Modifications to TCEQ Guidelines for Systematic Review in Support of Development of Toxicity Factors

White Paper

August 30, 2023
Update: September 30, 2024

TABLE OF CONTENTS

BLE OF FIGURES

TABLE OF TABLES

Acronyms and Abbreviations

Acronym / Abbreviation	Definition
ADME	absorption, distribution, metabolism, and excretion
AI	artificial intelligence
ATSDR	Agency for Toxic Substances and Disease Registry
BMDS	benchmark dose software
d	day
DART	developmental and reproductive toxicity
DSD	development support document
EFSA	European Food Safety Authority
HAWC	Health Assessment Workspace Collaborative
hr	hour(s)
IRIS	Integrated Risk Information System (USEPA)
LOAEL	lowest-observed-adverse-effect level
MeSH	medical subject headings
ML	machine learning
MOA	mode of action
NRC	National Resource Council
NOAEL	no-observed-adverse-effect level
NRC	National Research Council
NTP	National Toxicology Program
OECD	Organization for Economic Co-operation and Development
OHAT	Office of Health Assessment and Translation (NTP)
ORD	Office of Research and Development
PBPK	physiologically-based pharmacokinetic model
PCC	Population, Concept, and Context
PECO	Populations, Exposure, Comparator/Control, and Outcomes
PECOTS	Population, Exposure, Comparator, Outcome, Timing, Setting
POD	point of departure
ppm	parts per million
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses

Acronym / Abbreviation	Definition
QC	quality control
ReV	reference value
RfD	reference dose
RG	regulatory guidance
SEM	systematic evidence map
SFo	oral slope factor
SR	systematic review
TCEQ	Texas Commission on Environmental Quality
TD	Toxicology, Risk Assessment, and Research Division (TCEQ)
TSCA	Toxic Substances Control Act
UF	uncertainty factor
URF	unit risk factor
USEPA	United States Environmental Protection Agency
WHO	World Health Organization
wk	week(s)

Introduction

A systematic review (SR) is a methodological approach to answering a research question in a manner that minimizes the risk of bias and error and maximizes transparency. The method involves identifying, appraising, and synthesizing all relevant studies on a particular topic (Uman, 2011; WHO, 2021). This method also provides a scientifically robust approach to the review and interpretation of complex and often contradictory evidence in relation to a research question (WHO, 2021). To achieve this goal, SR methods are defined *a priori* in a comprehensive plan developed during problem formulation exercises. This *a priori* initiative supports answering the relevant research question in a transparent and unbiased fashion (WHO, 2021).

Adapted from the field of evidence-based medicine, SR in toxicology was proposed as a critical practice in determining causation (Guzelian et al., 2005). Since 2005, several organizations have adapted the concept and proposed best practices in the conduct of SR for the purposes of hazard and risk assessment. Notably, the European Food Safety Authority (EFSA), U.S. Environmental Protection Agency (USEPA), and the National Toxicology Program's former Office of Health Assessment and Translation (NTP OHAT) have published methods for conducting SR and evidence integration (EFSA, 2010; USEPA, 2022; NTP OHAT, 2019). The TCEQ has also previously published guidelines for performing SR for the purpose of developing toxicity factors (Schaefer and Myers, 2017). Among the peer-reviewed literature, Wikoff et al. (2020) proposes a framework that combines and builds on the aforementioned existing guidance for use by a practitioner in hazard and risk assessment. The World Health Organization (WHO, 2021) offers similar guidance on using systematic review to facilitate the chemical risk assessment process.

The overall objective of the TCEQ SR guidance is to provide a flexible, yet structured, framework for conducting an SR in the context of developing chemical-specific toxicity factors based on evidence from human and/or animal studies, along with supporting available mode-of-action (MOA) studies (when necessary). This white paper reflects an update to the previous TCEQ Guidance on Systematic Review and Evidence Integration, originally finalized in 2017 (TCEQ, 2017), which was developed to supplement the TCEQ's 2015 Guidelines to Develop Toxicity Factors (RG-442). As stated in those guidelines, toxicity factors are developed on an as-needed basis. This may include chemicals for which there are no existing toxicity factors or for which toxicity factors are outdated. The toxicity factors developed by the TCEQ are derived to protect the general public, as well as potentially sensitive populations such as children, pregnant women, and the elderly; thus, all available health endpoints and various types of studies are considered to determine the most sensitive adverse endpoint (i.e., critical effect) in the most relevant or sensitive species. This SR guidance, in principle, must also be applicable to chemicals for which limited toxicity data are available.

The TCEQ documents the development of a chemical-specific toxicity factor in a development support document (DSD). Briefly, the DSD process begins with the selection of a chemical, followed by a review of the physical and chemical properties and a critical review of dose-response data for all the available health endpoints. The empirical evidence is examined thoroughly to determine the no-observed-adverse-effect level (NOAEL) and/or the lowest-observed-adverse-effect level (LOAEL). When data are available, Benchmark Dose Software (BMDS) can be used to characterize dose-reponse relationships and to establish a point of departure (POD). To the extent possible, an evaluation of the MOA(s) for the most sensitive (i.e., critical) adverse endpoint is also included in the analysis. An MOA analysis is important in understanding the potential for toxicity and the most scientifically defensible extrapolation to lower exposures (USEPA, 2005a).

This update builds on previous SR guidance (TCEQ, 2017) with available existing methods in conducting SRs and integrating evidence for the purpose of developing reference values (ReVs), unit risk factors (URFs), oral slope factors (SFos), and reference doses (RfDs). A significant revision in the workflow presented herein compared to previous guidance is the addition of a potential systematic evidence map (SEM) workflow (**Figure 1**). Similar to SR, an SEM uses robust and transparent methods to systematically explore and describe the literature on a given topic. As stated by WHO (2021), this method can be used to help with prioritization, and the addition of this technique is anticipated to provide guidance when a narrow, more specific SR question cannot be formulated. In contrast to the specific SR question, an open-framed question (or one that lacks specification of some key elements) is posed instead. Practitioners have found that stepwise use of these evidence-based tools is helpful to facilitate the risk assessment process when large evidence bases involving many different outcomes and different evidence streams need to be assessed, as is common in risk assessment and development of toxicity factors.

