



ORD Staff Handbook for Developing IRIS Assessments

Office of Research and Development
Center for Public Health & Environmental Assessment



EPA/600/R-22/268
www.epa.gov/ord

ORD Staff Handbook for Developing IRIS Assessments

December 2022

Center for Public Health and Environmental Assessment
Office of Research and Development
U.S. Environmental Protection Agency
Washington, DC

Disclaimer

This document has been reviewed by the U.S. Environmental Protection Agency, Office of Research and Development and approved for publication. Any mention of trade names, products, or services does not imply an endorsement by the U.S. government or the U.S. Environmental Protection Agency. EPA does not endorse any commercial products, services, or enterprises.

ACKNOWLEDGMENTS

The following individuals and groups were instrumental in developing this handbook:

Primary Authors (currently or formerly EPA/ORD)

[Michelle Angrish](#)

[Xabier Arzuaga](#)

Vincent Cogliano

Glinda Cooper

[Allen Davis](#)

[Laura Dishaw](#)

[Catherine Gibbons](#)

Barbara Glenn

Karen Hogan

Samantha Jones

[Andrew Kraft](#)

April Luke

[Elizabeth Radke](#)

[Alan Sasso](#)

[Rachel M. Shaffer](#)

[Michele M. Taylor](#)

[Kristina Thayer](#)

Teneille Walker

[George Woodall](#)

[Erin Yost](#)

Additional Contributors (current or former EPA/ORD)

Norman Birchfield

[Francesca Branch](#)

Johanna Congleton

Jeffrey Dean

[Ingrid Druwe](#)

John Fox

Jason Fritz

John Lipscomb

Lucina Lizarraga

Roman Mezencev

[Margaret Pratt](#)

Susan Rieth

[Paul Schlosser](#)

[Ravi Subramaniam](#)

Production Team (EPA/ORD)

Maureen Johnson
Ryan Jones
Dahnish Shams
Andy Shapiro
Jessica Soto-Hernandez
Vicki Soto
Samuel Thacker
Sean Watford

Reviewers

EPA thanks the following reviewers for their thoughtful comments and suggestions on earlier drafts of this document:

John Bucher, National Toxicology Program
Barbara Buckley, EPA/ORD
Ila Cote, formerly EPA/ORD
Kathryn Guyton, International Agency for Research on Cancer
Ruth Lunn, National Toxicology Program, Report on Carcinogens
Jonathan Samet, Colorado School of Public Health

CONTENTS

PREFACE	xiii
OVERVIEW AND INTRODUCTION TO THE HANDBOOK FOR DEVELOPING INTEGRATED RISK INFORMATION SYSTEM (IRIS) ASSESSMENTS	xvi
1. SCOPING AND INITIAL PROBLEM FORMULATION	1-1
1.1. SCOPING	1-1
1.1.1. Key Considerations That Determine the Scope of an Assessment.....	1-2
1.2. INITIAL PROBLEM FORMULATION	1-3
1.2.1. Survey of Existing Assessments and Toxicity Values	1-4
1.2.2. Systematic Evidence Maps (SEMs)	1-5
1.2.3. Identification of Key Science Issues.....	1-7
1.3. KEY DOCUMENTS: IRIS ASSESSMENT PLAN (IAP) AND SYSTEMATIC REVIEW PROTOCOL.....	1-7
1.3.1. IRIS Assessment Plan (IAP)	1-8
1.3.2. Systematic Review Protocol	1-9
2. LITERATURE SEARCH, SCREENING, AND INVENTORY	2-1
2.1. POPULATIONS, COMPARATORS, EXPOSURES, OUTCOMES (PECO) CRITERIA	2-1
2.2. SUPPLEMENTAL MATERIAL SCREENING CRITERIA.....	2-3
2.3. LITERATURE SEARCH STRATEGIES.....	2-8
2.3.1. Health and Environmental Research Online (HERO).....	2-8
2.3.2. Core Database Searches	2-11
2.3.3. Additional Database Searches	2-15
2.3.4. Removing Duplicates	2-20
2.3.5. Updating the Literature Search	2-21
2.3.6. Documenting Search Results.....	2-21
2.4. LITERATURE SCREENING	2-22
2.4.1. Title and Abstract Screening.....	2-22
2.4.2. Full-Text Screening	2-23
2.4.3. Multiple Publications of the Same Data	2-25
2.4.4. Systematic Review Software and Artificial Intelligence (AI) Tools	2-25
2.4.5. Literature Flow Diagrams	2-30
2.5. LITERATURE INVENTORIES.....	2-33

ORD Staff Handbook for Developing IRIS Assessments

2.5.1. Literature Inventory of Studies Meeting Populations, Exposures, Comparators, and Outcomes (PECO) Criteria	2-34
2.5.2. Supplemental Content Literature Inventories	2-35
3. REFINE PROBLEM FORMULATION AND SPECIFY ASSESSMENT APPROACH.....	3-1
3.1. ASSESSMENT POPULATIONS, EXPOSURES, COMPARATORS, AND OUTCOMES (PECO) CRITERIA.....	3-1
3.2. DEFINING UNITS OF ANALYSIS	3-2
3.3. CONSIDERATION OF SUPPLEMENTAL MATERIAL	3-5
4. STUDY EVALUATION	4-1
4.1. STUDY EVALUATION OVERVIEW FOR HEALTH EFFECT STUDIES.....	4-1
4.1.1. Evaluation Ratings	4-4
4.1.2. Documenting Study Evaluations.....	4-6
4.2. EVALUATION OF EPIDEMIOLOGICAL STUDIES	4-9
4.2.1. Development of Evaluation Considerations.....	4-10
4.2.2. Final Observations	4-30
4.3. EVALUATION OF CONTROLLED HUMAN EXPOSURE STUDIES	4-30
4.4. EVALUATION OF EXPERIMENTAL ANIMAL TOXICOLOGICAL STUDIES	4-31
4.5. PRIORITIZATION AND EVALUATION OF NON-POPULATIONS, EXPOSURES, COMPARATORS, AND OUTCOMES (PECO) STUDIES	4-40
4.5.1. Prioritization of Non-Populations, Exposures, Comparators, and Outcomes (PECO) Studies	4-40
4.5.2. Evaluation of Non-Populations, Exposures, Comparators, and Outcomes (PECO) Studies	4-41
4.5.3. Evaluation of In Vitro Studies	4-42
4.6. EVALUATION OF EXISTING COMPUTATIONAL PHYSIOLOGICALLY BASED PHARMACOKINETIC MODELS	4-52
5. EXTRACTION AND DISPLAY OF STUDY RESULTS FROM EPIDEMIOLOGICAL AND TOXICOLOGICAL STUDIES.....	5-1
5.1. DATA EXTRACTION.....	5-2
5.1.1. Health Assessment Workspace Collaborative (HAWC)	5-3
5.1.2. Quality Control during Data Extraction	5-4
5.1.3. Best Practices for Data Extraction in Health Assessment Workspace Collaborative (HAWC) and Tabular Presentation	5-5
5.2. STANDARDIZING REPORTING OF OUTCOME MEASURES	5-9
5.3. STANDARDIZING ADMINISTERED DOSE LEVELS/CONCENTRATIONS.....	5-11
5.4. GENERAL PRINCIPLES FOR PRESENTING EVIDENCE.....	5-11

ORD Staff Handbook for Developing IRIS Assessments

5.5. GRAPHICAL AND TABULAR DISPLAY	5-12
5.5.1. GRAPHICAL DISPLAY	5-12
5.5.2. TABULAR DISPLAY.....	5-19
6. EVIDENCE SYNTHESIS AND INTEGRATION.....	6-1
6.1. EVIDENCE SYNTHESIS.....	6-6
6.1.1. Considerations for Developing the Human and Animal Evidence Syntheses	6-16
6.1.2. Approaches to Facilitate Evidence Synthesis of Mechanistic Studies	6-21
6.2. EVIDENCE INTEGRATION.....	6-24
6.2.1. Considerations That Inform Evidence Integration	6-25
6.2.2. Evidence Integration Judgment.....	6-32
7. HAZARD CONSIDERATIONS AND STUDY SELECTION FOR DERIVING TOXICITY VALUES.....	7-1
7.1. HAZARD CONSIDERATIONS FOR DOSE-RESPONSE	7-2
7.2. SELECTION OF STUDIES.....	7-4
7.2.1. SYSTEMATIC ASSESSMENT OF STUDY ATTRIBUTES TO SUPPORT DERIVATION OF TOXICITY VALUES.....	7-4
7.2.2. COMBINING DATA FOR DOSE-RESPONSE MODELING	7-9
8. DERIVATION OF TOXICITY VALUES	8-1
8.1. SELECTING BENCHMARK RESPONSE VALUES FOR DOSE-RESPONSE MODELING.....	8-2
8.2. CONDUCTING DOSE-RESPONSE MODELING.....	8-3
8.2.1. Exposure-Response Modeling of Human Data.....	8-4
8.2.2. Exposure-Response Modeling of Animal Data	8-5
8.2.3. Composite Risk	8-8
8.2.4. Tools and Documentation to Support Dose-Response Modeling.....	8-9
8.3. DEVELOPING CANDIDATE TOXICITY VALUES	8-10
8.3.1. Linear Low-Dose Extrapolation	8-10
8.3.2. Nonlinear Low-Dose Extrapolation	8-11
8.4. CHARACTERIZING UNCERTAINTY AND CONFIDENCE IN TOXICITY VALUES	8-16
8.4.1. Uncertainty in Toxicity Values	8-16
8.4.2. Characterizing Confidence.....	8-17
8.5. SELECTING FINAL TOXICITY VALUES.....	8-18
8.5.1. Organ/System-specific Toxicity Values.....	8-18
8.5.2. Overall Toxicity Values	8-19
REFERENCES.....	R-1

ORD Staff Handbook for Developing IRIS Assessments

APPENDIX A. GLOSSARYA-1

APPENDIX B. SURVEY OF EXISTING ASSESSMENTS AND TOXICITY VALUES.....B-1

APPENDIX C. EXAMPLE ISSUES FROM EXISTING INTEGRATED RISK INFORMATION SYSTEM (IRIS)
ASSESSMENTS C-1

APPENDIX D. SUPPLEMENTAL DATABASES..... D-1

APPENDIX E: ESTIMATING TIME TO CONDUCT THE ASSESSMENT E-1

TABLES

Table O-1. Orientation to Integrated Risk Information System (IRIS) assessment development	xix
Table 2-1. Components of populations, exposures, comparators, and outcomes (PECO) and potential types of evidence	2-2
Table 2-2. Example categories of “Potentially Relevant Supplemental Material” (from the Integrated Risk Information System [IRIS] Assessment Plan template)	2-4
Table 2-3. Core databases of published studies (searched by Health and Environmental Research Online [HERO] or contractors)	2-11
Table 2-4. Summary of search term development strategies for core databases	2-13
Table 2-5. Additional strategy resources for literature identification	2-16
Table 2-6. Example summary template of literature search results documentation.....	2-22
Table 2-7. Summary of commonly used specialized software applications for literature screening and visualization	2-27
Table 3-1. Example units of analysis	3-3
Table 4-1. Example question specification for evaluation of exposure measurement in epidemiological studies	4-12
Table 4-2. Example question specification for evaluation of outcome in epidemiological studies	4-15
Table 4-3. Example question specification for evaluation of participant selection in epidemiological studies	4-17
Table 4-4. Example question specification for evaluation of confounding in epidemiological studies.....	4-20
Table 4-5. Example question specification for evaluation of analysis in epidemiological studies	4-24
Table 4-6. Example question specification for evaluation of selective reporting in epidemiological studies.....	4-27
Table 4-7. Example question specification for evaluation of sensitivity in epidemiological studies.....	4-29
Table 4-8. Domains, questions, and general considerations to guide the evaluation of animal studies.....	4-32
Table 4-9. Domains, questions, and general considerations to guide the evaluation of in vitro studies.....	4-43
Table 5-1. Example epidemiological summary table of selected data on exposure antibody response to vaccines in children.....	5-20
Table 5-2. Example animal summary table showing percent change of liver weight	5-22
Table 6-1. Generalized evidence profile table showing the relationship between evidence synthesis and evidence integration to reach a judgment of certainty in the evidence for hazard	6-4
Table 6-2. Generalized evidence profile table to show the key findings and supporting rationale from mechanistic analyses.....	6-5
Table 6-3. Considerations that inform evidence synthesis judgments of the certainty in the animal or human evidence for hazard for each unit of analysis.....	6-8
Table 6-4. Framework for evidence synthesis judgments from studies in humans	6-13
Table 6-5. Framework for evidence synthesis judgments from studies in animals.....	6-14
Table 6-6. Considerations that inform the evidence integration judgment.....	6-25
Table 6-7. Framework for summary evidence integration judgments in the evidence integration narrative.....	6-33
Table 7-1. Factors that can increase susceptibility to exposure-related health effects	7-4
Table 7-2. Attributes used to evaluate studies for derivation of toxicity values.....	7-6

ORD Staff Handbook for Developing IRIS Assessments

Table A-1. Terms used in the Integrated Risk Information System (IRIS) HandbookA-1

Table B-1. Sources that can be queried for existing assessments and toxicity values, with
example search resultsB-1

Table C-1. Examples of key science issues in Integrated Risk Information System (IRIS)
assessmentsC-1

Table D-1. Supplemental databases that may be searched by the assessment team depending on
the topic..... D-1

Table E-1. Time estimates per study..... E-1

FIGURES

Figure O-1. Integrated Risk Information System (IRIS) assessment draft development process.....	xviii
Figure O-2. Stages in Integrated Risk Information System (IRIS) assessment development process.....	xix
Figure 1-1. Integrated Risk Information System (IRIS) systematic review problem formulation and method documents.....	1-8
Figure 1-2. Organization of the IRIS Assessment Plan (IAP).	1-9
Figure 2-1. Workflow for Health and Environmental Research Online (HERO)-facilitated literature searchers.....	2-10
Figure 2-2. Literature flow diagram: No machine learning (ML) software used.	2-31
Figure 2-3. Literature flow diagram: Machine learning (ML) software.	2-32
Figure 2-4. Health Assessment Workspace Collaborative (HAWC) literature tree.....	2-33
Figure 4-1. Overview of Integrated Risk Information System (IRIS) study evaluation approach	4-3
Figure 4-2. Examples of study evaluation displays at the individual level.....	4-7
Figure 4-3. Examples of study evaluation displays looking across studies	4-9
Figure 5-1. Examples of dose-response graphical displays for single endpoint created in Health Assessment Workspace Collaborative (HAWC) (for illustrative purposes only).	5-13
Figure 5-2. Examples of dose-response graphical displays across endpoints and studies created in Health Assessment Workspace Collaborative (HAWC) (for illustrative purposes only)	5-14
Figure 5-3. Examples of forest plots used for epidemiological evidence (for illustrative purposes only)	5-16
Figure 5-4. Examples of forest plots used for epidemiological evidence (for illustrative purpose only).	5-17
Figure 5-5. Examples of exposure-response arrays	5-19
Figure 6-1. Trichloroethylene (TCE) and kidney cancer: stratification by exposure level (U.S. EPA, 2011b).	6-17
Figure 8-1. Example summary of candidate toxicity values (for reference dose [RfD] derivation).....	8-15

ABBREVIATIONS

ADAF	age-dependent adjustment factors	NRC	National Research Council
ADME	absorption, distribution, metabolism, and excretion	NTP	National Toxicology Program
AI	Artificial Intelligence	OECD	Organisation for Economic Co-operation and Development
AOP	adverse outcome pathway	OHAT	Office of Health Assessment and Translation
AOPKB	adverse outcome pathway knowledge base	OR	odds ratio
APROBA	Approximate Probabilistic Analysis	ORD	Office of Research and Development
ATSDR	Agency for Toxic Substances and Disease Registry	PBPK	physiologically based pharmacokinetic
BMD	benchmark dose	PECO	populations, exposures, comparators, and outcomes
BMDL	benchmark dose lower confidence limit	PK	pharmacokinetic
BMDS	Benchmark Dose Software	POD	point of departure
BMR	benchmark response	PPRTV	Provisional Peer-Reviewed Toxicity Value
CASRN	Chemical Abstracts Service registry number	QA	quality assurance
CEBS	Chemical Effects in Biological Systems	QC	quality control
CI	confidence interval	RfC	reference concentration
CPHEA	Center for Public Health and Environmental Assessment	RfD	reference dose
DNA	deoxyribonucleic acid	RfV	reference value
EHV	Environmental Health Vocabulary	RIS	Restriction Information Systems
EPA	U.S. Environmental Protection Agency	RoB	risk of bias
GRADE	Grading of Recommendations Assessment, Development, and Evaluation	ROBINS-I	Risk of Bias in Nonrandomized Studies of Interventions
HAWC	Health Assessment Workspace Collaborative	SciRAP	Science in Risk Assessment and Policy
HDMI	the human dose associated with magnitude M of an adverse effect and incidence I in the population (also known as a “risk-specific dose”)	SD	standard deviation
HEC	human equivalent concentration	SE	standard error
HED	human equivalent dose	SEM	systematic evidence map
HERA	Health and Environmental Risk Assessment	SR	systematic review
HERO	Health and Environmental Research Online	StRAP	strategic research action plan
IAP	IRIS Assessment Plan	SVG	Scalable Vector Graphics
IPCS	International Programme on Chemical Safety	TCE	trichloroethylene
IRIS	Integrated Risk Information System	TIAB	title and abstract
LOAEL	lowest-observed-adverse-effect level	UF	uncertainty factor
MeSH	Medical Subject Headings	UF _H	intraspecies uncertainty factor
ML	machine learning	UF _L	LOAEL to NOAEL uncertainty factor
MOA	mode of action	UMLS	Unified Medical Language System
NAS	National Academy of Sciences	WHO	World Health Organization
NASEM	National Academy of Sciences, Engineering, and Medicine	WoS	Web of Science
NMD	normalized mean difference		
NOAEL	no-observed-adverse-effect level		

PREFACE

- The IRIS Program develops evidence-based, scientific human health assessments that focus on hazard identification and dose-response analyses for chemicals found in the environment.
- The IRIS Handbook provides operating procedures for developing assessments (Step 1 of the IRIS 7-step process: [IRIS process](#)) including incorporation of systematic review, hazard identification, and dose-response methods.
- The handbook does not supersede existing EPA risk assessment guidelines and does not serve as guidance for other EPA programs.

This *ORD Staff Handbook for Developing IRIS Assessments*, or IRIS Handbook, provides operating procedures to the scientists in the Integrated Risk Information System (IRIS) Program. Operating procedures are necessary for an efficient, productive, and consistent IRIS Program, which spans multiple organizational divisions and geographic locations. The handbook does not supersede existing U.S. Environmental Protection Agency (EPA) guidance and does not serve as guidance for other EPA programs. The handbook relies on and references a number of EPA guidelines and other recommendations. Please note that some of the URLs in this document are internal EPA sites and are not available publicly.

THE INTEGRATED RISK INFORMATION SYSTEM (IRIS) PROGRAM

The mission of EPA is to protect human health and the environment. EPA's IRIS Program plays an important role in helping EPA accomplish this mission through the development of human health hazard and dose-response assessments of potential health effects from exposure to environmental contaminants,¹ such as chemicals in drinking water, pollutants in air, and contaminants in soil. IRIS assessments are not regulations, but they may be considered influential scientific information that provide a critical part of the scientific foundation for decision-making to protect public health across EPA under an array of environmental laws (e.g., Clean Air Act; Safe Drinking Water Act; Comprehensive Environmental Response, Compensation, and Liability Act). IRIS assessments provide high quality, publicly available information on the toxicity of chemicals to which the public might be exposed and typically include human health hazard identification and

¹Although substances other than chemicals are assessed within the IRIS Program, "chemical" will be used as a shorthand throughout the remainder of this Handbook.

evaluation of dose-response² for those potential hazards, the first two steps in the risk assessment paradigm. IRIS assessments are used by EPA Programs and Regional Offices to complete the chemical-specific risk assessment process by factoring in exposure and conducting risk characterization. IRIS assessments may also be used by state regulators, tribes, and international entities.

SYSTEMATIC REVIEW IN INTEGRATED RISK INFORMATION SYSTEM (IRIS) ASSESSMENTS

IRIS assessments use the best available scientific information to answer the question(s) that are the focus of the review and strive to draw the conclusions that are best supported by the currently available data, even when the science is limited or incomplete. This general principle is consistent with the EPA Cancer Guidelines, which describes approaches for drawing judgments regarding the “available data” ([U.S. EPA, 2005a](#)), and supports the need for EPA customers to receive timely products from the IRIS Program. The transparency and scientific rigor of the IRIS process is enhanced through the application of systematic review.

The principles of systematic review have been well developed in the context of evidence-based medicine (e.g., evaluating efficacy in clinical trials) and more recently have been adapted for use across a more diverse array of scientific fields. The IRIS Program is helping advance the science of systematic review by improving the application of systematic review methods in the field of environmental health, which involves review of different types of studies compared with clinical medicine. The human studies available for IRIS assessments may cover diverse populations and exposure scenarios while varying in sensitivity. Animal studies generally include different experimental systems that may not be comparable. One challenge is to develop structured, reproducible procedures for aspects of IRIS assessments that are outside the usual domain of systematic review: evaluating mechanistic data and hypotheses, modeling pharmacokinetics and exposure-response relationships, and deriving toxicity values.

The IRIS Handbook implements recommendations from the National Research Council (NRC)/National Academy of Sciences (NAS), EPA’s Science Advisory Board (primarily during their review of IRIS assessment products³), and workshops involving expert practitioners of systematic review. In their 2014 review of the IRIS Program ([NRC, 2014](#)), NAS recommended the explicit inclusion of the principles of systematic review as a sequential process during Step 1 of the IRIS process. The IRIS Handbook has adapted the sequential process recommended by the National

²The IRIS Handbook uses the term “dose-response” generically to describe the relationship between an exposure and a health effect, regardless of the source or route of exposure, including internal dose as it impacts a target tissue. This term and others—including “low-dose extrapolation,” “dose-related trend,” “dose metric,” and “benchmark dose”—evolved in this more generic sense, most often in the context of laboratory animal experiments. The IRIS Handbook uses these terms as they originated, without limiting their use to oral exposures. Otherwise, the IRIS Handbook uses the term “exposure” to refer to any type of exposure pertinent to evaluating the impact of environmental exposure on human health.

³The Science Advisory Board also provided the following letter in response to a briefing encompassing the evolving handbook approaches in 2017: [2017 SAB Letter](#).

Academy of Sciences, Engineering, and Medicine (NASEM) as its underlying structural organization, which is described in the Overview chapter. This approach was further supported in a follow-up review of the IRIS Program’s systematic review methods by NASEM in 2018 ([NASEM, 2018](#)). The IRIS Program is committed to continued advancements of assessment methods; NASEM continues to hold workshops on topics pertinent to IRIS assessments, such as workshops on “Triangulation of Evidence in Environmental Epidemiology” and “Artificial Intelligence and Open Data Practices in Chemical Hazard Assessment” in 2022 (<https://www.epa.gov/iris/iris-and-national-academies-sciences-nas>). Advancements to IRIS assessment methodologies and tools are addressed in EPA’s Health and Environmental Risk Assessment (HERA) *Strategic Research Action Plan 2019–2022* (StRAP) (https://www.epa.gov/sites/default/files/2021-01/documents/hera_fy19-22_strap_final_2020_0.pdf).

In addition to implementing NASEM recommendations, the IRIS Handbook also reflects the IRIS Program’s experience with trying alternative approaches in many past and current assessments of varying scope and complexity. The IRIS Handbook clarifies and improves IRIS operating procedures in accordance with (and without changing) EPA guidance. The overall process of assessment development has not changed but is now supplemented by improved systematic review approaches that will help IRIS scientists retrieve, organize, evaluate, synthesize, integrate, and present scientific information in a more structured and transparent manner.

An overarching goal of these procedures is to promote an efficient and productive IRIS Program. An IRIS assessment is developed in alignment with the emphasis on tailoring risk assessments to inform the decision-making process in a meaningful way that is described in EPA’s *Framework for Human Health Risk Assessment to Inform Decision Making* ([U.S. EPA, 2014a](#)). The specific needs of a particular assessment are determined on the basis of scoping and problem formulation activities. The assessment objectives are focused to address the identified research question(s) and may include a modular approach (e.g., restrictions in scope or sequential development of assessments of specific health effects). The IRIS Handbook is intended to be a “living document”; the IRIS Program will update the IRIS Handbook as needed for major shifts in approaches based on emerging science and experience gained through its application to a broader spectrum of assessments.

OVERVIEW AND INTRODUCTION TO THE HANDBOOK FOR DEVELOPING INTEGRATED RISK INFORMATION SYSTEM (IRIS) ASSESSMENTS

The IRIS Handbook is intended to provide operating procedures for the development of Integrated Risk Information System (IRIS) assessments to promote consistency and ensure all contributors understand the methods used to develop the assessments, which include systematic review, dose-response, and application of U.S. Environmental Protection Agency (EPA) human health guidelines. The eight chapters in the IRIS Handbook describe each of the sequential stages involved in preparing a draft assessment (Step 1 of the IRIS process), as described at: <https://www.epa.gov/iris/basic-information-about-integrated-risk-information-system#process>).

ASSESSMENT DEVELOPMENT TASKS

A multidisciplinary assessment team develops each IRIS assessment and is responsible for all analyses and conclusions. The tasks of an assessment team include:

- Formulating the questions and key issues the assessment will address (e.g., scoping and problem formulation).
- Designing and implementing a systematic review process (i.e., systematic review protocol) that includes:
 - Populations, exposures, comparators, and outcomes (PECO) criteria that define the focus of the assessment.
 - Comprehensive literature search and screening strategies to address the identified questions and issues.
 - Evaluation of the studies that meet the PECO criteria using a systematic approach to identify strengths and limitations with regard to individual attributes for each study that can affect the confidence in the study results.
 - Development of syntheses of evidence for each evidence stream (i.e., human, animal, and specified questions about mechanisms).
 - Integration of the separate evidence streams to identify health hazards plausibly associated with the agent.
 - Selection of the data most informative for dose-response assessment.

- Deriving and characterizing toxicity values, when possible, for identified hazards of concern.
- Considering and addressing comments as the assessment moves through the review process.

Assessment teams generally comprise Office of Research and Development (ORD) scientists but can also include scientists from elsewhere in the EPA or expert consultants. Assessment teams also receive services from contractors on standardized tasks such as executing literature searches, creating a database of study details and results, and fitting standard dose-response models to data sets.

