Texas Commission on Environmental Quality Response to Public Comments Received on the July 14, 2017 on Proposed White Paper on Guidelines for Systematic Review and Evidence Integration

The public comment period for the April 2017 Proposed White Paper on Guidelines for Systematic Review and Evidence Integration ended in July 2017. The Toxicology Division (TD) received comments from Dr. Ivan Rusyn of Texas A&M University on May 1, 2017 and the American Chemical Council (ACC) on July 14, 2017. The TD of the Texas Commission on Environmental Quality (TCEQ) appreciates the efforts put forth by Dr. Rusyn and the ACC to provide technical comments on the proposed Guidelines for Systematic Review and Evidence Integration. The goal of the TD and the TCEQ is to protect human health and welfare based on the most scientifically-defensible approaches possible, and evaluation of these comments furthered that goal. A summary of the comments from the ACC and TCEQ responses are provided below. The full comments are provided in the Appendix. TCEQ responses indicate what changes, if any, were made to the White Paper in response to the comments.

Dr. Ivan Rusyn – Texas A&M University

Comment 1:

Dr. Rusyn**:**

Table 3 provides good examples of study questions and potential exclusion criteria. The granularity of the exclusion criteria is somewhat alarming as a more nuanced approach may need to be taken to each chemical or assessment. For example, many of these are too vague to be broadly applicable: "Significantly high concentrations used", "Endpoint not relevant to human health," "Endpoint not applicable to toxicity factor development," or not concurrent with the future developments in toxicology: "Study used non-mammalian animal models". I would advise against spelling out the exclusion criteria in the guidelines, rather leaving this up to the assessors making it transparent in each assessment.

TCEQ Response:

The TCEQ agrees that the determination of the inclusion/exclusion criteria is ultimately up to the toxicologists writing the Development Support Document and conducting the systematic review. Table 3 provides examples of inclusion/exclusion criteria that are often used in the development of toxicity factors. Defining one set of inclusion and exclusion criteria for all chemicals is difficult since often the criteria will be chemical- and/or purpose-specific. Therefore, as the guidance states, inclusion and exclusion criteria may be modified as needed and will be documented accordingly.

Comment 2:

Dr. Rusyn:

Step 3, page 16: "or can be created in commercially available databases such as HAWC" is incorrectly stating that HAWC is a commercial software.

TCEQ Response:

The TCEQ appreciates your comment and has corrected the text.

Comment 3:

Dr. Rusyn:

Section 4.1.3. This section does not provide clarity as to what mechanisms may be evaluated. While the complexity of the biological effects of chemicals in humans and animals is appreciable, several recent efforts have attempted to delineate and streamline these, at least for cancer (Smith et al., 2016). TCEQ may wish to incorporate these approaches into the systematic review using the search terms and examples published by the IARC Monographs program (Instructions for authors) and elsewhere (Chappell et al., 2016).

TCEQ Response:

Thank you for your feedback. The following information has been added to Section 4.1.3: "Computational systems biology toxicity pathway models must be further developed and validated to reliably distinguish non-adverse responses (or levels of responses) for in vitro endpoints (e.g., adaptive) from those that should be deemed adverse at the cellular level (e.g., produce progressive toxicity pathway perturbations sufficient to cause adverse effects *in vivo*) (TCEQ 2015). When available and appropriate, the TCEQ will use *in vitro – in vivo* extrapolation (IVIVE) tools to predict *in vivo* effects.

Mechanistic data may be used to evaluate toxicokinetics, metabolism, structure-activity relationships, susceptibility, carcinogenic mechanisms, and target-organ toxicity (IARC 2017). As stated in TCEQ (2015) Guidelines, once the POD for each key study is determined, adjustments must be made to account for differences between experimental and desired exposure durations and/or differences in anatomy and physiology in experimental animals and humans. A comprehensive biologically-based dose-response model links mechanistic determinants of chemical disposition, toxicant-target interactions, and tissue responses into an overall model of pathogenesis. The proposed stages between exposure and response include processes relating exposure to consequent tissue dose (i.e., toxicokinetics) and processes that determine response to the tissue dose (i.e., toxicodynamics). If empirical data are not available to construct a comprehensive biologically-based dose-response model for a chemical, then response can be related to exposure by incorporating and integrating as much mechanistic data as possible to allow a more accurate characterization of the pathogenic process (TCEQ 2015). When possible, the TCEQ uses verified physiologically-based pharmacokinetic (PBPK) compartmental models to

characterize pharmacokinetic (a.k.a. toxicokinetic) behavior of a chemical and to perform dosimetric adjustments (TCEQ 2015)."

Comment 4:

Dr. Rusyn:

Section 6 refers to a publication by Beck et al. 2015. This publication is listed as "submitted" and this reviewer could not find it in PubMed yet. As the whole section appears to rely on the information in this publication, TCEQ may wish to verify it public availability and content. This reviewer wasn't able to find it.

TCEQ Response:

Thank you for your comment. The reference section has been updated to reflect the current citation:

Beck, N., Wise, K., Becker, R., Dourson, M., Nancy, P., Erraguntla, N., Grant, R.L., Shirley, S., Gray, G., Farland, B., Lakind, J., Simon, T., Santos, S., Kirman, C., Lewis, R.J., Pottenger, L. (2016). Approaches for Communicating Overall Uncertainty in Hazard Assessment and Dose-Response Assessment: U.S. Environmental Protection Agency's Integrated Risk Information System (IRIS) as a Case Study. Environmental International 89-90:110-128.

Comment 5:

Dr. Rusyn:

Related to data integration is the approach recommended by the National Academies in several reports (PERC, formaldehyde, etc). Specifically, NAS has recommended that candidate toxicity values, at least for the non-cancer effects, are developed for multiple studies and then arrayed on the exposure continuum. These reports also recommended development of a range for the toxicity values, rather than one number. Similar approach has been implemented in HAWC whereby candidate tox values, uncertainty factors and other information can be visualized according to the NAS recommendations.

TCEQ Response:

The TCEQ appreciates the suggestion of the developing a range for toxicity values, rather than one number. The current systematic review guidelines are based on the 2015 TCEQ Guidelines to develop effects screening levels, reference values, and unit risk factors (RG-442). These guidelines outline identifying the study that results with the lowest POD_{HEC} for an adverse effect when multiple studies are available for toxicity factor derivation, and these guidelines will be used in conjunction with the proposed systematic review guidelines. Under the TCEQ guidelines, in some instances more than one toxicity factor has been developed for the same assessment (i.e., key and supporting toxicity factors), as determined appropriate on a chemical- and assessment-specific basis. Consequently, while not adopted as standard practice, candidate

toxicity values based on more than one study have been, and can be, developed on a limited basis as determined to be appropriate.

American Chemistry Council

Comment 1:

ACC:

2.1 Step 1 Problem Formulation and Protocol Development:

Overall, TCEQ's problem formulation and protocol development guidance is in line with best practices for the first phase of a systematic review (Rhomberg et al., 2013). The guidance should explicitly state that the problem formulation and protocol development steps may be iterative, because, in many cases, as the review progresses, new information may be found that requires changing the protocol. In addition, the guidelines should also require that changes to the review questions and protocol be documented, justified, and agreed upon by all research team members.

TCEQ Response:

The TCEQ agrees that the guidance should explicitly state that the problem formulation and protocol development steps may be altered and should be documented accordingly. Text was added to the guidelines for this purpose.

Comment 2:

ACC:

Literature Search Strategy and Study Selection:

Three minor changes could improve Step 2 of the TCEQ systematic review guidelines. First, the guidelines should explicitly specify that the literature search and study selection process is iterative, because new information is often identified in the early stages of the systematic review that may necessitate additional supplemental literature searches. Second, given that numerous reviewers are involved in any given systematic review, the guidelines should state that staff must fully document how any disagreements between them regarding the eligibility of studies for inclusion in the analysis are resolved. Third, while the specific study selection criteria prescribed in the TCEQ guidelines are generally relevant and in line with those outlined in many other frameworks, one criterion that should be updated is: "Exposure concentration is environmentally relevant" (TCEQ, 2017). TCEQ should focus on exposures encountered by the general population, but toxicity studies that include doses of chemicals that are well above concentrations of those chemicals found in the environment may still be relevant. Rather than immediately exclude these studies from further review, TCEQ should specify that further analysis of such studies be conducted to determine whether and how their results can be

extrapolated to humans, and then determine what priority to give these studies if other studies with more relevant doses are available.

TCEQ Response:

1. The TCEQ agrees that the literature review may be updated as new information becomes available and that changes made to the initial literature review should be documented accordingly. Text has been added to the Guidelines.

2. Disagreements between staff regarding the eligibility of studies for inclusion in the analysis are likely to be temporary and resolved through discussion to arrive at team consensus. The TCEQ considers this information deliberative and does not see the need to document any initial disagreements in the Systematic Review Framework.

3. The TCEQ agrees that toxicity studies that include doses of chemicals that are well above concentrations of those chemicals found in the environment may still be relevant. This example was used in the case of ethylene glycol due to the high number of studies examining intentional ingestions/poisonings and the extremely high exposures associated with them. However, the TCEQ recognizes that this is a chemical-specific case, and therefore this example of an exclusion criterion has been removed from the guidance.

Comment 3:

ACC:

Step 3: Data Extraction:

Data extraction is Step 3 of TCEQ's systematic review framework (TCEQ, 2017). In this phase of the review, studies meeting the review's inclusion criteria are critically reviewed and summarized in evidence tables, so that the investigators can identify trends in the available evidence about a chemical. TCEQ notes that these tables can be created in Microsoft Word or Excel or generated using the HAWC software. The example table provided in the TCEQ guidance is quite simple (see reproduction below). The guidelines do not include a requirement for staff toxicologists to create more detailed tables later in the review process (*e.g.*, to evaluate study quality). While these tables will be useful for quickly comparing exposures and no/lowest observed adverse effect levels (NOAELs/LOAELs) across studies, TCEQ's guidelines should require that more detailed tables be developed at this stage (and provide examples), because study quality and relevance should be considered when comparing these values. Evidence tables should also include information on study design, study size, exposure characterization and/or tested levels, the type of statistical analyses performed, and results.
Characterizing the evidence more completely during the data extraction phase of the systematic review will allow for a more transparent assessment overall and will allow others to clearly see important information from all of the studies considered without having to obtain all of the original publications.