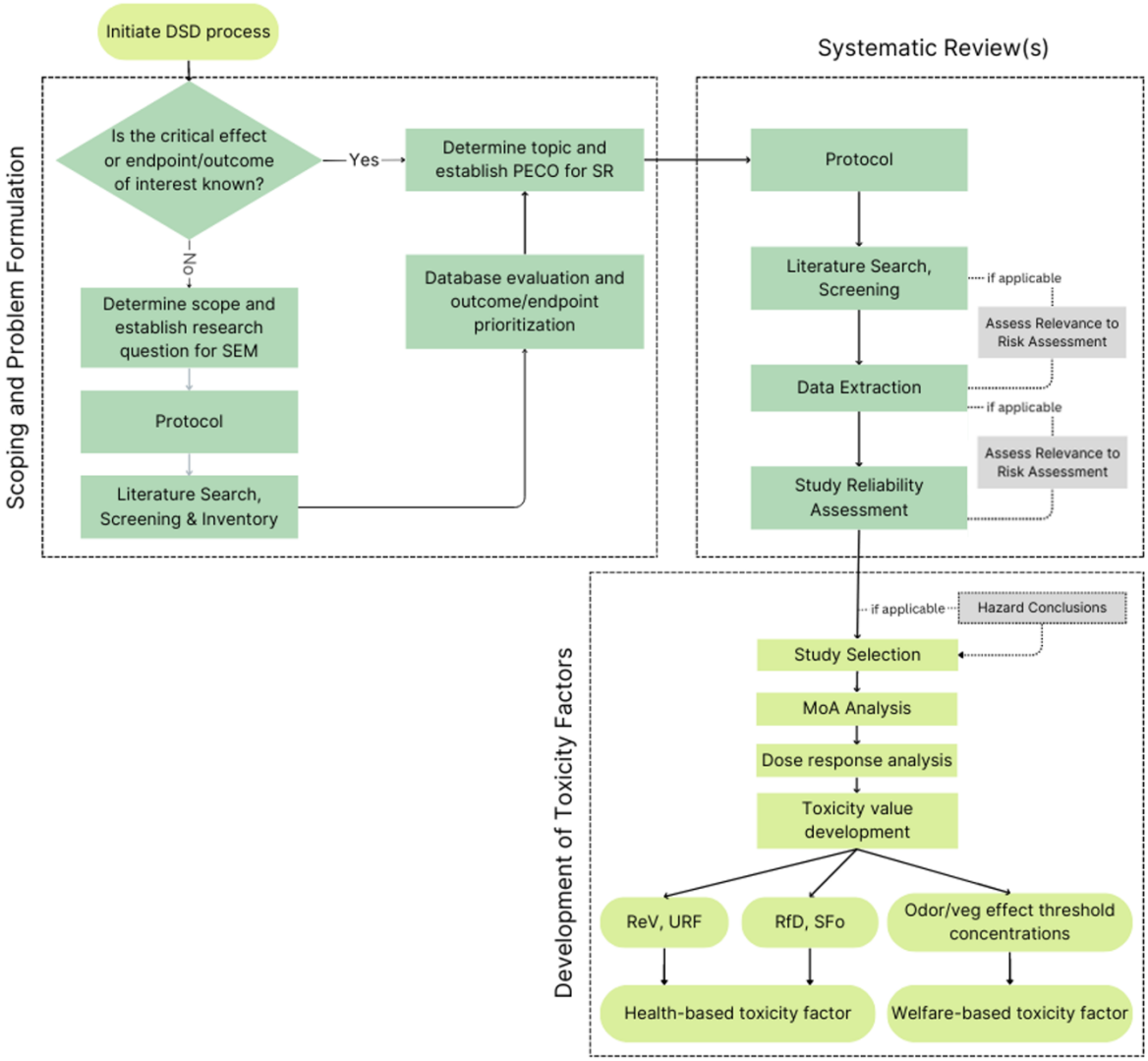

Figure 1. Framework for implementing systematic methods in support of developing toxicity factors

Scoping and Problem Formulation

Following the decision to develop a DSD for a given chemical(s), characterizing the scope of the assessment will begin with a planning phase. In evidence-based methods, this is known as Problem Formulation. Importantly, in the context of risk assessment, “assessment planning” is also incorporated into problem formulation. This includes designing and stating the methods

for components of risk assessment that the SR could inform (e.g., hazard identification, toxicity factor derivation, exposure assessment, MOA analysis, toxicokinetics, etc.), recognizing that a systematic review method can be applied to facilitate multiple aspects of risk assessment.

Scoping exercises to aid in this effort are performed during Problem Formulation. Prompting questions may be useful in guiding these exercises. These may include:

- What is the specific context of the assessment?
- What is the timeline, and what resources are available?
- What is the required output to meet the overall goal of the assessment?
- What are the physical and chemical properties of the chemical?
- Are there existing systematic reviews or agency evaluations?
- What is the data availability?
- Are the critical effects known?
- Are there known potentially sensitive subpopulations?
- Are the toxicokinetics known, and does route of exposure play a role in toxicity?
- Is the chemical carcinogenic? If so, is the chemical carcinogenic only by a specific route of exposure or when a biologically plausible threshold is exceeded?

Key exercises performed in this phase will include:

- Identification and review of assessments conducted by other organizations or in the peer-reviewed literature
- Scoping the volume and nature of evidence to determine the need for a SEM or SR
- Definition of risk assessment question and structured PECO or PCC elements (defined below)
- Protocol development, including determination of inclusion/exclusion criteria
- Piloting[a]

Scoping

In support of developing toxicity factors, it is standard practice for the TCEQ to review all available relevant data for a particular chemical. Based on the identified database, a toxicologist then identifies the critical effect that occurs at the lowest human equivalent concentration or dose. As described in RG-442 (TCEQ, 2015), evaluation and selection of key

[a] Piloting is an exercise in which the inclusion and exclusion criteria are tested and refined for clarity and reproducibility, and in which the forms, study tags, information fields, and notes related to the templates are defined for each step in the review (Wikoff et al., 2020).

studies follows the guidelines detailed by USEPA (1994, 2005b) and the National Research Council (NRC) (2001). In some circumstances, the state of knowledge collated by the TCEQ or other relevant agencies (e.g., USEPA, Agency for Toxic Substances and Disease Registry [ATSDR]) may provide insight to the critical effect of interest without the need for a *de novo* assessment of the literature. Scoping and problem formulation exercises at the start of the DSD process will determine whether the assessment team should first undertake an SEM or SR based on this knowledge, recognizing that one or the other—or both (sequentially)—may be appropriate. Typically, the question posed in the diamond text box of Figure 1, "*Is the critical effect or endpoint/outcome of interest known?*", will facilitate the determination of whether an SEM and/or SR is most appropriate based on the state of knowledge and the subject-matter expertise of the TCEQ risk assessment team.

Based on the output of scoping exercises, the assessment team will determine whether the state of knowledge is sufficiently defined to develop a narrow risk assessment question (or questions) appropriate for SR. If, however, it is not sufficiently defined, an SEM will help clarify the broad topics and key concepts and prioritize the assessment needs. If an SEM is most appropriate, the output will be used to identify information to be prioritized further and evaluated by SR.

As indicated in Figure 1, the SEM portion of problem formulation consists of an *a priori* protocol and literature search, screening, and inventory. These first two steps are similar to the SR process, with the primary difference being the broad nature of literature relevant to the SEM. For this reason, the first two steps of both an SEM and an SR will be discussed together, with any differences highlighted. In practice, these first steps of the SEM will not be replicated during the SR effort; rather, they will be updated as needed to account for the more refined topic of evaluation. **Table 1**, below, displays key concepts for the two evidence-based methods.