STAGES IN DEVELOPING A DRAFT ASSESSMENT

Figure O-1 summarizes the assessment development process from initial scoping through the derivation of toxicity values for identified hazards. The process builds from the recommended process described in Figure 1-2 in the National Academy of Sciences, Engineering, and Medicine (NASEM) review ([NRC, 2014](#)). The IRIS process applies a systematic review approach from the literature identification stage through the selection of studies for dose-response assessment. In addition to presenting stages ancillary to “systematic review,” including scoping, problem formulation, and dose-response assessment, this figure highlights that a single IRIS assessment typically involves multiple systematic reviews (e.g., different human health effects; different routes of exposure), each of which may involve different considerations and procedures.

The chapters in this IRIS Handbook follow the sequential stages in developing a draft IRIS assessment, as indicated by the schematic in Figure O-2. The result is a draft IRIS assessment that undergoes review in accordance with the 7-step [IRIS Process](https://www.epa.gov/iris/basic-information-about-integrated-risk-information-system#process) (<https://www.epa.gov/iris/basic-information-about-integrated-risk-information-system#process>). The topic of each chapter with a brief description is provided in Table O-1.

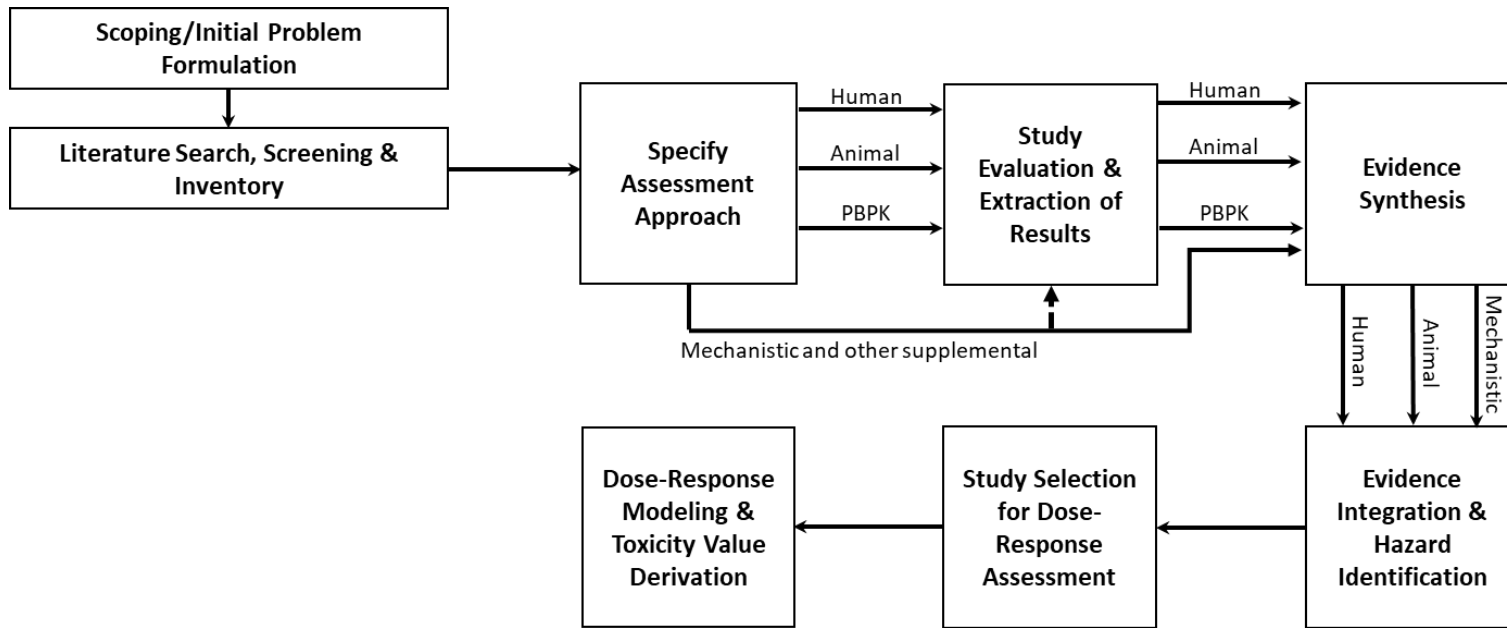


Figure O-1. Integrated Risk Information System (IRIS) assessment draft development process.

PBPK = physiologically based pharmacokinetic.

Building on the National Academy of Sciences, Engineering, and Medicine (NASEM) illustration for considering systematic review in the context of the IRIS process [see Figure 1-2 in [NRC \(2014\)](#)], the IRIS assessment development process outlined in this IRIS Handbook can be similarly depicted, with minor modifications (as shown). Steps in the IRIS Handbook process that may differ from the NASEM process are emphasized in red. The IRIS Handbook process encompasses all the steps in the figure; only those steps in the box are considered part of the systematic review. Mechanistic evidence may be incorporated at multiple stages of the process; this complexity is described in Chapter 6.

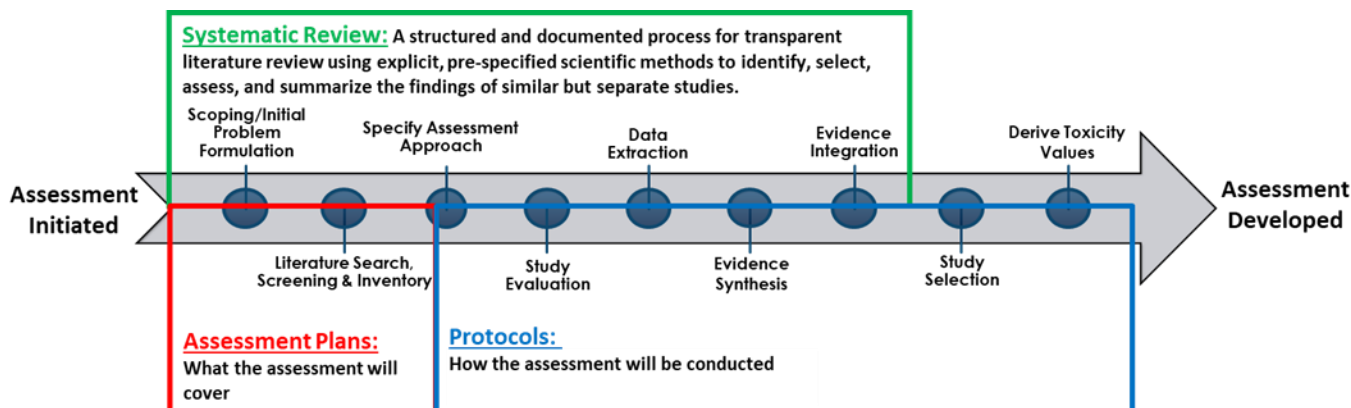


Figure O-2. Stages in Integrated Risk Information System (IRIS) assessment development process.

Table O-1. Orientation to Integrated Risk Information System (IRIS) assessment development

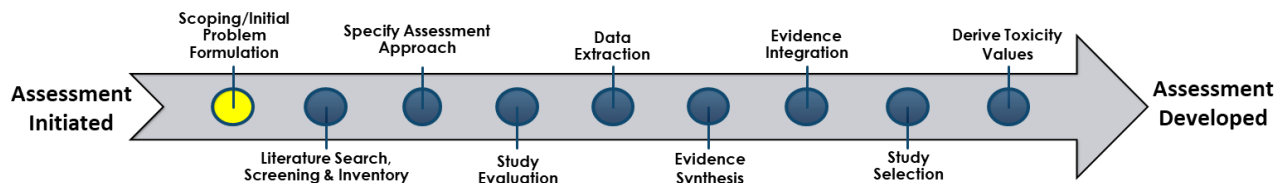
Assessment development stage	Chapter	Purpose and other useful details
Scoping, initial problem formulation <ul style="list-style-type: none"> Develop IRIS assessment plan Define problem formulation PECO 	1	Define the parameters of the assessment to meet EPA decision-making needs. Develop a systematic evidence map (using the literature search, screening, and inventory methods described in Chapter 2) and describe health effects of potential interest and key science issues. Develop problem formulation PECO criteria. Output: IRIS IAP
Literature search, screening, and inventory <ul style="list-style-type: none"> Identify health effect studies Identify other informative studies relevant to evaluating potential health effects 	2	Perform comprehensive literature search(es). Use problem formulation PECO criteria to identify and select relevant human and animal health effect studies. Identify potentially relevant supplemental material (e.g., mechanistic, pharmacokinetic). Create literature inventories that capture basic study design and health effect information. The initial inventories are used for problem formulation. The literature search, screening, and inventory methods are also used in the draft assessment, post-problem formulation. Output: Can appear in IRIS IAP or Systematic Review Protocol
Refine problem formulation and specify assessment approach <ul style="list-style-type: none"> Define the assessment PECO and units of analysis Refine the IAP on the basis of information gained during scoping and problem formulation 	3	Refine the IAP on the basis of the literature Inventory and other relevant information. Define the assessment PECO and units of analysis. Units of analysis are related endpoints considered together to inform hazard identification. The protocol also specifies plans for consideration of supplemental material. There are circumstances where specific mechanistic evidence (typically biological precursors) can be included in the units of analysis for human or animal evidence synthesis. Output: Systematic Review Protocol

ORD Staff Handbook for Developing IRIS Assessments

Assessment development stage	Chapter	Purpose and other useful details
Study evaluation <ul style="list-style-type: none"> Evaluate health effect studies for risk of bias and insensitivity Evaluate PBPK models and other information as needed 	4	Evaluate individual human and animal health effect studies, considering risk of bias and sensitivity. Evaluate pharmacokinetic (PK) and physiologically based pharmacokinetic (PBPK) models and other information (e.g., mechanistic) as needed.
Extraction and display of study results from epidemiological and toxicological studies <ul style="list-style-type: none"> Human and animal health effect studies 	5	Collect key health effect study information in a database and prepare graphical and tabular displays.
Evidence synthesis <ul style="list-style-type: none"> Human and animal health effect studies Mechanistic studies 	6	<p>Analyze results incorporating the strengths and weaknesses of the sets of human and animal health effect studies by health effect or other selected grouping. Develop separate syntheses for human and animal evidence.</p> <p>Determine if mechanistic information will be considered. Conduct focused, stepwise analyses of the most relevant mechanistic evidence and summarize results of the analyses by health effect or other selected grouping. The degree of analysis depends on the unique needs of the assessment. Although mechanistic information is typically presented separately (in narrative format, tables, or figures), the conclusions may inform within stream evidence synthesis judgments related to coherence, directness, and biological significance/adversity.</p>
Evidence integration <ul style="list-style-type: none"> Summarizes the strength of each evidence stream as part of the evidence integration narrative Overall evidence integration across evidence streams (hazard identification, including review of susceptibility) 	6	Prepare evidence integration for hazard identification and overall summary conclusions. Summarize the strength of the evidence from the available human and animal health effect studies. Incorporate mechanistic evidence important for evidence integration judgments of human relevance of the animal evidence, cross-evidence stream coherence, biological plausibility and mode-of-action understanding, potential susceptibility, and other critical inferences (e.g., read-across analyses).
Hazard considerations and study selection for dose-response assessment <ul style="list-style-type: none"> Study selection for dose-response analysis 	7	Select the most informative studies and outcomes for dose-response analysis on the basis of study confidence and other predefined considerations including hazard identification decisions and susceptibility.
Derivation of toxicity values <ul style="list-style-type: none"> Cancer, noncancer 	8	Model studies and develop a quantitative estimate for each hazard of concern. Consider uncertainty and susceptibility and describe confidence in the estimates.

IAP = IRIS Assessment Plan; PBPK = physiologically based pharmacokinetic; PECO = populations, exposures, comparators, and outcomes; PK = pharmacokinetic.

1. SCOPING AND INITIAL PROBLEM FORMULATION



Purpose

- Define scope of the IRIS assessment
- Publish assessment-specific documents:
 - IRIS Assessment Plan (IAP) to summarize scoping and initial problem formulation
 - Systematic review protocol that describes the strategy for implementation of systematic review and dose-response analyses

This chapter describes the scoping and initial problem formulation process and general points for Integrated Risk Information System (IRIS) staff to follow as they initiate an assessment. It also provides an overview of the two preassessment documents published for each IRIS assessment as part of the scoping and problem formulation process: (1) the IRIS Assessment Plan (IAP) and (2) the systematic review protocol.

1.1. SCOPING

Scoping is the first stage in the development of an IRIS assessment. The purpose of scoping is to ensure the IRIS assessment meets the human health chemical toxicity assessment needs of U.S. Environmental Protection Agency (EPA) Program⁴ and Regional Offices,⁵ the primary requestors of IRIS assessments. In contrast to assessments developed by individual EPA programs under specific statutory mandates (e.g., Toxic Substances Control Act), IRIS assessments are typically not focused on specific chemical uses or cleanup scenarios.

The IRIS assessment manager takes an active role in scoping and works with other IRIS and Office of Research and Development (ORD) staff and contractors assigned to provide support in this process. Scoping involves consultation and close coordination with EPA Program and Regional Offices and typically involves one or more meetings to discuss their expectations and the specific components of an IRIS assessment most important for addressing their needs ([NRC, 2009](#)). EPA

⁴<https://www.epa.gov/aboutepa>.

⁵<https://www.epa.gov/aboutepa/regional-and-geographic-offices>.

Program and Regional Offices may be aware of other federal, state, and tribal needs related to the chemical(s) of interest and can relay this information to the IRIS Program during scoping.

Regular follow-up communications throughout scoping, as well as the initial problem formulation process (see Section 1.2), allow for the assessment team and interested Program and Regional Offices to share any changes or new information relevant to the scope or timing of the assessment. Assessment teams should document scoping decisions, for example, via a project-specific decision tracker or in their meeting agendas and minutes. The results of scoping are communicated across EPA and with other federal or state agencies, tribes, and the public via the IAP (see Section 1.3.1), which is released for public comment to provide additional opportunities for input early in the assessment process.

1.1.1. Key Considerations That Determine the Scope of an Assessment

The following are examples of questions that may be addressed during scoping communications with EPA Program and Regional Offices. These considerations help inform the project management timelines and specific aims for the assessment.

- What are the decision-making needs (including statutory authority, regulatory decisions, or health effects of public concern) of the requesting Program or Regional Offices that will be informed by the IRIS assessment, and what is the time frame for those needs?
- What are the exposure scenarios of primary concern or most immediate need? Is there a need for an assessment of particular routes (e.g., oral, inhalation, dermal) or durations (e.g., chronic, subchronic, short-term, acute)? Do exposure levels from scenarios of concern fluctuate significantly over time?
- What form(s) of the chemical are most relevant for EPA Program and Regional Offices (e.g., elemental forms or certain oxidation states or salts for metals)?
- How is the chemical measured in environmental samples: by itself, transformed (e.g., methyl mercury), or as part of a mixture? Is there a need to address individual components of a mixture or the mixture as a whole? Would it be useful to develop a single assessment for a group of chemicals (e.g., all vanadium compounds)?
- Are there early indications that other issues might affect dose-response [e.g., chemicals with a potential mutagenic mode of action (MOA)], potentially impacting risk management decisions?
- Is dose-response information that enables cost-benefit analysis needed? What type of outcomes would be useful for cost-benefit analysis?
- Are there relevant occupational risks or other exposures that may be at ranges above typical environmental exposures?
- Are there susceptible populations that might be at increased risk from exposure to the chemical of interest?

- Is there a strong risk communication or decision-making need to characterize toxicity at exposures above a reference value?
- Are there available or in-progress assessment products from other federal, state, or international agencies that may be informative? A list of agencies that may be relevant is available in Section 1.2.1.

1.2. INITIAL PROBLEM FORMULATION

During initial problem formulation, the IRIS Program identifies health effects that have been studied in relation to exposure to the chemical, as well as key science issues that may need to be considered for hazard evaluation or deriving toxicity values. This is an iterative process that often begins with the development of a systematic evidence map (SEM), which is a formal analysis that uses systematic review methods to develop literature inventories that characterize the extent of the evidence base for a topic ([Wolffe et al., 2019](#); [Miake-Lye et al., 2016](#); [Bragge et al., 2011](#)). An SEM may not be needed if initial problem formulation needs can be addressed from recent assessments available from other health agencies or other analogous sources (e.g., review articles, state-of-the-science workshops). Initial problem formulation also considers stakeholder feedback received during the public comments period for the IAP, which is released early in the assessment development process.

The general steps for initial problem formulation in an IRIS assessment are as follows.

- A survey of existing assessments from federal, state, and international health agencies is conducted (see Section 1.2.1). If a recent assessment is available, it may be used as the starting point for the literature search (i.e., the literature search conducted by IRIS may focus on studies that have been published since the development of the existing assessment) or may be used in lieu of an SEM to inform initial problem formulation.
- If warranted, an SEM (see Section 1.2.2) is developed. The SEM identifies health effects that have been studied in conjunction with exposure to the chemical or substance, as well as key pharmacokinetic (PK)⁶ and MOA issues, susceptible populations and lifestages, and differences in scientific interpretation or controversies the assessment may need to address. The populations, exposures, comparators, and outcomes (PECO) criteria used to develop the SEM (referred to hereafter as the “problem formulation PECO”) are typically broad to identify all studies that are potentially relevant to the scope of the assessment.
- The results of the SEM (or literature inventory from a recent existing assessment) are considered in the context of the needs identified by EPA during scoping (see Section 1.1) to prepare the IAP. The IAP provides a summary of the Agency need for the assessment; objectives and specific aims; problem formulation PECO criteria, literature search and screening methods, and literature inventory (when the SEM was conducted); identification of key areas of scientific complexity (see Section 1.2.3); and proposed refinements to the

⁶The terms “toxicokinetic” and “pharmacokinetic” are often used interchangeably. Pharmacokinetic is more aptly used for pharmacologically active compounds, while toxicokinetic would cover toxic compounds. By convention, however, pharmacokinetic is commonly used in EPA, including in the description of physiologically based pharmacokinetic (PBPK) models.

problem formulation PECO criteria (referred to hereafter as the “assessment PECO”) that identify the evidence considered most pertinent to the assessment. The contents of the IAP are outlined in Section 1.3.1.

- The IAP undergoes Agency review, is revised as needed, and is presented at a public science meeting to solicit scientific and stakeholder input. Typically, external experts are identified to provide feedback on the IAP at the public science meeting, especially the key science issues. Similar to standards set for external peer review, these experts must have no financial conflict of interest with the chemical(s) presented in the IAP. A financial conflict of interest (or other interest that conflicts) with the service of an individual could impair the individual’s objectivity or could create an unfair competitive advantage for a person or an organization. Financial conflicts may include, but are not limited to significant investments, consulting arrangements, employer affiliation grants/contracts, expert witness, consulting arrangements, and honoraria. Information related to conflicts of interest is outlined in EPA’s Peer Review Handbook ([U.S. EPA, 2015b](#)).
- After stakeholder input is received on the IAP, any revisions to the specific aims or PECO criteria resulting from the public comments will be reflected in the systematic review protocol, which also undergoes public comment. This document includes methodological details on the process for study evaluation, evidence synthesis and integration, and the steps beyond the systematic review such as toxicity value derivation. The contents of the systematic review protocol are outlined in Section 1.3.2.

Further details on the survey of existing assessments, SEM development, and identification of key science issues during initial problem formulation are provided in the respective subsections below.

1.2.1. Survey of Existing Assessments and Toxicity Values

Existing health assessments for the chemical(s) of interest provide a source of previously evaluated health effects evidence and toxicity values.⁷ The availability of other assessments and reviews (especially when recent) provides context for the IRIS assessment and may mitigate the need to develop an SEM. For instance, the assessment team may use existing assessments as the starting point for an SEM (see Building Search Strategies from Prior Assessments in Section 2.3.2), or in some cases may be able to base their assessment plan on the health effects identified in existing assessments. Decisions on whether to use prior assessments for these purposes tend to be very assessment specific, based on both the nature of the prior assessments in relation to the planned IRIS assessment (e.g., age of assessment, assessment methodology, nature of peer review, overlap in content with EPA scoping needs), as well as anticipated decision-making usage of the

⁷“Toxicity value” is a broad term that encompasses reference values and cancer risk estimates (i.e., slope factors and unit risk estimates). The term reference value applies to values designed to provide a “benchmark” or exposure limit, below which adverse effects on human health are not expected to occur ([U.S. EPA, 2002b](#)). Reference values are the most common final output from the dose-response assessment component of the risk assessment paradigm set forth by the National Research Council ([NRC, 1983](#)); ([NRC, 2009](#)) and are based on an observed or estimated threshold for an effect, usually noncancer.

IRIS assessment by EPA. For these reasons, decisions on how prior assessments have been used is presented in the chemical-specific SEM.

When existing health assessments are identified, the assessment team compiles the reported toxicity values and pertinent details about their derivation (e.g., health effect, point of departure, uncertainty factors). The available toxicity values can be summarized graphically as an array or in a tabular format that includes derivation details. Toxicity values without derivation details or that are derived secondarily from another agency's value are considered less informative, and therefore are not shown in the graphical array and are summarized in a separate table from those values that have derivation details available. Tables and figures indicate the month and year the searches were conducted. A more detailed explanation of the development and evolution of the graphical toxicity value arrays used by the IRIS Program, along with chemical-specific examples, can be found in reports ([U.S. EPA, 2013](#), [2009](#)). A list of organizations that may be queried to conduct this survey is presented in Appendix B.

1.2.2. Systematic Evidence Maps (SEMs)

Unless a very recent assessment or review exists for the chemical(s) of interest, it is beneficial for the assessment team to develop an SEM. SEMs have been defined as “A comprehensive summary of the characteristics and availability of evidence as it relates to broad issues of policy or management relevance. SEMs do not seek to synthesize evidence or draw conclusions but instead to catalogue the available evidence, utilizing systematic search and selection strategies to produce searchable databases of studies along with detailed descriptive information” [[Elsevier \(2017\) <https://www.elsevier.com/journals/environment-international/0160-4120/guidance-notes>](#)]. SEMs are used as tools to refine the focus of health effects assessments, inform future research, and identify data gaps ([Wolfe et al., 2019](#); [Miake-Lye et al., 2016](#); [Bragge et al., 2011](#)). SEMs are a relatively recent addition to IRIS assessments and are formalized extensions of the literature inventories that have been traditionally produced by IRIS. Examples of IRIS assessments that include SEMs are mercury salts ([U.S. EPA, 2021d](#)) and vanadium and vanadium compounds ([U.S. EPA, 2021e](#)).

In the context of IRIS assessments, SEMs have proven useful for developing the a priori analysis plans typically presented in systematic review protocols ([Thayer et al., 2022a](#); [Thayer et al., 2022b](#)). The development of an SEM early in the IRIS assessment process provides the assessment team with summary-level, sortable lists that can be used to assess the amount and type of health effects data available. This information may be used to refine (narrow or broaden) the scope of the assessment to capture the information considered most pertinent to the analysis. A template is available presenting core methods for producing SEMs in the IRIS Program ([Thayer et al., 2022a](#)).

SEMs in IRIS assessments are developed on the basis of an initial broad search of the literature, which is conducted using the methods for literature search, screening, and inventory described in Chapter 2. Studies that meet the problem formulation PECO criteria are inventoried

according to study design and health effects evaluated (see Section 2.5.1). Studies that provide potentially relevant supplemental information (e.g., mechanistic, PK, susceptible populations) are identified and tagged to the broad categories shown in Section 2.2. Supplemental information may also be organized into inventories as part of the SEM, although this is more commonly done later in the assessment as part of finalizing the protocol (see Chapter 3). Examples of specific areas where SEMs in the IRIS Program have been helpful include ([Thayer et al., 2022b](#)):

- Identifying data gaps:
 - Identifying knowledge gaps early in the assessment process, especially those that could be addressed by new research within the time frame for conducting the assessment.
 - Determining whether or to what extent alternative methodologies need to be considered (such as read-across and other new approach methods). This information can be used to supplement other streams of evidence, or in some cases, it may be the only available method or evidence.
- Determining need for an updated assessment:
 - Identifying new key evidence published since a previous assessment was completed.
- Informing assessment priorities and refining the scope of the review:
 - Identifying health outcomes likely to be included or prioritized in the assessment.
 - Identifying key science issues and understanding the availability of evidence to address those issues under the desired time frame for conducting the assessment.
 - Determining the extent to which the evidence base is likely to inform conclusions on susceptible populations.
- Informing development of analysis plans for mechanistic, ADME (absorption, distribution, metabolism, and excretion), PK, and similar types of evidence.
- Informing development of study evaluation considerations for health outcomes included in the assessment:
 - Understanding the strengths and limitations of the exposure assessment methods used in a set of epidemiological studies.
 - Developing considerations for assessing study sensitivity (e.g., whether exposure in a developmental study covers the appropriate time frame).
- Understand the level of effort and expertise required for an upcoming assessment, which informs timelines for conducting the assessment.

- SEM methods can also be used in later stages of the assessment development process to determine the impact of new studies published during later stages of the assessment review process on overall assessment conclusions (e.g., during external peer review).

1.2.3. Identification of Key Science Issues

Key science issues are scientific questions or uncertainties that are important to address during the assessment process. Key science issues that must be considered during IRIS assessment development may be identified during the SEM process, through consideration of topics identified in existing health assessments or analogous information (e.g., review articles), and through public comments. Draft key science issues are presented for public comment in the IAP and are discussed at the public science meeting for the IAP. Any revisions to the key science issues following the IAP public comment period are presented in the systematic review protocol.

Examples of science issues include the following (see Appendix C for examples of these key science issues in existing IRIS assessments):

- Human relevance of findings in animals
- Whether an endpoint is considered adverse or adaptive
- Issues where conflicts in the evidence are known, including hypothesized MOAs that lack scientific consensus
- Issues relating to PK used to identify susceptible groups
- Identification of published physiologically based pharmacokinetic (PBPK) models that have no or limited in vivo validation data
- Identification of PBPK models that use novel modeling methods or calculations, including pharmacodynamic components, not previously reviewed for use in an IRIS assessment
- Complex chemistry issues that may affect toxicity, PK, or applicability of toxicity values to different forms of the chemical
- Confounding by coexposures in epidemiological studies
- Issues where missing chemical-specific information can be informed using analogues.

1.3. KEY DOCUMENTS: IRIS ASSESSMENT PLAN (IAP) AND SYSTEMATIC REVIEW PROTOCOL

EPA releases two documents for public comment during the scoping and problem formulation phase of each IRIS assessment: (1) the IAP and (2) the systematic review protocol. Figure 1-1 summarizes the purposes of the IAP and protocol. When an assessment needs to be conducted under an especially expedited time frame, the IAP and protocol may be released concurrently.