TCEQ Response:

The TCEQ appreciates the suggestion of adding study design, study size, exposure characterization and/or tested levels, the type of statistical analyses performed, and results to the data extraction tables. The example table provided in Step 3 of the TCEQ's systematic review framework is a very simple example and was used for a chemical with very little data available. Data-rich chemicals would require more extensive and likely chemical-specific data extraction tables, and the suggestions provided here have been added to the text.

Comment 4:

ACC:

Step 4: Study Quality Analysis:

TCEQ has developed a very detailed and thorough study quality assessment system. However, several of TCEQ's quality assessment criteria reflect issues regarding the applicability and external validity (i.e., relevance) of a study, rather than its internal validity (i.e., quality). These include the criteria for "original data," "applicable route of exposure," "single route" (which is relevant to ReV development), and "health effects relevant to ReV development," in the general study quality criteria (Table 5 of TCEQ, 2017), as well as a number of other criteria presented in the realm-specific and reproductive- and developmental-specific criteria (see Table 3.2, below).

While the applicability and external validity of a study are critical factors in all systematic reviews, and particularly in those intended to derive toxicity factors, these considerations are independent of a study's internal validity, and thus, should be considered independently.

As shown in Table 3.2, TCEQ should remove questions of relevance and external validity from the quantitative scoring procedure and instead incorporate them into the inclusion and exclusion criteria and/or other phases of the analysis, as appropriate.

TCEQ Response:

The TCEQ appreciates the discussion of internal versus external validity. As stated in the guidelines, although some agencies assess the various aspects of study quality separately, the TCEQ has chosen to follow something more similar to the NTP OHAT review, which defines study quality more broadly. Due to the wide range of chemicals that the TCEQ evaluates, it is difficult to develop single set of study quality criteria that would work for every review. The examples provided in the guidelines were developed for a chemical with very little data, so external validity criteria were evaluated as part of the study quality. A review for a data-rich chemical, however, may use some of these same external validity aspects as exclusion criteria in order to more efficiently narrow down a much larger literature pool. The TCEQ guidelines are written in a way that leaves it up to the study review team to make these types of decisions based on the chemical-specific data and available literature pool.

Comment 5:

ACC:

Realm-Specific Study Quality Criteria:

The TCEQ guidelines contain a set of 16 general criteria for assessing study quality (Table 5), and additional criteria specific to reproductive and developmental studies (Table 6), epidemiology studies (Table 8), animal studies (Table 9), and mechanistic studies (Table 10) (TCEQ, 2017). The intention is to use the criteria listed in Table 5 in conjunction with the other relevant criteria relevant to each realm of evidence (e.g., consider the criteria in both Table 5 and Table 8 when evaluating epidemiology studies).

Rather than having one table of general study quality criteria and tables of realm-specific criteria for each study type, it would be better to have only tables specific to each realm of evidence, with the general study quality criteria currently in Table 5 incorporated into each of these tables, as appropriate. This is because some of the criteria currently in Table 5 are more applicable to specific realms of evidence, and some criteria cannot be readily compared across realms. For example, one general study quality criterion relates to the control of confounding (e.g., smoking and other behavioral patterns that may be associated with both exposure and disease), which is clearly applicable primarily to human studies. This approach will also keep all of the quality criteria for each type of study in one place, so that investigators do not have to reference multiple tables when reviewing a study.

In addition to some of the general criteria in Table 5 relevant to animal studies (e.g., sample size calculation, blinded study), TCEQ should also add some additional considerations to its proposed animal study scoring system. These criteria should encompass those that are used in other study quality assessments of animal evidence (e.g., OHAT and US EPA's IRIS program [US EPA, 2013, 2014; NTP, 2015a), such as the use of appropriate control animals and sufficient quality control measures (Lynch et al., 2016). A list of expanded study quality criteria for animal studies is shown in Appendix B, Table B.2.

TCEQ's discussion of evaluating the quality of epidemiology studies is thorough and useful and raises many of the salient issues that arise when reviewing such studies. In particular, Table 7 in the TCEQ guidelines is a helpful visual aid and provides a rough overview of the inherent limitations of epidemiology study design, which often make it difficult to use these studies for deriving toxicity factors. A number of other criteria could be added to this table, including some of the general study quality criteria from Table 5 (see Appendix B, Table B.3).

Finally, with regard to the in vitro study quality criteria tables, the guidelines currently only include relevance-related criteria (i.e., relevance to human exposure, ReV development, and whether a positive dose-response was observed in the study). These criteria should be considered elsewhere in the systematic review (as shown in Table 3.2). Instead, the study quality criteria for in vitro studies should be similar to those applied to animal studies, with

additional criteria specific to in vitro assays, including the validity and precision of the chosen assays (see Appendix B, Table B.4).

TCEQ Response:

The TCEQ appreciates the suggestion regarding combining the general study criteria and the realm specific criteria into a single table. As stated in previous responses, the tables and criteria presented in the guidelines are simply examples of what can be used, and study authors have the discretion to develop what works best for their specific review. Appendix A of the guidelines shows an example of how these proposed tables were used in the ethylene glycol DSD, and as this comment suggests, there is a single table for each data stream (Table 24 for human studies, Table 25 for animal studies, and Table 25 for mechanistic studies). Each of these tables incorporates the general study criteria, which the TCEQ feels are applicable to each data stream (i.e. confounding factors such as unrelated disease or early death can also occur in animal studies), and criteria specific to that realm. And all of the scoring criteria for an individual study can be found in a single table, as is seen in Tables 24-26 in the Appendix of the guidelines.

The TCEQ also appreciates the suggestions regarding additional study quality criteria, and some of these have been added to the example tables in the TCEQ Systematic Review Guidelines. The number and extent of the criteria used in a systematic review will be dependent upon chemical-specific data and the available literature, and as such these guidelines are just simple suggestions for the review authors to build upon.

Comment 6:

ACC:

<u>Interpreting Study Quality and Selecting Studies for Quantitative Risk Assessment</u>

In its guidelines for the hazard assessment phase of systematic reviews, TCEQ should use total quality scores to assign studies to general quality tiers (e.g., Tier I and Tier II). Tier I studies represent those with more strengths than limitations (i.e., the positives outweighed the negatives), and Tier II studies represent those with more limitations than strengths (i.e., the positives did not outweigh the negatives). While a researcher should not totally exclude studies based on quality for the purposes of hazard assessment (except in the case of quantitative risk assessment [QRA], as discussed below), these quality tiers allow the researcher to get a better idea of the relative quality of the (often voluminous) body of evidence for a chemical. A study being of higher quality increases confidence in its results, and thus, Tier I studies can be given more weight in the review (Lynch et al., 2016). However, a narrative qualitative assessment is needed for each study to determine how issues relating to study quality impact the interpretation of each study's results. Because the scores for each quality criteria are not necessarily equivalent, and because there could be a low score for a specific category (e.g., exposure characterization) that makes a study unusable for the purposes of QRA and toxicity factor derivation, TCEQ's systematic review guidelines should indicate that these scores are intended for hazard assessment and/or as a first pass to more generally assess study quality.

The guidelines should also provide additional detail regarding how supporting and informative studies should be factored into its qualitative assessments of hazard. Additional guidance should also be provided regarding how to select a study for ReV derivation (after study quality assessment), e.g., by considering both study quality and the appropriateness of the study data (i.e., study relevance) for calculating risk and ReVs.

Selecting epidemiology studies for QRA can be particularly challenging, and therefore, TCEQ should include additional guidance on the use of epidemiology for risk assessment. Several frameworks have developed specific guidelines on this issue, most notably, the "Guidelines to Evaluate Human Observational Studies for Quantitative Risk Assessment" (Vlaanderen et al., 2008). The Vlaanderen et al. (2008) framework includes criteria for assessing overall study quality, criteria for selecting studies suitable for QRA, and guidance for the final selection of a study (or studies) for QRA based on the study quality ranking. The set of criteria that to determine whether the study is suitable for QRA include both general quality considerations as well as those required for calculations (e.g., "is the exposure expressed on a ratio scale and specific to the agent of interest?") (see Table 3.3). TCEQ could adapt this approach for assessing epidemiology studies for inclusion in the QRA portion of the systematic review. First, investigators would assess all the identified epidemiology studies using its study quality criteria, then, from the higher-quality studies, select a study (or studies) to use for the QRA using the QRA-specific set of questions put forth by Vlaanderen et al. (2008). While some of the Vlaanderen et al. (2008) criteria overlap with the epidemiology study quality criteria provided in Appendix B, Table B.3, in the case of QRA, these criteria would be used as exclusion criteria, rather than for a general rating of overall study quality.

TCEQ Response:

Although many of the chemicals that the TCEQ evaluates lack a sufficient data pool to make assigning quality tiers a worthwhile exercise, the TCEQ agrees that in the case of a data-rich chemical, a tiered approach would be useful. The following text has been added to the guidelines to provide the study review authors some guidance on assigning study quality tiers:

“Assigning study quality tiers can be a useful tool when evaluating data-rich chemicals, especially when the data are primarily from a single stream (e.g., animal studies, human inhalation chamber studies, epidemiology studies). Total quality scores within each stream can be divided into two tiers, with Tier 1 studies having higher overall scores, suggesting more positive attributes, while Tier 2 studies with lower overall scores suggest more limitations. These tiers would not be used to exclude studies, but rather to present a better idea of the overall quality of the study in relation to other studies in the data stream. Text should be added along with the data tables to explain how the tiers were chosen and what role the aspects of study quality played in the overall selection of the key studies, especially when lower scoring studies are chosen. Studies can be identified at this step as key, supporting, and informative based on their ability to be used in the derivation of a toxicity factor. Since the end goal of the review is the derivation of a toxicity factor, studies that have low quality scores but are amenable to this process may be selected over studies that score higher but that lack the

necessary detailed to derive a POD. Supporting studies may be used to support the use of an MOA, a route of exposure, or an exposure concentration (e.g., POD), while an informative study may have information on MOA or the critical effect, but lacks any exposure information."