Table 1. Key concepts typical of systematic evidence maps and systematic reviews

Concept	Systematic Evidence Map	Systematic Review
Objective	Characterize landscape to determine key topics	Answer specific, defined risk assessment question
Question	Open-framed, often broad	Closed-framed, narrow
Evidence Identification	Systematic search and screening based on broad inclusion criteria	Systematic search and screening based on narrow inclusion criteria
Data Type	Study characteristics*	Data for dose-response analysis
Critical appraisal of individual studies	Optional (often no)**	Yes
Synthesis	Identification of key concepts, gaps, and clusters	Qualitative and/or quantitative synthesis to answer question

** In cases where SEM is not followed by SR, it may be necessary to also capture data for dose-response analysis.*
*** In cases where SEM is not followed by SR, a general descriptive qualitative assessment of the study could be included.*

Both SEMs and SRs typically are conducted by multidisciplinary teams that could include subject-matter experts, an information specialist or librarian, methodologists, and/or evidence assessors/analysts. However, the team at the TCEQ will consist of at least two toxicologists, with possible assistance from a librarian. After establishing the assessment team, the first stage of problem formulation is to understand the current state of knowledge as it pertains to chemical risk assessment. Scoping searches will be performed to establish the state of knowledge, as well as identify data gaps. These searches can be implemented by any member of the assessment team, with the potential for collaboration with an information specialist or librarian. This scoping exercise will also allow the assessment team to evaluate the potential volume of available literature.

Literature of high relevance will be noted for reference during the literature identification phase. For example, notable experimental animal, human, or MOA studies can be labeled as "high relevance" for use during search validation, piloting, or further problem formulation efforts. Authoritative assessments and reviews previously conducted should also be noted.

PECO Development

Another critical component of problem formulation is development of a specific risk assessment question that is to be addressed by the evidence-based methods. In SR, this structured statement defines the Population, Exposure, Comparator, and Outcome (PECO) of

interest to the chemical risk assessment. Other, similar structured statements such as PECOTS (Population, Exposure, Comparator, Outcome, Timing, Setting) may be considered. Alternatively, systematic maps may consider a broader model such as Population, Concept, and Context (PCC). Morgan et al. (2019) provides guidance for the formulation of these informative questions and explores how to develop them in the context of environmental exposures and health outcomes.

A structured PECO question (or questions) based on the specific scenario at hand will be developed. Examples of PECO elements commonly considered in a TCEQ DSD are shown in **Table 2**, along with potential refinements to narrow the scope and help focus the assessment.

Table 2. Components, data elements and focusing aspects of an example PECO question: In humans (*population*), what concentration of chemical A (*exposure*) is associated with significantly increased hepatotoxicity (*outcome*) when compared to controls (*comparator*)?

Component	Potentially relevant elements	Focusing aspects
Population	• Human evidence • Experimental animal evidence • Mechanistic evidence	Sensitive or target populations, such as pregnant women or children (and experimental models of such)
Exposure	• Chemical of interest • Route(s) of interest • Exposure duration of interest	Exposure scenario (e.g., occupational), route (e.g., inhalation), dose ranges (e.g., relevance to environmental exposures) or timing (e.g., developmental window)
Comparator	• No or low exposure to chemical of interest	Only non-exposed or a specific level associated with low exposure
Outcome	• Adverse* non-cancer outcomes • Cancer outcomes • Mode of action • Toxicokinetics	Subset of outcomes (e.g., cancer, developmental effects, specific cancer types, specific developmental endpoints).

*The TCEQ defines an adverse effect as a biochemical change, functional impairment, or pathologic lesion that affects the performance of the whole organism or reduces an organism's ability to respond to an additional environmental challenge (TCEQ, 2015). Consistent with the goal of protecting public health, the TCEQ calculates conservative health-based toxicity factors to protect against adverse health effects. More information is available in Section 3.6.1, Determination of Adverse Effect (TCEQ, 2015).

The risk assessment team will determine which aspects of the assessment will fall under the scope of the SR. It is standard to consider hazard data (which are then used in dose-response) in an SR supporting risk assessment, and most often, this will focus on apical endpoints that characterize potential adversity (and thus are candidate endpoints for toxicity factor development). However, additional risk assessment topics—such as exposure, MOA, and vegetation data, as examples—may be relevant to a given SR as well. In this scenario, a PECO will be developed for each component. The structured assessment question for both SEM and SR (e.g., PECO, PECOTS, PCC) will then be used to inform the development of inclusion and exclusion criteria and the literature search strategy.

During scoping, the need to identify and evaluate mechanistic or toxicokinetic data will be identified. If these data are important to drawing conclusions for the toxicity factors, assessors may elect to include these data types in the SR process (which involves full appraisal, synthesis, and integration). In other cases, the assessment team may elect to pragmatically use the systematic search to identify these data for subsequent use (either iteratively if needed in the SR or for contextual use in the DSD portion of toxicity factor development). For the latter scenario, these data types will be described in the protocol and will not be included in the formal SR but will be documented during the SR. The iterative consideration of mechanistic data, as an example, reflects that, in practice, risk assessors may not know *a priori* whether an MOA assessment is necessary—and, more often, do not know the parameters of an MOA assessment until the adverse-outcome data are assessed. Inclusion of mechanistic data in systematic review is still an evolving practice; however, best practices involve evaluation and integration of data in pathway-based constructs, such as MOA (Meek & Wikoff, 2023). . The assessment team will determine through scoping (and through the evaluation process) whether mechanistic data are needed to help inform toxicity factor derivation decisions as part of the risk assessment process, and if so, the specific data that are needed to inform the assessment, as well as the most appropriate construct in which to utilize such data.

Inclusion and Exclusion Criteria

A strength of the SR approach is the documentation of clear study inclusion/exclusion criteria. This step is useful in documenting why particular studies were chosen as potential key studies and the reasons for excluding other studies (i.e., excluding them as potential key studies or completely excluding studies from the review). These criteria enhance transparency and subsequently improve risk communication to a wide range of stakeholders. Clear and direct inclusion/exclusion criteria based on the stated structured risk assessment question should be specified to identify the initial study database from which key and supporting studies are selected. These criteria may include adverse health outcomes, exposures, durations, and the types of studies relevant to the toxicity factor being developed. During screening, studies that meet inclusion criteria are retained for further review. Developing explicit criteria *a priori* to select or omit studies helps to balance scientific judgment by providing clear and transparent

documentation. This documentation allows the literature identification process to be reproduced easily by other assessors if needed, which in turn can improve confidence in the TCEQ's derivation of toxicity factors.

Defining one set of inclusion and exclusion criteria for all chemicals is difficult, because the criteria often will be specific to a chemical and/or purpose. Therefore, inclusion and exclusion criteria will be modified to make them fit for purpose for each assessment. For example, if the purpose of a particular assessment is to develop an inhalation reference value, oral studies may be excluded. However, if the inhalation database is lacking and the effects are not route dependent, oral studies may be included. More stringent exclusion criteria may be required for data-rich chemicals to identify data most relevant to the specific assessment being conducted. **Table 3** represents the inclusion and exclusion criteria developed during a systematic review of vanadium in support of a DSD.