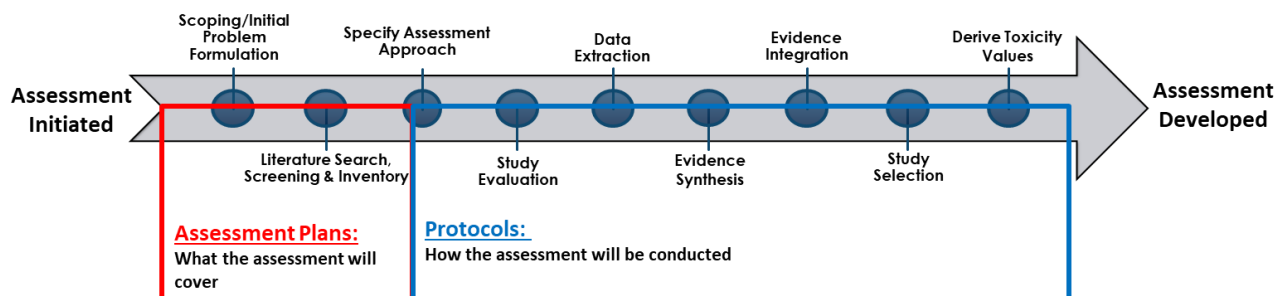


Figure 1-1. Integrated Risk Information System (IRIS) systematic review problem formulation and method documents.

1.3.1. IRIS Assessment Plan (IAP)

The IAP provides a summary of the Agency need for the assessment; objectives and specific aims; problem formulation PECO criteria; SEM (or literature inventory from a recent existing assessment); and identification of key science issues. Brief background information on uses and potential for human exposure is provided for context. The IAP may also include proposed assessment PECO criteria that identify the evidence considered most pertinent to the assessment. Additionally, on the basis of needs identified during scoping, the IAP should also indicate any proposed modularity or interim products (e.g., separation of noncancer and cancer conclusions into separate assessments, narrowed focus to specific route of administration or lifestage).

The IAP is released following scoping and initial problem formulation for the IRIS assessment. Prior to public release, the IAP is shared for Agency input. After the IAP is publicly posted to the EPA website, there is a 30-day public comment period prior to holding a public science meeting to present and discuss the document (<https://www.epa.gov/iris/basic-information-about-integrated-risk-information-system#process>). The public science meeting may be held in person or virtually. Stakeholder input received during the comment period on the IAP is considered as part of preparing the assessment’s systematic review protocol, and any revisions to the specific aims and problem formulation PECO are reflected in the assessment’s systematic review protocol.

An outline of the contents of the IAP is shown in the text box below. Examples of public IAPs are available on the IRIS website (<https://www.epa.gov/iris>), and a template version is available on the [IRIS resource page in HAWC](#).

ORGANIZATION OF THE IRIS ASSESSMENT PLAN (IAP)

1. Introduction
2. Scoping and initial problem formulation
 - 2.1 Background (brief, provided for context)
 - 2.2 Scoping summary—summarize Agency needs and anticipated uses in tabular format
 - 2.3 Survey of existing assessments and toxicity values
 - 2.4 Key science issues
3. Overall objectives and specific aims
4. Literature search, screening, and inventory*
 - 4.1 Problem formulation PECO criteria
 - 4.2 Supplemental screening criteria
 - 4.3 Literature search strategies
 - 4.4 Literature screening
 - 4.5 Literature inventory
5. (Optional) Proposed assessment PECO criteria
6. References
7. Appendices (e.g., literature search strings)

*May be based on an SEM or literature inventory from a recent existing assessment.

Figure 1-2. Organization of the IRIS Assessment Plan (IAP).

PECO = populations, exposures, comparators, and outcomes.

1.3.2. Systematic Review Protocol

The protocol is a central component of a systematic review. Protocols improve transparency and reduce bias in the conduct of the review by prespecifying the review question and methods([CRD, 2013](#); [Higgins and Green, 2011a](#); [IOM, 2011](#)).

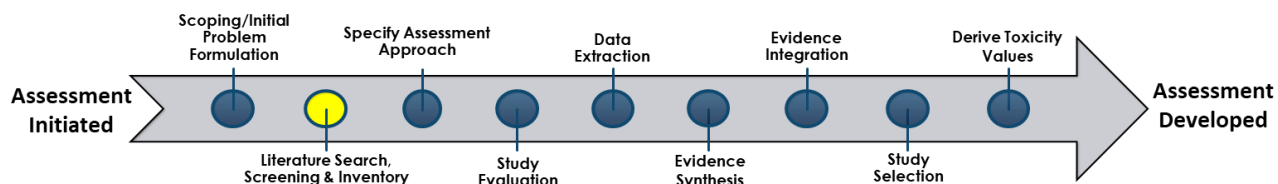
In the IRIS Program, the initial sections of the protocol are identical to the IAP but have been revised to reflect any adjustments made in response to public input. The protocol also includes the assessment PECO, the units of analyses used for evidence synthesis, and any prioritized analyses of supplemental evidence (see Chapter 3). A unit of analysis is an outcome or

group of related outcomes within a health effect category considered together during evidence synthesis. Finally, the protocol includes methodological details on the process that will be used for study evaluation, the structured frameworks used during evidence synthesis and integration, dose-response, and toxicity value derivation. The methods shown in the protocol are an abridged version of what is presented in the subsequent chapters of the IRIS Handbook, adjusted for the specific chemical(s) being assessed.

The IRIS Program posts assessment protocols and protocol updates for a chemical publicly on the IRIS website at www.epa.gov/iris. The protocol is released after the IAP and prior to the publication of the draft IRIS assessment. After the protocol is posted, there is typically a 30-day public comment period. If any changes are made to the protocol following release of the draft IRIS assessment, these changes will be documented in the finalized version of the IRIS assessment.

Examples of public protocols are available on the IRIS website (<https://www.epa.gov/iris>) and a template version is available on the IRIS resource page in HAWC. The template includes the elements described in checklists for peer-reviewed systematic review protocol reporting ([Haddaway et al., 2018](#); [Moher et al., 2009](#)).

2. LITERATURE SEARCH, SCREENING, AND INVENTORY



Purpose

- Identify the relevant citations for use in the assessment, document the search and screening process, and categorize citations in a literature inventory.

This chapter describes the elements and tasks involved in developing a literature search strategy, screening identified references, and creating an inventory of citations. The search, screening, and inventory workflows noted here are employed in both the initial search (e.g., to generate a systematic evidence map [SEM] during the problem formulation phase) and in all subsequent literature searches conducted to develop the assessment. The [Health and Environmental Research Online](#) (HERO) database is typically used to conduct and document literature searches.

2.1. POPULATIONS, COMPARATORS, EXPOSURES, OUTCOMES (PECO) CRITERIA

The populations, exposures, comparators, and outcomes (PECO), along with the supplemental tagging structure (see Section 2.2), are used to identify the evidence that addresses the specific aims of the assessment and to focus the search terms and inclusion criteria. Depending on the assessment specific aims, the PECO may be broad, encompassing multiple health effects and exposure routes, or more specific, targeting specific susceptible populations and lifestages (e.g., pregnant women and their fetuses, infants, and children), health effects, exposures, etc. (see Table 2-1).

The problem formulation PECO criteria used to prepare the SEM and to develop the literature inventory are typically broad to identify all literature that is potentially relevant to the assessment. For instance, problem formulation PECO criteria generally specify that all health outcomes will be included but may be later refined (narrowed or broadened) in the assessment PECO to encompass the health outcomes that are considered most pertinent to the assessment. The

refinement of problem formulation PECO criteria to the ultimate assessment PECO is described in Section 3.1, Assessment PECO Criteria.

Table 2-1. Components of populations, exposures, comparators, and outcomes (PECO) and potential types of evidence

PECO element	Evidence
Populations	<p>Human: Any population and lifestage (occupational or general population, including children and other sensitive populations).</p> <p>Animal: Nonhuman mammalian animal species (whole organism) of any lifestage (including fetal, early postnatal, adolescents and adults). <i>[Note: Full-text retrieval is recommended for studies of transgenic model systems. These studies typically provide mechanistic evidence tracked as “potentially relevant supplemental material” but may present phenotypic information in wildtype animals that meets PECO criteria but is not mentioned in the abstract].</i></p> <p><i>[For both human and animal studies, any study that includes populations that address susceptibility should also be tagged with that supplemental content tag.]</i></p>
Exposures	<p>Relevant forms:</p> <p><i>[chemical X] (CASRN)</i></p> <p><i>Other forms of [chemical X] that readily dissociate (e.g., list any salts)</i></p> <p><i>Metabolites of interest, including metabolites used to estimate exposures to [chemical X]</i></p> <p><i>Occupations that may be considered surrogates of exposure</i></p> <p>Human: Any exposure to <i>[chemical X]</i> via <i>[oral or inhalation]</i> route[s] if applicable. Citations will also be included if biomarkers of exposure are evaluated (e.g., measured chemical or metabolite levels in tissues or bodily fluids) but the exposure route is unclear or likely from multiple routes. Other exposure routes, such as those that are clearly dermal, will be tracked during title and abstract screening and tagged as “potentially relevant supplemental material.”</p> <p><i>[Specify if certain exposure assessment matrices or methods will NOT be included. Also, although many epidemiological studies measure multiple chemicals, they are not considered mixture studies (a type of supplemental content) as long as the study presents health outcome analyses specific to the chemical(s) of interest. For most assessments that include hazard assessment, exposure does not have to be quantitated, e.g., low versus high is considered to meet PECO criteria.]</i></p> <p>Animal: Any exposure to <i>[chemical X]</i> via <i>[oral or inhalation]</i> route[s] of >1 d duration, or any duration assessing exposure during reproduction or development. Studies involving exposures to mixtures will be included only if they include an experimental arm with exposure to <i>[chemical X]</i> alone. Other exposure routes, including <i>[dermal or injection]</i>, will be tracked during title and abstract as “potentially relevant supplemental material.”</p> <p><i>[Typically, IRIS assessments to develop chronic toxicity values consider acute nondevelopmental exposure studies (<1 d duration) as supplemental content, but this can be adjusted depending on assessment needs. Also, the assessment-specific protocol should specify if certain exposures/study designs will NOT be included, or if a minimum number of dose or concentration levels tested in experimental animal studies is indicated.]</i></p>

PECO element	Evidence
<p>Comparators <i>[Note: PECO element name can be adjusted toward SEM-specific aims, e.g., Comparisons or Comparison group may be more precise for SEMs targeted toward identifying dose-response suitable studies.]</i></p>	<p>Human, Example A (general SEM): A comparison or referent population exposed to lower levels (or no exposure/exposure below detection limits), or exposure for shorter periods, or cases versus controls, or a repeated measures design. However, worker surveillance studies are considered to meet PECO criteria even if no statistical analyses using a referent group are presented. Case reports or case series of >3 people will be considered to meet PECO criteria, while case reports describing findings in 1–3 people will be tracked as “potentially relevant supplemental material.”</p> <p>Human, Example B (targeted SEM to identify studies suitable for dose-response): Studies reporting effect measures (e.g., relative risk, standardized mortality ratio, beta coefficients) based on a comparison or referent population exposed to lower levels (or no exposure/exposure below detection limits), or cases versus controls, or a repeated measures design.</p> <p><i>[Notes: Studies based exclusively on duration of exposure analyses (i.e., longer versus shorter exposure duration) are not likely to be informative for SEMs focused on identifying studies plausibly suitable for dose-response.]</i></p> <p>Animal: A concurrent control group exposed to vehicle-only treatment or untreated control (control could be a baseline measurement, e.g., acute toxicity studies of mortality, or a repeated measure design).</p>
<p>Outcomes</p>	<p>All health outcomes (cancer and noncancer). In general, endpoints related to clinical diagnostic criteria, disease outcomes, biochemical, histopathological examination, or other apical/phenotypic outcomes are considered to meet PECO criteria</p> <p><i>[Note: Studies meeting PECO criteria may also contain supplemental mechanistic content that describes biological or chemical events associated with phenotypic effects. When this occurs, these studies are also tagged as having supplemental mechanistic information. This typically happens during full-text review or when doing the literature inventory. Full-text retrieval is recommended for studies of transgenic model systems that meet E and C criteria because they may present phenotypic information in wildtype animals that meet P and O criteria but is not reported in the abstract.]</i></p>

PECO = populations, exposures, comparators, and outcomes; SEM = systemic evidence map.

2.2. SUPPLEMENTAL MATERIAL SCREENING CRITERIA

In addition to citations meeting the PECO criteria, citations containing supplemental material that are potentially relevant to the specific aims of the assessment are tracked during the literature screening process. Table 2-2 presents major categories and potential uses of supplemental material. Because the major health effect categories and units of analysis will not have been fully identified when screening is initially conducted, the broad tagging categorization is used to characterize the available evidence base and helps identify the key science issues presented in the IRIS Assessment Plan (IAP). The categories are designed to help the assessment team prioritize citations for consideration in the assessment on the basis of the likelihood they will impact assessment conclusions.

Table 2-2. Example categories of “Potentially Relevant Supplemental Material” (from the Integrated Risk Information System [IRIS] Assessment Plan template)

Category (Tag)	Description	Typical assessment use
Pharmacokinetics data potentially informative to assessment analyses		
<p>Classical pharmacokinetic (PK) or physiologically based pharmacokinetic (PBPK) model studies</p>	<p>Classical pharmacokinetic or dosimetry model studies: Classical PK or dosimetry modeling usually divides the body into just one or two compartments, which are not specified by physiology, where movement of a chemical into, between, and out of the compartments is quantified empirically by fitting model parameters to ADME (absorption, distribution, metabolism, and excretion) data. This category is for papers that provide detailed descriptions of PK models but are not PBPK models. The data are typically the time-course concentration in blood or plasma after oral and or intravenous exposure, but other exposure routes can be described.</p> <p>Physiologically based pharmacokinetic or mechanistic dosimetry model studies: PBPK models represent the body as various compartments (e.g., liver, lung, slowly perfused tissue, richly perfused tissue) to quantify the movement of chemicals or particles into and out of the body (compartments) by defined routes of exposure, metabolism, and excretion, and thereby estimate concentrations in blood or target tissues.</p> <p>A defining characteristic is that key parameters are determined from a substance’s physicochemical parameters (e.g., particle size and distribution, octanol-water partition coefficient) and physiological parameters (e.g., ventilation rate, tissue volumes).</p>	<p>PBPK and PK model studies are included in the assessment and evaluated for possible use in conducting quantitative extrapolations. PBPK/PK models are categorized as supplemental material with the expectation that each will be evaluated for applicability to address assessment extrapolation needs and technical conduct. Specialized expertise is required for their evaluation.</p> <p>Standard operating procedures for PBPK/PK model evaluation and the identification, organization, and evaluation of ADME studies are outlined in <i>An umbrella Quality Assurance Project Plan (QAPP) for PBPK Models</i> (U.S. EPA, 2018e).</p>
<p>Pharmacokinetic (ADME)</p>	<p>Pharmacokinetic (ADME) studies are primarily controlled experiments, where defined exposures usually occur by intravenous, oral, inhalation, or dermal routes, and the concentration of particles, a chemical, or its metabolites in blood or serum, other body tissues, or excreta are then measured.</p> <p>These data are used to estimate the amount absorbed (A), distributed (D), metabolized (M), or excreted (E).</p> <p>ADME data can also be collected from human subjects who have had environmental or workplace exposures that are not quantified or fully defined.</p>	<p>ADME studies are inventoried and prioritized for possible inclusion in an ADME synthesis section on the chemical’s PK properties and for conducting quantitative adjustments or extrapolations (e.g., animal-to-human). Specialized expertise in PK is necessary for inventory and prioritization.</p> <p>Standard operating procedures for PBPK/PK model evaluation and the identification, organization, and evaluation of ADME</p>

Category (Tag)	Description	Typical assessment use
	<p>ADME data, especially metabolism and tissue partition coefficient information, can be generated using in vitro model systems. Although in vitro data may not be as definitive as in vivo data, these studies should also be tracked as ADME. For large evidence bases it may be appropriate to separately track the in vitro ADME studies.</p> <p><i>*Studies describing environmental fate and transport or metabolism in bacteria or model systems that are not applicable to humans or animals should not be tagged.</i></p>	<p>studies are outlined in <i>An umbrella Quality Assurance Project Plan (QAPP) for PBPK Models</i> (U.S. EPA, 2018e).</p>
Supplemental evidence potentially informative to assessment analyses		
Mechanistic endpoints	<p>Studies that do not meet PECO criteria but report measurements that inform the biological or chemical events associated with phenotypic effects related to a health outcome. Experimental design may include in vitro, in vivo (by various routes of exposure; includes all transgenic models), ex vivo, and in silico studies in mammalian and nonmammalian model systems. Studies using new approach methodologies (NAMs; e.g., high throughput testing strategies, read-across applications) are also categorized here. Studies where the chemical is used as a laboratory reagent (e.g., as a chemical probe used to measure antibody response) generally should not be tagged.</p> <p>Mechanistic evidence can also help identify factors contributing to susceptibility; these studies should also be tagged “susceptible populations.”</p> <p><i>[Notes: During screening, especially at the title and abstract (TIAB) level, it may not be readily apparent for studies that meet P, E, and C criteria if the endpoint(s) in a study are best classified as phenotypic or mechanistic with respect to the O criteria. In these cases, the study should be screened as “unclear” during TIAB screening, and a determination made based on full-text review (in consultation with a content expert as needed). Full-text retrieval is performed for studies of transgenic model systems that meet E and C criteria to determine if they include phenotypic information in wildtype animals that meet P and O criteria that is not reported in the abstract.]</i></p>	<p>Prioritized studies of mechanistic endpoints are described in the mechanistic synthesis sections; subsets of the most informative studies may become part of the units of analysis used to structure evidence synthesis. Mechanistic evidence can provide support for the relevance of animal effects to humans and biological plausibility for evidence integration judgments [including MOA analyses, e.g., using the MOA framework in the EPA Cancer Guidelines (U.S. EPA, 2005a)].</p>
Non-PECO animal model	<p>Studies reporting outcomes in animal models that meet the outcome criteria but do not meet the “P” in the PECO criteria.</p> <p>Depending on the endpoints measured in these studies, they can also provide mechanistic information (in these cases studies should also be tagged “mechanistic endpoints”).</p>	<p>Studies of non-PECO animals, exposures, or durations can be summarized to inform evaluations of consistency (e.g., across species or routes or durations), coherence, or adversity; subsets of the most informative</p>

Category (Tag)	Description	Typical assessment use
	<p><i>*This categorization generally does not apply to studies that use species with limited human health relevance (e.g., ecotoxicity-focused studies are typically excluded).</i></p>	<p>studies may be included in the unit of analysis. These studies may also be used to inform evidence integration judgments of biological plausibility or MOA analyses and thus may be summarized as part of the mechanistic evidence synthesis.</p>
<p>Non-PECO route of exposure</p>	<p>Epidemiological or animal studies that use a non-PECO route of exposure, (e.g., injection studies or dermal studies if the dermal route is not part of the exposure criteria).</p> <p><i>*This categorization generally does not apply to epidemiological studies where the exposure route is unclear; such studies are considered to meet PECO criteria if the relevant route(s) of exposure are plausible, with exposure being more thoroughly evaluated at later steps.</i></p>	
<p>Non-PECO exposure duration (optional)</p>	<p>For assessments that focus on chronic exposure, acute exposure durations (defined as animal studies of less than 1 d in duration) are generally considered supplemental. In rare cases and for very large evidence bases, short-term (i.e., less than subchronic) exposure durations could also be categorized as supplemental.</p> <p><i>*Some assessment teams might prefer to keep these studies as PECO relevant and summarize them in the literature inventory rather than track them as supplemental.</i></p>	
<p>Susceptible populations</p>	<p>Studies that help identify potentially susceptible subgroups, including citations investigating how intrinsic factors such as sex, lifestage, genotype, or other factors (e.g., health status) that can influence toxicity. These are often co-tagged with other supplemental material categories, such as mechanistic or ADME. Studies meeting PECO criteria that also address susceptibility should be co-tagged as supplemental.</p> <p><i>*Susceptibility based on most extrinsic factors, such as increased exposure due to residential proximity to exposure sources, is not considered an indicator of susceptible populations for the purposes of IRIS assessments.</i></p>	<p>Provides information on factors that might predispose sensitive populations or lifestages to a higher risk of adverse health effects following exposure to the chemical. This information is summarized during evidence integration for each health effect and is considered during dose-response, where it can directly impact modeling decisions.</p>
<p>Background information potentially useful to problem formulation and protocol development (These studies fall outside the scope of IRIS assessment analyses.)</p>		
<p>Human exposure and biomonitoring (no health outcome)</p>	<p>Information regarding exposure monitoring methods and reporting that are unrelated to health outcomes but provide information on the following: methods for measuring human exposure, biomonitoring (e.g., detection of chemical in blood, urine, hair), defining exposure sources, or modeled estimates of exposure (e.g., in occupational settings). Studies that compare exposure levels to a reference value, risk threshold, or assessment points of departure are also included in this category. Studies related to</p>	<p>This information might be useful for developing exposure criteria for study evaluation or refining problem formulation decisions.</p>

ORD Staff Handbook for Developing IRIS Assessments

Category (Tag)	Description	Typical assessment use
	environmental fate and transport are typically tagged as background materials unless otherwise described in the assessment-specific protocol. <i>*Assessment teams might want to subtag studies that describe or predict exposure levels versus those that present exposure assessment methods.</i>	Notably, providing an assessment of typical human exposures (e.g., sources, levels) falls outside the scope of an IRIS assessment.
Mixture study	Mixture studies use methods that do not allow investigation of the health effects of exposure to the chemical of interest by itself [e.g., animal studies that lack exposure to chemical of interest alone or epidemiological studies that do not evaluate associations of the chemical of interest with relevant health outcome(s)]. <i>*Methods used to assess investigation of the exposure by itself might not be clear from the abstract, in particular for epidemiological studies. When unclear, the study is advanced to full-text review to determine eligibility.</i>	Mixture studies are tracked to help inform cumulative risk analyses, which could provide useful context for risk assessment but fall outside the scope of an IRIS assessment.
Case reports or case series	Human studies that present an investigation of a single exposed individual or group of ≤3 subjects that describe health outcomes after exposure but lack a comparison group (i.e., do not meet the “C” in the PECO criteria) and typically do not include reliable exposure estimates.	Tracking case studies can facilitate awareness of potential human health issues missed by other types of studies during problem formulation.
Other background information	Chemical-specific studies that might provide introductory information on chemical and physical properties (note: assessors typically will separately consult the EPA CompTox Dashboard); sources, production, and use; and environmental occurrence and fate. Additional groupings of information can be determined on an assessment-specific basis and some assessments might decide to separately tag different subsets of information (e.g., tag chemical properties studies separately from those on environmental occurrence and fate).	Although formal analyses on these general background topics are not part of an IRIS assessment, this information can be useful to protocol development (e.g., chemical property information for evaluating PK or exposure in animal studies). In addition, brief summary overviews are typically provided in the introductory materials.
Reference materials		
Records with no original data	Records that do not contain original data, such as other agency assessments, informative scientific literature reviews, editorials, or commentaries.	Studies tracked for potential use in identifying missing studies, background information, or current scientific opinions (e.g., hypothesized MOAs).
Posters or conference abstracts	Records that do not contain sufficient documentation to support study evaluation and data extraction.	

ADME = absorption, distribution, metabolism, and excretion); MOA = mode of action; NAMs = new approach methodologies; PBPK = physiologically based pharmacokinetic; PECO = populations, exposures, comparators, and outcomes; PK = pharmacokinetic; QAPP = Quality Assurance Project Plan; TIAB = title and abstract.

The inventory of supplemental material can assist in the development of focused analyses that come later in the assessment process (i.e., targeted analyses related to mechanistic, ADME, pharmacokinetic (PK)/physiologically based pharmacokinetic [PBPK], and susceptibility evidence); this inventory and customized screening instructions developed to reflect chemical-specific evidence base considerations are described in the assessment's systematic review protocol. Certain studies categorized and tagged as supplemental evidence could emerge during the course of assessment development as being critically important to the assessment. Therefore, some supplemental studies will undergo extensive analysis at the individual study level. Some examples of this include PBPK models supporting dose-response modeling and mechanistic evidence considered integral to the interpretation of other evidence (e.g., genotoxicity studies for a potentially mutagenic agent when conducting a cancer mode-of-action [MOA] analysis). Citations tagged as supplemental that contribute to a well-accepted scientific conclusion typically do not need to be evaluated and summarized at the individual study level (e.g., dioxin as an AhR [aryl hydrocarbon receptor] agonist).

A single study might have multiple tags to identify all reported content and potential applications. The tagging occurs at both the title/abstract (TIAB) and full text screening steps and facilitates preparation of the literature inventory during problem formulation. Assessment teams might identify additional categories specific to their chemical assessment needs; a stable and customizable subtagging structure for supplemental content beyond the broad categories listed here can be found in the template protocol. See Section 2.5.2 for more information on selecting subcategories to characterize the evidence base for creating the literature inventory.

2.3. LITERATURE SEARCH STRATEGIES

The following sections discuss key components in a literature search process, including using HERO, selecting core databases, developing the literature search terms, searching other resources, and documenting literature searches.

2.3.1. Health and Environmental Research Online (HERO)

HERO is a database that serves as a repository of scientific citations, including literature search results and other references that are cited in many U.S. Environmental Protection Agency (EPA) assessments. HERO is developed and managed by staff in EPA's Office of Research and Development (ORD) Center for Public Health and Environmental Assessment (CPHEA). HERO staff include information scientists who specialize in developing and conducting literature searches (broad and targeted), software programming experts who work to expand HERO's capabilities and interoperability with other software applications, and researchers who focus on incorporating use of machine learning (ML) and artificial intelligence (AI) in the assessment development process. It is highly recommended that the assessment team work closely with the HERO information specialists throughout the literature search process. Some useful tips and links for using HERO are described in Figure 2-1.

Using HERO for Literature Searches	
Create HERO project page	<ul style="list-style-type: none"> • Use of HERO databases (https://hero.epa.gov/heronet/index.cfm/litsearch/manual). • Complete a project page request form and initiate a collaboration with a HERO information specialist. Instructions for establishing a project page are available at https://hero.epa.gov/heronet/index.cfm/project/requestassessment. • Requests for HERO literature searches (https://hero.epa.gov/heronet/index.cfm/litsearch/request).
Develop search strategy	<ul style="list-style-type: none"> • Most searches will be based on the chemical name and synonyms. • When a more targeted search is needed, test and refine database-specific literature search results (BEFORE using HERO).
Retrieve references in HERO	<p>Retrieve results from each database using HERO in this order:</p> <ul style="list-style-type: none"> • PubMed • Web of Science • SCOPUS • Other resources (e.g., NTP, ECHA, TSCATS)
Automated duplicate review	<ul style="list-style-type: none"> • Screening mechanisms in HERO will “deduplicate” (remove duplicate) references as each database is searched and references are retrieved. • Remaining duplicates can be identified in screening software (e.g., DistillerSR) or manually.
Import references into screening software	<ul style="list-style-type: none"> • Obtain references in RIS file format. The RIS file can be obtained from HERO either by using the “Tools” link or directly from a project page. • From the “Tools” link, select “Export to RIS using a List of HERO IDs” and select the button “Distiller, etc. (With PDF links).” Input HERO IDs separated by a line or comma and retrieve the RIS file. • A list of HERO IDs or a RIS file can also be obtained from the project page by selecting all references or specific references. • Alternatively, HERO staff can directly provide the RIS file, when necessary. • Import the RIS file into problem formulation or screening software (e.g., DistillerSR, SWIFT Review, SWIFT Active). Make sure the bibliographic format includes HERO URLs with links to the full text PDFs in the URL fields (this will facilitate full text review).