Epidemiological studies do present a unique challenge in both the systematic review process and in the derivation of toxicity factors. As mentioned in Comment 5, the TCEQ appreciates the suggestions regarding additional study quality criteria, and some of these have been added to the example tables in the TCEQ Systematic Review Guidelines. As far as guidance on the use of epidemiological studies in risk assessment, Chapter 7 of the TCEQ Guidelines to Develop Toxicity Factors, titled "Hazard Characterization and Exposure-Response Assessment Using Epidemiology Data," will be used in conjunction with the Systematic Review Guidelines.

Comment 7:

ACC:

Step 5: Evidence Integration:

Step 5 of the TCEQ systematic review framework (evidence integration) provides a very brief discussion of the importance of evidence integration in a systematic review (TCEQ, 2017). Evidence integration is the phase of the systematic review in which the assessor considers the results of all realms of evidence to determine where they agree and where they disagree, as well as to identify any critical data gaps in the body of evidence. In other words, evidence integration allows each realm of evidence to inform the interpretation of the evidence from the other realms. The guidelines reference several publications that propose best practices for evidence integration (Rhomberg et al., 2013; NRC, 2014; Rooney et al., 2014), however, the guidance provided in these publications differs, and the methods presented in these documents are not necessarily reflected in the draft TCEQ guidelines. This section of TCEQ's guidelines should be more prescriptive, and should include a broader discussion of what evidence integration entails as well as more detailed guidelines on how to accomplish this task.

The lack of explicit guidance on evidence integration appears to be the result of TCEQ's concern that, "given that chemicals differ in amount and quality of each data stream, prescribing universally applicable rules for evidence integration is difficult" (TCEQ, 2017). Although it is challenging to make general guidelines applicable to all assessments, TCEQ's guidelines should be expanded to provide more explicit guidance on how to conduct a thorough evidence integration, while remaining flexible enough to allow for modifications, as needed, for different types of datasets and chemicals (Goodman et al., 2013).

The guidelines should provide a general "baseline" framework for performing the evidence integration step of a systematic review. While there are several options for methods of evidence integration, we recommend the method presented by Goodman et al. (2013, 2015), which incorporates many of the best practices for evidence integration assembled by others, particularly Rhomberg et al. (2013). The basis of the evidence integration step ("Phase 3") of

the Goodman et al. framework are the Bradford Hill postulates, as well as additional considerations of bias and confounding (see Table 4.1).

TCEQ Response:

The TCEQ agrees that the guidelines should provide a general "baseline" framework as part of the evidence integration step and has adapted the basic steps from Goodman et al. (2013, 2015). The following information has been added to the text:

"As a general guideline, the following steps should be considered during the evidence integration step (adapted from Goodman et al., 2013, 2015):

- Integrate data across all realms of evidence (*e.g*., animal, epidemiology, and mechanistic;
- Assess all data;
- Assign less weight to the results of studies that are of lower quality;
- Incorporate peer and public comments;
- Formulate conclusions.

The TCEQ provides evidence integration tables to summarize the available data for toxicity factor derivation in its DSDs. Information on the type of POD (e.g., free-standing NOAEL, minimal LOAEL) or exposure method (e.g., single dose, data amenable to benchmark dose modeling) are provided as a means to measure a study's strength for toxicity factor development. Some additional considerations when developing evidence integration tables include strength and consistency of association, biological plausibility and dose-response, coherence across data streams, and biological and clinical relevance (Goodman et al., 2013, 2015). These tables are also indicative of the considerations behind designating studies as key, supporting, or informative (See section A.5). Examples of evidence integration tables used for the ethylene glycol DSD can be found in Tables 27-29 in the Appendix. Due to the variety of chemicals and toxicity factors that are developed, these tables may be altered by TCEQ as needed."

The TCEQ also appreciates the suggestion for more explicit evidence integration guidance, and will keep this suggestion in mind during future reviews. The guidance is currently being tested in conjunction with several DSDs, and may ultimately be updated, revised, and refined to reflect more prescriptive guidance as the comment suggests.

Comment 8:

ACC:

<u>Step 6: Confidence Rating for the Body of Evidence:</u>

Step 6 in TCEQ's systematic review framework is to "rate the confidence in the body of evidence" (TCEQ 2017). The guidelines provide a helpful general discussion of some of the key considerations when integrating evidence within and across realms (e.g., high-quality studies

provide greater confidence and lower uncertainty that the key study findings accurately depict the relationship between the exposure and effect of interest). There are no specific instructions, however, for rating the confidence in the body of evidence.

The guidelines should indicate that, after working through all of the considerations in Table 4.1 above, the risk assessor should formulate an overall hazard/causality conclusion in a narrative discussion, with consideration of study quality, uncertainty, variability, and sensitivity analyses, and how animal, human, MoA, dose-response relationships, and all relevant data are integrated as part of the conclusion. Working through a narrative assessment that attempts to reconcile differences across the realms of evidence assessed will likely be less burdensome than the more formal approaches to rating confidence (as presented by TCEQ and in other frameworks, such as OHAT [NTP, 2015a]) and will lead to more interpretable results, rather than just presenting the derived toxicity factor without sufficient explanation. This narrative should consider the overall quality and uncertainty in each of the realms of evidence and bring them together to form a causal conclusion regarding the hazard of the chemical under evaluation. In this step of the systematic review, TCEQ should require that investigators summarize uncertainties and data gaps across all the realms of evidence and provide a discussion of whether the data are sufficient for performing a dose response analysis and toxicity factor derivation.

Finally, a table or figure that illustrates the evidence integration across the different realms of evidence may be helpful for presenting the results of this step clearly and coherently. See, for example, the set of evidence tables provided in US EPA's recent Toxicological Review of Benzo(a)pyrene (US EPA, 2017).

TCEQ Response:

The TCEQ chose the Beck et al. (2016) uncertainty assessment tool because it is a clear and concise method to portray the overall uncertainty and provide a rapid visualization of the confidence scoring for the overall toxicity assessment. The use of a single table to display the overall assessment allows the user to quickly view the overall confidence in the assessment without having to search through text for specific points. Although the Beck et al. (2016) is the primary tool used in the assessment, a narrative may be added in addition to the table to strengthen the assessment, and this has been added to the text:

“In addition to the confidence table, narrative discussion of the overall uncertainty may be added to strengthen the assessment, including details on study quality, existing data gaps, uncertainty, variability, and sensitivity analyses, and how animal, human, MoA, dose-response relationships, and all relevant data are integrated as part of the conclusion.”

In regard to discussion on whether the data are sufficient for performing a dose-response analysis and toxicity factor derivation, the TCEQ is responsible for developing toxicity factors for all chemicals used and produced in the state of Texas, whether or not the database would be deemed sufficiently complete by some other regulatory agencies to develop similar values (e.g., USEPA’s IRIS program, ATSDR MRLs). A DSD is generally not developed unless there is the

requisite amount of information necessary for performing a dose-response analysis and toxicity factor derivation under TCEQ guidelines, something determined upstream of actually writing the DSD. In deriving the toxicity factors, the agency appropriately considers associated uncertainty and data gaps, such as in the narrative justification for the UF_D.

Comment 9:

ACC:

Rating Confidence in the Toxicity Factor:

Although Step 6 of TCEQ's framework for conducting systematic reviews indicate that the body of evidence is being rated, in actuality, the guidelines have the investigator rate the confidence in the derived toxicity factor. Therefore, Step 6 should be renamed accordingly. However, the guidance it provides on rating the confidence in the derived toxicity factor is explicit and useful. Specifically, TCEQ points to tables in a publication by Beck et al. (2016, Tables 3 and 4). Beck et al. (2016) include 10 elements for toxicity assessment confidence scoring, which include determining whether the assessment uses a systematic review approach, and others elements that involve different steps in the process, focused on the confidence in the relevance and quantitative derivation of the toxicity value. The Beck et al. (2016) approach to assessing confidence and uncertainty in toxicity assessments and, more specifically, toxicity factors, is scientifically sound and will be useful for TCEQ and other agencies to use as a final step to rate their overall confidence in their toxicity assessments.

TCEQ Response:

The TCEQ appreciates your comment and has altered the title of Step 6 to “Rate the Confidence in the Toxicity Assessment”.

Comment 10:

ACC:

Evidence Integration for Ethylene Glycol (TCEQ Appendix A.5)

Examples of completed TCEQ systematic reviews will be critical for clearly illustrating how TCEQ's framework should be implemented. Appendix A.5 (i.e., the ethylene glycol DSD) of the guidelines provides examples of the current evidence integration tables for each realm of evidence (e.g., animal, human, and mechanistic studies). These tables are useful illustrations, although we would still suggest updating these tables according to the recommendations outlined in Section 3. In addition, the narrative of the assessment and, specifically, the reasoning behind these tables is almost entirely missing. The written portion of the assessment should be expanded to include some of the key features described in the above comments. Most notably, as discussed above, the evidence integration step of the systematic review needs to go beyond just listing the evidence for each realm of investigation and should be explicit about how these realms of evidence are brought together to form conclusions. Further, TCEQ

should clearly discuss how it decided whether the evidence overall was sufficient for deriving quantitative toxicity values for ethylene glycol.

Further developing the example systematic review for ethylene glycol by incorporating the suggestions made in these comments will aid in clearly demonstrating how TCEQ staff should work through systematic reviews to arrive at scientifically sound, transparent assessments that can be reproduced by other risk assessors. An expanded example assessment will also allow others to see what a TCEQ DSD that follows the new guidelines looks like.

TCEQ Response:

Thank you for the suggestion. The TCEQ has added additional language to the evidence integration step (See Comment 7).

Attachments: Comments received from Dr. Ivan Rusyn of Texas A&M University on May 1, 2017 and the American Chemical Council (ACC) on July 14, 2017

From: Ivan Rusyn [mailto:ivan.rusyn@gmail.com]
Sent: Monday, May 01, 2017 11:01 AM
To: TOX <TOX@tceq.texas.gov>
Subject: Public comment on TCEQ The Guidelines for Systematic Review and Evidence Integration

These guidelines are a very important component of human health assessments and will assist TCEQ in its mission to protect human health and the environment. Overall, the document is concise, clear and straightforward. Appropriate points are covered in sufficient detail. The following are several comments that TCEQ may wish to consider as they proceed to finalize these guidelines:

Table 3 provides good examples of study questions and potential exclusion criteria. The granularity of the exclusion criteria is somewhat alarming as a more nuanced approach may need to be taken to each chemical or assessment. For example, many of these are too vague to be broadly applicable: "Significantly high concentrations used", "Endpoint not relevant to human health," "Endpoint not applicable to toxicity factor development," or not concurrent with the future developments in toxicology: "Study used non-mammalian animal models". I would advise against spelling out the exclusion criteria in the guidelines, rather leaving this up to the assessors making it transparent in each assessment.