Table 3. Examples of study inclusion and exclusion criteria

Category	Include	Exclude
Population	Humans Experimental animals (mammalian species)	Non-mammalian species Ecological field studies (e.g., ecotoxicity) Mechanistic evidence[a]
Exposure	Inhalation exposure to any of the Vanadium compounds listed in Table 1 Any exposure duration **Human-specific**: Exposure metrics provided as actual measured air concentrations (e.g., µg/m^3) **Animal-specific**: Controlled/known exposures	Any route other than inhalation (e.g. injection, oral, dermal) Biological biomarker studies (e.g., vanadium in toenail clippings) Exposure metrics not provided as an actual measured air concentration Exposure concentration unknown or does not have a control group for comparison
Outcomes	Any adverse cancer or non-cancer outcomes such as respiratory, immune, developmental and reproductive toxicology (DART), hepatic, renal, cardiometabolic, hematologic, nervous compared to a control population or experimental group.	Any non-apical outcomes such as Mode-of-Action[a] or toxicokinetics[a]
Reference type	Primary experimental or observational studies that report empirical research	Any case studies, reviews (including SR), meta-analyses, commentaries, editorials, or any other secondary reporting[a]

Category	Include	Exclude
Additional criteria	See exclusion criteria	Publications not available in English Study reports or publications that are not available for review in full (e.g., only the abstract is available)

[a] These categories will be excluded from systematic review and toxicity factor development; however, they will still be categorized for potential contextual information and data used to support interpretation of eligible studies.

Other criteria that may be considered include considerations for any relevant aspect of the risk assessment, such as the examples provided below in **Table 4.** Examples of additional considerations for inclusion and exclusion criteriaIt is key that these are discussed by the assessment team at the problem formulation stage, so they are outlined in the protocol. However, if at any point during the review it becomes apparent that a certain type of data cannot be further considered for the development of toxicity factors, the assessment team should revise the inclusion criteria and document the deviation from the *a priori* protocol.

Table 4. Examples of additional considerations for inclusion and exclusion criteria

Inclusion Criteria	Exclusion Criteria
Exposure concentration is relevant to developing toxicity factors	• Study focused on overdose/poisoning or mortality • Exposure concentration unknown
Study evaluates relevant endpoints assessed per individual	• Study reports all-cause mortality at the population level
Study focused on the chemical of concern or active metabolites	• Study examined multiple chemicals not of interest • Study on beneficial treatment following chemical exposure
Structured evaluations of mode of action for outcomes specific to toxicity factor development	Assessments of mechanistic data that do not use pathway-based constructs (e.g., key characteristics)

Tool Selection

Consideration and selection of tools to facilitate the SR is a critical component of problem formulation. The concept of tools, as discussed here, may be a construct or framework (e.g., NTP OHAT Handbook for Conducting a Literature-Based Health Assessment) or software

used to facilitate the mechanics (e.g., Endnote, HAWC). An example of the tools selected to facilitate the case study on Vanadium is shown in **Figure 2**.

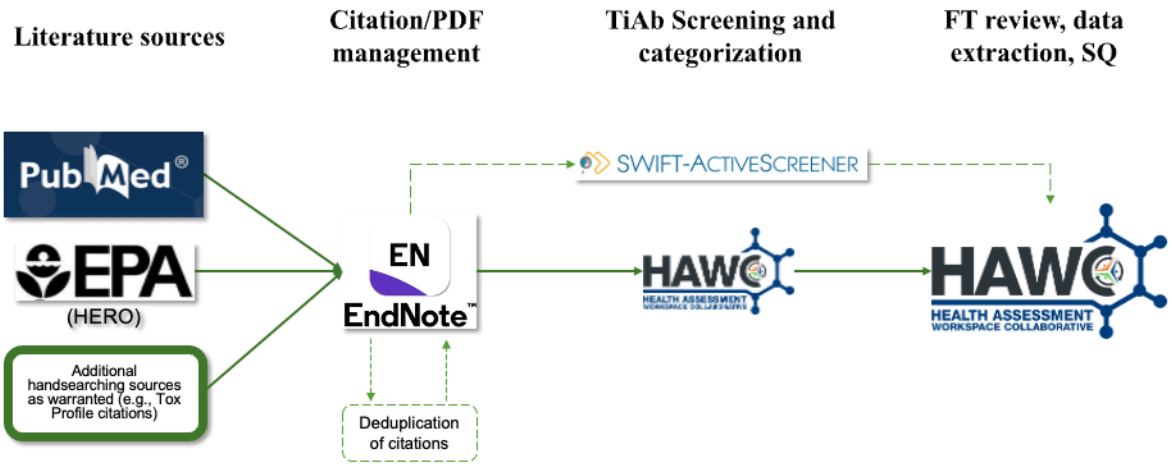

Figure 2. Literature databases and software selected for the vanadium case study workflow.

The TCEQ will base tool selection on feasibility, cost, and complexity of the assessment. The landscape of both frameworks and software specific to SR in the context of risk assessment, in particular, is rapidly changing. Thus, it is not appropriate for this guidance to dictate the specific tools to be used in the process. The TCEQ will use expert judgement and consult relevant resources to make decisions related to tool selection based on priorities and nature of the assessment being undertaken. Anticipated tools will be described in the protocol and included in piloting exercises. Iterative refinement of tools may occur throughout piloting and implementation of the SR.

Protocol Development

Building on the problem formulation exercises, the assessment team will document the anticipated approach for either the SEM or SR, *a priori*, in a review protocol.. The protocol should be developed based on the needs of the assessment and as such, fields may vary depending on the topic of interest. During the course of an assessment, deviations from the protocol may occur as new information becomes available; these deviations should be documented accordingly as they occur. While the format and content of a review protocol will vary based on the needs of a particular assessment, at a minimum, the following will be described:

- Assessment team and anticipated role of each member.
- The objective of the assessment and structured question(s) via PECO, PCC, or similar.

- Search strategy for identification of evidence, including databases, search syntax, software tools, artificial intelligence (AI) methods, etc. (documented at a level that ensures transparency and reproducibility).
- Inclusion/exclusion criteria for selection of evidence, typically built around the PECO or PCC components.[b]
- Piloting and reviewer calibration exercises.
- Fields or information to be extracted from individual studies.
- Individual study appraisal methods and tools, including any refinements.

In cases where a protocol is developed initially for an SEM, this protocol will be updated to include aspects of SR (e.g., PECO, updated inclusion criteria, study quality assessment) based on the database evaluation and prioritization. The updates and decisions leading to the development of the refined risk assessment question will be documented to facilitate transparency.

Evidence Identification

Literature Search

The general objective of the literature search strategy for a specific chemical risk assessment is to identify all relevant studies, which may include both published and unpublished literature. Data relevant to the risk assessment will be acquired from three main sources: data solicitation procedures for DSDs; traditional literature searches, including querying publicly available databases ; and targeted searching of reference lists. The assessment team may work closely with an information specialist or research librarian as needed to develop and implement the search strategy and assist with reference management and workflow facilitation. To this end, it may be useful for the assessment team to use a reference manager (e.g., EndNote) to organize the results of the literature search and facilitate the exchange of citation metadata between software tools, if necessary.