<p>Request removal of duplicate records or related reference linking in HERO</p>	<ul style="list-style-type: none"> • Duplicate removal: If duplicate references are identified during screening, send a list of duplicate HERO IDs to HERO@epa.gov for removal, indicating which to delete and which to keep (e.g., 5678 keep 1234). HERO convention is to retain the smaller HERO ID number; HERO IDs are found in the label field in the RIS file. Removal of duplicates can also be requested as a reference correction request submitted through the reference details page. Requests are sent to HERO@epa.gov. • Reference linking: At times, retrieved citations might be linked. For example, HERO ID 5400977 is a peer-reviewed technical report that was preceded in time by related citations 4309149, 4309651, 4450232 describing the experimental protocol and original study data. When linked citations are identified, send the citation information (HERO IDs) to HERO@epa.gov indicating their linkage/relationship.
<p>Setting up tagging</p>	<ul style="list-style-type: none"> • Tagging references: Tags provide a means to organize references into groups and are typically based on the literature search (e.g., search date, database searched), or the categorical tagging structure (e.g., PECO relevance, evidence type) established in the screening forms. • References are tagged in HERO and Health Assessment Workspace Collaborative (HAWC), a web-based content managements system used to organize information for developing human health assessments. • A description of tagging in HERO can be found using the following link: https://hero.epa.gov/heronet/files/support/HEROtagging.pdf. HERO contains all references associated with an assessment, including all references that were identified in the literature searches. • References containing data relevant to the assessment are also tagged in HAWC. In general, tagging in HAWC should be consistent with HERO. Depending on assessment needs, additional tags can be applied in HAWC beyond what is presented in HERO (e.g., subtagging for mechanistic or PK studies).

Figure 2-1. Workflow for Health and Environmental Research Online (HERO)-facilitated literature searchers.

ECHA = European Chemicals Agency; HAWC = Health Assessment Workspace Collaborative; HERO = Health and Environmental Research Online; NTP = National Toxicology Program; PECO = populations, exposures, comparators, and outcomes; PK = pharmacokinetic; RIS = Research Information Systems; TSCATS = Toxic Substances Control Act Test Submissions.

The assessment team is responsible for initiating the literature search request and working with information specialists and librarians through EPA (e.g., HERO staff) or contractors to devise and execute the search. Both HERO and contractor information specialists offer extensive experience with database searching and information management. In either case, the process of

developing, testing, and implementing a comprehensive literature search strategy is expected to be an iterative, collaborative effort between the IRIS assessment team and the information specialist. Regardless of who devises the search strategy (EPA staff or a contractor), HERO should be used to perform the literature search and serve as the repository of the identified references. It is critical that the reference files provided from this search, typically shared in a Research Information System (RIS) format, include the HERO Uniform Resource Locator (URL) link in the URL field. The HERO URL should be the one that provides direct access to the PDF. This promotes interoperability between HERO and other software platforms used to help screen citations, especially at the full-text level. When a full-text version is requested and procured through HERO, inclusion of the HERO URL link in the record will enable the full-text version to be automatically accessible for EPA staff in the literature screening software application.

2.3.2. Core Database Searches

The goal of the search process is to identify full reports of **primary studies** (i.e., original data sources of health effects) pertaining to the key assessment question(s). IRIS uses multiple strategies to identify primary studies, either published papers or unpublished reports, that provide sufficient detail to allow evaluation of the study methods. The core databases used to search for published studies are described in Table 2-3.

Core Databases

Table 2-3. Core databases of published studies (searched by Health and Environmental Research Online [HERO] or contractors)

Database	Description and Notes
PubMed	Approximately 5,600 medical, biology, and other life sciences journals (through MEDLINE), with coverage back to 1946. Includes some conference abstracts, typically through entry for the proceedings of the entire conference. Uses MeSH terms. Can access through HERO. Test page for developing searches: http://www.ncbi.nlm.nih.gov/pubmed/advanced .
Web of Science	12,000 science and social science journals, back to 1970; includes conference abstracts. Maintained by Thompson Reuters: http://apps.webofknowledge.com , select Web of Science Core databases, advanced search. Can do citation mapping searching (searching for publications that cite a specified reference). Can access through HERO. Test page for developing searches requires subscription.
SCOPUS	35,000 scientific journals, books, and trade publications, including title, abstract, citation, as well as bibliographic analysis tools such as impact measurement.

MeSH = Medical Subject Headings.

Developing Search Terms

Search string design and other aspects of the literature identification strategy should involve information specialists, either with HERO or with a contractor working on the assessment.

Developing and refining search strategies, applying limits in the search strategy, and correctly using Boolean operators (e.g., [AND]/[OR]/[NOT]) requires a high level of training and experience.

Typically, the literature search focuses on the chemical name (and synonyms, trade names, and metabolites/degradants of interest) without additional limits or language restrictions. Chemical synonyms are identified by searching the EPA CompTox Chemicals Dashboard ([U.S. EPA, 2021a](#); [Williams et al., 2021](#)) and selecting those indicated as “valid” or “good” in the Chemicals Dashboard. The preferred chemical name (as presented in the CompTox Chemicals Dashboard), CASRN, and synonyms are used to identify existing toxicity values (see Section 1.2.1) and shared with information specialists who use these inputs to develop search strategies that are specifically formulated to match database specific structure (see Table 2-4 for details). For some assessments, it might be useful to expand the chemical-specific search terms. For example, specification of chemical form(s), active metabolite(s), mixtures, or valence/oxidation state (for metals) can be drawn from work in the scoping and problem formulation stages of the assessment workflow. If studies based in occupational settings are anticipated, expertise in industrial hygiene or occupational epidemiology should be sought to create a list of industries, job categories, and titles that should be included in the search. Full details of the search strategy for each database are presented with the SEM and assessment protocol.

Table 2-4. Summary of search term development strategies for core databases

Search Term Recommendations for PubMed, Web of Science, and Scopus			
	PubMed (http://www.ncbi.nlm.nih.gov/pubmed)	Web of Science (http://apps.webofknowledge.com)	SCOPUS (URL)
What fields are searched by default?	All fields. ^a	Topic, which includes title, abstract, and keyword fields.	Title, abstract, keywords. Other fields searchable as needed.
Can I limit by publication date?	Yes—can refine by publication month and year.	Yes—can refine by publication year only; if possible, schedule search updates to beginning of calendar year.	Yes—can refine by publication year only; if possible, schedule search updates to beginning of calendar year.
Can I limit by language?	Yes—although IRIS does not limit based on language, it is helpful to import foreign language results separately into HERO so that they can be screened based on database-specific metadata that might be available.		
Can I search by CASRN?	Use quotation marks around CASRNs; CASRNs not widely found in Web of Science records.		CASRNs are searchable as a separate field in Scopus
Can I truncate terms?	Use with caution; truncated terms could explode to hundreds of terms. Truncated terms are treated as wildcards and will return up to 600 variations of the truncated term.	Yes.	Yes.
Should I include synonyms in my search strategy?	Yes—include synonyms and alternative spellings; use the EPA CompTox Chemicals Dashboard (U.S. EPA, 2021a ; Williams et al., 2021) to identify synonyms, selecting those indicated as “valid” or “good.” When a chemical is referred to by various names, use the preferred chemical name (in addition to synonyms for the preferred name) as presented in the CompTox Chemicals Dashboard.		
Does the database include “gray” literature?	PubMed and Web of Science are predominantly populated with peer-reviewed publications. However, TOXLINE, once a resource for gray literature from multiple sources, has now been integrated into other National Library of Medicine (NLM) resources, including PubMed. ^b		Scopus primarily indexes peer-reviewed publications, but their broad coverage also includes gray literature sources.

Search Term Recommendations for PubMed, Web of Science, and Scopus			
	PubMed	Web of Science	SCOPUS
	(http://www.ncbi.nlm.nih.gov/pubmed)	(http://apps.webofknowledge.com)	(URL)
Other tips	<p>Reviewing the search details window is highly recommended.</p> <p>Recently published articles might be in PubMed, but not indexed for Medline for several weeks or months.^c</p>	<p>Use research areas to limit search results; recommend choosing research areas to include instead of excluding areas.</p>	<p>Citation coverage is extensive, but the data indexed for each citation is basic, and often benefits from supplementation by other databases.</p>

CASRN = Chemical Abstracts Service registry number; NLM = National Library of Medicine.

^aMedical Subject Headings (MeSH) is the NLM controlled vocabulary thesaurus used for indexing articles for PubMed. If a MeSH or entry term is used in the search strategy, the MeSH field is automatically searched. Using truncation will prevent the MeSH field from being searched—avoid if possible.

^bThe records previously available at TOXLINE, which was phased out in 2019, include citations to Toxic Substances Control Act Test Submissions (TSCATS) records through approximately 2002; these records include health and safety studies, substantial risk notices, and voluntary information submitted to EPA under the Toxic Substances Control Act (TSCA). See <https://www.nlm.nih.gov/toxnet/index.html> for more information. Some studies are available through the National Technical Reports Library (<https://ntrl.ntis.gov/NTRL/>). EPA's website ChemView (<https://chemview.epa.gov/chemview>) contains copies of the actual studies and reports for these types of TSCA submissions.

^cTo search for a term only in the MeSH field, repeat the search in all fields for the most recent 6 months to capture records not yet indexed for Medline.

Building Search Strategies from Prior Assessments

Existing assessment(s) from IRIS and other sources (e.g., EPA, Agency for Toxic Substances and Disease Registry [ATSDR], National Toxicology Program [NTP], or other federal, state, or international health agencies) can be used as a starting point for the literature search. For instance, the assessment team might use the references identified in the existing assessment and conduct database searches that are date limited to begin after the existing assessment was developed. More specifically, the updated search starts one year prior to publication date of the existing assessment to identify studies published during the late stages of finalization of the existing assessment. This approach leverages existing assessments and decreases the time and effort to develop a search strategy for the IRIS assessment. Although the possibility exists that the literature searches conducted for existing assessments might have missed studies, the IRIS process provides overlapping workflows to ensure key literature is identified, including use of additional search strategies (see Table 2-5) and multiple opportunities for public input. References cited in human health chapters of a prior assessment(s) are retrieved by HERO information specialists, assigned a HERO ID (if the record is not already in HERO), imported into a screening application, and screened according to PECO and supplemental material criteria (see Section 2.4 below). If the date of the last literature search is not known for the prior assessment, the new IRIS search should start at least two years before the release date. When the date is specified, the IRIS search should start at the beginning of the calendar year (January 1) in the relevant year the prior assessment was initiated.

2.3.3. Additional Database Searches

Gray Literature and Other Resources Consulted

In addition to searches of the core databases (PubMed, Web of Science, and SCOPUS), the resources listed in Table 2-5 can be consulted to identify studies that could have been missed with searches of the core database. The utility of searching some of the resources in Table 2-5 is assessment specific, and therefore some are indicated as optional, with the expectation that the specific list selected by the assessment team would be indicated in the SEM or protocol. For example, some of the resources are most pertinent to data-poor chemicals with no or very limited animal bioassay or epidemiological evidence available. In other cases, publications can be missed because they are not indexed correctly; the databases searched do not include those journals; or the relevant data in the paper are not mentioned in the title, abstract, or indexing terms. In addition, many older papers (e.g., published before 1970) do not include an abstract and are therefore more difficult to find during the initial literature search process. There might also be gray literature such as technical reports from government agencies or scientific research groups, unpublished laboratory studies conducted by industry, working papers from research groups or committees, white papers, and some foreign language studies that would not be captured by the core database search. Note that although information from the gray literature can be used in IRIS assessments, if

the results are unpublished and are influential to the decisions made in the assessment (e.g., key for hazard characterization, used in dose-response modeling), the studies should be peer reviewed as described in the section below.

A training guide for conducting the gray literature searches below is available in the Health Assessment Workspace Collaborative (HAWC) project “[IRIS PPRTV SEM Template Figures and Resources](#) (2021)” (see “Gray Literature Training” attachment).

Table 2-5. Additional strategy resources for literature identification

Database	Description and notes
Other resources consulted (searched by assessment team or contractors)	
References identified by technical experts or public	References identified by technical consultants, during peer-review, and during public comment periods.
Reference list from studies that meet PECO criteria (Optional)	Manual review (at the title/abstract level) of reference list in studies screened as PECO-relevant after full-text review.
References identified from “backward/forward” literature searching (Optional)	“Backward” searches (to identify articles cited by included studies, reviews, or prior assessments by other agencies) and “forward” searches (to identify articles that cite those studies). <i>This type of searching is done on a case-by-case basis depending on factors such as whether the PECO has a targeted evidence stream or health outcome focus, extent of the evidence, and use of other assessments to serve as a starting point. In general, the feasibility of conducting backward and forward searches is reduced when the PECO is broad, and the number of included studies is large. These searches might be more appropriate to conduct when other assessments are used as the starting point for a review and those other assessments were not conducted using systematic review methods.</i>
EPA CompTox Chemicals Dashboard ToxVal (Optional)	Retrieval of references from EPA CompTox Chemicals Dashboard ToxVal database (U.S. EPA, 2021a ; Williams et al., 2021) to identify studies or assessments that present point of departure (POD) information. ToxValDB collates publicly available toxicity dose-effect related summary values typically used in risk assessments. These include POD data collected from data sources within ACToR (Aggregated Computational Toxicology Resource) and ToxRefDB, and no-observed and lowest-observed (adverse) effect levels (NOEL, NOAEL, LOEL, LOAEL) data extracted from repeated dose toxicity studies submitted under REACH (Registration, Evaluation, Authorisation and Restriction of Chemicals). Also included are reference dose and concentration values (RfDs and RfCs) from EPA’s Integrated Risk Information System (IRIS) and dose descriptors from EPA’s Provisional Peer-Reviewed Toxicity Value (PPRTV) documents. Acute toxicity information was extracted from a number of different sources, including OECD eChemPortal, ECHA (European Chemicals Agency), NLM (National Library of Medicine) HSDB (Hazardous Substances Data Bank), ChemIDplus via EPA TEST (Toxicity Estimation Software Tool), and the EU JRC (European Union Joint Research Centre) AcutoxBASE. Data from the EU COSMOS database project have also been included in ToxValDB. Many of the PODs presented in ToxValDB are based on gray literature studies or assessments not available in databases such as PubMed, WoS, etc. Although many of the resources included in the “Additional Sources” list are represented in ToxValDB, they are also manually searched

ORD Staff Handbook for Developing IRIS Assessments

Database	Description and notes
	<p>because most of the ToxValDB entries have not undergone quality control to ensure accuracy or completeness and might not include recent studies. Searching ToxValDB can be helpful to provide an indication of how much gray literature might be available for a chemical. Searches can be launched from: https://comptox.epa.gov/dashboard/.</p>
<p>ChemView (Searched by assessment team)</p>	<p>Under TSCA, companies that manufacture, process, or commercially distribute a chemical might be required to submit results of chemical monitoring, exposure, and health and safety studies to EPA. Submissions of information made to EPA electronically can be found through EPA’s ChemView online database (U.S. EPA, 2019a). There is no requirement that these studies also be submitted for publication, so this database might be the only source of the data contained in these studies. EPA ChemView database might contain primary hazard studies and summaries such as the following:</p> <ul style="list-style-type: none"> • Unpublished studies, information submitted to EPA under TSCA Section 4 (chemical testing results); Section 8(d) (health and safety studies); Section 8(e) (substantial risk of injury to health or the environment notices); and For Your Information (FYI) submissions (voluntary or third party submitted substantial risk information documents). • Other databases accessible via ChemView include EPA’s High Production Volume (HPV) Challenge database) and the Toxic Release Inventory database. <p>Additional information in ChemView includes EPA actions (such as TSCA Section 5 orders or Significant New Use Rules) and manufacturing, processing, use, and release data submitted to EPA.</p> <p>Searches by chemical and CASRN and a User’s Guide^a can be launched from: https://chemview.epa.gov/chemview. To search ChemView, enter the chemical name(s) or identifier(s), such as CASRN in the left panel of the screen. Scroll down to the bottom of the left panel to check “Select All [/Deselect All] Outputs” under “Show Output Selection.” Click the green button at the bottom left of the screen that says, “Generate Results.” The results will appear on the right side of the screen. Click on the chemical name or colored box to view more specific information. Refer to the User’s Guide on the ChemView website for more details regarding searches.</p>
<p>European Chemicals Agency (ECHA) registration dossiers</p>	<p>European Chemicals Agency (ECHA) registration dossiers to identify data submitted by registrants. Registration dossiers contain data on substances such as hazardous properties, safe uses, classifications, environmental fate, and ecotoxicological and toxicological information. The amount of information provided for each substance varies and is obtained directly from companies’ REACH registrations. ECHA gives no guarantees or warranties regarding the quality and correctness of the published information. The information in the portal is published ‘as provided’ by industry, and its accuracy has not been verified by ECHA (https://echa.europa.eu/information-on-chemicals/registered-substances).</p>
<p>National Toxicology Program (NTP)</p>	<p>NTP Chemical Effects in Biological Systems (CEBS) database of study results and research projects (https://cebs.niehs.nih.gov/cebs/).</p>
<p>eChemPortal</p>	<p>The Organisation for Economic Cooperation and Development (OECD) eChemPortal to retrieve results for OECD Screening Information DataSet (SIDS) and HPV chemicals (https://www.chemportal.org/chemportal/).</p>
<p>AEGLs</p>	<p>AEGLs represent threshold exposure limits of airborne concentrations for the general public applicable to emergency exposures ranging in duration from 10 min to 8 h. AEGL-1</p>

ORD Staff Handbook for Developing IRIS Assessments

Database	Description and notes
	<p>is the concentration above which individuals could experience notable discomfort, irritation, or certain asymptomatic nonsensory effects. AEGL-2 is the concentration above which individuals could experience irreversible or other serious, long-lasting adverse health effects. AEGL-3 is the concentration above which individuals could experience life-threatening adverse health effects or death.</p> <p>AEGLs and their technical support documents are available from the following website: https://www.epa.gov/aegl/access-acute-exposure-guideline-levels-aegls-values#chemicals.</p>
Agricola	Use for U.S. Department of Agriculture-related compounds. Available through HERO. Test page for developing searches: http://agricola.nal.usda.gov/ .
ChemIDPlus	<p>Includes links to resources from a variety of sources in the United States (e.g., ATSDR; Registry of Toxic Effects of Chemical Substances) and other countries (OECD member country assessments of HPV chemicals, summaries of studies submitted to ECHA under REACH, International Uniformed Chemical Information database, IUCLID): http://chem.sis.nlm.nih.gov/chemidplus/.</p> <p>Note that although IUCLID houses similar data, the OECD HPV assessments, or SIAPs and SIARs, do have some government review/oversight. IUCLID summaries can simply house study summaries provided by industry without review by government.</p> <p>OECD SIARs/SIAPs are available through the eChemPortal (https://www.chemportal.org/chemportal/index.action, listed as OECD HPV).</p>
DTIC	Contains government-funded (primarily Department of Defense) research, studies, and other materials relevant to the defense community. Advance search options available through the R&E gateway. Requires government sponsor to access advanced search options: https://www.dtic.mil/DTICOnline/ .
ECOTOX (Optional)	Review of the list of references in the ECOTOX database for the chemical(s) of interest (https://cfpub.epa.gov/ecotox/).
Japan CHEMicals Collaborative Knowledge database (J-CHECK)	Japan CHEMicals Collaborative Knowledge database (J-CHECK, http://www.safe.nite.go.jp/jcheck/top.action) is a database developed to provide the information regarding “Act on the Evaluation of Chemical Substances and Regulation of Their Manufacture, etc.” (CSCL) by the authorities of the law, Ministry of Health, Labour and Welfare, Ministry of Economy, Trade and Industry, and Ministry of the Environment. J-CHECK provides the information regarding CSCL, such as the list of CSCL, chemical safety information obtained in the existing chemicals survey program, risk assessment, etc., in cooperation with eChemPortal by OECD.
OPP, EPA^b IHAD	Contains DERs (reviews of toxicological study reports), memoranda, cancer reports, metabolism reports, etc. for all of OPP. Accessible to any EPA employee with FIFRA confidential business information access authorization.

Database	Description and notes
OPP, EPA^b PRISM Documentum	<p>Contains GLP guideline toxicological study reports for all pesticides from 1996 to present. Study reports older than 1996 can be acquired within a few days. Accessible to any EPA employee with FIFRA confidential business information access authorization. Go to: OPP@Work—http://intranet.epa.gov/opp00002/ (can require permission).</p> <p>OPP Applications (under popular sites in green box on left).</p> <p>e-Registration Workflow (Documentum Login).</p>

AEGL = acute exposure guideline level; ACToR = Aggregated Computational Toxicology Resource; ATSDR = Agency for Toxic Substances and Disease Registry; CASRN = Chemical Abstracts Service registry number; CEBS = Chemical Effects in Biological Systems; CSCL = Chemical Substances Control Law; DER = data evaluation record; DTIC = Defense Technical Information Center; ECHA = European Chemicals Agency; EU = European Union; FIFRA = Federal Insecticide, Fungicide, and Rodenticide Act; FYI = for your information; GLP = Good Laboratory Practice; HSDB = Hazardous Substances Data Bank; HPV = high production volume; IHAD = Integrated Hazard Assessment Database; IRIS = Integrated Risk Information System; IUCLID = International Uniformed Chemical Information Database; JRC = Joint Research Centre; LOAEL = lowest-observed-adverse-effect level; LOEL = lowest-observed-effect level; NLM = National Library of Medicine; NOAEL = no-observed-adverse-effect level; NOEL = no-observed-effect level; NTP = National Toxicology Program; OECD = Organisation for Economic Co-operation and Development; OPP = Office of Pesticide Program; PECO = populations, exposures, comparators, and outcomes; POD = point of departure; PPRTV = Provisional Peer-Reviewed Toxicity Value; PRISM = Pesticide Registration Information System; R&E = research and engineering; REACH = Registration, Evaluation, Authorisation and Restriction of Chemicals; RfC = reference concentration; RfD = reference dose; SIAP = SIDS Initial Assessment Profile; SIAR = SIDS Initial Assessment Report; SIDS = Screening Information DataSet; TEST = Toxicity Estimation Software Tool; TSCA = Toxic Substances Control Act; WoS = Web of Science.

^aTo search for EPA hazard characterizations of high production volume chemicals, use the following steps: Enter chemical identifiers and choose all results (bottom left of page), but make sure the box associated with “EPA Assessments” is checked. Results of this search will appear under the column headed “EPA assessments.” Click on the small dark green/black box to open a page with links to summaries of individual studies. Click on any of the links to view the study summary. On any summary page, there is a link at the top right that says, “View Hazard Characterizations Summary.” Clicking there will bring up another summary box that has a link at the top right to “View Hazard Characterizations.” That will pull up the full hazard characterization written by EPA, which includes an executive summary of all information (physicochemical properties, environmental fate, human health data, and ecotoxicity data). If the chemical has a risk-based prioritization (with a hazard characterization as an appendix), that information will include very preliminary risk information along with some information on uses.

^bContractors do not have access to PRISM Documentum or IHAD; other pesticide databases, such as the National Pesticide Information Retrieval System through Purdue University, can also be assessed for relevance.

Use of Non-Peer-Reviewed Data

IRIS assessments rely mainly on publicly accessible, peer-reviewed information. However, it is possible that unpublished data directly relevant to the PECO criteria are identified during assessment development. On rare occasions, considering the type of report and whether it is expected to have a substantial impact on major assessment conclusions, EPA might obtain external peer review if the owners of the data are willing to have the study details and results made publicly accessible ([U.S. EPA, 2015b](#)). This independent peer review managed by a contractor external to EPA would include an evaluation of the study similar to the peer review done for a journal publication. The contractor would identify and select two or three scientists knowledgeable in

scientific disciplines relevant to the topic as potential peer reviewers. Persons invited to serve as peer reviewers would be screened for conflict of interest prior to confirming their service. In most instances, the peer review would be conducted by letter review. The study authors would be informed of the outcome of the peer review and given an opportunity to clarify issues or provide missing details. The study and its related information, if used in the IRIS assessment, would become publicly available. In the assessment, EPA would acknowledge the document underwent external peer review managed by EPA, and the names of the peer reviewers would be identified. In certain cases, especially when the assessment is time sensitive, the IRIS Program will conduct an assessment for utility and data analysis based on having access to a description of study methods and raw data that have undergone rigorous quality assurance/quality control review (e.g., ToxCast/Tox21 data, results of NTP studies) but that have not yet undergone external peer review.

Unpublished data from personal author communication can supplement a peer-reviewed study, provided the information is made publicly available (typically through documentation in HERO).

Targeted Literature Searches

In later stages of the assessment development process, more refined sets of focused searches might be required. These targeted searches generally fall outside the scope of the initial assessment search strategy. The following bullets provide additional examples of possible scenarios for which a supplemental targeted literature search could be developed.

- A specific health effect question (e.g., reproductive toxicity, cancer, pulmonary function, or even finer divisions such as autoimmunity within the broader area of immunotoxicity); a particular exposure scenario of interest (e.g., exposure during pregnancy, exposure to a specific formulation of the agent); or potentially susceptible subpopulations and lifestages.
- Mechanistic data informing biological pathways that might not involve exposure to the specific agent of interest, but are informative to, for example, the human relevance or adversity of the biological effect.
- Studies using descriptions of exposure to the agent of interest that do not include the chemical name (e.g., epidemiological studies of a broad chemical class or occupation might provide useful information).
- ADME and mechanistic studies, or studies of PBPK models; searches using the parent chemical name and CASRN alone might be too limiting for these types of data.

2.3.4. Removing Duplicates

The literature search strategy includes searching across multiple bibliographic databases. These databases have much of the same content, but often with slight variations in bibliographic format. Removing duplicate references can be a labor intensive process but is important. Failure to

remove duplicates causes problems in tracking the literature results (e.g., the number in the database will change when duplicates are later identified and removed). HERO automatically removes duplicates as searches from individual databases (e.g., PubMed) are added to the HERO Project Page (see Figure 2-1). HERO uses five automated duplicate checking screens while importing references; however, some duplicates might persist and will require human review to identify and resolve. Many software applications used to screen studies for relevance (e.g., DistillerSR) have features to facilitate identifying duplicates that are not exact matches. Duplicates identified during screening should be sent to HERO@epa.gov for removal, indicating which HERO ID to delete and which to keep (e.g., 5678 keep 1234). HERO convention is to retain the smaller HERO ID number; HERO IDs are found in the label field in the Research Information Systems (RIS) file.