Step 3, page 16: "or can be created in commercially available databases such as HAWC" is incorrectly stating that HAWC is a commercial software.

Section 4.1.3. This section does not provide clarity as to what mechanisms may be evaluated. While the complexity of the biological effects of chemicals in humans and animals is appreciable, several recent efforts have attempted to delineate and streamline these, at least for cancer (Smith et al, 2016). TCEQ may wish to incorporate these approaches into the systematic review using the search terms and examples published by the IARC Monographs program (Instructions for authors) and elsewhere (Chappell et al., 2016).

Section 6 refers to a publication by Beck et al 2015. This publication is listed as "submitted" and this reviewer could not find it in PubMed yet. AS the whole section appears to rely on the information in this publication, TCEQ may wish to verify it public availability and content. This reviewer wasn't able to find it.

Related to data integration is the approach recommended by the National Academies in several reports (PERC, formaldehyde, etc). Specifically, NAS has recommended that candidate toxicity values, at least for the non-cancer effects, are developed for multiple studies and then arrayed on the exposure continuum. These reports also recommended development of a range for the toxicity values, rather than one number. Similar approach has been implemented in HAWC whereby candidate tox values, uncertainty factors and other information can be visualized according to the NAS recommendations.

Sincerely,
Ivan Rusyn

MICHAEL P. WALLS
VICE PRESIDENT
REGULATORY & TECHNICAL AFFAIRS

July 14, 2017

Toxicology Division, MC 168
Texas Commission on Environmental Quality
P.O. Box 13087
Austin, TX 78711-3087
Submitted via e-mail to: tox@tceq.texas.gov

RE: Comments on the Texas Commission on Environmental Quality's White Paper on Guidelines for Systematic Review and Evidence Integration, April 13, 2017

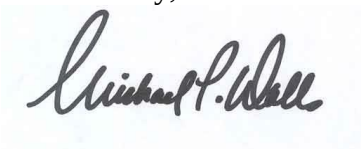
Dear Sir or Madam:

I am pleased to submit the attached comments of the American Chemistry Council (ACC) on the Texas Commission for Environmental Quality's Guidelines for Systematic Review and Evidence Integration.

In general, ACC supports the Guidelines as drafted. In our comments, ACC provides several recommendations for strengthening systematic review under this framework.

If you have any questions regarding ACC's comments, please contact me at mike_walls@americanchemistry.com, or at 202 249 6400.

Sincerely,



Before the
Texas Commission on Environmental Quality

Comments of the American Chemistry Council on the TCEQ Guidelines for Systematic Review and Evidence Integration

Michael P. Walls
Vice President
Regulatory & Technical Affairs
American Chemistry Council
700 2nd Street, N.E.
Washington, D.C.
July 14, 2017

Table of Contents

Page

1 Introduction 1

2 Problem Formulation, Literature Searches, and Study Selection 2
- 2.1 Step 1: Problem Formulation and Protocol Development 2
- 2.2 Step 2: Literature Search Strategy and Study Selection 3

3 Data Extraction and Study Quality Analysis 4
- 3.1 Step 3: Data Extraction 4
- 3.2 Step 4: Study Quality Analysis 4
  - 3.2.1 Study Relevance *vs*. Quality 5
  - 3.2.2 Realm-specific Study Quality Criteria 6
  - 3.2.3 Interpreting Study Quality Scores and Selecting Studies for Quantitative Risk Assessment 6

4 Evidence Integration and Confidence Rating 9
- 4.1 Step 5: Evidence Integration 9
- 4.2 Step 6: Confidence Rating for the Body of Evidence 10
- 4.3 Rating Confidence in the Toxicity Factor 11
- 4.4 Evidence Integration for Ethylene Glycol (TCEQ Appendix A.5) 11

References 13

Appendix A Suggested References for Systematic Review and Evidence Integration
Appendix B Suggested Tables and Study Quality Criteria

List of Tables

Table 3.1 Example TCEQ Data Extraction Table

Table 3.2 Study Relevance Criteria Currently Classified as Study Quality Criteria in the TCEQ Guidelines

Table 3.3 Criteria for Selecting Epidemiology Studies for Quantitative Risk Assessments

Table 4.1 Considerations for Evidence Integration

List of Figures

Figure 2.1 Systematic Review Framework

Abbreviations

DSD	Development Support Document
ESL	Effect Screening Level
GLP	Good Laboratory Practice
HAWC	Health Assessment Workspace Collaboration
IRIS	Integrated Risk Information System
LOAEL	Lowest Observed Adverse Effect Level
MoA	Mode of Action
NOAEL	No Observed Adverse Effect Level
NTP	National Toxicology Program
OHAT	Office of Health Assessment and Translation
PECO	Population, Exposure, Comparator/Control, and Outcome
QRA	Quantitative Risk Assessment
ReV	Reference Value
RfD	Reference Dose
RoB	Risk of Bias
SF	Slope Factor
SOP	Standard Operating Procedure
TCEQ	Texas Commission on Environmental Quality
URF	Unit Risk Factor
US EPA	US Environmental Protection Agency
WoE	Weight of Evidence

1 Introduction

The Texas Commission on Environmental Quality (TCEQ) has been in the process of updating and improving its approach to conducting systematic reviews for a number of years. In 2014, it published "TCEQ Recommendations for Systematic Review and Evidence Integration" (TCEQ, 2014), a position paper that provides general information on the key steps of a systematic review. On April 13, 2017, TCEQ published its draft "Guidelines for Systematic Review and Evidence Integration" (TCEQ, 2017). These guidelines provide a more comprehensive framework for TCEQ to use in developing chemical-specific toxicity factors, including chemical-specific reference values (ReVs) such as unit risk factors (URFs), reference doses (RfDs), slope factors (SFs), and effect screening levels (ESLs). The 2017 guidelines are intended to supplement TCEQ's existing published regulatory guidance for deriving toxicity factors (TCEQ, 2012). TCEQ's new systematic review framework incorporates concepts from several other agencies' systematic review frameworks, including those developed by the US Environmental Protection Agency (US EPA) Integrated Risk Information System (IRIS) and the National Toxicology Program's (NTP) Office of Health Assessment and Translation (OHAT) (US EPA, 2013, 2014; NTP, 2015a).[1]

The new guidelines are much more detailed than the previous position paper in that they provide more explicit step-by-step guidance on conducting a systematic review (*i.e.*, instructions for how to perform each of six steps), additional recommendations for best practices (*e.g.*, requiring an *a priori* review protocol), and examples (*e.g.*, a case study of ethylene glycol). These guidelines will help facilitate more scientifically sound and transparent toxicology reviews both within and outside of TCEQ. Herein, we provide recommendations that could further strengthen the TCEQ systematic review framework.

In general, ACC supports the guidelines as drafted. In these comments, we provide recommendations that could further strengthen the TCEQ systematic review framework.

ACC's comments were prepared by Gradient (20 University Road, Cambridge, MA 02138) under contract to the Council, and reflect the positions of the American Chemistry Council.

[1] The TCEQ guidelines cite the peer-reviewed summary of the OHAT guidance by Rooney *et al.* (2014); however, more details can be found in the OHAT handbook (NTP, 2015).

2 Problem Formulation, Literature Searches, and Study Selection

2.1 Step 1: Problem Formulation and Protocol Development

Step 1 in TCEQ's systematic review framework is problem formulation (TCEQ, 2017). The problem formulation phase of a systematic review is critical to determining key factors and potential challenges from the outset. The problem formulation phase includes a review of chemical and physical properties, dose-response data, critical effect(s) (*i.e.*, endpoints occurring at the lowest exposures), issues of route-specific toxicity, and, to the extent possible, the most likely mode of action (MoA) for the critical endpoint(s). TCEQ's new systematic review guidelines provide concise, yet thorough, guidance on the problem formulation phase of systematic review. It states that the problem formulation phase should be clearly documented and formulated around a PECO statement (*i.e.*, a statement about the populations, exposure, comparator/control, and outcomes of interest), which is consistent with NTP's OHAT framework (NTP, 2015b). The "output" of the problem formulation step is a set of questions that identify the important concepts relevant to all phases of the systematic review. The TCEQ guidelines also explicitly requires staff toxicologists to develop a written protocol at this stage of the process. While it is expected that the systematic review protocol will only be general, it should include the major critical phases of a systematic review (see Figure 2.1) as well as the steps required to calculate a toxicity factor from the information identified in the review.

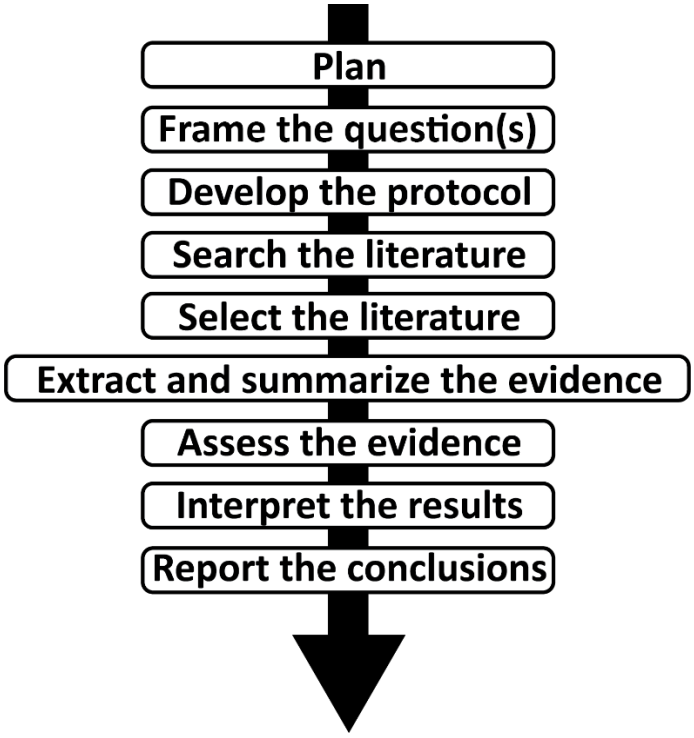


Figure 2.1 Systematic Review Framework. Adapted from Hoffman *et al*. (2017).