Data solicitation

Several months prior to the start of work on a DSD, the TCEQ staff will perform a scoping exercise to identify all available toxicity information for the chemical. This process is announced using its email listserve to solicit information for a particular chemical or class of chemicals; interested parties are encouraged to provide citations or toxicological information. Chapter 1, Sections 1.9 and 1.10 of the TCEQ (2015) Guidelines provides more detailed information on the selection of chemicals and data solicitation for DSDs. The literature review can be updated as

[b] Pending the state of knowledge around a given chemical and toxicity factor workflow, relevance to risk assessment (described subsequently) may also be considered as part of inclusion/exclusion criteria.

new information becomes available or additional supplemental literature searches are warranted. Changes made to the initial literature review should be documented accordingly.

Database searching

The TCEQ conducts thorough literature searches of relevant databases and takes other prudent steps to identify relevant studies during the literature review. The TCEQ Toxicology, Risk Assessment, and Research Division (TD) trains its toxicology staff to conduct their own systematic literature searches and will consult its internal research librarian when necessary. For example, in addition to relevant guidance (e.g., Section 3.3.2 of TCEQ 2015), the TCEQ staff have access to the National Library of Medicine's resources for training on advanced uses of the various databases (e.g., PubMed), and/or to train in person with an Instructional Services Librarian. The TCEQ staff also uses other resources such as webinars and/or in-person training (as available).

Development of search syntax for bibliographic databases will vary depending on the focus of the SEM or SR. The search strategy and syntax will be informed by scoping searches performed during problem formulation, the volume of potentially relevant data, and the available resources (e.g., timeline, staff). The strategy will also consider previously published systematic reviews and may adapt as appropriate for the risk assessment question. It may also be necessary to perform supplementary searches for contextual considerations such as toxicokinetic or mechanistic data. Collaboration with an information specialist or research librarian may be critical to this phase and can provide input on use of controlled language for bibliographic databases (e.g., medical subject headings [MeSH] terms for PubMed) and Boolean operators, which are recommended in conducting a systematic literature search. Below are concepts that can be considered in the development of the syntax:

- "AND" is used to group *keywords or ideas* together in the search (e.g., benzene AND cancer)
- "OR" is used to search for multiple *synonyms* (e.g., inhalation OR air OR aerosol)
- "NOT" is used to exclude keywords (e.g., ethylene NOT diethylene); note that this must be used with caution so as not to unintentionally exclude articles of overlapping concepts
- Quotation marks (" ") are used when multiple keywords are searched together (e.g., "ethylene glycol")
- Asterisks (*) are used to search all of the forms of a root word (truncation) to get all derivatives of the term (e.g., a search for carcinogenic effects can include the term carc*, which will search carcinogen, carcinogenic, carcinoma, etc.)
- MeSH terms are used in PubMed to look for the search term in a specified heading group, rather than just key words, to return more relevant results.

These terms can be grouped together to narrow a literature search that otherwise may produce an excess of irrelevant results. For example, the search syntax for ethylene glycol may look like this:

"ethylene glycol" [mesh] NOT "ethylene oxide" AND (inhal OR air OR carc* OR onco*)*

This search string identifies studies with the keywords ethylene and glycol together in a medical subject heading, excludes studies referring to ethylene oxide, and includes only the studies that use a form of inhal* (inhale, inhalation), air, carc* (carcinogenic, carcinogen), or onco* (oncogenesis, oncogenicity).

Search validation may be performed as a preventive measure to ensure a comprehensive search. This process compares the highly relevant publications noted during problem formulation to the bibliographic database results. If publications are not identified in the results of the search, the syntax should be revised to capture missing concepts that expand the search and lead to the inclusion of these relevant publications. This process will be documented by including citations of publications used for validation, and any resulting changes to the syntax or overall search strategy.

Targeted searching

Through the course of the review, additional publications, studies, or data may be identified while screening publications, in authoritative reviews, etc. Another consideration for targeted searching is the use of resources such as Connected Papers, Research Rabbit, or PubMed's Similar Article feature. Citations closely related to the specified publication are generated based on similarity in metrics such as title, abstract, key words, and overlapping citations. In this process, the additional publications will be added to the workflow and screened using the same criteria as the literature identified from other sources.

Literature Screening

The literature screening process will be performed at the title/abstract and full-text levels using screening software (e.g., SWIFT ActiveScreener, Health Assessment Workspace Collaborative [HAWC]). Prior to initiating the formal screening phase, piloting exercises will be undertaken to gauge the adequacy of the review forms and inclusion criteria. Iterative refinements to the process may be made based on the outcome of piloting exercises. This may include developing additional inclusion or exclusion criteria, or guidance to support the reviewer in determining eligibility. As mentioned previously, deviations from the original SR protocol will be documented, including the modification of criteria. Piloting also contributes to reviewer calibration, by facilitating discussion of aspects of the risk assessment question and protocol that may be interpreted differently by reviewers.

In both title/abstract and full-text screening, each study will be evaluated in accordance with the specified inclusion criteria to determine eligibility for review. For example, can it be determined from the title/abstract that the chemical of interest was investigated under the specific inclusion/exclusion criteria? If the answer is yes, the study may move on to full-text review but if not, the study is excluded due to a lack of relevance to the PECO, PECOTA, or PCC outlined. Some criteria may be uncertain at the title/abstract level, in which case the publication should be advanced to full text for further review. Two reviewers will perform screening; conflicts between the reviewers will be addressed in discussion. Should reviewers be unable to come to a resolution, a third reviewer will review the study for consideration in conflict resolution. Justification for studies excluded at full text will be documented. The results of the screening process will also be documented in the form of a literature flow diagram, such as a PRISMA chart (Moher et al., 2009). An example of this is provided in **Figure 3.**