2.3.5. Updating the Literature Search

The literature search is updated periodically to identify new literature published during assessment development and review. The frequency of updates varies across assessments and is related to factors such as the size of the evidence base and any insight the assessment team has on volume of new research being released. Studies identified in literature search updates are screened according to the problem formulation and assessment PECO criteria, which allows the SEM to be continually updated throughout assessment development; however, depending on the size and scope of the assessment, the team might elect to screen the literature search updates only according to assessment PECO criteria. The last full literature search update is conducted several months prior to the planned release of the draft document for public comment. The assessment team will manage the literature update process with HERO information specialists, including (but not limited to) when and how often an update should be performed, updating search strategies, etc. If the search string(s) are altered for an update, the dates for this search should include the years encompassed by the original literature search and previous updates for the assessment. Subsequent updates should use the latest search string. Studies identified after peer review begins are considered for inclusion only if they are directly relevant to the assessment PECO criteria and are expected to potentially impact assessment conclusions or address key uncertainties.

2.3.6. Documenting Search Results

Accurate documentation of the search strategy is essential to the systematic review process. Documentation of literature searches should include the database(s) and date range covered by the search, search terms used and the filters (e.g., matching specific article types or PubMed Medical Subject Headings [MeSH] terms, matching topic areas in Web of Science) that were applied, and date(s) the searches were performed (see an example template for documentation in Table 2-6). Documentation of gray literature and other additional resources (see Table 2-5) should also be summarized to include the date(s) of search(es), source type or name, the search string (when

applicable), the URL (when available and applicable), number of results, and number of unique references not otherwise identified from database searching.

Table 2-6. Example summary template of literature search results documentation

Database	Terms	# Citations
PubMed Date range	CHEMICAL TERMS; ADDITIONAL TERMS Search strings should include use of Boolean operators, wildcard, and punctuation.	Citation count should be presented for each search date
Web of Science Date range	CHEMICAL TERMS; ADDITIONAL TERMS Search strings should include use of Boolean operators, wildcard, and punctuation.	
Other database Date range	CHEMICAL TERMS; ADDITIONAL TERMS Search strings should include use of Boolean operators, wildcard, and punctuation.	
Merged reference set	(After removal of duplicates.)	

2.4. LITERATURE SCREENING

The literature screening process focuses on categorizing (or “tagging”) studies by those that provide data relevant to the PECO criteria or supplemental information. It is important to emphasize that during the screening process neither the quality nor the results of the study are considered. Although a contractor can help facilitate this process, the assessment manager and assessment team should be directly involved in the literature screening process. A variety of software applications can be used for screening. All screening applications make use of structured forms to guide the process, which can be tailored to meet assessments needs.

The literature screening results are released to the public in the IAP, protocol, and draft assessment. Screening results released as part of the IAP and protocol reflect screening that was done to support problem formulation (i.e., SEM screening results) and used to define the focus of the assessment. Any additional studies identified during public comment will be screened for adherence to the PECO criteria (see Section 2.1).

2.4.1. Title and Abstract Screening

The citations identified from the searches described above are imported into screening software application(s) that might or might not use ML (see Section 2.4.4 for details on software applications used by IRIS). Following a pilot phase to calibrate screening guidance, two screeners independently perform a TIAB screen using a structured form. Citations considered “relevant” or “unclear” on the basis of the PECO criteria at the TIAB level are considered for inclusion and advanced to full-text screening. Any screening conflicts must be resolved between the two

independent reviewers, with consultation by a third reviewer if needed. Other approaches can be used in circumstances where time frames and resource availability make use of two screeners impractical. For example, it is acceptable to require only one screener to screen a citation as “include” but require two screeners to screen a citation as “exclude.” This is acceptable because those studies marked as included would be confirmed relevant at the full-text level. It is not necessary for screeners to annotate the rationale for excluding studies at the TIAB level, since studies are frequently excluded for failing to meet multiple PECO criteria and this becomes cumbersome to track.

TIAB screening should serve to quickly remove most nonpertinent studies from consideration (excluded studies). To ensure that all relevant studies are included, it is best to err on the side of including studies for full-text review when potential relevance is unclear. Also, during TIAB screening, studies not meeting the PECO criteria but identified as supplemental content can be identified and categorized (i.e., tagged) as such. It is possible that studies meeting the PECO criteria also contain supplemental material content and should be additionally tagged as “relevant” *and* “supplemental.” For example, a citation might examine health effect-related endpoints in exposed humans, but also test endpoints related to potential mechanisms and metabolism of the test agent. In this case, the citation would be considered as meeting the PECO criteria but should also be tagged for having supplemental mechanistic *and* ADME content. Conflicts between screeners in applying the supplemental tags, which primarily occur at the TIAB level, are resolved similarly, erring on the side of over-tagging based on TIAB content. Note that more granular subtagging of supplemental material occurs during preparation of the literature inventory as described in Section 2.5.2. In addition, during preparation of the literature inventory, supplemental content can be identified (and tagged as such) in studies that meet the PECO criteria.

For citations with no abstract, articles are initially screened on the basis of the following: title relevance (title should indicate clear relevance), and page length (articles two pages in length or less are assumed to be conference reports, editorials, or letters). Eligibility status of non-English studies is assessed using the same approach with online translation tools or engagement with a native speaker.

Prior to importing gray literature and literature from additional search strategies into a screening software application (e.g., DistillerSR), a unique reference citation for each identified citation is generated in HERO. This process includes a step to verify the reference was not already identified from the core database searches (e.g., PubMed, Web of Science [WoS]). Unique references are then screened according to the PECO criteria using the same methods applied to the core database search results.

2.4.2. Full-Text Screening

Full-text references are sought through EPA’s HERO database for citations identified as meeting PECO criteria or “unclear” based on the TIAB screening. Full-text copies of these records are retrieved, stored in the HERO database, and independently assessed by two screeners to

confirm eligibility for inclusion according to the PECO criteria. Screening conflicts are resolved by discussion among the primary screeners with consultation by a third reviewer or technical advisor as needed to resolve any remaining disagreements. Rationales for excluding studies are documented, e.g., citation did not meet PECO, full-text not available, critical reporting/analysis limitations. Approaches for screening non-English studies include online translation tools and engagement of a native speaker. Use of fee-based translation services to generate reports for public dissemination are costly and would be prioritized for non-English studies likely to be informative on hazard conclusions or dose-response analysis. Otherwise, non-English studies are tracked as supplemental material.

Other Exclusions based on Full-text Content

In addition to failure to meet PECO criteria (described above), epidemiological and toxicological studies could be excluded at the full-text level due to critical reporting limitations. Reporting limitations can be identified during full-text screening but are more commonly identified during subsequent phases of the assessment (e.g., literature inventory, study evaluation). Regardless of when the limitation is identified, exclusions based on full-text content are documented at the level of full-text exclusions in literature flow diagrams with a rationale of “critical reporting limitation.”

A similar approach is taken for in vitro studies prioritized for focused analysis during assessment development (i.e., the critical reporting deficiency might preclude them from consideration). Critical reporting information for different study types are summarized below. For each piece of information, if the information can be inferred (when not directly stated) for an exposure/endpoint combination, the citation should be included.

Epidemiological studies

- Sample size
- Exposure characterization or measurement method. Note, studies for which the chemical of interest was not detectable in the study population would be tagged as excluded for not meeting PECO with respect to the exposure component.
- Outcome ascertainment method
- Study design
- Quantitative or qualitative (e.g., author-reported lack of an effect on the outcome, graphical display) results for at least one endpoint of interest

Animal studies

- Species

- Test article name
- Levels and duration of exposure
- Route of exposure
- Quantitative or qualitative results for at least one endpoint of interest

In vitro studies prioritized for focused analysis

- Cell/tissue type(s) or test system
- Test article name
- Concentration and duration of treatment
- Quantitative or qualitative results for at least one endpoint of interest

2.4.3. Multiple Publications of the Same Data

When there are multiple publications using the same or overlapping data, all publications are included, with one selected for use as the primary citation; the others are considered as secondary publications with annotation in HAWC and HERO indicating their relationship to the primary record during data extraction. For epidemiological studies, the primary publication is generally the one with the longest follow-up, the largest number of cases, or the most recent publication date. For animal studies, the primary publication is typically the one with the longest duration of exposure, the largest sample size, or with the outcome(s) most informative to the PECO. For both epidemiological and animal studies, the assessments include relevant data from all publications of the study, although if the same data are reported in more than one citation, the data are extracted only once (see Chapter 5). For corrections, retractions, and other companion documents to the included publications, a similar approach to annotation is taken and the most recently published data are incorporated into the assessments.

2.4.4. Systematic Review Software and Artificial Intelligence (AI) Tools

Table 2-7 describes software applications commonly used for IRIS assessments as of 2022 to facilitate screening (DistillerSR, SWIFT-Review, SWIFT-Active Screener); literature inventory (DistillerSR); data extraction and study evaluation (HAWC); and data visualization (HAWC, Tableau). The use of systematic review software tools is documented in the assessment's systematic review protocol.

Some systematic review (SR) literature screening software (e.g., SWIFT-Active, DistillerSR) incorporate AI features to streamline TIAB screening and identify studies most likely to be relevant to the assessment. The decision to apply AI during assessment development should consider the number of studies that need to be screened, the size of the screening team, whether training data/models are available, projected assessment delivery date, etc. For example, manual screening

at the TIAB level using DistillerSR is relatively fast (typically 10–20 seconds per citation), so for smaller screening projects of fewer than 2,000 studies, there might not be a significant time savings by using AI approaches. For projects with more than 2,000 citations, AI tools might save time and be considered for screening using machine learning (ML). In some instances, a large database can be prioritized using ML approaches that leverage a seed set (training data) that includes studies previously screened and passed through a quality-control (QC) process as a validation data set. Care should be taken when seed studies are used to provide sufficient coverage of studies that meet PECO and supplemental criteria. If seed studies are not available, then active learning approaches such as SWIFT-Active Screener can be considered ([Howard et al., 2020](#)).

The availability of specialized software applications for conducting literature assessments is expanding rapidly, especially for screening studies for relevance ([Tsafnat et al., 2014](#)), and it is likely that the SR software and AI tools used within the IRIS Program will evolve and expand over time. The SR Toolbox (<http://systematicreviewtools.com/>) is a repository of available tools that has advanced search features to help a user find tools tailored to specific task(s) of systematic review. Before using a new software tool, the assessment team should be prepared to confirm its methodological documentation, performance capabilities, audit functions, and availability of technical support. Preferred software applications are publicly available, free (when possible), interoperable with other software applications used behind EPA firewalls, and have access to technical support and documentation provided by the developer. Note that HERO IDs are the key, unique identifiers for references retrieved and assessed, and it is therefore essential the software application maintain HERO ID provenance.

Users are encouraged to use training materials provided by the developer when using these tools. One-on-one or small group training sessions—both internal and external to EPA—can be organized upon request by contacting IRIS Program staff. When methodological documentation for software applications that use ML is not available from the developer, the performance should be evaluated internally prior to implementation.

Table 2-7. Summary of commonly used specialized software applications for literature screening and visualization

Software	Key features	Use in IRIS assessments
DistillerSR	<ul style="list-style-type: none"> • Web-based. • Subscription required. • Artificial intelligence features added in 2018. • Easy to add screeners, including from outside EPA. • Help instructions available from within the software. • Full-text articles can be uploaded as attachments (individually or in batch) or accessed via HERO URLs. For IRIS purposes, URLs are preferred to PDFs to address issues related to copyright restrictions. • Form customization options are extensive and can be done by the user (i.e., do not require programmer support). Forms can be used for screening or for data extraction. • Mail merge features in Word can be used to create tables based on DistillerSR Excel input files. • Interoperable with HERO and other software applications such as, SWIFT-Active Screener, and HAWC. 	<ul style="list-style-type: none"> • Used in IRIS assessments for title-abstract screening, full-text screening, and to conduct the data extraction used to create literature inventories. • The IRIS Program generally uses HAWC for full data extraction of studies that meet assessment PECO criteria. DistillerSR does not have the visualization capabilities of HAWC.
SWIFT-Review	<ul style="list-style-type: none"> • Must be downloaded for installation. • Free. • Preset literature search filters can be used to automatically tag and identify different types of study populations (human, animal, in vitro) and health outcomes (Howard et al., 2016) <ul style="list-style-type: none"> ○ The search strategies used in the filters were developed by professional information scientists and are available from within the software and documented online. The search strategies can be customized by the user. • Machine learning (ML) module prioritizes documents based on title, abstract, and keyword information, given a user-defined training set. • Prioritized records must be exported into another software application for screening. 	<ul style="list-style-type: none"> • The search filters are widely used in IRIS assessments during problem formulation and to prioritize records for screening in another software application (e.g., Figures 2-3 and 2-4 in Section 2.4.5).

Software	Key features	Use in IRIS assessments
	<ul style="list-style-type: none"> • Help instructions available from within the software. • Interoperable with HERO and other software applications such as DistillerSR, SWIFT-Active Screener, and HAWC. 	
SWIFT-Active Screener	<ul style="list-style-type: none"> • Web-based and free (upon request). • Easy to add screeners, including from outside EPA. • Incorporates “Active Learning” ML methods that continuously update a prioritization model during screening, pushing the articles most likely to be relevant to the top of the list (Howard et al., 2020). • Incorporates a statistical model that estimates recall (percentage of relevant articles found so far), allowing users to make an educated decision about when to stop screening. • ML and recall estimation models have been validated using a large corpus comprising 26 systematic review data sets varying in size, percentage of relevant studies, and overall topic area. • Help instructions available from within the software. • Full-text articles can be uploaded as attachments (individually or in batch) or accessed via HERO URLs. For IRIS purposes, URLs are preferred to PDFs to address issues related to copyright restrictions. • Form creation and customization can be done by the user (i.e., does not require programmer support). • Interoperable with HERO and other software applications such as DistillerSR, SWIFT Review, and HAWC. 	<ul style="list-style-type: none"> • Widely used in IRIS assessments for title and abstract screening, especially when there are many studies to screen (e.g., 2,000+) or there is time urgency. Under rapid time frames, use of one screener can be considered for title and abstract screening. Full-text screening is not typically done in SWIFT Active because of the extensive tagging that occurs at this level, which is easier to conduct in DistillerSR.
HAWC	<ul style="list-style-type: none"> • Web-based, free, open-source application (Shapiro et al., 2018) • Literature screening, tagging, data extraction and study evaluation capabilities, interactive visualizations, literature tag trees, exposure-response arrays, study evaluation visualization, and evidence maps. • Integration with National Institutes of Health and EPA software • Limited artificial intelligence capabilities; flexibility in building custom, interactive visualizations, and tabular summaries. 	<ul style="list-style-type: none"> • IRIS uses HAWC extensively for study evaluation and data extraction (see Chapter 5), but not currently for literature screening (does not support multiple screeners and conflict identification/resolution tracking). Screening decisions from other software applications can be imported into HAWC

Software	Key features	Use in IRIS assessments
	<ul style="list-style-type: none"> • Interoperable with HERO and other software applications such as DistillerSR, SWIFT Review, and SWIFT Active Screener. 	for subsequent study evaluation and data extraction.
Tableau	<ul style="list-style-type: none"> • Tableau is not a screening tool but can be used to create web-based interactive literature inventories. • Free version available to read, but subscription required to generate visuals. • Help instructions available from within the software. • Allows user to create many different visual displays. 	<ul style="list-style-type: none"> • The input Excel file is typically based on literature inventories developed in DistillerSR.

IRIS = Integrated Risk Information System; ML = machine learning; PECO = populations, exposures, comparators.

DistillerSR: <https://www.evidencepartners.com/products/distillersr-systematic-review-software/>.

SWIFT Review: <https://www.sciome.com/swift-review/>.

SWIFT Active: <https://www.sciome.com/swift-activescreener/>.

HAWC: EPA version <https://hawcprd.epa.gov/portal/>; Public version <https://hawcproject.org/>.

Tableau: <https://public.tableau.com/en-us/s/>.

2.4.5. Literature Flow Diagrams

The results of the screening process are posted on the project page for the assessment in the HERO database. Results are also summarized in a literature flow diagram and interactive HAWC literature tree (where additional subtagging beyond what is presented in HERO is documented and visualized, e.g., more details on the nature of mechanistic or ADME studies). Figures 2-2 and 2-3 present example literature flow diagrams for displaying search and screening results, including for projects that used ML.

Results of the search and screening process are imported into HAWC to create interactive literature tree visualizations using the “Literature Review” module. An example is presented in Figure 2-4.

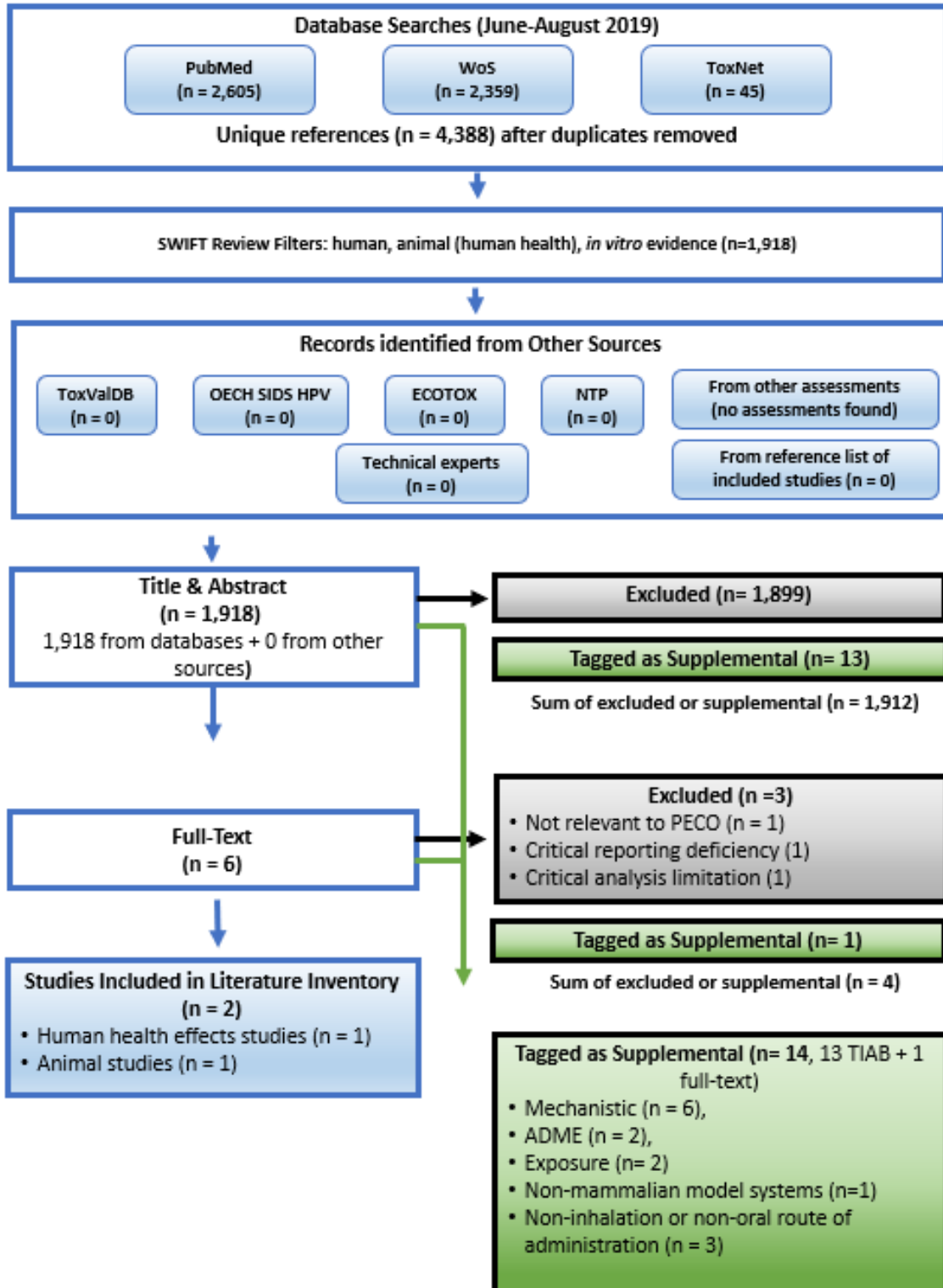


Figure 2-2. Literature flow diagram: No machine learning (ML) software used.

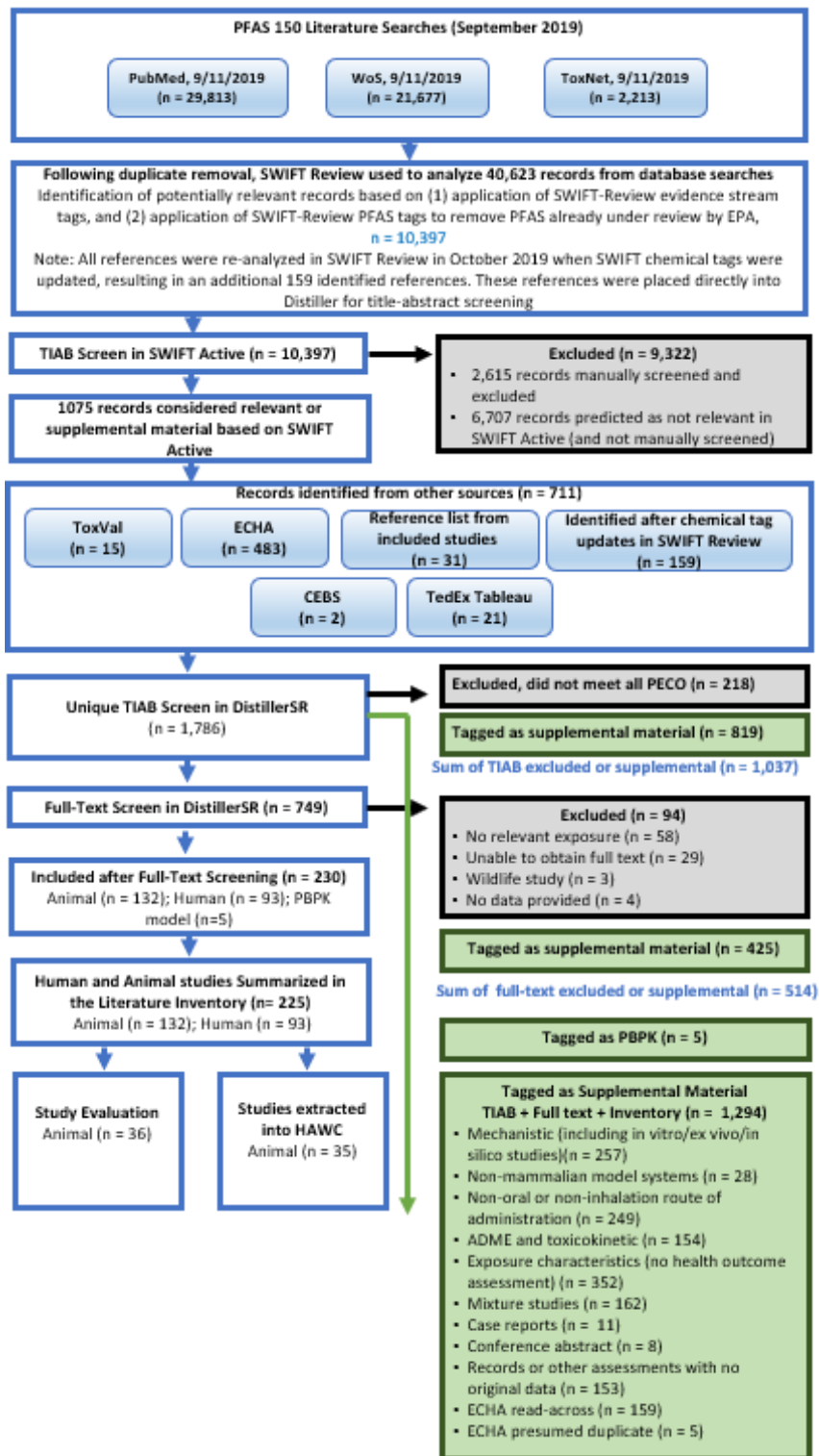


Figure 2-3. Literature flow diagram: Machine learning (ML) software.

ADME = absorption, distribution, metabolism, and excretion; CEBS = Chemical Effects in Biological Systems; ECHA = European Chemicals Agency; PBPK = physiologically based pharmacokinetic; PECO = population, exposure, comparators, outcome; TIAB = title and abstract.

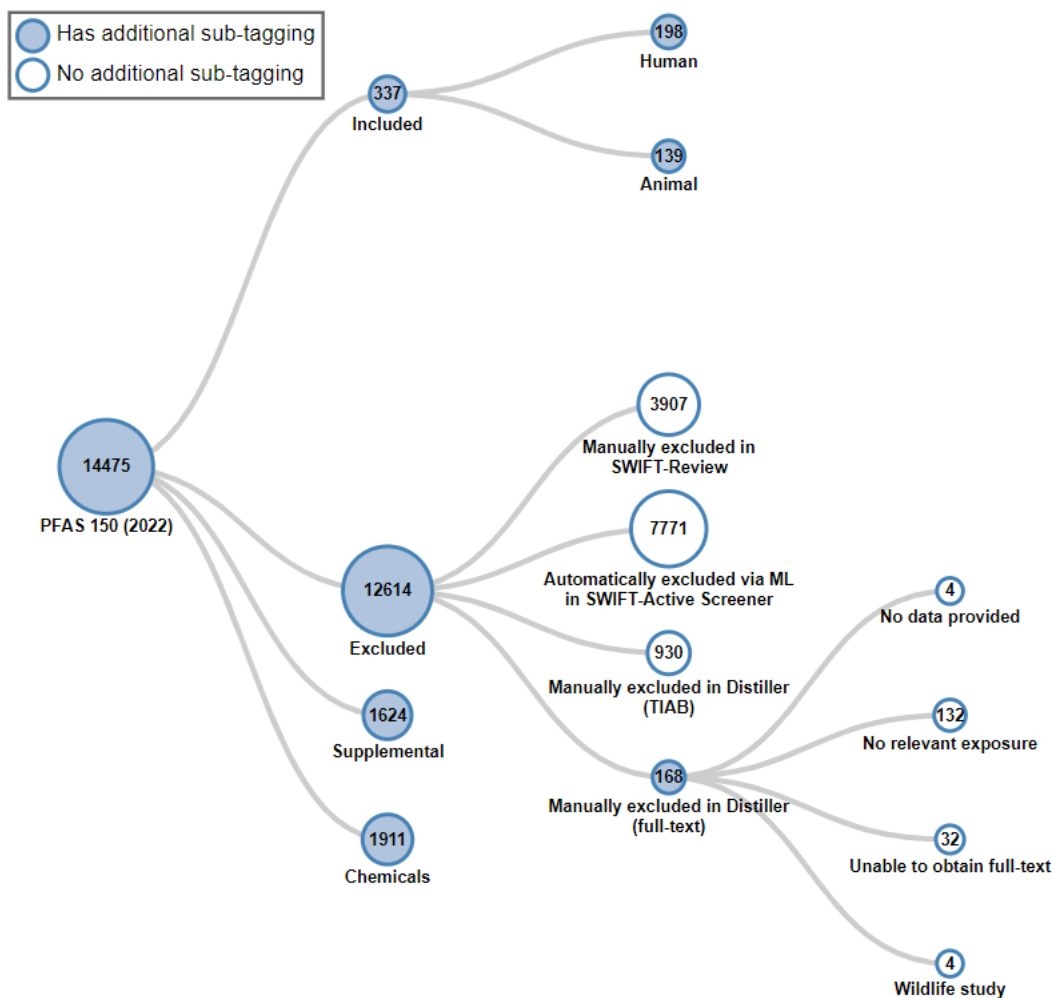


Figure 2-4. Health Assessment Workspace Collaborative (HAWC) literature tree.