Overall, TCEQ's problem formulation and protocol development guidance is in line with best practices for the first phase of a systematic review (Rhomberg *et al*., 2013). The guidance should explicitly state that the problem formulation and protocol development steps may be iterative, because, in many cases, as the review progresses, new information may be found that requires changing the protocol. In addition, the guidelines should also require that changes to the review questions and protocol be documented, justified, and agreed upon by all research team members.

Although TCEQ cannot create a "one-size-fits-all" protocol that can be applied to all the chemicals it will review, the generic review protocol provided in the TCEQ guidance could be considered a general standard operating procedure (SOP) for conducting systematic reviews, similar to Good Laboratory Practice (GLP) standards. TCEQ can then require that staff expand upon the general SOP to create chemical-specific protocols for each development support document (DSD) or systematic review. A chemical-specific protocol should include the planned literature search strategy, inclusion/exclusion criteria, process for data extraction, criteria for evaluating study quality, potential analyses (*e.g*., if it is known that epidemiology evidence predominates, and a meta-analysis will be performed), and any confidence rating system (*i.e*., method for determining data gaps, limitations, and uncertainties in the evidence and the overall systematic review). Developing a detailed *a priori* protocol for the systematic review will limit potential biases in the review and help ensure that its results can be reproduced by others.

2.2 Step 2: Literature Search Strategy and Study Selection

TCEQ's guidelines on the literature search strategy and study selection processes (Step 2 of its framework; TCEQ, 2017) are thorough, clear, and consistent with the recommendations of other agencies and researchers (*e.g*., NTP, 2015a; Hoffman *et al*., 2017). The guidance indicates that the TCEQ Toxicology Department should announce the scoping process for a particular chemical using its email listserve and solicit information on relevant toxicological information about that chemical. Then, toxicologists must conduct and maintain a record of the literature searches performed during the systematic review using several databases. TCEQ also states that OHAT's Health Assessment Workspace Collaboration (HAWC) software should be used to compile literature and document the decision-making process when moving from the initial literature search to data extraction. This level of documentation, which is often overlooked in systematic reviews, is critical for analyses to be transparent and reproducible. TCEQ's guidelines note that at least two people should review each study identified in the literature review for relevance, and the HAWC software can also be used to track these reviews more easily.

Three minor changes could improve Step 2 of the TCEQ systematic review guidelines. First, the guidelines should explicitly specify that the literature search and study selection process is iterative, because new information is often identified in the early stages of the systematic review that may necessitate additional supplemental literature searches. Second, given that numerous reviewers are involved in any given systematic review, the guidelines should state that staff must fully document how any disagreements between them regarding the eligibility of studies for inclusion in the analysis are resolved. Third, while the specific study selection criteria prescribed in the TCEQ guidelines are generally relevant and in line with those outlined in many other frameworks, one criterion that should be updated is: "Exposure concentration is environmentally relevant" (TCEQ, 2017). TCEQ should focus on exposures encountered by the general population, but toxicity studies that include doses of chemicals that are well above concentrations of those chemicals found in the environment may still be relevant. Rather than immediately exclude these studies from further review, TCEQ should specify that further analysis of such studies be conducted to determine whether and how their results can be extrapolated to humans, and then determine what priority to give these studies if other studies with more relevant doses are available.

3 Data Extraction and Study Quality Analysis

3.1 Step 3: Data Extraction

Data extraction is Step 3 of TCEQ's systematic review framework (TCEQ, 2017). In this phase of the review, studies meeting the review's inclusion criteria are critically reviewed and summarized in evidence tables, so that the investigators can identify trends in the available evidence about a chemical. TCEQ notes that these tables can be created in Microsoft Word or Excel or generated using the HAWC software. The example table provided in the TCEQ guidance is quite simple (see reproduction below). The guidelines do not include a requirement for staff toxicologists to create more detailed tables later in the review process (*e.g.*, to evaluate study quality).

Table 3.1 Example TCEQ Data Extraction Table

Reference	Species/n/ Sex	Exposure Concentration	Exposure Duration	NOAEL	LOAEL	Notes
Smith *et al*. (1973)	Humans/10/ males	0, 50, 100 ppm	6 hours	50 ppm	100 ppm	Respiratory irritation in 9/10 volunteers

Notes:
LOAEL = Lowest Observed Adverse Effect Level; NOAEL = No Observed Adverse Effect Level; ppm = Parts Per Million; TCEQ = Texas Commission on Environmental Quality.
Reproduced from TCEQ (2017).

While these tables will be useful for quickly comparing exposures and no/lowest observed adverse effect levels (NOAELs/LOAELs) across studies, TCEQ's guidelines should require that more detailed tables be developed at this stage (and provide examples), because study quality and relevance should be considered when comparing these values. Evidence tables should also include information on study design, study size, exposure characterization and/or tested levels, the type of statistical analyses performed, and results. Characterizing the evidence more completely during the data extraction phase of the systematic review will allow for a more transparent assessment overall and will allow others to clearly see important information from all of the studies considered without having to obtain all of the original publications.

The evidence tables can be attached to a DSD as supplemental material, so as not to interrupt the flow of the narrative discussion. For examples of these types of tables, see recent OHAT evaluations of perfluorinated chemicals (for example, Tables 10 and 11 in NTP, 2016) or recent peer-reviewed articles on particulate air pollution (*e.g.*, Supplementary Tables 5-7 in Lynch *et al*., 2016). We have also provided an example of an expanded evidence table in Appendix B, Table B.1.

3.2 Step 4: Study Quality Analysis

After extracting the data from the identified studies into evidence tables, TCEQ's systematic review framework directs the investigators to assess the quality of these studies (Step 4 of the framework; TCEQ, 2017). An assessment of study quality evaluates the extent to which a study's researchers conducted their research to the highest possible standards, how thoroughly they considered and attempted to control study design characteristics that introduce systematic error (*i.e.*, internal validity), and whether they provided complete reporting of methods and results (NRC, 2014). Such an analysis is a key element of the systematic

review process, and developing a study quality evaluation system that can be applied consistently and objectively can prove challenging. However, there are numerous available systems that can be adapted and applied, several of which include considerations for all or most realms of evidence used in systematic review (*i.e.*, animal, human, and mechanistic studies), including the NTP OHAT and US EPA IRIS risk of bias (RoB) frameworks (US EPA, 2013, 2014; NTP, 2015a).

3.2.1 Study Relevance *vs.* Quality

TCEQ has developed a very detailed and thorough study quality assessment system. However, several of TCEQ's quality assessment criteria reflect issues regarding the applicability and external validity (*i.e.*, relevance) of a study, rather than its internal validity (*i.e.*, quality). These include the criteria for "original data," "applicable route of exposure," "single route" (which is relevant to ReV development), and "health effects relevant to ReV development," in the general study quality criteria (Table 5 of TCEQ, 2017), as well as a number of other criteria presented in the realm-specific and reproductive- and developmental-specific criteria (see Table 3.2, below).

While the applicability and external validity of a study are critical factors in all systematic reviews, and particularly in those intended to derive toxicity factors, these considerations are independent of a study's internal validity, and thus, should be considered independently.

As shown in Table 3.2, TCEQ should remove questions of relevance and external validity from the quantitative scoring procedure and instead incorporate them into the inclusion and exclusion criteria and/or other phases of the analysis, as appropriate.

Table 3.2 Study Relevance Criteria Currently Classified as Study Quality Criteria in the TCEQ Guidelines

TCEQ Study Quality Category	Proposed Phase of Analysis for Consideration
General Criteria	
Original data	Hazard assessment (exclude from study quality review) and/or Criteria for selection of critical study
Applicable route of exposure	Criteria for selection of critical study
Single route	Criteria for selection of critical study
Range of doses/exposures	Criteria for selection of critical study
Health effects relevant to ReV development	Exclusion criteria
Developmental and Reproductive Effects	
Critical window for effects	Criteria for selection of critical study
Maternal and fetal toxicity	Criteria for selection of critical study
Human Studies	
Study results consistent with other available evidence	Confidence in body of evidence/evidence integration
Animal Studies	
Multiple species	Confidence in body of evidence/evidence integration (single species in one study is not an indicator of quality)
Exposure regimes (repeated *vs*. continuous)	Criteria for selection of critical study
Concentration relevant to human exposure	Evidence integration and/or Criteria for selection of critical study
Dose applicable to ReV Development	Criteria for selection of critical study
[Presence of] dose-response relationship	Confidence in body of evidence/evidence integration and/or Criteria for selection of critical study

Notes:
ReV = Reference Value; TCEQ = Texas Commission on Environmental Quality.

3.2.2 Realm-specific Study Quality Criteria

The TCEQ guidelines contain a set of 16 general criteria for assessing study quality (Table 5), and additional criteria specific to reproductive and developmental studies (Table 6), epidemiology studies (Table 8), animal studies (Table 9), and mechanistic studies (Table 10) (TCEQ, 2017). The intention is to use the criteria listed in Table 5 in conjunction with the other relevant criteria relevant to each realm of evidence (*e.g.*, consider the criteria in both Table 5 and Table 8 when evaluating epidemiology studies).

Rather than having one table of general study quality criteria and tables of realm-specific criteria for each study type, it would be better to have only tables specific to each realm of evidence, with the general study quality criteria currently in Table 5 incorporated into each of these tables, as appropriate. This is because some of the criteria currently in Table 5 are more applicable to specific realms of evidence, and some criteria cannot be readily compared across realms. For example, one general study quality criterion relates to the control of confounding (*e.g.*, smoking and other behavioral patterns that may be associated with both exposure and disease), which is clearly applicable primarily to human studies. This approach will also keep all of the quality criteria for each type of study in one place, so that investigators do not have to reference multiple tables when reviewing a study.

In addition to some of the general criteria in Table 5 relevant to animal studies (*e.g.*, sample size calculation, blinded study), TCEQ should also add some additional considerations to its proposed animal study scoring system. These criteria should encompass those that are used in other study quality assessments of animal evidence (*e.g.*, OHAT and US EPA's IRIS program [US EPA, 2013, 2014; NTP, 2015a), such as the use of appropriate control animals and sufficient quality control measures (Lynch *et al.*, 2016). A list of expanded study quality criteria for animal studies is shown in Appendix B, Table B.2.