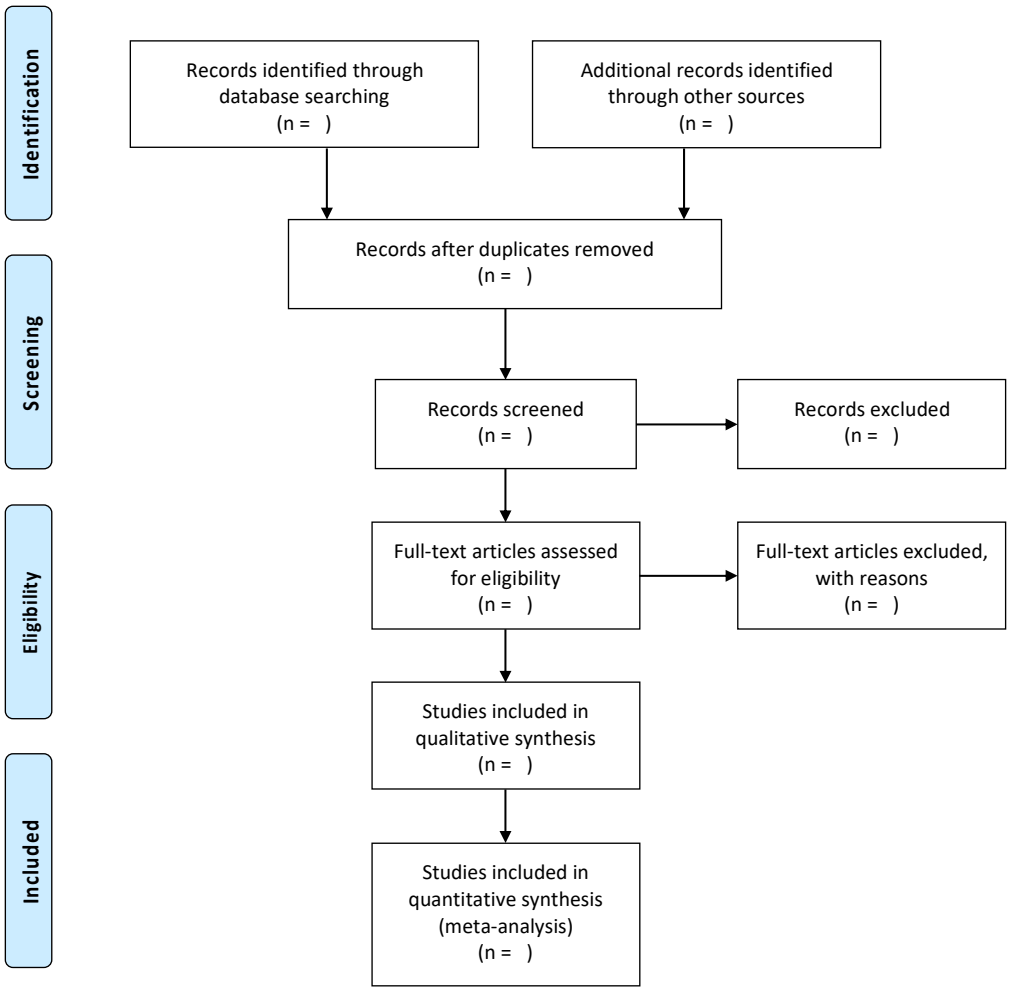

Figure 3. PRISMA flow diagram adapted from Moher et al. (2009) used in systematic review reporting to increase transparency and reproducibility of the results of evidence identification

In the case of an SEM, the literature may also be inventoried by study characteristics at the literature screening stage. Depending on the assessment needs, this inventory may include categories such as evidence stream, health outcome, exposure duration, route of exposure, or other aspects of study design. This will allow the assessment team to evaluate the available database and prioritize the outcome(s) or endpoint(s) of relevance to carry forward to the SR.

Use of AI/ML-based literature review tools

Tools that integrate AI and machine learning (ML) models to facilitate literature review are becoming increasingly common in evidence-based toxicology. As confidence and validation of these technologies increase, they may be used by the agency to assist with SR efforts. For

example, software to help prioritize specific citations for consideration based on real-time screening and development of a training set can be used to conserve resources during title/abstract screening of a large body of literature. This active learning feature is available in two commonly used SR literature screening tools (SWIFT ActiveScreener and DistillerSR) that have been used by TCEQ to facilitate the timely screening of large literature title and abstract datasets for relevant studies. The tool iteratively updates its model as reviewers screen articles and labels them as included or excluded. Based on this, citations most likely to be relevant (as predicted by the algorithm) are prioritized for review. During the course of manual review, the tool determines when reviewers have reached the appropriate threshold indicating that all relevant studies have likely been identified. The agency will consider use of these tools on a case-by-case basis and may consider such tools to be of greater utility for scoping vs. assessment (e.g., in determining volume of evidence base or general nature of mechanistic data).

Assessing Relevance to Risk Assessment

Once screened, it may become evident that the potential body of literature is data rich and/or diverse. In such a case, it is commonly understood that evidence may vary in its applicability to the development of a toxicity factor, but that inclusion/exclusion criteria may not sufficiently differentiate such elements. In this (or similar) scenarios, it may be appropriate for risk assessors to identify specific relevance criteria to further narrow the focus of the assessment and allow for the prioritization of studies most relevant for review and consideration in the toxicity factor derivation process. As described in RG-422 (TCEQ, 2015), the risk assessor should prioritize studies that correspond to the type of toxicity value under development. For example, if broad inclusion/exclusion criteria result in the identification of studies that evaluate all study durations, and the goal of the assessment is a 1-h acute ReV, it would be appropriate to narrow the evidence base carried forward to extraction and evaluation of acute exposure studies, once it was confirmed with the systematic search that sufficient studies were available to do so. Other considerations for assessing the relevance of a study duration in context of the particular toxicity value are described in Section 3.2 of RG-422.

As further examples, at this juncture in the SR process, the assessment team may also elect to narrow the evidence base using study attributes that are important to dose-response assessment and toxicity factor development. This may include prioritizing epidemiological studies that have used study designs that sufficiently limit the influence of chance, bias, and confounding, and/or that limit studies carried forward to those in which causation between the specific exposure and the outcome can be established sufficiently with the underlying evidence base. This may also include prioritizing outcomes based on adversity or on study type or reliability (e.g., prioritization of guideline-based studies).

Data Extraction

Studies that meet the inclusion criteria are advanced to data extraction, where adverse health endpoint data are summarized into evidence tables (e.g., HAWC). Each study will undergo extraction by a single reviewer and will be reviewed for quality control (QC) by a second reviewer for accuracy. Data extraction may be performed independently of, or simultaneously during full-text inclusion screening and/or study quality assessment, to expedite the review process. The strategy for data extraction, including fields to be considered for the evidence table(s), will be developed during problem formulation and described in the protocol. Some SR tools, such as HAWC, have standardized fields in pre-populated forms to facilitate the data extraction.

Prior to initiating the full data extraction effort, the review team will perform several rounds of piloting with personnel assigned to perform the data extraction. During pilot exercises, a small but diverse set of publications across evidence streams, exposure scenarios, and outcomes (where applicable) will be extracted. Comparison of data extraction tables and discussion of consensus responses will be performed. Issues raised during piloting and reviewer calibration should be addressed with revisions to the forms, guidance, or process.

The expected output for the data extraction phase are evidence tables of all relevant, extracted data including study design, exposure parameters, and study findings. **Table 5** and **Table 6** are simple examples of data extracted from a human and animal toxicology study, respectively. The purpose of these tables is to briefly summarize the available data in the literature, identify potential trends in PODs, and provide a basis for using the data to select key and supporting studies. More extensive data extraction tables may be required for data-rich chemicals to fully characterize the available data, including columns for study design, study size, exposure characterization and/or tested levels, outcome categories, the type of statistical analyses performed, and results. These more extensive fields can be retained in the database, while a simplified table such as those below can provide details necessary for purposes of the final DSD.