ML = machine learning; PFAS = per-and polyfluoroalkyl substances; TIAB = title and abstract.

2.5. LITERATURE INVENTORIES

Literature inventories are summary level, sortable lists of the available citations that include additional basic study design elements (beyond the tags applied during screening) to be used by the subject matter experts to organize and prioritize studies for further review. IRIS assessments typically include inventories of studies meeting PECO criteria (see Section 2.5.1). Depending on assessment needs, separate inventories may also be developed for supplemental material to facilitate review by subject matter experts and identify supplemental information that could be informative to the assessment (see Section 2.5.2).

The purpose of a literature inventory is to help the assessment team organize and prioritize their review of studies by subject matter experts. Importantly, the inventories should not include a detailed extraction of study details; rather, they should be limited to a few key pieces of information that can be quickly extracted. Literature inventories provide insight into the evidence base characteristics (such as health outcome, exposure duration, dosing regimen, etc.) useful for prioritization decisions that are further described in Chapter 3, Assessment PECO and Evaluation Plan, and can be useful for organizing the hazard review and identifying areas of expertise that will be needed to appraise studies and develop the synthesis.

Literature inventories are typically prepared using DistillerSR, with one reviewer extracting each study and a second reviewer providing QC. Template forms are available in DistillerSR in the “IRIS Template Form” project and can be customized as needed by the assessment team. Inventories can be developed by a contractor or by in-house personnel; however, decisions regarding the groupings of study types and the basic study information to be extracted should be made by the assessment team, in consultation with disciplinary workgroups as needed.

2.5.1. Literature Inventory of Studies Meeting Populations, Exposures, Comparators, and Outcomes (PECO) Criteria

An initial inventory of studies meeting PECO criteria is prepared during problem formulation as part of the SEM, as described in Section 1.2, and is updated with all subsequent literature search updates. This inventory is based on full text and generally encompasses basic aspects of study design and endpoints. For epidemiological studies, the inventory includes information on study design (e.g., cross-sectional, cohort, case-control); study population (e.g., adults, children, occupational); major route of exposure if known; and method of exposure measurement (e.g., biomarker, air, water, food, occupational). For animal toxicological studies, it includes information on exposure duration and timing (e.g., acute, chronic, developmental); administered exposure levels; route of exposure; lifestage of exposure; species, strain, and sex. The endpoints evaluated in each study are extracted into the inventory and are categorized according to health system (e.g., cancer, neurological, immune). The Environmental Health Vocabulary (EHV) database available in HAWC (<https://hawc.epa.gov/vocab/ehv/>) can be used as a resource to map specific endpoints to the most pertinent health outcome. The EHV endpoint term list is also available for download in the “IRIS PPRTV SEM Template Figures and Resources (2021)” project in HAWC. A brief description of key study findings (e.g., the direction of effect for each endpoint) can also be extracted. Extracting more detailed study information at this stage is typically not recommended as the literature inventory is intended to aid problem formulation for the draft assessment, where more detailed extraction is conducted for studies that meet assessment PECO criteria.

2.5.2. Supplemental Content Literature Inventories

The analysis of supplemental evidence will be shaped by assessment needs and the PECO criteria, and therefore the decision to develop inventories of studies tagged as “potentially relevant supplemental material” is typically made in the systematic review protocol (see Chapter 3). The literature inventory identifies the main sources of available supplemental evidence and serves as the starting point for organizing the evidence to be analyzed. The categorization of mechanistic, ADME/PK, and other supplemental information helps identify the available evidence to address key assessment questions. This knowledge allows the assessment team to provide more details on the analysis of mechanistic evidence in the assessment protocol compared to what can be done at assessment initiation. When no or minimal evidence is available, this may indicate an area of uncertainty that is not possible to address during the assessment. When evidence is available, focused analyses can then be considered and prioritized. Focused analyses that could be pursued based on evidence availability, but that do not address key issues pertinent to the assessment, may not be prioritized.

The early identification of predefined mechanistic analyses and key science issues during the scoping and problem formulation phase described in Chapter 1 helps frame the approach used for organizing the literature inventory. It is important to consider the likely impact of potentially controversial mechanistic issues (e.g., evidence a chemical is mutagenic, the human relevance of α 2u globulin) on assessment conclusions early in the process. This involves an initial review of existing mechanistic analyses and information regarding the ADME/PK of the chemical and possibly other related chemicals in the same class. Even a cursory mechanistic understanding of how a health outcome develops can help identify susceptible population groups. The early identification of lifestages or groups likely at greatest risk can clarify hazard descriptions, including whether the most susceptible populations and lifestages have been adequately tested.

The basic process for developing literature inventories for supplemental evidence is similar to that described above for studies meeting PECO criteria. Supplemental literature inventories should be derived from base forms available in DistillerSR but are customizable (with tags selected depending on the evidence base) and can be tailored to evaluate key issues specific to the agent (e.g., regarding ADME, mechanistic pathways, susceptibility, human relevance) that arise during assessment development. The data extraction forms and overall organizational schema shaping the supplemental inventories are informed by the literature inventories of studies meeting PECO criteria and the developing assessment. Note that literature inventories are intended to provide a high level “snapshot” of the available evidence. Therefore, forms should facilitate rapid data extraction and efficient analysis and synthesis of the summary level information. It is important during these initial screening stages for the assessment team to become familiar with chemical-specific issues and potentially relevant toxicity pathways. This preliminary work could include clearly defining the chemical(s) of interest, active metabolites, and applicable chemical formulations. Importantly, supplemental literature inventories help the assessment team narrow

the available evidence to those studies most relevant to informing hazard evaluations and assessment conclusions. Inventories of supplemental material are most commonly developed for mechanistic and ADME/PBPK information, which are described in the sections below. Depending on assessment needs, it might also be useful to develop inventories for other supplemental categories such as non-PECO routes of exposure and susceptible populations. In addition, the screening tools and inventories provide a decision record that increases transparency in the process for analyzing supplemental information.

Mechanistic Information Inventories

Many IRIS assessments identify a large number of studies that report mechanistic information, so typically a tiered approach to the tagging and data extraction workflow is used when developing the mechanistic literature inventory. During the initial literature inventory, mechanistic evidence is tagged as supplemental content, as described in Section 2.2. Next, at the assessment-specific protocol level, prioritized analyses of mechanistic evidence resulting from refinements made to the PECO criteria and the decisions made on health effects that will be the focus for the human and animal evidence syntheses are identified (see Chapter 3) and described in the protocol. Based on these prioritized analyses, a more granular plan for the tagging of mechanistic studies is formulated by disciplinary experts, selecting studies pertinent to address those analyses. The specific mechanistic tagging structure will evolve as the assessment needs are identified and might not be known at the time of the protocol release.

Once the mechanistic literature inventory is developed, it is useful not only for determining whether to include a supplemental study in an assessment, but also for deciding whether study evaluation, typically only conducted for PECO studies, is warranted. Mechanistic studies are not included in the PECO because they typically include nontraditional routes of exposure or study endpoints upstream of apical effects that are not suitable for the derivation of toxicity values; however, because mechanistic considerations could still influence the shape of the dose-response curve, in rare instances they might require study evaluation. Since study evaluations are time-intensive, the assessment team should carefully consider the utility of full study evaluation with respect to addressing key assessment analyses and uncertainties, resource constraints, and time frames.

The stepwise selection of more detailed mechanistic categorization and tagging is approached in a variety of ways. For instance, even with minimal decisions on the prioritization of studies reporting mechanistic information, studies can initially be organized with respect to study design and results on the basis of relevance to broad categories, e.g., health system, organ, and outcome (note that studies can be added to more than one category). The inventory might also capture any known issues relating to chemical purity and mixtures of isomers, valence, or oxidation state (for metals), or concerns regarding solubility or volatility. Additional, targeted categories corresponding to mechanistic or key events (i.e., as part of an MOA or adverse outcome pathway [AOP]) or biological pathways can be added to the base screening and data extraction forms.

The supplemental literature inventories may help highlight database deficiencies for chemicals that have little if any mechanistic information reported in the literature, or, conversely, deficiencies in the animal and human health effect literature where only mechanistic studies are available to inform hazard. These categories are typically based on the health effects indicated by the human and animal evidence or existing MOA hypotheses, and, where available, on key characteristics ([Smith et al., 2016](#)), an objective organizational approach based upon common properties of known toxic agents that can facilitate the grouping of studies reporting mechanistically related endpoints and assays. Key characteristics approaches have been identified for carcinogens ([Smith et al., 2016](#)), male reproductive toxicants ([Arzuaga et al., 2019](#)), female reproductive toxicants ([Luderer et al., 2019](#)), endocrine disrupting chemicals ([La Merrill et al., 2020](#)), cardiovascular toxicants ([Lind et al., 2021](#)), hepatotoxicants ([Rusyn et al., 2021](#)), and immunotoxic agents ([Germolec et al., 2022](#)).

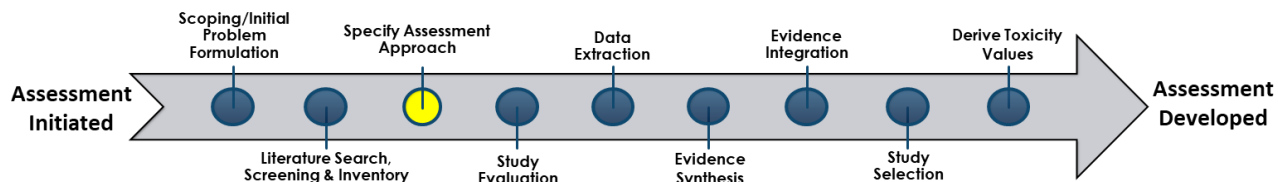
Ultimately, once the prioritized mechanistic categories have been identified, the inventory should capture high level information from these studies that will facilitate further analysis of the studies. Examples of easily extractable information include the test article, vehicle, and method of exposure (including exposure levels tested); experimental design (e.g., in vivo, in vitro, in silico); the species, strain, and sex of the experimental model; the tissue, region, or cell type studied; and the endpoints or outcomes measured, the assays used, and results. This information will facilitate prioritization decisions that will help focus the mechanistic syntheses.

Absorption, Distribution, Metabolism, and Excretion (ADME), Pharmacokinetic (PK), or Physiologically Based Pharmacokinetic (PBPK) Study Inventories

Similar to mechanistic evidence, a tiered approach is used when organizing and summarizing ADME and PK/PBPK model content. During the initial literature inventory, ADME and PK/PBPK model evidence is tagged as supplemental content as described in Section 2.2. Next, more granular tagging of studies is conducted by disciplinary experts. This more granular tagging can be presented in the IAP but is more commonly presented in protocol. Common approaches for this more granular tagging are to categorize primary data ADME studies as absorption, distribution, metabolism, or excretion (using a “tag all that apply” approach). PK/PBPK models are often tagged according to species applicability, i.e., animal, human, or multiple species (to include human).

Almost all ADME studies provide information that is at least qualitatively useful. Because ADME studies vary widely in study design and reported details, flexible Microsoft Excel-based inventory table structures have been developed. The inventory may include information on the type of evidence (human, animal, in vitro/ex vivo); type of ADME; route of exposure; parent compound or metabolite; and range of exposures and timepoints studied. This inventory can also be used to summarize publications describing PBPK/PK computational models, which may or may not include unique ADME data. The identification of existing PBPK models warrants the immediate initiation of model scoping efforts (see Section 4.6).

3. REFINE PROBLEM FORMULATION AND SPECIFY ASSESSMENT APPROACH



Purpose

- Define refinements to the assessment approach based on the Literature Inventory.

The purpose of this section is to refine the assessment approach, including problem formulation decisions, on the basis of the extent and nature of the evidence identified during scoping and initial problem formulation. This includes defining the assessment populations, exposures, comparators, and outcomes (PECO) criteria (i.e., refinements to the problem formulation PECO criteria), defining the unit(s) of analysis of health effect categories to be used during evidence synthesis, and presenting any assessment-specific analysis approaches for mechanistic, absorption, distribution, metabolism, and excretion (ADME), or other types of supplemental material content. These refinements are documented in the systematic review protocol. As mentioned previously (see Section 1.3.2), the protocol also includes methodological details on the process that is used for literature search and screening, study evaluation, the structured frameworks used during evidence synthesis and integration, dose-response, and toxicity value derivation. Any adjustments to assessment methods that occur after the protocol is released will be documented as an amendment to the protocol when the draft assessment is released. These adjustments may occur as a result of complex analyses that often happen during assessment development or the various review steps in the Integrated Risk Information System (IRIS) [assessment development process](#) prior to public release of the draft assessment (i.e., Division/ORD review, Agency review, interagency review).

3.1. ASSESSMENT POPULATIONS, EXPOSURES, COMPARATORS, AND OUTCOMES (PECO) CRITERIA

On the basis of information identified during scoping and problem formulation, including the initial literature inventory, the assessment PECO criteria may be developed from the initial

problem formulation PECO to focus on the studies most likely to be informative to assessment conclusions, such as:

- A subset of health effects, e.g., to focus on those with enough evidence to support developing hazard conclusions
- Toxicity studies employing an exposure route(s) that is the primary focus of the assessment
- Studies with more specific or objective measures of toxicity (e.g., functional endpoints), rather than studies with nonapical, broad, or nonspecific measures (e.g., self-reported symptoms)
- Studies that address critical lifestage or exposure duration based on specific knowledge of the development of the health outcome (e.g., for endpoints relating to organ development or cancer, respectively)
- Inclusion of mechanistic precursors or biomarkers that can be measured upstream of an apical outcome

3.2. DEFINING UNITS OF ANALYSIS

A unit of analysis is an outcome or group of related outcomes within a health effect category considered together during evidence synthesis. Specification of these groupings is needed for transparency of the evidence synthesis and integration phases of systematic review. The planned units of analysis might differ across individual assessments, depending on the nature and extent of the available evidence, and are described in the systematic review protocol.

Some types of evidence not meeting PECO criteria may be included in a unit of analysis. For example, there might be nonapical evidence evaluated in a unit of analysis when there is a strong, biologically plausible rationale (e.g., upstream precursors or biomarkers of exposure or effect known to precede an apical outcome). Supplemental evidence providing support to the human and animal evidence might also be included in a unit of analysis to provide direct support to the unit of analysis judgment, for example, by including evidence from animal bioassays using a non-PECO route of exposure. More typically, mechanistic and other supplemental evidence will be prioritized and synthesized with the goal of informing coherence, biological significance, or directness of outcome measures in the within-stream human and animal evidence synthesis judgments (see Section 6.1.2), or to inform considerations during evidence integration (e.g., human relevance of findings; biological plausibility) (see Section 6.2).

Identifying units of analysis for evidence synthesis is informed by understanding the available evidence for the chemical regarding routes of exposure, metabolism and distribution, health categories evaluated, and number of studies within each evidence stream pertaining to each health category. Thus, for some databases, the available evidence might be sufficient to draw separate conclusions for subcategories of evidence within an organ system. For example, within the overall category of respiratory effects, the evidence could be synthesized separately for biomarkers

of effect in bronchoalveolar lavage fluid, asthma, respiratory infection, pathological endpoints in the upper and lower respiratory tract, and findings in noninvasive tests of pulmonary function. These decisions might differ across the human and animal evidence syntheses, particularly when the effects evaluated in the available studies do not easily align (e.g., spontaneous abortion observed in human studies might relate to endpoints in female reproductive or developmental studies in animal studies). Such decisions can sometimes be informed by specific mechanistic evaluations, for example, analyses of the extent of the biological linkage between related outcomes. If a mechanistic pathway is known to be pertinent to multiple outcomes, consideration might be given to organizing those related outcomes or hazards together. At this point, enough information might be available to begin to determine which mechanistic studies will best inform mechanistic pathways relevant to observations in human or animal health effect studies. Therefore, it might be possible to begin the prioritization process for the mechanistic analyses, including which mechanistic studies need to be evaluated at the individual level, concurrently with the synthesis of the human and animal health effect studies. Considerations for grouping related outcomes into a unit of analysis could also be directly informed by studies describing the pharmacokinetics of the chemical (ADME) or presenting pharmacokinetic (PK) or physiologically based pharmacokinetic (PBPK) models, e.g., if outcomes are expected to differ based on route of exposure, or whether the agent is expected to reach the target organ system via that route of exposure.

An example of how the units of analysis would be presented in an assessment protocol is shown in Table 3-1. It is important to note that the units of analysis for a given chemical are selected on the basis of the available endpoints and outcomes in the human and animal evidence identified from the literature inventory. As additional examples are developed, they are posted to the HAWC project “IRIS PPRTV SEM Template Figures and Resources (2021)” (see “Example Units of Analysis” attachment).

Table 3-1. Example units of analysis

Health effect categories for evidence integration	Examples of units of analysis for evidence synthesis that inform evidence integration (each bullet represents a unit of analysis based on the endpoints and outcomes available for an example chemical)	
	Human evidence	Animal evidence
Developmental	<ul style="list-style-type: none"> • Fetal viability/pregnancy outcomes (spontaneous abortion) • Fetal structural alterations (neural tube defects) 	<ul style="list-style-type: none"> • Fetal viability/survival or other birth parameters (e.g., resorptions, number of pups per litter) • Fetal growth (e.g., weight or length) • Fetal structural alterations (e.g., external, soft tissue, or skeletal findings) <p>*An analysis of dam health (e.g., weight gain, food consumption) is also conducted to support conclusions of specificity of the effects</p>

ORD Staff Handbook for Developing IRIS Assessments

Health effect categories for evidence integration	Examples of units of analysis for evidence synthesis that inform evidence integration (each bullet represents a unit of analysis based on the endpoints and outcomes available for an example chemical)	
	Human evidence	Animal evidence
		as being developmental (versus derivative of maternal toxicity).
Respiratory	<ul style="list-style-type: none"> No studies available 	<ul style="list-style-type: none"> Histopathology and cell proliferation (e.g., nasal lesions, atrophy, respiratory metaplasia, osseous metaplasia)
Hepatic	<ul style="list-style-type: none"> Clinical chemistry, clinical effects (e.g., jaundice, hepatomegaly) 	<ul style="list-style-type: none"> Organ weight (liver) Histopathology and cell proliferation (e.g., total altered cell foci, central cell atypia, central collapse, central deposit of ceroid, central vacuolic change, fatty change, central necrosis, focal necrosis, total necrosis, liver cell proliferation) Clinical chemistry (serum or tissue liver enzymes—e.g., ALT and AST)
Renal/Urinary	<ul style="list-style-type: none"> Clinical chemistry (e.g., BUN in workers) 	<ul style="list-style-type: none"> Organ weight (kidney) Histopathology and cell proliferation (e.g., nuclear enlargement, basophilia, tubular dilation, atypical tubular hyperplasia, kidney cell proliferation) Clinical chemistry (serum and urinary markers- e.g., BUN, creatinine, urinary protein, urinary glucose, urinary occult blood)
Endocrine	<ul style="list-style-type: none"> Thyroid hormone and antibodies 	<ul style="list-style-type: none"> Histopathology (pituitary) Clinical chemistry (e.g., serum glucose)
Immune	<ul style="list-style-type: none"> Sensitization and allergic response (multiple chemical sensitivity, basophil levels) 	<ul style="list-style-type: none"> Immunosuppression Immunostimulation, sensitization, and allergic response <p>*Some immune outcomes (spleen weight, thymus weight, antibody response, etc.) can reflect changes in multiple functional outcomes depending on direction of effect and study design. Thus, these outcomes will be considered as part of both units of analysis.</p>
Musculoskeletal	<ul style="list-style-type: none"> Clinical effects (osteoarthritis) 	<ul style="list-style-type: none"> Histopathology (bone)
Carcinogenicity	<ul style="list-style-type: none"> Blood cancer Brain tumors Kidney cancer Thyroid adenoma Mortality due to cancer. 	<ul style="list-style-type: none"> Kidney tumors (e.g., adenomas, carcinomas) or dysplasia Liver tumors (e.g., adenomas, carcinomas) or dysplasia

Health effect categories for evidence integration	Examples of units of analysis for evidence synthesis that inform evidence integration (each bullet represents a unit of analysis based on the endpoints and outcomes available for an example chemical)	
	Human evidence	Animal evidence
		<ul style="list-style-type: none"> • Pituitary tumors (e.g., adenomas, carcinomas) or dysplasia

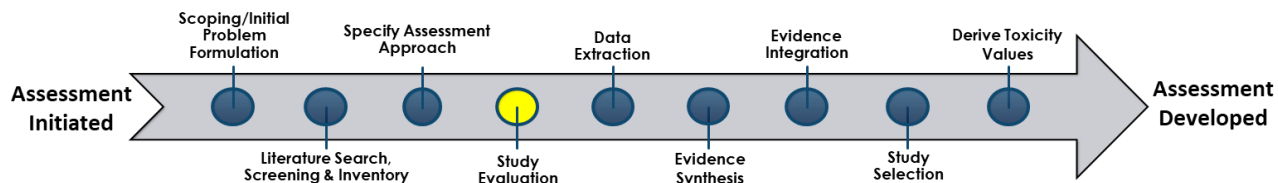
ALT = Alanine amino-transferase; AST = Aminotransferase ; BUN = blood urea nitrogen.

3.3. CONSIDERATION OF SUPPLEMENTAL MATERIAL

In the systematic review protocol, the assessment team should describe how supplemental material will be considered in the assessment. At a minimum, this should include a description of constructs used to organize publications with supplemental studies (e.g., key characteristics of carcinogens, subtagging of PK content). The assessment protocol should also describe any planned analyses of supplemental content that are considered essential to developing the draft assessment. The IRIS Program uses a stepwise approach for identifying prioritized analyses of supplemental content. This process includes considering assessment-specific key science issues in the IRIS Assessment Plan (IAP). Other analyses can be reasonably predicted across assessments (e.g., mode-of-action analysis for carcinogenicity hazard identification and dose-response analyses when chemical exposure-induced tumors are observed; synthesizing evidence on susceptibility for health effects that are more likely to be identified as hazards; scientific evaluation of any PBPK models). The literature inventory is used to characterize the extent of information available to address these topics. This knowledge allows the assessment team to provide more details on the analysis of supplemental content in the assessment protocol than can be done at assessment initiation. Development of this stepwise approach is based on staff experience in conducting assessments and several U.S. Environmental Protection Agency (EPA)-sponsored workshops [e.g., [NAS \(2018\)](#) and [NAS \(2019\)](#)].

The utility of supplemental materials can vary greatly depending on the type of materials available and the specific needs and uncertainties of the assessment. Decisions on if and how to prioritize subsets of supplemental materials typically warrants additional consideration. For example, use of data from non-PECO routes of exposure considers both the type of outcomes being informed (e.g., portal-of-entry versus systemic) and the level of understanding of the chemical’s ADME properties. Similarly, decisions on how useful non-PECO animal model evidence will be to an assessment considers the extent to which those models have been established for use in evaluating the health effects and specific endpoints of interest. Any such considerations applied to justify how subsets of supplemental materials will be organized, prioritized, or analyzed for use in the assessment should be documented in the protocol.

4. STUDY EVALUATION



Purpose

- Ensure the studies used in the assessment were conducted in such a manner that the results are credible.

4.1. STUDY EVALUATION OVERVIEW FOR HEALTH EFFECT STUDIES

The purpose of this stage is to evaluate the studies for their internal validity and utility in assessing a potential change in the health effect outcome or endpoint under consideration, independent of the direction or magnitude of the study findings. These evaluations result in a summary confidence judgment about the reliability of the study for assessing the outcome(s) of interest. Although the evaluation of physiologically based pharmacokinetic (PBPK) models primarily focuses on transparency and validation of the model itself, key concerns for the review of epidemiological, controlled human exposure, animal, and in vitro studies are risk of bias (RoB), which is the assessment of internal validity (factors that might affect the magnitude or direction of an effect in either direction), and sensitivity (factors that limit the ability of a study to detect a true effect; low sensitivity is a bias toward the null when an effect exists). The Integrated Risk Information System (IRIS) approach includes the sensitivity domain to capture certain aspects of study design that do not strictly fall under RoB defined as “a systematic error, or deviation from the truth, in results or inferences” (Cooper et al., 2016) but that are important for interpretation of study findings. Additional detail on these concerns is provided below.

- *Risk of bias*: Assesses the internal validity of the study, which reflects the extent to which the authors controlled for factors in the design and conduct of the study that could bias the results.
- *Study sensitivity*: Assesses whether there are factors in the design and conduct of the study that might reduce its ability to observe an effect if present. Study sensitivity is an important consideration in the interpretation of null findings (i.e., whether a null finding can be confidently interpreted as evidence of a lack of association) and in consideration of heterogeneity across studies (i.e., studies with greater sensitivity might be more likely to observe an effect, which could explain apparent inconsistency). A study might have been well conducted with minimal RoB but have reduced sensitivity due to population

characteristics (e.g., limited exposure contrast or few observed cases of the outcome of interest in epidemiological studies).

Study evaluation, as defined herein, is a broad term encompassing interpretation of a variety of methodological features (e.g., study design and conduct, exposure measurement or characterization, selective reporting bias). The study evaluations are aimed at discerning limitations that could substantively change a result presented in the study or the interpretation of that result, also considering the expected direction of the bias. The overall goal of the study evaluation approaches discussed in this chapter is to evaluate the extent to which the results are likely to represent a reliable, sensitive, and informative presentation of a true response. The use of scientific expertise and judgment is an inherent part of the process.