TCEQ's discussion of evaluating the quality of epidemiology studies is thorough and useful and raises many of the salient issues that arise when reviewing such studies. In particular, Table 7 in the TCEQ guidelines is a helpful visual aid and provides a rough overview of the inherent limitations of epidemiology study design, which often make it difficult to use these studies for deriving toxicity factors. A number of other criteria could be added to this table, including some of the general study quality criteria from Table 5 (see Appendix B, Table B.3).

Finally, with regard to the *in vitro* study quality criteria tables, the guidelines currently only include relevance-related criteria (*i.e.*, relevance to human exposure, ReV development, and whether a positive dose-response was observed in the study). These criteria should be considered elsewhere in the systematic review (as shown in Table 3.2). Instead, the study quality criteria for *in vitro* studies should be similar to those applied to animal studies, with additional criteria specific to *in vitro* assays, including the validity and precision of the chosen assays (see Appendix B, Table B.4).

3.2.3 Interpreting Study Quality Scores and Selecting Studies for Quantitative Risk Assessment

Although not discussed in the main text of the TCEQ guidelines, brief guidance on interpreting study quality scores is presented in the ethylene glycol systematic review example provided in the Appendix of the guidelines (TCEQ, 2017). TCEQ indicates that study quality scores are to be summed for each study in a realm and compared across studies within that realm of evidence, but not across realms. In Tables 24-26 in the guidelines, TCEQ provides the study's scores for each criterion and the study's total score (*i.e.*, when all the criteria scores are summed), then indicates whether a study was selected as a "key," "supporting," or "informative" study (TCEQ, 2017). However, the guidelines do not discuss which scores/ranges of scores

correspond to each of these three classifications. For instance, in Table 24 in the guidelines, the epidemiology studies by Bond (1985) and Wills (1974) were both given a total score of 9, but Wills (1974) was selected as the key study for toxicity factor derivation, while Bond (1985) was judged to be only an informative study (TCEQ, 2017).

In its guidelines for the hazard assessment phase of systematic reviews, TCEQ should use total quality scores to assign studies to general quality tiers (*e.g.*, Tier I and Tier II). Tier I studies represent those with more strengths than limitations (*i.e.*, the positives outweighed the negatives), and Tier II studies represent those with more limitations than strengths (*i.e.*, the positives did not outweigh the negatives). While a researcher should not totally exclude studies based on quality for the purposes of hazard assessment (except in the case of quantitative risk assessment [QRA], as discussed below), these quality tiers allow the researcher to get a better idea of the relative quality of the (often voluminous) body of evidence for a chemical. A study being of higher quality increases confidence in its results, and thus, Tier I studies can be given more weight in the review (Lynch *et al*., 2016). However, a narrative qualitative assessment is needed for each study to determine how issues relating to study quality impact the interpretation of each study's results.

Because the scores for each quality criteria are not necessarily equivalent, and because there could be a low score for a specific category (*e.g.*, exposure characterization) that makes a study unusable for the purposes of QRA and toxicity factor derivation, TCEQ's systematic review guidelines should indicate that these scores are intended for hazard assessment and/or as a first pass to more generally assess study quality. The guidelines should also provide additional detail regarding how supporting and informative studies should be factored into its qualitative assessments of hazard. Additional guidance should also be provided regarding how to select a study for ReV derivation (after study quality assessment), *e.g.*, by considering both study quality and the appropriateness of the study data (*i.e.*, study relevance) for calculating risk and ReVs.

Selecting epidemiology studies for QRA can be particularly challenging, and therefore, TCEQ should include additional guidance on the use of epidemiology for risk assessment. Several frameworks have developed specific guidelines on this issue, most notably, the "Guidelines to Evaluate Human Observational Studies for Quantitative Risk Assessment" (Vlaanderen *et al*., 2008). The Vlaanderen *et al*. (2008) framework includes criteria for assessing overall study quality, criteria for selecting studies suitable for QRA, and guidance for the final selection of a study (or studies) for QRA based on the study quality ranking. The set of criteria that to determine whether the study is suitable for QRA include both general quality considerations as well as those required for calculations (*e.g.*, "is the exposure expressed on a ratio scale and specific to the agent of interest?") (see Table 3.3). TCEQ could adapt this approach for assessing epidemiology studies for inclusion in the QRA portion of the systematic review. First, investigators would assess all the identified epidemiology studies using its study quality criteria, then, from the higher-quality studies, select a study (or studies) to use for the QRA using the QRA-specific set of questions put forth by Vlaanderen *et al*. (2008). While some of the Vlaanderen *et al*. (2008) criteria overlap with the epidemiology study quality criteria provided in Appendix B, Table B.3, in the case of QRA, these criteria would be used as exclusion criteria, rather than for a general rating of overall study quality.

Table 3.3 Criteria for Selecting Epidemiology Studies for Quantitative Risk Assessments[a]

Evaluation Criteria	Relevant Study Type		
	CC	COH	CR
Exposure route relevant to ReV development	x	x	x
Range of exposures relevant to general population	x	x	x
Exposure is expressed on a ratio scale and specific for the agent of interest	x	x	x
Criteria for inclusion of subjects are described with sufficient detail	x	x	x
The assessment of the health effect was performed according to standard practice	x	x	x
Relevant potential strong confounding factors were considered in the study design	x	x	x
Adequate response rate[b]	x	x	x
Loss to follow-up sufficiently minimized[b]		x	
Minimum follow-up time achieved[b]		x	
Exposure measurements sufficiently representative of the exposure of interest (*e.g.*, proxy exposure is highly correlated to the exposure of interest)	x	x	x
Exposure assessment blinded	x	x	x
Health outcome assessment blinded	x	x	x

Notes:
CC = Case Control; COH = Cohort; CR = Cross-sectional; QRA = Quantitative Risk Assessment.
Adapted from Vlaanderen *et al.* (2008).
(a) All studies carried forth for consideration in the QRA should meet these inclusion criteria.
(b) Risk assessors should *a priori* define minimum requirements for inclusion in QRA for these categories (*e.g.*, acceptable levels of loss to follow-up).

4 Evidence Integration and Confidence Rating

4.1 Step 5: Evidence Integration

Step 5 of the TCEQ systematic review framework (evidence integration) provides a very brief discussion of the importance of evidence integration in a systematic review (TCEQ, 2017). Evidence integration is the phase of the systematic review in which the assessor considers the results of all realms of evidence to determine where they agree and where they disagree, as well as to identify any critical data gaps in the body of evidence. In other words, evidence integration allows each realm of evidence to inform the interpretation of the evidence from the other realms. The guidelines reference several publications that propose best practices for evidence integration (Rhomberg *et al*., 2013; NRC, 2014; Rooney *et al*., 2014), however, the guidance provided in these publications differs, and the methods presented in these documents are not necessarily reflected in the draft TCEQ guidelines. This section of TCEQ's guidelines should be more prescriptive, and should include a broader discussion of what evidence integration entails as well as more detailed guidelines on how to accomplish this task.

The lack of explicit guidance on evidence integration appears to be the result of TCEQ's concern that, "given that chemicals differ in amount and quality of each data stream, prescribing universally applicable rules for evidence integration is difficult" (TCEQ, 2017). Although it is challenging to make general guidelines applicable to all assessments, TCEQ's guidelines should be expanded to provide more explicit guidance on how to conduct a thorough evidence integration, while remaining flexible enough to allow for modifications, as needed, for different types of datasets and chemicals (Goodman *et al*., 2013).

The guidelines should provide a general "baseline" framework for performing the evidence integration step of a systematic review. While there are several options for methods of evidence integration, we recommend the method presented by Goodman *et al*. (2013, 2015), which incorporates many of the best practices for evidence integration assembled by others, particularly Rhomberg *et al*. (2013). The basis of the evidence integration step ("Phase 3") of the Goodman *et al*. framework are the Bradford Hill postulates, as well as additional considerations of bias and confounding (see Table 4.1). The general steps of the evidence integration phase of the Goodman weight-of-evidence (WoE) framework include:

- Integrate data across all realms of evidence (*e.g*., toxicology, epidemiology, and MoA), so interpretation of one will inform interpretation of the other(s);
- Assess all data, including negative, null, and positive results;
- Assign less weight to the results of studies that are of lower relative quality;
- Incorporate peer and public comment and advice; and
- Formulate WoE conclusions.

Table 4.1 Considerations for Evidence Integration

Category	Considerations
Strength of Association	Investigators should determine *a priori* how a ''strong'' association should be defined (*e.g*., risk estimate or changes in biological measures of a specific magnitude).
Consistency of Association	Associations should be consistent both within and across studies, particularly those with different study designs and in different populations (or different animal species). Even if every statistical model does not generate statistically significant results, the results should be relatively consistent.
Coherence	Results across all lines of evidence should be coherent (*i.e*., the interpretation of evidence does not conflict with what is known about the biology of the endpoint in question; if it does, the species closest to humans should be considered to have more relevance to humans).
Biological Plausibility	Both evidence indicating biological plausibility and a lack of biological plausibility should be considered. A known, biologically plausible MoA increases the likelihood that an association is causal, and *vice versa*.
Biological Gradient (Dose-Response)	If increasing effects are observed with increasing exposures or duration of exposures, this is evidence for a causal relationship; a lack of such an association is evidence against a causal relationship.
Experimental Evidence	Quasi-natural experiments that provide information on the causal association (*e.g*., for air pollution – epidemiology studies during the Beijing Olympics, a time when air pollutant emissions were reduced substantially) provide strong evidence for a causal association.
Temporality	The exposure should occur before the effect under study, and also within an appropriate time frame.
Confounding	Using information collected during the study quality assessment, determine whether associations are likely attributable to a confounder. Determine whether consistent findings are likely the results of a consistent confounder.
Bias	Using information collected during the study quality assessment, investigators should determine the likelihood of bias (*e.g*., selection bias and publication bias), its potential impact on specific realms of evidence, and subsequent causal conclusions.
Clinical Relevance	Investigators should consider the biological and clinical relevance of effects across realms of evidence to determine the likelihood of adverse effects in the target population.

Notes:
Adapted from Goodman *et al*. (2013, 2015).
(a) Other Bradford Hill postulates (*e.g*., analogy, specificity) may be used, as applicable, although they generally are not as informative with regard to causality.