Table 5. Example data extraction table for epidemiology studies

Reference	Study Type	Species/n/Sex	Exposure Concentration (µg/m^3)	Exposure Duration	Health Outcome Examined	Highest level of no statistically significant effect or association	Lowest level of statistically significant effect or association	Notes
Ostro et al. (2007)	Ecological	Humans/ approx. 8,700,000/ male and female	0.002 (mean); 46.91 (95th percentile)	4 years	Mortality (all cause, CVD, or respiratory)	0.002 µg/m^3 (mean)	None observed	Percent change in mortality per 1 µg/m^3 increase reported, no statistically significant excess risk reported.

Abbreviation: CVD, cardiovascular disease

Table 6. Example data extraction table for animal toxicology studies

Reference	Study Type	Species/n/Sex	Exposure Concentrations (µg/m^3)	Most Sensitive Health Outcome Examined	NOAEL	LOAEL	Notes
NTP et al. (2002)	Subchronic Inhalation (13 weeks - 6 hr/d, 5 d/wk)	Mouse/10/ male and female	0, 1, 2, 4, 8, 16	Non-neoplastic lesions	1 mg/m^3	2 mg/m^3	Statistically significant changes also reported for organ weights, mean body weight, male and female reproductive endpoints and survival (males more sensitive than females)

Data extraction will differ for each data stream (e.g., experimental animal vs. epidemiological) because of differences in study design and methods. NTP OHAT's Handbook (2019) provides a comprehensive list of key data extraction elements that are typically recorded for studies; resources such as this, and previous experience, will be considered during problem formulation to inform the assessment team's decisions for data extraction.

It is standard to develop multiple data extraction templates to fit the various evidence streams included in the assessment. Epidemiology studies may include experimental and observational (analytical and descriptive) data. Experimental animal toxicity studies are conducted to determine dose response, and are usually conducted for specific durations (i.e., acute, subchronic, chronic), or to study a specific effect (e.g., carcinogencity, reproductive, developmental, neurological). Mechanistic studies, which may be based on *in vivo* or *in vitro* model designs, are often conducted to determine genotoxic potential, cell transformation, or cytotoxicity, or to understand the MOA. These studies, particularly the *in vitro* studies, are often difficult to extrapolate to human-relevant exposures. Inclusion of mechanistic studies in data extraction is dependent on the PECO, and mechanistic data may be used to support the overall risk assessment.

Study Reliability Assessment

Studies that meet the inclusion criteria for the SR will be critically evaluated for study reliability (or "quality"). This is a broad term that covers concepts of validity, sensitivity, and reporting quality (NRC, 2014; WHO, 2021). Validity in the context of this guidance is intended to include internal validity (i.e., risk of bias), external validity (i.e., generalizability), and construct validity (i.e., "fitness for purpose"), as is appropriate for each toxicity factor derivation process. Currently, there is no consensus on the best practice for evaluating study quality for risk assessment (Wikoff et al., 2020; WHO, 2021). Many of the tools developed and published to assess study quality in evidence-based toxicology have been adapted from tools specific to evidence-based medicine. However, well-studied domains in the field of evidence-based medicine can differ significantly from aspects of studies important to risk assessment. As a result, these tools are limited in their ability to assess study quality concepts for risk assessment. Wikoff et al. (2020) recommends developing a "fit-for-purpose" approach for each assessment by addressing the concepts of study quality in the protocol. Importantly, the method will be applied consistently and transparently to eligible studies. Section 3.3.3.1 of TCEQ RG-442 (2015) briefly describes study quality assessments that will include data quality evaluations, considering method validity, reproducibility, study reliability, dose-response relationships, temporal associations between exposures and adverse health effects, and whether critical effects are relevant to humans. Study reliability frameworks, or tools, commonly used for chemical risk assessment are presented below in **Table 7**. These frameworks can be used as standalone approaches for study reliability or in combination with each other, depending on the needs of the assessment.

Table 7. Common frameworks used for assessing study reliability concepts in chemical risk assessment

Framework	Evidence Types	Quality Concepts	Output
USEPA IRIS	Human, experimental animal	Internal validity; reporting + sensitivity	Heat map; categorical assignments (including uninformative)
USEPA TSCA	Human health, ecotoxicity, physiologically-based pharmacokinetic model (PBPK), exposure, and more	Combines some aspects of internal, external, and construct (includes reporting)	Categories (based on numerical assignments)
OHAT Risk of Bias	Human, experimental animal	Internal validity; reporting (limited)	Heat map by domain; optional categorical assignments of studies into Tiers
ToxRTool	Experimental animal, in vitro	Focuses on reporting; indirect consideration of some validity concepts	Klimisch categories
SciRAP	Experimental animal, in vitro	Combines some aspects of internal, external, and construct (includes reporting)	Visualized percentages of criteria fulfilled

Another scoring system frequently used in regulatory risk assessment is that developed by Klimisch et al. (1997), which describes four categories of reliability. In this framework, studies that were conducted and reported according to accepted test guidelines (e.g., USEPA, OECD) and in compliance with good laboratory practice (GLP) are considered to have the highest reliability. This conceptual framework is particularly useful for prioritizing and/or differentiating studies for further review, as well as in categorizing studies as key and supporting studies, when sufficient evidence is available to do so. When such categories do not apply, such as in the use of observational studies in humans, tools and approaches such as that discussed by LaKind et al. (2023) may be considered.

Because of the limitations of tools to assess study quality at the time of writing this guidance, no specific method is recommended. Rather, it is recommended that the assessment team apply guidance for the tool most relevant to the data types being evaluated. In practice, the study quality evaluation should also include a pilot phase during which assessors apply the

refined tool, provide feedback and discussion, and ensure judgements are aligned across the assessment team. The standard output of the evaluation includes a heat map of all studies, judgements for individual metrics, and the overall study confidence score**..** As in other phases of the SR process, the assessment team should undertake a QC of the study quality evaluations. Judgements that the QC lead does not agree with should be discussed and documented.

Considerations that may be prioritized for study reliability assessments are presented in **Table 8** and include concepts linked to internal, external, and construct validity, sensitivity, and reporting quality. The assessment team will discuss these considerations in the context of the evaluation's priorities while considering the topic of the assessment, included evidence streams, volume of literature to be assessed, and resources such as timeline needs. The determined approach or selected framework for assessing these topic areas will be described clearly in the protocol.

Table 8. General concepts for assessment of study reliability in systematic review and ideal study attributes for each

Reliability Concept	Ideal Study Attribute
Range of doses/exposures	• Study examines more than two dose/exposure concentrations. • Doses/concentrations are biologically and environmentally relevant.
Selection bias	• Acceptable methods of randomization are reported. • Study specifically states that blind testing was used, when appropriate. • Baseline characteristics are comparable for exposed and comparator groups.
Exposure confidence	• Valid, reliable, and sensitive methods were used to measure exposure. • Appropriate characterization of the chemical is reported.
Outcome assessment	• Valid, reliable, and sensitive methods were used to assess the outcome.
Selective reporting bias	• All measured outcomes were reported with adequate detail to perform an independent analysis.
Reporting quality	• Study design clearly defined and detailed in methods. • Study provides enough detail to assume quality, uniformity, consistency, and reproducibility.
Confounding factors	• Study eliminates or controls for any possible confounding factors or covariates, including outcome-specific variables, as well as exposure variables.
Control group	• Concurrent vehicle controls are reported, including sham treatments where relevant.