IRIS uses a domain-based approach for evaluating studies, consistent with best practices in systematic review ([Kase et al., 2016](#); [Segal et al., 2015](#); [Beronius et al., 2014](#); [NRC, 2014](#); [Higgins and Green, 2011b](#); [IOM, 2011 p. 132](#); [Juni et al., 1999](#); [Moher et al., 1996](#); [Schulz et al., 1995](#); [Emerson et al., 1990](#)). Examination of specific methodological features for each exposure-outcome combination is accomplished by applying prespecified considerations to a set of domains. These domains differ for epidemiological and animal studies (see Figure 4-1), which have the most well-developed tools for use in human health assessment of environmental chemicals and are discussed below in their respective sections (see Sections 4.2 and 4.4). Domains for the evaluation of in vitro studies are adapted from the animal study evaluation domains, although in vitro and other non-populations, exposures, comparators, and outcomes (PECO) studies will undergo a prioritization step before reaching the level of study evaluation (see Section 4.5). The core and prompting questions provided for each domain are meant to guide the reviewer to seek and consider relevant information pertaining to specific aspects of the study. Prespecified considerations and refinements are documented in the study evaluation component of the assessment's systematic review protocol.





Additional chemical-, outcome-, or exposure-specific considerations for evaluating studies are developed as needed in consultation with topic-specific technical experts and with use of existing guidance documents when available, including U.S. Environmental Protection Agency (EPA) guidelines for carcinogenicity, neurotoxicity, reproductive toxicity, and developmental toxicity ([U.S. EPA, 2005a, 1998, 1996, 1991a](#)). Some prespecified considerations (e.g., the validity of the methods used to ascertain a specific outcome) might be used for an evaluation of that outcome in any assessment, whereas others could be assessment specific. For example, evaluation of exposure measures in epidemiological studies will often need to be developed for each chemical. When possible, criteria should identify high priority issues that would be expected to result in substantial bias or insensitivity and thus a reduced rating for overall confidence. As reviewers examine a group of studies, additional chemical-specific knowledge or methodological concerns might emerge and a second pass could become necessary. Once developed, the reviewers and assessment managers must ensure that each criterion is applied consistently across studies. This process is undertaken for all studies evaluated.

(a) Individual evaluation domains

Epidemiology	Animal	In vitro
<ul style="list-style-type: none"> • Exposure measurement • Outcome ascertainment • Participant selection • Confounding • Analysis • Selective reporting • Sensitivity 	<ul style="list-style-type: none"> • Allocation • Observational bias/blinding • Confounding • Attrition • Chemical administration and characterization • Endpoint measurement • Results presentation • Selective reporting • Sensitivity 	<ul style="list-style-type: none"> • Observational bias/blinding • Variable control • Selective reporting • Chemical administration and characterization • Endpoint measurement • Results presentation • Sensitivity

(b) Domain level judgments and overall study rating

Domain judgments

Judgment	Interpretation
 Good	Appropriate study conduct relating to the domain and minor deficiencies not expected to influence results.
 Adequate	A study that may have some limitations relating to the domain, but they are not likely to be severe or to have a notable impact on results.
 Deficient	Identified biases or deficiencies interpreted as likely to have had a notable impact on the results or prevent reliable interpretation of study findings.
 Critically Deficient	A serious flaw identified that makes the observed effect(s) uninterpretable. Studies with a critical deficiency are considered "uninformative" overall.

Overall study rating for an outcome

Rating	Interpretation
High	No notable deficiencies or concerns identified; potential for bias unlikely or minimal; sensitive methodology.
Medium	Possible deficiencies or concerns noted but they are unlikely to have a significant impact on results.
Low	Deficiencies or concerns were noted, and the potential for substantive bias or inadequate sensitivity could have a significant impact on the study results or their interpretation.
Uninformative	Serious flaw(s) makes study results uninterpretable but may be used to highlight possible research gaps.

Figure 4-1. Overview of Integrated Risk Information System (IRIS) study evaluation approach. (a) individual evaluation domains organized by evidence type, and (b) individual evaluation domain judgments and definitions for overall ratings (i.e., domain and overall judgments are performed on an outcome-specific basis).

As part of quality assurance, each study evaluation is conducted independently by at least two reviewers, with a process for comparing and resolving differences (typically, a third independent reviewer is consulted when two reviewers do not reach consensus). For studies that examine more than one outcome, the evaluation process should be outcome-specific, as the utility of a study could vary for the different outcomes. If a study examines multiple endpoints for the same outcome, evaluations could be performed at a more granular level if appropriate, but these measures might still be grouped for evidence synthesis. These evaluations could require additional reviewers with expertise in these endpoints. The evaluation provides a transparent means to convey the study's methodological strengths and limitations, and, thus, the ability to rely on the results to reach conclusions about the potential hazard of an exposure.

Study authors might be queried to obtain missing information that could inform domain judgments or additional analyses that could address major limitations. The decision on whether to seek missing information is largely based on the likelihood that such information would affect the overall confidence in the study. Outreach to study authors is documented in Health Assessment Workspace Collaborative (HAWC) or Health and Environmental Research Online (HERO) and considered unsuccessful if researchers do not respond to an email or phone request within 1 month of the attempt to contact. Study evaluation is completed with currently available information but will be updated to reflect any additional data received from the study authors before the end of the response period.

4.1.1. Evaluation Ratings

For each outcome in a study,⁸ reviewers reach a consensus judgment of *good*, *adequate*, *deficient*, *not reported*, or *critically deficient* in each domain. It is important to stress that these evaluations are performed in the context of the study's utility for identification of individual hazards, rather than the usability of a study for dose-response analysis (noting that study confidence is one consideration used in selecting studies for dose-response analysis; see Chapter 7). Although study design features specific to the usability of the study for dose-response analysis can be noted, they do not contribute to the study confidence classifications. The following judgment ratings are applied to each evaluation domain for each study.

- *Good* represents a judgment that the study was conducted appropriately in relation to the evaluation domain, and any minor deficiencies noted are not expected to influence the study results or interpretation of the study findings.

⁸Note: "study" is used instead of a more accurate term (e.g., "experiment") throughout these sections owing to an established familiarity within the field for discussing a study's risk of bias or sensitivity, etc. However, all evaluations discussed herein are explicitly conducted at the level of an individual outcome or group of outcomes tested within a matched group (e.g., exposed and unexposed) of animals or humans.

- *Adequate* indicates a judgment that methodological limitations related to the evaluation domain are (or are likely to be) present, but that those limitations are unlikely to be severe or to notably impact the study results or interpretation of the study findings.
- *Deficient* denotes identified biases or deficiencies interpreted as likely to have had a notable impact on the results, or that limit interpretation of the study findings.
- *Not reported* indicates the information necessary to evaluate the domain was not available in the study and could not be inferred. Depending on the expected impact, the domain may be interpreted as *adequate* or *deficient* for the purposes of the study confidence rating.
- *Critically deficient* reflects a judgment that the study design or conduct relating to the evaluation domain introduced a serious flaw that is interpreted to be the primary driver of any observed effect(s) or makes the study findings uninterpretable. Studies with *critically deficient* judgments in any evaluation domain are almost always considered overall *uninformative* for the relevant outcome(s).

Once the evaluation domains are rated, the identified strengths and limitations are considered to reach a study confidence rating of high, medium, or low confidence, or uninformative for each specific health outcome(s). This classification is based on the reviewer judgments across the evaluation domains and considers the likely impact the noted deficiencies in bias and sensitivity have on the outcome-specific results. There are no defined weights for the domains, and the reviewers are responsible for applying expert judgment to make this determination. The study confidence classifications, which reflect a consensus judgment among reviewers, are defined as follows.

- *High* confidence: No notable deficiencies or concerns identified; the potential for bias is unlikely or minimal, and the study used sensitive methodology. *High* confidence studies generally reflect judgments of *good* across all or most evaluation domains.
- *Medium* confidence: Possible deficiencies or concerns are identified, but the limitations are unlikely to have a significant impact on the study results or their interpretation. Generally, *medium* confidence studies include *adequate* or *good* judgments across most domains, with the impact of any identified limitation not being judged as severe.
- *Low* confidence: Deficiencies or concerns are identified, and the potential for bias or inadequate sensitivity is expected to have a significant impact on the study results or their interpretation. Typically, *low* confidence studies have a *deficient* evaluation for one or more domains, although some *medium* confidence studies may have a *deficient* rating in domain(s) considered to have less influence on the magnitude or direction of effect estimates. *Low* confidence results are given less weight compared to *high* or *medium* confidence results during evidence synthesis and integration (see Section 6.1, Table 6-3), and are generally not used as the primary sources of information for hazard identification or derivation of toxicity values unless they are the only studies available (in which case this significant uncertainty would be emphasized during dose-response analysis). Studies rated *low* confidence only because of sensitivity concerns are asterisked or otherwise noted because they often require additional consideration during evidence synthesis. Effects

observed in studies biased toward the null may increase confidence in the results, assuming the study is otherwise well conducted (see Section 6.1).

- *Uninformative*: Serious flaw(s) are judged to make the study results uninterpretable for use in the assessment. Studies with *critically deficient* judgments in any evaluation domain are almost always rated *uninformative* (see explanation above). Given the findings of interest are considered uninterpretable based on the identified flaws (see above definition of *critically deficient*) and do not provide information of use to the assessment interpretations, these studies have no impact on evidence synthesis or integration judgments and are not useable for dose-response analyses but may be used to highlight research gaps.

After the initial evaluation of the studies by level of overall confidence, each group (confidence level) of studies is examined for quality and consistency of judgments across studies. In this stage, the reviewer rereads the studies and asks the following.

- Does the separation between the levels of confidence make sense (i.e., are the *high* confidence studies distinct from the *low* confidence studies, and do the *medium* confidence studies fall in between these two groups)?
- Have the evaluation judgments been consistently applied across the set of studies? (For example, if a specific limitation was identified in one study and may be applicable to other studies, the reviewers should go back and make sure the judgment was applied in the same way.)
- Do the flaws identified in studies classified as *uninformative* truly warrant exclusion?

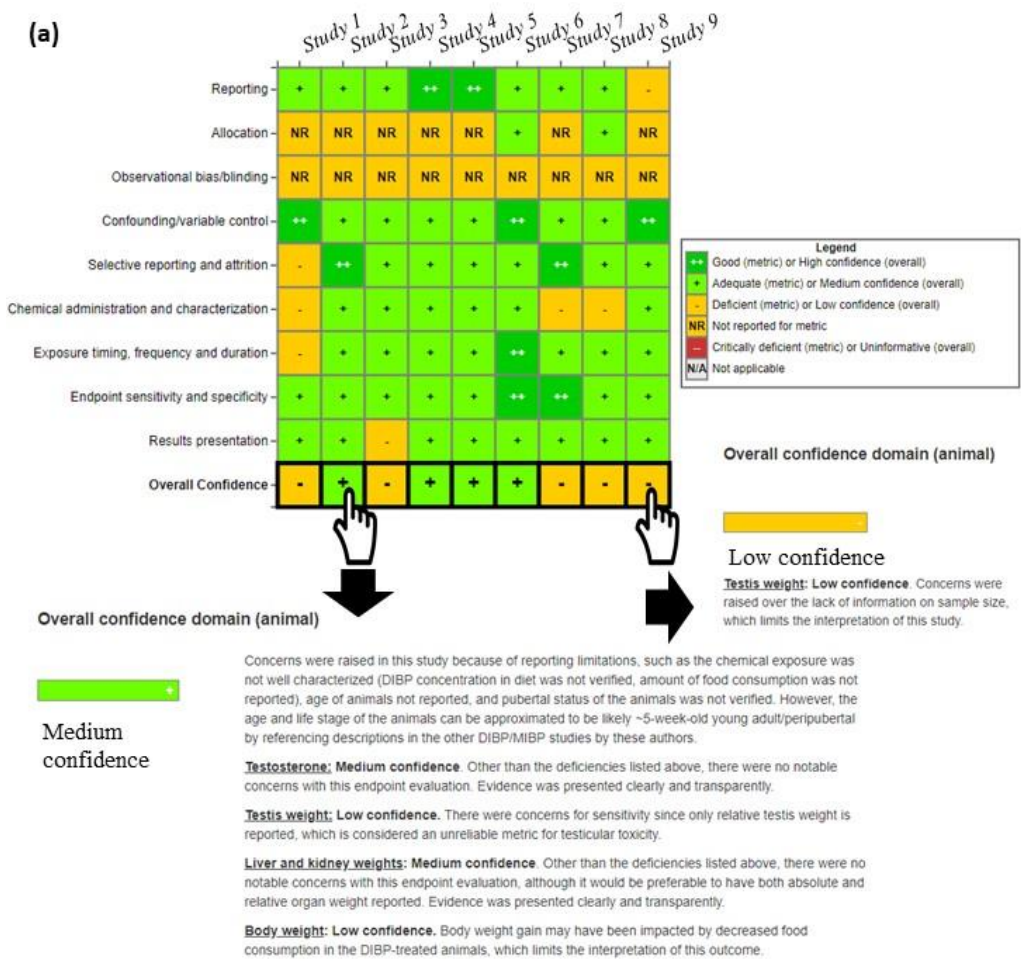
4.1.2. Documenting Study Evaluations

Study evaluation determinations reached by each reviewer and the consensus judgments across reviewers are recorded in EPA's version of HAWC (<https://hawcprd.epa.gov/>) or documented in another format. Tutorials for using HAWC for study evaluation are available at <https://hawcproject.org/resources> (Note: the tutorials are not IRIS specific). The final study evaluations are anonymized and reflect the consensus review. They are made available when the draft is publicly released. There are several options for displaying study evaluation results in the assessment, typically using visualizations created automatically in HAWC (see Figures 4-2 and 4-3). Note: All HAWC visualizations have "click to see more" functionality, where the user can click a domain to see the rationale—see Figure 4-2 (c) and (d). The study confidence classifications and their rationales are carried forward and considered as part of evidence synthesis (see Section 6.1), to aid in the interpretation of results across studies.



Figure 4-2. Examples of study evaluation displays at the individual level. (a) A “doughnut” visualization. (b) A “caterpillar” visualization. (c) Study evaluation rationale for a domain. (d) Overall study confidence evaluation rationale. All the above visualizations are created automatically in Health Assessment Workspace Collaborative (HAWC) after the final rating has been entered. Clicking on a domain in (a) or (b) will display the rationale for the rating, similar to examples in (c) and (d).

ORD Staff Handbook for Developing IRIS Assessments



(b)

	Reference	Study description			Includes metabolites of:					Study evaluation						
		Population	Exposure	Outcome	DEHP	DINP	DBP	DIBP	BBP	DEP	Exposure	Outcome	Selection	Confounding	Analysis	Overall confidence
Included	Study 1	Preconception cohort in U.S. (N=221 women)	Three urine samples from cycle of conception, pooled	Early pregnancy loss, identified via hCG	✓	✓	✓	✓	✓	✓	G	G	A	G	G	High
	Study 2	Cohort of women receiving assisted reproductive technology in U.S. (N=256)	Two urine samples per conception cycle	Total pregnancy loss identified prospectively	✓	✓	✓	✓	✓	✓	G	G	A	A	G	High
	Study 3	Case-control in China (N=132 cases, 172 controls) of women receiving ultrasound	Single morning urine sample at 5-13 wks gestation	Clinical pregnancy loss identified by ultrasound	✓	✓	✓	✓	✓	✓	A	A	D	D	A	Low
	Study 4	Preconception cohort in Denmark (N=242 women)	Single urine sample from cycle of conception	Early and clinical pregnancy loss identified through urine samples or medical provider	✓	✓	✓	✓	✓	✓	A	G	D	A	A	Low
	Study 5	Case-control in China (N=150 cases, 172 controls) of women	Single morning urine sample at admission to hospital	Clinical pregnancy loss identified by ultrasound	✓	✓	✓	✓	✓	✓	A	A	D	A	A	Low
		Total Studies per Phthalate			5	2	5	4	3	5						

Excluded studies (N): list with reasons
G=good; A=adequate; D=deficient

(c)

Author (year)	Species (strain)	Exposure life stage and duration	Exposure route	Female reproductive*				Developmental*		
				Morphological development	Maternal body weight	Gestation length	Reproductive organ weight	Survival	Growth	Malformations/ variations
Study 1	Rat (Wistar)	GD 6-20	Diet	-	H	-	H	H	H	H
Study 2	Rat (Wistar)	GD 7-19	Gavage	H	H	-	-	H	H	-
Study 3	Rat (Sprague-Dawley)	GD 6-20	Gavage	H	H	-	H	H	H	H
Study 4	Rat (Sprague-Dawley)	GD 8-18	Gavage	-	M	-	-	H	-	-
Study 5	Rat (Sprague-Dawley)	GD 12-21	Gavage	H	M	H	-	H	H	-
Study 6	Rat (Sprague-Dawley)	GD 13-19	Gavage	-	H	-	H	H	H	H
Study 7	Rat (Sprague-Dawley)	GD 14-18	Gavage	-	L	-	-	M	-	-
Study 8	Rat (Sprague-Dawley)	GD 14-18	Gavage	-	L	-	-	M	-	-
Study 9	Rat (Sprague-Dawley)	GD 14-18	Gavage	-	L	-	-	L	-	-
Study 10	Mouse (ICR)	GD 0-21; GD 0-PND 21	Diet	-	L	-	-	H	M	-
Study 11	Rat (Wistar)	PND 21-23; PND 21-40	Gavage	M	-	-	M	-	M	-
Study 12	Rat (JCL:Wistar)	~PND 35-42	Diet	-	-	-	-	-	L	-
Study 13	Mouse (JCL:ICR)	~PND 35-42	Diet	-	-	-	-	-	L	-
Study 14	Mouse (JCL:ICR)	~PND 35-42	Diet	-	-	-	-	-	L	-
Study 15	Rat (albino; strain not reported)	Weaning to 4 months post-weaning	Diet	-	-	-	-	-	L	-

Figure 4-3. Examples of study evaluation displays looking across studies. (a) Heat map created in Health Assessment Workspace Collaborative (HAWC). (b) Heat map created in Microsoft Word with study details. (c) Heat map created in Microsoft Word with overall confidence presented for multiple health effects.

GD = gestation day; PND = postnatal day.

Across-study heat maps are a visualization option in HAWC that need to be created by the user (see the creating visualization tutorial at <https://hawcprd.epa.gov/about/>). Clicking on any cell in a Health Assessment Workspace Collaborative (HAWC) heat map will display the rationale for the rating. An interactive version of this figure with rationales is available at <https://hawcprd.epa.gov/summary/visual/100000096/>.

4.2. EVALUATION OF EPIDEMIOLOGICAL STUDIES

The principles and framework used for the evaluation of epidemiological studies examining chemical exposures are adapted from the principles in the Risk of Bias in Nonrandomized Studies of Interventions (ROBINS-I), modified for use with the types of studies more typically encountered in environmental and occupational epidemiology rather than clinical interventions ([Sterne et al. 2016](#)). The RoB evaluation domains for IRIS’s adapted approach are exposure measurement, outcome ascertainment, participant selection, confounding, analysis, and selective reporting. In addition, the IRIS approach includes a domain for study sensitivity. For each domain, “core,” “prompting,” and follow-up questions are provided below, and are used to guide the development of assessment specific considerations. Ratings may be lowered when information needed to evaluate a domain is not reported in the study and cannot be obtained by author correspondence (as described above).

4.2.1. Development of Evaluation Considerations

A distinctive aspect of a systematic review is the process of developing considerations to be used across studies to make judgments (e.g., define good vs. deficient) for each domain. This requires a familiarity with the exposure and outcome being reviewed and with the studies to be evaluated; it cannot be conducted in the absence of knowledge of the study designs, measurements, and analytic issues encompassed within the set of studies ([Higgins et al., 2022a](#); [Sterne et al., 2016](#)). The process used to develop these specific considerations typically involves research into the issues identified in the set of studies; consultation with additional subject area experts might be needed as described in the previous section. The considerations should provide different reviewers with a common basis for reaching decisions ([Sterne et al., 2016](#)).

The purpose of the evaluation considerations is to:

- i) Specify attributes of the study that would impact your confidence in the study results;
- ii) Differentiate between those attributes that would be likely to have a large effect, compared to a small effect, on confidence in the study results;
- iii) Anticipate, if possible, the likely direction of effect on the study results;
- iv) Provide a guide to the evaluation process that can be documented and followed by others; and
- v) Ensure consistency in evaluations across studies and across reviewers.

The evaluation strategy might define an “ideal” design (i.e., a study design with no RoB and high sensitivity) for the review question. This is defined based on the specific exposure and outcome being evaluated. What type of measurement would be needed to accurately capture the exposure? What type of outcome ascertainment would optimize sensitivity and specificity? How would participants be identified? What information on other risk factors would you want to have? What kind of analyses would you want to see? From this reference point, considerations for each of the rating levels (good, adequate, deficient, not reported, critically deficient) are developed and specified. The decisions regarding ratings are judgments, considering severity and consequences of the noted deficiency or bias ([Sterne et al., 2016](#)). As stated previously, the potential direction of bias (i.e., leading to an inflated or attenuated effect estimate) and magnitude of bias are also noted in situations in which it can be reasonably anticipated. For complex topics, the considerations could be pilot tested on three to five studies; this testing process will improve consistency in applying the considerations and reduce the potential for conflicts in the evaluations. Any revisions to the considerations resulting from this testing process should be incorporated in the revised protocol and applied uniformly across all evaluated studies.

The following discussion summarizes the considerations for each of the evaluation domains. The core questions represent the key concepts, while the prompting questions help the reviewer focus on relevant details when developing and applying the evaluation considerations specific to

the exposure and outcome (as described above). Some considerations have been developed for participant selection, confounding, analysis, and study sensitivity that generally apply to all exposures and outcomes and are listed in the tables for each domain below. Assessment teams develop exposure- and outcome-specific considerations as needed for each assessment.

Exposure Measurement

This domain concerns the ability of the exposure measures to correctly classify exposure status and exposure level. Nondifferential exposure misclassification is likely to lead to attenuated risk estimates and attenuated dose-response, but differential exposure misclassification can result in either attenuated or inflated risk estimates. The core, prompting, and follow-up questions are provided in Table 4-1.

A concern is how well the exposure measure represents the exposure in an etiologically relevant time window. IRIS does not make this evaluation strictly on the basis of general study design (e.g., cohort is always better than cross-sectional); rather, IRIS bases this decision on knowledge of the relationship between a specific disease process and the expected relevant timing for exposure measure under review, and what study designs are appropriate for the research question. The reason for this distinction is there can be situations in which the exposure assessment conducted by a prospective design does not adequately represent the etiologically relevant time (i.e., exposure is not measured during a relevant time window), while in other situations, a cross-sectional design does provide an adequate representation of the etiologically relevant time (e.g., outcomes with potential for a short-term response, chemicals with long half-lives). Research into the reliability and interpretation of various exposure measures and into the biological processes involved in the effect(s) under study is a key stage in the process of customizing the study evaluation considerations for exposure measurement. This research should also include information pertaining to the possibility that the effect under study could influence the exposure measure (e.g., through effects on lipid mobilization or kidney function for biomarker measures or through differential recall for measures based on self-report).

Information relevant to evaluation of exposure measures includes, but is not limited to, source(s) of exposure (consumer products, occupational, an industrial accident) and source(s) of exposure data, blinding to outcome, level of detail for job history data, when measurements were taken, type of biomarker(s), assay information (including measurement accuracy and precision), reliability data from repeat measures studies, and validation studies.

The decisions regarding confidence in different types of exposure measures are documented in the protocol.

Table 4-1. Example question specification for evaluation of exposure measurement in epidemiological studies

Domain and core question	Prompting questions	Follow-up questions	Considerations that apply to most exposures and outcomes
<p><u>Exposure measurement</u></p> <p>Does the exposure measure reliably distinguish between levels of exposure in a time window considered most relevant for a causal effect with respect to the development of the outcome?</p>	<p>For all:</p> <ul style="list-style-type: none"> • Does the exposure measure capture the variability in exposure among the participants, considering intensity, frequency, and duration of exposure? • Does the exposure measure reflect a relevant time window? If not, can the relationship between measures in this time and the relevant time window be estimated reliably? • Was the exposure measurement likely to be affected by a knowledge of the outcome? • Was the exposure measurement likely to be affected by the presence of the outcome (i.e., reverse causality)? 	<p>Is the degree of exposure misclassification likely to vary by exposure level?</p> <p>If the correlation between exposure measurements is <i>moderate</i>, is there an adequate statistical approach to ameliorate variability in measurements?</p> <p>If there is a concern about the potential for bias, what is the predicted direction or distortion of the bias on the effect estimate (if there is enough information)? (continued)</p>	<p>These considerations require customization to the exposure and outcome (relevant timing of exposure)</p> <p>Good</p> <ul style="list-style-type: none"> • Valid exposure assessment methods used, which represent the etiologically relevant time period of interest. • Exposure misclassification is expected to be minimal. <p>Adequate</p> <ul style="list-style-type: none"> • Valid exposure assessment methods used, which represent the etiologically relevant time period of interest. • Exposure misclassification might exist but is not expected to greatly change the effect estimate.

Domain and core question	Prompting questions	Follow-up questions	Considerations that apply to most exposures and outcomes
<p>Exposure measurement</p> <p>Does the exposure measure reliably distinguish between levels of exposure in a time window considered most relevant for a causal effect with respect to the development of the outcome? (continued)</p>	<p>For case-control studies of occupational exposures:</p> <ul style="list-style-type: none"> Is exposure based on a comprehensive job history describing tasks, setting, period, and use of specific materials? <p>For biomarkers of exposure and other analytic measures of exposure:</p> <ul style="list-style-type: none"> Is a standard assay used? Is the measure valid and precise? What are the intra- and interassay coefficients of variation? Is the assay likely affected by contamination? Are values less than the limit of detection dealt with adequately? What exposure time period is reflected by the biomarker? If the half-life is short, what is the correlation between serial measurements of exposure? 	<p>Is the degree of exposure misclassification likely to vary by exposure level?</p> <p>If the correlation between exposure measurements is <i>moderate</i>, is there an adequate statistical approach to ameliorate variability in measurements?</p> <p>If there is a concern about the potential for bias, what is the predicted direction or distortion of the bias on the effect estimate (if there is enough information)?</p>	<p>Deficient</p> <ul style="list-style-type: none"> Valid exposure assessment methods used, which represent the etiologically relevant time period of interest. Specific knowledge about the exposure and outcome raises concerns about reverse causality, but there is uncertainty whether it is influencing the effect estimate. Exposed groups are expected to contain a notable proportion of unexposed or minimally exposed individuals, the method did not capture important temporal or spatial variation, or there is other evidence of exposure misclassification that would be expected to notably change the effect estimate. <p>Critically deficient</p> <ul style="list-style-type: none"> Exposure measurement does not characterize the etiologically relevant time period of exposure or is not valid. There is evidence that reverse causality is very likely to account for the observed association. Exposure measurement was not independent of outcome status.

