A Practical Approach,* 2006, Informa Healthcare.
14. Newman, L.M., E.M. Johnson, and R.E. Staples, *Assessment of the Effectiveness of Animal Developmental Toxicity Testing for Human Safety.* Reproductive Toxicology, 1993. **7**(4): p. 359-390.
15. National Research Council (U.S.) Committee on Developmental Toxicology and National Research Council (U.S.) Commission on Life Sciences, *Scientific frontiers in developmental toxicology and risk assessment,* 2000, Washington, DC: National Academy Press. xviii, 327 p.
16. Guyatt, G.H., et al., *GRADE guidelines: 7. Rating the quality of evidence-inconsistency.* J Clin Epidemiol, 2011.
17. Guyatt, G., et al., *GRADE guidelines: 6. Rating the quality of evidence--imprecision.* J Clin Epidemiol, 2011. **64**(12): p. 1283-93.

18. Guyatt, G.H., et al., *GRADE guidelines: 9. Rating up the quality of evidence.* J Clin Epidemiol, 2011.
19. International Agency for Research on Cancer, *Preamble to the IARC Monographs (amended January 2006)*, 2006, World Health Organization: Lyon.
20. US EPA, *Guidelines for Reproductive Toxicity Risk Assessment. Fed Reg 61:56274-56322*, U.S. Environmental Protection Agency, 1996.
21. US EPA, *Guidelines for Developmental Toxicity Risk Assessment. Fed Reg 56:63798-63826*, U.S. Environmental Protection Agency, 1991.