Evaluations of study reliability and validity are of particular importance in determining confidence in the epidemiological evidence. Three key steps for characterizing reliability of epidemiological findings include: 1) characterizing and understanding the study design, or construct; and 2) use of critical appraisal to qualitatively identify the potential for bias or limitations that might impact study conclusions (risk of bias assessment), and 3) when appropriate, further characterizing the potential impact (direction, magnitude) of potential biases.

Evidence Synthesis, Integration, and Derivation of Chemical Toxicity Values

Evidence synthesis is summarizing the evidence collected during the prior steps of the SR into a format that facilitates the integration and/or interpretation of the available data. This includes the interpretation of evidence within each data stream (animal, human, mechanistic), and ultimately, interpretation of the combined evidence. As a general matter, formal hazard assessments such as that described by NTP OHAT (2019) are not part of the TCEQ's DSD process, which is focused on the derivation of chemical toxicity values. However, this may be incorporated into the assessment on a case-by-case basis.

Ultimately, the synthesis and integration process is guided by the TCEQ's Guidelines to Develop Toxicity Factors (2015), and typically results in a quantitative estimate of risk (e.g., ReV, URF, RfD, SFo). Using the available data identified and assembled in the SR, key and supporting studies will be identified based on attributes of study design, study reliability, and needs of dose-response modeling. The TCEQ DSD process provides guidance on evaluating the weight of evidence—including factors such as toxicokinetics and MOA that can affect human relevance and quantitative dose response, as well as study quality and database confidence (TCEQ, 2015). Although outside the SR framework, the decisions made in the development of chemical toxicity values based on these factors are determined using expert judgment and the TCEQ's DSD guidance.

The anticipated method, consistent with the TCEQ's 2015 Guidelines, will be defined *a priori* during the problem formulation stage (Wikoff et al., 2020; WHO, 2021). This process will ensure identification of evidence-based and defensible toxicity values supported by sufficient documentation of these decision points. Consequently, this approach will support the development of chemical risk assessments that are transparent, consistent, and reliable, conferring confidence in the DSD process.

Reporting

This framework is established to support developing toxicity factors and the DSD process; therefore, reporting of the SEM and/or SR will accompany the associated DSD. This will include the protocol, explicit methods (including deviations from the protocol), outcome of the evidence identification stage, data extraction tables, and the study quality assessment.

References

EFSA (European Food Safety Authority). (2010). Application of systematic review methodology to food and feed safety assessments to support decision making. *EFSA 8*(6), 1637. doi:10.2903/j.efsa.2010.1637.

Guzelian PS, Victoroff MS, Halmes NC, James RC, & Guzelian CP. (2005). Evidence-based toxicology: A comprehensive framework for causation. *Human & Experimental Toxicology*, *24*(4), 161–201. https://doi.org/10.1191/0960327105ht517oa.

Klimisch HJ, Andreae M, & Tillmann U. (1997). A systematic approach for evaluating the quality of experimental toxicological and ecotoxicological data. *Regul Toxicol Pharmacol, 25*(1), 1-5. https://doi.org/10.1006/rtph.1996.1076.

LaKind JS, Burns CJ, Johnson GT, & Lange SS. (2023). Epidemiology for risk assessment: The US Environmental Protection Agency quality considerations and the Matrix. *Hygiene and Environmental Health Advances, 6*, 100059. https://doi.org/10.1016/j.heha.2023.100059.

Meek MEB & Wikoff D. (2023, Jun 28). The need for good practice in the application of mechanistic constructs in hazard and risk assessment. *Toxicol Sci, 194*(1), 13-22. https://doi.org/10.1093/toxsci/kfad039.

Moher D, Liberati A, Tetzlaff J, & Altman DG. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA Statement. *PLoS Medicine*, 6(7):e1000097. https://doi.org/10.1371/journal.pmed.1000097.

Morgan RL, Whaley P, Thayer KA, & Schunemann HJ. (2019). Identifying the PECO: A framework for formulating good questions to explore the association of environmental and other exposures with health outcomes. *Environment International, 121*(Pt 1), 1027-1031. https://doi.org/10.1016/j.envint.2018.07.015.

NRC (National Research Council). (2001). Standing Operating Procedures (SOP) of the National Advisory Committee on Acute Exposure Guideline Levels for Hazardous Substances (PDF). www.epa.gov/opptintr/aegl/pubs/sop.pdf

NRC (National Research Council). (2014). Review of the Environmental Protection Agency's Integrated Risk Information System (IRIS) process. Washington (DC): National Academies Press; 2014.

NTP OHAT (National Toxicology Program, Office of Health Assessment and Translation). (2019). Handbook for conducting a literature-based health assessment using OHAT approach from systematic review and evidence integration. OHAT, Division of the National Toxicology Program, National Institute of Environmental Health Sciences. Available at: https://ntp.niehs.nih.gov/sites/default/files/ntp/ohat/pubs/handbookmarch2019_508.pdf

Schaefer HR & Myers JL. (2017). Guidelines for performing systematic reviews in the development of toxicity factors. *Regulatory Toxicology and Pharmacology*, 91, 124–141. https://doi.org/10.1016/j.yrtph.2017.10.008.

TCEQ (Texas Commission on Environmental Quality). (2015). Guidelines to develop toxicity factors (RG-442). Chief Engineer's Office. Available at: https://www.tceq.texas.gov/toxicology/esl/guidelines/about

TCEQ (Texas Commission on Environmental Quality). (2017). TCEQ Guidelines for systematic review and evidence integration. White Paper. Toxicology Division.

Uman L. (2011). Systematic reviews and meta-analyses. Journal of the Canadian Academy of Child & Adolescent Psychiatry, 20(1): 57-59. PMID: 21286370.

USEPA (U.S. Environmental Protection Agency). (1994). *Methods for Derivation of Inhalation Reference Concentrations and Application of Inhalation Dosimetry.* EPA/600/8-90/066F. Washington, DC. Available at: https://www.epa.gov/risk/methods-derivation-inhalation-reference-concentrations-and-application-inhalation-dosimetry.

USEPA (U.S. Environmental Protection Agency). (2005). United States Environmental Protection Agency. Science Issue Paper: Mode of Carcinogenic Action for Cacodylic Acid (Dimethylarsinic Acid, DMA) and Recommendations for Dose Response Extrapolation. Washington, DC. Available at: https://archive.epa.gov/pesticides/reregistration/web/pdf/dma_moa-2.pdf

USEPA (U.S. Environmental Protection Agency). (2022). *Office of Research and Development (ORD) Staff Handbook for Developing IRIS Assessments.* EPA/600/R-22/268. Washington, DC. Available at: https://cfpub.epa.gov/ncea/iris_drafts/recordisplay.cfm?deid=356370

WHO (World Health Organization). (2021). *Framework for the use of systematic review in chemical risk assessment*. ISBN 978-92-4-003448-8. Geneva: World Health Organization. Available at: https://www.who.int/publications/i/item/9789240034488