4.2 Step 6: Confidence Rating for the Body of Evidence

Step 6 in TCEQ's systematic review framework is to "rate the confidence in the body of evidence" (TCEQ, 2017). The guidelines provide a helpful general discussion of some of the key considerations when integrating evidence within and across realms (*e.g*., high-quality studies provide greater confidence and lower uncertainty that the key study findings accurately depict the relationship between the exposure and

effect of interest). There are no specific instructions, however, for rating the confidence in the body of evidence.

The guidelines should indicate that, after working through all of the considerations in Table 4.1 above, the risk assessor should formulate an overall hazard/causality conclusion in a narrative discussion, with consideration of study quality, uncertainty, variability, and sensitivity analyses, and how animal, human, MoA, dose-response relationships, and all relevant data are integrated as part of the conclusion. Working through a narrative assessment that attempts to reconcile differences across the realms of evidence assessed will likely be less burdensome than the more formal approaches to rating confidence (as presented by TCEQ and in other frameworks, such as OHAT [NTP, 2015a]) and will lead to more interpretable results, rather than just presenting the derived toxicity factor without sufficient explanation. This narrative should consider the overall quality and uncertainty in each of the realms of evidence and bring them together to form a causal conclusion regarding the hazard of the chemical under evaluation. In this step of the systematic review, TCEQ should require that investigators summarize uncertainties and data gaps across all the realms of evidence and provide a discussion of whether the data are sufficient for performing a dose-response analysis and toxicity factor derivation.

Finally, a table or figure that illustrates the evidence integration across the different realms of evidence may be helpful for presenting the results of this step clearly and coherently. See, for example, the set of evidence tables provided in US EPA's recent Toxicological Review of Benzo(a)pyrene (US EPA, 2017).

4.3 Rating Confidence in the Toxicity Factor

Although Step 6 of TCEQ's framework for conducting systematic reviews indicate that the body of evidence is being rated, in actuality, the guidelines have the investigator rate the confidence in the derived toxicity factor. Therefore, Step 6 should be renamed accordingly. However, the guidance it provides on rating the confidence in the derived toxicity factor is explicit and useful. Specifically, TCEQ points to tables in a publication by Beck *et al*. (2016, Tables 3 and 4). Beck *et al*. (2016) include 10 elements for toxicity assessment confidence scoring, which include determining whether the assessment uses a systematic review approach, and others elements that involve different steps in the process, focused on the confidence in the relevance and quantitative derivation of the toxicity value. The Beck *et al*. (2016) approach to assessing confidence and uncertainty in toxicity assessments and, more specifically, toxicity factors, is scientifically sound and will be useful for TCEQ and other agencies to use as a final step to rate their overall confidence in their toxicity assessments.

4.4 Evidence Integration for Ethylene Glycol (TCEQ Appendix A.5)

Examples of completed TCEQ systematic reviews will be critical for clearly illustrating how TCEQ's framework should be implemented. Appendix A.5 (*i.e*., the ethylene glycol DSD) of the guidelines provides examples of the current evidence integration tables for each realm of evidence (*e.g.*, animal, human, and mechanistic studies). These tables are useful illustrations, although we would still suggest updating these tables according to the recommendations outlined in Section 3. In addition, the narrative of the assessment and, specifically, the reasoning behind these tables is almost entirely missing. The written portion of the assessment should be expanded to include some of the key features described in the above comments. Most notably, as discussed above, the evidence integration step of the systematic review needs to go beyond just listing the evidence for each realm of investigation and should be explicit about how these realms of evidence are brought together to form conclusions. Further, TCEQ should clearly discuss how it decided whether the evidence overall was sufficient for deriving quantitative toxicity values for ethylene glycol.

Further developing the example systematic review for ethylene glycol by incorporating the suggestions made in these comments will aid in clearly demonstrating how TCEQ staff should work through systematic reviews to arrive at scientifically sound, transparent assessments that can be reproduced by other risk assessors. An expanded example assessment will also allow others to see what a TCEQ DSD that follows the new guidelines looks like.

References

Beck, NB; Becker, RA; Erraguntla, N; Farland, WH; Grant, RL; Gray, G; Kirman, C; LaKind, JS; Jeffrey Lewis, R; Nance, P; Pottenger, LH; Santos, SL; Shirley, S; Simon, T; Dourson, ML. 2016. "Approaches for describing and communicating overall uncertainty in toxicity characterizations: U.S. Environmental Protection Agency's Integrated Risk Information System (IRIS) as a case study." *Environ. Int.* 89-90:110-128. doi: 10.1016/j.envint.2015.12.031.

Goodman, JE; Prueitt, RL; Sax, SN; Bailey, LA; Rhomberg, LR. 2013. "Evaluation of the causal framework used for setting National Ambient Air Quality Standards." *Crit. Rev. Toxicol.* 43(10):829-849. doi: 10.3109/10408444.2013.837864.

Goodman, JE; Prueitt, RL; Sax, SN; Pizzurro, DM; Lynch, HN; Zu, K; Venditti, FJ. 2015. "Ozone exposure and systemic biomarkers: Evaluation of evidence for adverse cardiovascular health impacts." *Crit. Rev. Toxicol.* 45(5):412-452. doi: 10.3109/10408444.2015.1031371.

Hoffmann, S; de Vries, RBM; Stephens, ML; Beck, NB; Dirven, HAAM; Fowle, JR III; Goodman, JE; Hartung, T; Kimber, I; Lalu, MM; Thayer, K; Whaley, P; Wikoff, D; Tsaioun, K. 2017. "A primer on systematic reviews in toxicology." *Arch. Toxicol.* doi: 10.1007/s00204-017-1980-3.

Lynch, HN; Loftus; CT; Cohen, JM; Kerper, LE; Kennedy, EM; Goodman, JE. 2016. "Weight-of-evidence evaluation of associations between particulate matter exposure and biomarkers of lung cancer." *Regul. Toxicol. Pharmacol.* 82:53-93. doi: 10.1016/j.yrtph.2016.10.006.

National Research Council (NRC). 2014. "Review of EPA's Integrated Risk Information System (IRIS) Process." National Academies Press, Washington, DC, 204p. Accessed at http://www.nap.edu/catalog.php?record_id=18764.

National Toxicology Program (NTP). 2015a. "Handbook for Conducting a Literature-Based Health Assessment Using OHAT Approach for Systematic Review and Evidence Integration." Office of Health Assessment and Translation (OHAT), 98p., January 9. Accessed at http://ntp.niehs.nih.gov/pubhealth/hat/noms/index-2.html.

National Toxicology Program (NTP). 2015b. "OHAT Risk of Bias Rating Tool for Human and Animal Studies." Office of Health Assessment and Translation (OHAT), 37p., January. Accessed at http://ntp.niehs.nih.gov/pubhealth/hat/noms/index-2.html.

National Toxicology Program (NTP). 2016. "NTP Monograph: Immunotoxicity Associated with Exposure to Perfluorooctanoic Acid or Perfluorooctane Sulfonate." 147p., September.

Rhomberg, LR; Goodman, JE; Bailey, LA; Prueitt, RL; Beck, NB; Bevan, C; Honeycutt, M; Kaminski, NE; Paoli, G; Pottenger, LH; Scherer, RW; Wise, KC; Becker, RA. 2013. "A survey of frameworks for best practices in weight-of-evidence analyses." *Crit. Rev. Toxicol.* 43(9):753-784. doi: 10.3109/10408444.2013.832727.

Rooney, AA; Boyles, AL; Wolfe, MS; Bucher, JR; Thayer, KA. 2014. "Systematic review and evidence integration for literature-based environmental health science assessments." *Environ. Health Perspect.* 122(7):711-718. doi: 10.1289/ehp.1307972.

Texas Commission on Environmental Quality (TCEQ). 2012. "TCEQ Guidelines to Develop Toxicity Factors (Revised)." Toxicology Division, RG-442, 346p, October. Accessed at http://www.tceq.texas.gov/publications/rg/rg-442.html.

Texas Commission on Environmental Quality (TCEQ). 2014. "TCEQ Recommendations for Systematic Review and Evidence Integration." Office of the Executive Director, 27p., November 20. Accessed at http://www.tceq.com/assets/public/implementation/tox/dsd/proposed/nov14/systematic_review.pdf.

Texas Commission on Environmental Quality (TCEQ). 2017. "TCEQ Guidelines for Systematic Review and Evidence Integration." Toxicology Division, 50p., April 13. Accessed at https://www.tceq.texas.gov/assets/public/implementation/tox/dsd/whitepaper/Proposed%20_Systematic_Review.pdf.

US EPA. 2013. "Applying Systematic Review to Assessments of Health Effects of Chemical Exposures." National Center for Environmental Assessment (NCEA), 220p., August 26.

US EPA. 2014. "Draft Development Materials for the Integrated Risk Information System (IRIS) Toxicological Review of Inorganic Arsenic [CASRN 7440-38-2]." National Center for Environmental Assessment (NCEA), EPA/630/R-14/101, 731p., April.

US EPA. 2017. "Toxicological Review of Benzo(a)pyrene (CAS No. 50-32-8) (Final)." National Center for Environmental Assessment (NCEA), EPA/635/R-17/003Fa. 234p., January.

Vlaanderen, J; Vermeulen, R; Heederik, D; Kromhout, H. 2008. "Guidelines to evaluate human observational studies for quantitative risk assessment." *Environ. Health Perspect.* 116(12):1700-1705.

Appendix A

Suggested References for Systematic Review and Evidence Integration

Systematic Review and Evidence Integration References

Bailey, LA; Nascarella, MA; Kerper, LE; Rhomberg, LR. 2015. "Hypothesis-based weight-of-evidence evaluation and risk assessment for naphthalene carcinogenesis." *Crit. Rev. Toxicol.* 1-42.

Dorne, JLCM; Bottex, B; Merten, C; Germini, A; Georgiadis, N; Aiassa, E; Martino, L; Rhomberg, L; Clewell, HJ; Greiner, M; Suter, GW; Whelan, M; Hart, ADM; Knight, D; Agarwal, P; Younes, M; Alexander, J; Hardy, AR. 2016. "Special issue: Weighing evidence and assessing uncertainties." *EFSA J.* 14(S1):S0511.

Goodman, JE; Prueitt, RL; Sax, SN; Pizzurro, DM; Lynch, HN; Zu, K; Venditti, FJ. 2015. "Ozone exposure and systemic biomarkers: Evaluation of evidence for adverse cardiovascular health impacts." *Crit. Rev. Toxicol*. 45(5): 412-452. doi: 10.3109/10408444.2015.1031371.

Hoffmann, S; de Vries, RBM; Stephens, ML; Beck, NB; Dirven, HAAM; Fowle, JR III; Goodman, JE; Hartung, T; Kimber, I; Lalu, MM; Thayer, K; Whaley, P; Wikoff, D; Tsaioun, K. 2017. "A primer on systematic reviews in toxicology." *Arch. Toxicol.*

Institute of Medicine (IOM). 2008. *Improving the Presumptive Disability Decision-Making Process for Veterans.* Committee on Evaluation of the Presumptive Disability Decision-Making Process for Veterans, Board on Military and Veterans Health. National Academies Press, Washington, DC. 781p.

Lynch, HN; Loftus; CT; Cohen, JM; Kerper, LE; Kennedy, EM; Goodman, JE. 2016. "Weight-of-evidence evaluation of associations between particulate matter exposure and biomarkers of lung cancer." *Regul. Toxicol. Pharmacol.* 82:53-93. doi: 10.1016/j.yrtph.2016.10.006.

Morgan, RL; Thayer, KA; Bero, L; Bruce, N; Falck-Ytter, Y; Ghersi, D; Guyatt, G; Hooijmans, C; Langendam, M; Mandrioli, D; Mustafa, RA; Rehfuess, EA; Rooney, AA; Shea, B; Silbergeld, EK; Sutton, P; Wolfe, MS; Woodruff, TJ; Verbeek, JH; Holloway, AC; Santesso, N; Schunemann, HJ. 2016. "GRADE: Assessing the quality of evidence in environmental and occupational health." *Environ. Int.* 92-93: 611-616.

Prueitt, RL; Rhomberg, LR; Goodman, JE. 2013. "Hypothesis-based weight-of-evidence evaluation of the human carcinogenicity of toluene diisocyanate." *Crit. Rev. Toxicol.* 43(5):391-435.

Rhomberg, L. 2014. " Hypothesis-based weight-of-evidence: An approach to assessing causation and its application to regulatory toxicology." *Risk Anal.* 35(6):1114-1124.

Vyskocil, A; Truchon, G; Leroux, T; Lemay, F; Gendron, M; Gagnon, F; El Majidi, N; Boudjerida, A; Lim, S; Emond, C; Viau, C. 2012. "A weight of evidence approach for the assessment of the ototoxic potential of industrial chemicals." *Toxicol. Ind. Health.* 28(9):796-819.

Appendix B

Suggested Tables and Study Quality Criteria

Table B.1 General Results Table

Endpoint	Reference	Species/Strain	*N* per Sex Group	Route	Exposure Duration (hrs)	Unit of Measurement	Dose (mg/kg)	Results	*p* Value
Endpoint 1	Ref 1	Rat/Wistar	5	Oral	2		0		
							50		
							100		
Endpoint 2	Ref 2								

Table B.2 Study Quality Criteria for Animal Studies

Category	Score	
	1	-1
Study Design Reporting	Study design was clearly defined and detailed in the methods.	Study design not adequately defined and detailed information not provided.
Protocol Reporting	Protocol defined and deviations described.	Protocol not described and/or deviations not reported.
General Experimental Conditions	Used identical experimental methods across study groups.	Study used experimental methods with minor differences. Use of identical experimental methods is unclear.
Randomization	Explicitly stated whether animals were randomized into treatment or control groups.	Animals not randomized or no discussion of randomization included.
Control Groups	Appropriate control group used.	No control group used or inappropriate control group used.
Sample Size	Sufficient number of animals used (n = 5/sex/group, or power calculation showing sufficient size).	Insufficient number of animals used.
Exposure or Test Substance Characterization	Details regarding source, composition, purity, and stability of test substance reported.	One or more details regarding test substance missing.
Exposure Maintenance (as applicable, *e.g.*, inhalation studies)	Measures taken to ensure consistent exposure, including continuous monitoring of concentration (in chamber studies), type of exposure method used (*e.g.*, chamber or nose-only), maintenance of adequate environmental conditions, and density of animals in each chamber.	Study does not provide sufficient information to verify proper exposure maintenance.
Animal Housing and Husbandry	Description of the animals used provided (*i.e.*, age, strain, and where purchased or bred), methods for feeding and housing of animals (including number of animals/cage, light/dark cycle, temperature, and humidity), treatment conditions (including ethical guidelines), acclimation period, age of animals, and sacrifice methods.	At least one of the animal housing and husbandry details is missing.
QA/QC Protocols	Provided details on any biological sample collection, handling, and storage methods (*e.g.*, temperature).	Any QA/QC protocol details missing.

Category	Score	
	1	-1
Assay Reproducibility	Details provided about the assays or kits (and their source) used to measure endpoints.	Assay details absent. A non-standardized or novel method was referenced, but not described in detail.
Attrition Bias	Details of study-related deaths provided.	Study-related deaths not reported/described.
Statistical Methods	Appropriate statistical methods used, given the type of exposure and outcome tested (*e.g.*, mixed effects models for outcomes with repeated measures).	Study did not use statistical methods appropriate for study design.

Table B.3 Study Quality Criteria for Epidemiology Studies[a]

Category	Score	
	1	-1
Study Design Reporting	Study design clearly defined and detailed in methods.	Study design not adequately defined and detailed information not provided.
Protocol Reporting	Protocol defined and deviations described.	Protocol not described and/or deviations not reported.
Study Size	Calculations conducted to determine appropriate sample size or sufficiency of sample size otherwise supported.	No calculation conducted to determine appropriate sample size.
Appropriate Comparison or Control Groups	Similar baseline characteristics between comparison groups (*e.g.*, for population-based case-control study, cases reasonably arise from the same population as controls and the case definition sufficient and independently validated)	Unclear if baseline characteristics are sufficiently similar between comparison groups. Case definition not reported, not validated or based on self-report (case-control studies)
Follow-up of Subjects (cohort studies)	Subject follow-up was thorough and sufficient to develop endpoint of interest (*e.g.*, cancer latency considered).	Subject follow-up was not well documented and/or was not sufficient for the endpoint of interest.
Blinded Study	Outcome assessors and participants blinded to exposure status. Most relevant for controlled human exposure studies, but outcome assessment can be blinded to exposure or case/control status for other designs.	Study was not blinded.
Selection and Response Bias	Low selection and response bias. For example, for panel studies, data completeness rates should be at least 70%, or if adherence is lower, authors addressed the problem of missing data (*e.g.*, by determining whether the pattern of missing data was random with respect to outcome and exposure). Cross-sectional studies should have relatively high response rates that are similar across groups. Adapt as needed for other study designs.	Study did not meet criteria for score of 1.
Exposure Methods	Exposure ascertainment are consistent across groups and appropriate exposure methods given the review question (*e.g.*, urinary arsenic measurements speciated	Exposure ascertainment not consistent across groups or exposure measurement not appropriate for review question (*e.g.*, urinary arsenic measurements not

Category	Score	
	1	-1
	for an analysis of the health effects of inorganic arsenic).	speciated for analysis of the health effects of inorganic arsenic).
Appropriate Outcome Assessment Methods	Study used validated and/or standard outcome assessment methods.	Outcome assessment methods not validated or standard.
QA/QC Protocols for Exposure and Outcome Assessment Involving Biological Samples	Clearly and completely described the procedures used for storing, handling, and processing biological specimens and listed the name and source of any kits used for bioassays (when appropriate).	Storage, handling, and assay kits for biological samples not described.
Assay Precision (when applicable)	Study evaluated the precision of repeated biomarker measurements on the same sample and reported a coefficient of variation (CV) of 10% or lower.	Study reported CV of >10%.
Statistical Modeling (adapt as needed)	Study used appropriate statistical analyses to evaluate associations between exposure and endpoint of interest (*e.g.*, received methods that account for within-subject correlation inherent in a repeated measurements study, such as generalized estimating equations [GEEs], linear mixed models [LMMs], or generalized linear models [GLMs]). Adapt as needed, based on standard practice for given exposure and outcome type.	Statistical methods not appropriate for tested associations or not fully executed (*e.g.*, no tests for multiple comparisons in such situations).
Confounding	Study assessed potential confounders relevant to the exposure and outcome being assessed. Assessment of confounding can be tiered based on critical confounders (*e.g.*, smoking, SES), and those that are important but with less potential to affect risk estimates.	Study did not assess relevant confounders identified by assessors.
Sensitivity Analyses (as appropriate)	Study assessed alternative model assumptions in sensitivity analyses.	No sensitivity analyses conducted.

Notes:
QA/QC = Quality Assurance/Quality Control; SES = Socioeconomic Status.
(a) Guidelines are general and should be tailored, as needed, per the systematic review question (*i.e.*, based on the specific exposure routes and outcome of interest).

Table B.4 Study Quality Criteria for *In Vitro* Studies

Category	Score	
	1	-1
Study Design Reporting	Study design and test systems clearly defined and detailed in methods.	Study design not adequately defined and detailed information not provided.
Protocol Reported	Protocol defined and deviations described.	Protocol not described and/or deviations not reported.
Control Groups	Appropriate control group(s) used.	No control group(s) used (positive or negative) or inappropriate control group(s) used.
Sample Size (replicates)	Replicates reported; sufficient number of replicates used given method/test kit specifications.	Insufficient number of replicates.
Test Substance Characterization	Details regarding source, composition purity, and stability of test substance reported.	One or more details regarding test substance missing
Blinding	Outcome assessment blinded.	Outcome assessment not blinded.
QA/QC Protocols	Details provided on precision of test system kits and any storage conditions for test materials.	Any QA/QC protocol details missing.
Assay Reproducibility	Details provided about the assays or kits (and their source) used to measure endpoints.	Assay details were absent, or a non-standardized or novel method was referenced but not described in detail.
Statistical Methods	Appropriate statistical methods used, given the type of exposure and outcome tested (*e.g.*, mixed effects models for outcomes with repeated measures).	Study did not use statistical methods appropriate for study design.

Note:
QA/QC = Quality Assurance/Quality Control.