Outcome Ascertainment

This domain concerns the ability of the outcome measure to correctly classify outcomes or effects. The inability to correctly classify individuals, if this misclassification is not related to exposure, can result in underestimation of effects. The core, prompting, and follow-up questions are provided in Table 4-2.

Outcome measures can involve a variety of sources including national databases (e.g., mortality data, cancer registries), medical records, pathology reports, self-report, assessment by study examiners, and biomarkers based on urine or blood samples. IRIS bases the evaluation decision on knowledge of the specific disease or outcome under review. Research into the reliability and validity of various outcome measures, and how this might vary in different populations or at different times, is a key stage in the evaluation process.

Information relevant to evaluation of outcome measures includes, but is not limited to, source of outcome (effect) measure, blinding to exposure status or level, how measured/classified, incident versus prevalent disease, evidence from validation studies, and prevalence (or distribution summary statistics for continuous measures) of outcome.

The decisions regarding confidence in different types of outcome measures will be documented in the protocol.

Table 4-2. Example question specification for evaluation of outcome in epidemiological studies

Domain and core question	Prompting questions	Follow-up questions	Considerations that apply to most exposures and outcomes
<p>Outcome ascertainment</p> <p>Does the outcome measure reliably distinguish the presence or absence (or degree of severity) of the outcome?</p>	<p>For all:</p> <ul style="list-style-type: none"> Is outcome ascertainment likely affected by knowledge of, or presence of, exposure (e.g., consider access to health care, if based on self-reported history of diagnosis)? <p>For case-control studies:</p> <ul style="list-style-type: none"> Is the comparison group without the outcome (e.g., controls in a case-control study) based on objective criteria with little or no likelihood of inclusion of people with the disease? <p>For mortality measures:</p> <ul style="list-style-type: none"> How well does cause of death data reflect occurrence of the disease in an individual? How well do mortality data reflect incidence of the disease? <p>For diagnosis of disease measures:</p> <ul style="list-style-type: none"> Is the diagnosis based on standard clinical criteria? If it is based on self-report of the diagnosis, what is the validity of this measure? <p>For laboratory-based measures (e.g., hormone levels):</p> <ul style="list-style-type: none"> Is a standard assay used? Does the assay have an acceptable level of interassay variability? Is the sensitivity of the assay appropriate for the outcome measure in this study population? 	<p>Is there a concern that any outcome misclassification is nondifferential, differential, or both?</p> <p>What is the predicted direction or distortion of the bias on the effect estimate (if there is enough information)?</p>	<p>These considerations require customization to the outcome</p> <p>Good</p> <ul style="list-style-type: none"> High certainty in the outcome definition (i.e., specificity and sensitivity), minimal concerns with respect to misclassification. Assessment instrument was validated in a population comparable to the one from which the study group was selected. <p>Adequate</p> <ul style="list-style-type: none"> <i>Moderate</i> confidence that outcome definition was specific and sensitive, some uncertainty with respect to misclassification but not expected to greatly change the effect estimate. Assessment instrument was validated but not necessarily in a population comparable to the study group. <p>Deficient</p> <ul style="list-style-type: none"> Outcome definition was not specific or sensitive. Uncertainty regarding validity of assessment instrument. <p>Critically deficient</p> <ul style="list-style-type: none"> Invalid/insensitive marker of outcome. Outcome ascertainment is very likely to be affected by knowledge of, or presence of, exposure. <p>Note: Lack of blinding should not be automatically construed to be <i>critically deficient</i>.</p>

Participant Selection

This domain concerns the process through which participants are selected for (or leave/are lost to attrition) a study; a biased selection (or follow-up) can result in effect estimates that are either attenuated or inflated. The core, prompting, and follow-up questions are provided in Table 4-3.

In occupational cohort studies, the selection into the workforce (or into specific jobs within a work setting) can be influenced by an individual's overall health ("healthy worker effect"); a comparison of workers to a referent population that includes people who cannot work could result in a biased (attenuated) risk estimate. This type of bias has been seen in outcomes relating to physical exertion (e.g., cardiovascular disease, asthma), and to a lesser degree, cancer. Similarly, the decision to stay in a job or at a worksite can also be influenced by overall health or by sensitivity or susceptibility of an individual to effects of an exposure ("healthy worker survivor effect"). The formation of the study population (e.g., were all workers entered at the time exposure began or was it a "prevalent" cohort, consisting of workers in the workplace at a given time?), extent of follow-up, and degree to which follow-up is related to exposure level, comparison group, and analytic approaches to address changes in exposures in relation to disease status are all considered within this domain.

Similar considerations could also be at play in population-based cohorts in which selection into the study, selection into a subgroup of the study used in an analysis, or attrition out of the study might be jointly related to exposure and to disease. Directed acyclic graphs might be useful for visualizing relationships between variables that could lead to a selection bias.

For case-control studies, controls are optimally selected to represent the population from which the cases were drawn (e.g., similar geographic area, socioeconomic status, period). The interest and motivation to participate is generally higher for cases than for controls, and some attributes (e.g., lower education level, smoking history) could also be associated with likelihood to participate. A low participation rate of either or both groups does not inherently indicate the occurrence of selection bias; a biased risk estimate is produced if exposure and disease are jointly related to participation but not if either is independently related to participation. For example, a bias is not produced if cases are more likely to participate than controls; a bias is produced, however, if cases with high exposure are more likely to participate than cases with low exposure. Considerations regarding selection bias for case-control studies include the catchment area and recruitment methods for cases and controls and the participants' knowledge of study hypotheses and of their own exposure status or level.

Table 4-3. Example question specification for evaluation of participant selection in epidemiological studies

Domain and core question	Prompting questions	Follow-up questions	Considerations that apply to most exposures and outcomes
<p><u>Participant selection</u></p> <p>Is there evidence that selection into or out of the study (or analysis sample) was jointly related to exposure and to outcome? (continued)</p>	<p>For longitudinal cohort:</p> <ul style="list-style-type: none"> • Did participants volunteer for the cohort based on knowledge of exposure or preclinical disease symptoms? Was entry into the cohort or continuation in the cohort related to exposure and outcome? <p>For occupational cohort:</p> <ul style="list-style-type: none"> • Did entry into the cohort begin with the start of the exposure? • Was follow-up or outcome assessment incomplete, and if so, was follow-up related to both exposure and outcome status? <p>Is there evidence that less healthy workers leave employment or experience changes in employment-related exposure status (e.g., “healthy worker survivor effect”)?</p> <p>For case-control study:</p> <ul style="list-style-type: none"> • Were controls representative of population and time periods from which cases were drawn? • Are hospital controls selected from a group whose reason for admission is independent of exposure? • Could recruitment strategies, eligibility criteria, or participation rates result in differential participation relating to both disease and exposure? 	<p>Were differences in participant enrollment and follow-up evaluated to assess bias?</p> <p>If there is a concern about the potential for bias, what is the predicted direction or distortion of the bias on the effect estimate (if there is enough information)?</p> <p>Were appropriate analyses performed to address changing exposures over time in relation to symptoms?</p> <p>Is there a comparison of participants and nonparticipants to address whether differential selection or study retention/continuation is likely? (continued)</p>	<p>These considerations may require customization to the outcome. This could include determining what study designs effectively allow analyses of associations appropriate to the outcome measures (e.g., design to capture incident vs. prevalent cases, design to capture early pregnancy loss).</p> <p>Good</p> <ul style="list-style-type: none"> • Minimal concern for selection bias based on description of recruitment process and follow-up (e.g., selection of comparison population, population-based random sample selection, recruitment from sampling frame including current and previous employees). • Exclusion and inclusion criteria for participants specified and would not induce bias. • Participation rate is reported at all steps of study (e.g., initial enrollment, follow-up, selection into analysis sample). If rate is not high, there is appropriate rationale for why it is unlikely to be related to exposure (e.g., comparison between participants and nonparticipants or other available information indicates differential selection is not likely).

Domain and core question	Prompting questions	Follow-up questions	Considerations that apply to most exposures and outcomes
<p><u>Participant selection</u></p> <p>Is there evidence that selection into or out of the study (or analysis sample) was jointly related to exposure and to outcome?</p>	<p>For population based-survey:</p> <ul style="list-style-type: none"> Was recruitment based on advertisement to people with knowledge of exposure, outcome, and hypothesis? 	<p>Were differences in participant enrollment and follow-up evaluated to assess bias?</p> <p>If there is a concern about the potential for bias, what is the predicted direction or distortion of the bias on the effect estimate (if there is enough information)?</p> <p>Were appropriate analyses performed to address changing exposures over time in relation to symptoms?</p> <p>Is there a comparison of participants and nonparticipants to address whether differential selection is likely?</p>	<p>Adequate</p> <ul style="list-style-type: none"> Enough of a description of the recruitment process to be comfortable that there is no serious risk of bias. Inclusion and exclusion criteria for participants specified and would not induce bias. Participation rate is incompletely reported but available information indicates participation is unlikely to be related to exposure. <p>Deficient</p> <ul style="list-style-type: none"> Little information on recruitment process, selection strategy, sampling framework or participation OR aspects of these processes raises the potential for bias (e.g., healthy worker effect, survivor bias). <p>Critically deficient</p> <ul style="list-style-type: none"> Aspects of the processes for recruitment, selection strategy, sampling framework, or participation result in concern that selection bias is likely to have had a large impact on effect estimates (e.g., convenience sample with no information about recruitment and selection, cases and controls are recruited from different sources with different likelihood of exposure, recruitment materials stated outcome of interest and potential participants are aware of or are concerned about specific exposures).

The more participants are asked to do, the more likely participation will decrease. For example, there can be a considerable difference between the number of people who complete a questionnaire (initial study enrollment), the number who provide a blood sample, and the number who complete a follow-up interview or clinical exam at a later age. Some studies define the sample on the basis of the availability of each of the key variables (exposure, outcome, and in some cases, covariates). If missing data are not random (i.e., if jointly related to exposure and disease), however, this sample definition can introduce a kind of selection bias. The topic of the extent and treatment of missing data is discussed in the analysis domain, but if used as inclusion criteria, it should be considered here.

It is also important to consider whether susceptible or vulnerable populations or lifestages have been investigated in the available studies, and the possibility of latency (e.g., a hazard might not be detected if an outcome is incorrectly assessed in young adults when it is more relevant to elderly individuals).

Information relevant to evaluation of participant selection includes, but is not limited to, study design, where and when the study was conducted, recruitment process, exclusion and inclusion criteria, type of controls, total eligible, comparison between participants and nonparticipants (or followed and not followed), final analysis group, and included vulnerable/susceptible groups or lifestages.

The decisions regarding confidence in different types of participant selection methods will be documented in the specific exposure-outcome component of the protocol used for an assessment.

Confounding

This domain concerns the potential for confounding; confounding can result in effect estimates that are either attenuated or inflated. Confounding refers to risk factors for the outcome that are also associated with the exposure in the study but are not intermediaries on the pathway between the exposure and the outcome. The association between the confounder and the outcome should be to a degree strong enough to explain the observed effect estimate for the exposure of interest, either individually or in conjunction with other confounders. The core, prompting, and follow-up questions are provided in Table 4-4.

Table 4-4. Example question specification for evaluation of confounding in epidemiological studies

Domain and core question	Prompting questions	Follow-up questions	Considerations that apply to most exposures and outcomes
<p>Confounding</p> <p>Is confounding of the effect of the exposure likely?</p>	<p>Is confounding adequately addressed by considerations in</p> <ul style="list-style-type: none"> • Participant selection (matching or restriction)? • Accurate information on potential confounders and statistical adjustment procedures? • Lack of association between confounder and outcome, or confounder and exposure in the study? • Information from other sources? <p>Is the assessment of confounders based on a thoughtful review of published literature, potential relationships (e.g., as can be gained through directed acyclic graphing), and minimizing potential overcontrol (e.g., inclusion of a variable on the pathway between exposure and outcome)? (continued)</p>	<p>If there is a concern about the potential for bias, what is the predicted direction or distortion of the bias on the effect estimate (if there is enough information)? (continued)</p>	<p>These considerations require customization to the exposure and outcome, but this could be limited to identifying key covariates.</p> <p>Good</p> <ul style="list-style-type: none"> • Conveys strategy for identifying key confounders, including coexposures. This could include a priori biological considerations, published literature, causal diagrams, or statistical analyses, with recognition that not all “risk factors” are confounders. • Inclusion of potential confounders in statistical models not based solely on statistical significance criteria (e.g., $p < 0.05$ from stepwise regression). • Does not include variables in the models likely to be influential colliders or intermediates on the causal pathway. • Key confounders are evaluated appropriately and considered unlikely sources of substantial confounding. This often will include <ul style="list-style-type: none"> ○ Presenting the distribution of potential confounders by levels of the exposure of interest or the outcomes of interest (with amount of missing data noted); ○ Consideration that potential confounders were rare among the study population, or were expected to be poorly correlated with exposure of interest; ○ Consideration of the most relevant functional forms of potential confounders; ○ Examination of the potential impact of measurement error or missing data on confounder adjustment; or ○ Presenting a progression of model results with adjustments for different potential confounders, if warranted.

Domain and core question	Prompting questions	Follow-up questions	Considerations that apply to most exposures and outcomes
<p><u>Confounding</u></p> <p>Is confounding of the effect of the exposure likely?</p>	<p>Is confounding adequately addressed by considerations in</p> <ul style="list-style-type: none"> • Participant selection (matching or restriction)? • Accurate information on potential confounders and statistical adjustment procedures? • Lack of association between confounder and outcome, or confounder and exposure in the study? • Information from other sources? <p>Is the assessment of confounders based on a thoughtful review of published literature, potential relationships (e.g., as can be gained through directed acyclic graphing), and minimizing potential overcontrol (e.g., inclusion of a variable on the pathway between exposure and outcome)?</p>	<p>If there is a concern about the potential for bias, what is the predicted direction or distortion of the bias on the effect estimate (if there is enough information)?</p>	<p>Adequate</p> <p>Similar to <i>good</i> but might not have included all key confounders, or less detail might be available on the evaluation of confounders (e.g., subbullets in <i>good</i>). It is possible that residual confounding could explain part of the observed effect, but concern is minimal.</p> <p>Deficient</p> <ul style="list-style-type: none"> • Does not include variables in the models likely to be influential colliders or intermediates on the causal pathway. <p>And any of the following</p> <ul style="list-style-type: none"> • The potential for bias to explain some of the results is high based on an inability to rule out residual confounding, such as a lack of demonstration that key confounders of the exposure-outcome relationships were considered; • Descriptive information on key confounders (e.g., their relationship relative to the outcomes and exposure levels) are not presented; or • Strategy of evaluating confounding is unclear or is not recommended (e.g., only based on statistical significance criteria or stepwise regression [forward or backward elimination]). <p>Critically deficient</p> <ul style="list-style-type: none"> • Includes variables in the models that are colliders or intermediates in the causal pathway, indicating that substantial bias is likely from this adjustment; or • Confounding is likely present and not accounted for, indicating that all results were most likely due to bias.

The potential for confounding is challenging to assess. It can be addressed in the design or the analysis of a study (or both), and requires consideration of participant selection, measurement of variables, relationships among variables, statistical analysis, and comparison of results (e.g., associations between confounder and exposure/outcome, effect estimates with and without adjustment), and can often require knowledge from other sources regarding risk factors and exposures in different types of settings. The background research for this domain includes information on risk factors for the outcome under study, information on exposures in specific industrial or occupational settings, and patterns of exposures in different populations, as well as specific data from each of the individual studies. Directed acyclic graphs can be useful for visualizing relationships between variables, and the potential impact of inadequate or inappropriate control of variables. A particular concern is the unnecessary adjustment for an intermediary between exposure and the outcome, which would result in a biased effect estimate.

Information relevant to evaluation of potential confounding includes, but is not limited to, background research on key confounders for specific populations or settings, participant characteristic data (by group), strategy/approach for consideration of confounding, strength of associations between exposure and potential confounders and between potential confounders and outcome, and degree of exposure to the confounder in the population. Coexposures should also be considered as potential confounders. Some exposures tend to be found together in the environment or in occupational settings and are highly correlated. For example, it might be difficult to distinguish the independent effects from exposure to specific phthalate or per- and polyfluoroalkyl substances in drinking water, isomers of polychlorinated biphenyls in fish, or volatile organic compounds generated by a common source (e.g., benzene, toluene, ethylbenzene, xylene in traffic emissions) due to confounding by these coexposures. While it might be possible to conclude that confounding by another coexposure is not a major concern if a study reports that the correlation between concentrations of some chemical species or isomers is low, if the correlation between pollutants is high (or expected to be high), confounding of effect estimates is likely to be an uncertainty across all the studies individually. In these cases, it is particularly important to not only consider confounding at the individual study level, but to also, during evidence synthesis, analyze potential confounding by comparing across studies in populations with exposure to different pollutant combinations where the correlation between these coexposures might vary, or focus on studies that used more robust analytical methods to explore potential confounding. The decisions regarding confidence in different approaches to addressing confounding will be documented in the specific exposure-outcome evaluation components of the protocol used for an assessment and will include lists of key confounders.

Analysis

Information relevant to evaluation of analysis includes, but is not limited to, the extent (and if applicable, treatment) of missing data for exposure, outcome, and confounders, approach to

modeling, classification of exposure and outcome variables (continuous vs. categorical), testing of assumptions, sample size for specific analyses, and relevant sensitivity analyses.

The decisions regarding confidence in different types of analytic procedures will be documented in the specific exposure-outcome evaluation components of the protocol used for an assessment. The core, prompting, and follow-up questions are provided in Table 4-5.

Table 4-5. Example question specification for evaluation of analysis in epidemiological studies

Domain and core question	Prompting questions	Follow-up questions	Considerations that apply to most exposures and outcomes
<p>Analysis Does the analysis strategy and presentation convey the necessary familiarity with the data and assumptions? (continued)</p>	<ul style="list-style-type: none"> • Are missing outcome, exposure, and covariate data recognized, and if necessary, accounted for in the analysis? • Does the analysis appropriately consider variable distributions and modeling assumptions? • Does the analysis appropriately consider subgroups of interest (e.g., based on variability in exposure level or duration or susceptibility)? • Is an appropriate analysis used for the study design? • Is effect modification considered based on considerations developed a priori? 	<p>If there is a concern about the potential for bias, what is the predicted direction or distortion of the bias on the effect estimate (if there is enough information)? (continued)</p>	<p>These considerations may require customization to the outcome. This could include the optimal characterization of the outcome variable and ideal statistical test (e.g., Cox regression).</p> <p>Good</p> <ul style="list-style-type: none"> • Use of an optimal characterization of the outcome variable. • Quantitative results presented (effect estimates and confidence limits or variability in estimates; i.e., not presented only as a <i>p</i>-value or “significant”/“not significant”). • Descriptive information about outcome and exposure provided (where applicable). • Amount of missing data noted and addressed appropriately (discussion of selection issues—missing at random vs. differential). • Where applicable, for exposure, includes limit of detection (LOD, and percentage below the LOD), and decision to use log transformation. • Includes analyses that address robustness of findings, e.g., examination of exposure-response (explicit consideration of nonlinear possibilities, quadratic, spline, or threshold/ceiling effects included, when feasible); relevant sensitivity analyses; effect modification examined only on the basis of a priori rationale with sufficient numbers. • No deficiencies in analysis evident. Discussion of some details might be absent (e.g., examination of outliers).

Domain and core question	Prompting questions	Follow-up questions	Considerations that apply to most exposures and outcomes
<p>Analysis</p> <p>Does the analysis strategy and presentation convey the necessary familiarity with the data and assumptions?</p>	<ul style="list-style-type: none"> Does the study include additional analyses addressing potential biases or limitations (i.e., sensitivity analyses)? 	<p>If there is a concern about the potential for bias, what is the predicted direction or distortion of the bias on the effect estimate (if there is enough information)?</p>	<p>Adequate</p> <p>Same as <i>good</i>, except:</p> <ul style="list-style-type: none"> Descriptive information about exposure provided (where applicable) but could be incomplete; might not have discussed missing data, cut-points, or shape of distribution. Includes analyses that address robustness of findings (examples in <i>good</i>), but some important analyses are not performed. <p>Deficient</p> <ul style="list-style-type: none"> Does not conduct analysis using optimal characterization of the outcome variable. Descriptive information about exposure levels not provided (where applicable). Effect estimates and <i>p</i>-value presented without standard error or confidence interval. Results presented as statistically “significant”/“not significant.” <p>Critically deficient</p> <ul style="list-style-type: none"> Analysis methods are not appropriate for design or data of the study.

LOD = Limit of detection.

Selective Reporting

This domain concerns the potential for misleading results that can arise from selective reporting (e.g., of only a subset of the measures or analyses that were conducted). The concept of selective reporting involves the selection of results from among multiple outcome measures, multiple analyses, or different subgroups, based on the direction or magnitude of these results (e.g., presenting “positive” results). This domain can have fewer than four levels of rating. The core and prompting questions are presented in Table 4-6.

A related topic is the issue of multiple comparisons, and whether adjustment for the number of independent analyses (e.g., different exposures) in a study should be used. For synthesizing results across studies, IRIS focuses on the effect estimate and its variability (e.g., a Beta and the standard error of a Beta) from each study. The purpose of the systematic review is to first describe the available evidence, and then to evaluate that evidence for any causal association. Adjustment for multiple comparisons within an individual study is not necessary for this purpose ([Rothman, 2010](#)).

Table 4-6. Example question specification for evaluation of selective reporting in epidemiological studies

Domain and core question	Prompting questions	Follow-up questions	Considerations that apply to most exposures and outcomes
<p>Selective reporting Is there reason to be concerned about selective reporting?</p>	<ul style="list-style-type: none"> • Were results provided for all the primary analyses described in the methods section? • Is there appropriate justification for restricting the amount and type of results that are shown? • Are only statistically significant results presented? 	<p>If there is a concern about the potential for bias, what is the predicted direction or distortion of the bias on the effect estimate (if there is enough information)?</p>	<p>These considerations generally do not require customization and could have fewer than four levels.</p> <p>Good</p> <ul style="list-style-type: none"> • The results reported by study authors are consistent with the primary and secondary analyses described in a registered protocol or methods paper. <p>Adequate</p> <ul style="list-style-type: none"> • The authors described their primary (and secondary) analyses in the methods section and results were reported for all primary analyses. <p>Deficient</p> <ul style="list-style-type: none"> • Concerns were raised based on previous publications, a methods paper, or a registered protocol indicating that analyses were planned or conducted that were not reported, or that hypotheses originally considered to be secondary were represented as primary in the reviewed paper. • Only subgroup analyses were reported, suggesting that results for the entire group were omitted. • Only statistically significant results were reported.

Sensitivity

The domain of study “sensitivity” concerns study features that affect the ability of a study to detect a true association ([Cooper et al., 2016](#)). An insensitive study will fail to show a difference that truly exists, leading to an underestimation of the effect estimate (a “false negative” result) or an inappropriate interpretation of the study results as support for “no effect.”

Some of the study features that can affect study sensitivity might have already been included in the outcome, exposure, or other domains, such as the validity of a method used to ascertain an outcome, ability to characterize exposure in a relevant time period for the outcome under consideration, selection of affected individuals out of the study population, or inclusion of intermediaries in a model. These features should not be double counted in the “sensitivity” domain. Other features might not have been addressed and, therefore, should be included here. Examples include the exposure contrast (e.g., the ability to distinguish between the low and high exposure groups within a study), duration of exposure, and length of follow-up (for outcomes with long latency periods). Sample size or number of observed cases might also be considered within this domain but is not used as a factor that would result in a rating of “critically deficient.” The age group under study could also be relevant within the context of study sensitivity, as the appropriate age group will depend on the outcome being examined; a population might be too young or too old to provide a meaningful analysis of the effect of interest.

A rating of “good” in this domain indicates that a reported lack of association in a study can be interpreted with high confidence (barring biases towards the null in other domains), while a rating of “critically deficient” indicates the study is unlikely to be able to detect a true association that exists, and thus the result is uninterpretable. This is uncommon; an example is if there are very few participants with measurable exposure (e.g., very high percentage below the limit of detection).

The core and prompting questions for this domain are presented in Table 4-7. The decisions regarding which attributes belong in this domain will be documented in the specific exposure-outcome component of the protocol used for an assessment.

Table 4-7. Example question specification for evaluation of sensitivity in epidemiological studies

Domain and core question	Prompting questions	Follow-up questions	Considerations that apply to most exposures and outcomes
<p>Sensitivity Is there a concern that sensitivity of the study is not adequate to detect an effect?</p>	<ul style="list-style-type: none"> • Is the exposure contrast adequate to detect associations and exposure-response relationships? • Was the appropriate population included? • Was the length of follow-up adequate? Is the time/age of outcome ascertainment optimal given the interval of exposure and the health outcome? • Are there other aspects of the study that raise concerns about sensitivity? 		<p>These considerations might require customization to the specific exposure and outcome and could have fewer than four levels. Recognizing that sources of bias captured in other domains can impact study sensitivity, this domain focuses on additional considerations not specifically captured elsewhere. Some considerations include:</p> <p>Good</p> <ul style="list-style-type: none"> • There is sufficient variability/contrast in exposure to evaluate primary hypotheses. • The study population was sensitive to the development of the outcomes of interest (e.g., ages, lifestage, sex). • The timing of outcome ascertainment was appropriate given expected latency for outcome development (i.e., adequate follow-up interval). • The study was considered adequately powered to detect an effect [based on factors such as sample size (overall and across subgroups), precision, prevalence of outcome, number of covariates in model]. • No other notable concerns raised regarding study sensitivity. <p>Adequate Same considerations as Good, except:</p> <ul style="list-style-type: none"> • There might be issues identified that could reduce sensitivity, but they are considered unlikely to substantially impact the overall findings of the study. <p>Deficient</p> <ul style="list-style-type: none"> • Concerns were raised about the considerations described for Good that are expected to notably decrease the sensitivity of the study to detect associations for the outcome. <p>Critically deficient</p> <ul style="list-style-type: none"> • Severe concerns were raised about the considerations described for Good such that a true association is unlikely to be detected (i.e., null results cannot be interpreted as a lack of association). Sample size should not be used to reach this rating.

4.2.2. Final Observations

As described in Section 4.1, once the considerations have been developed and tested, the reviewers perform the study evaluations and assign ratings for each domain (good, adequate, deficient, critically deficient) and for the overall study confidence (high, medium, low, or uninformative). The results are documented as described in Section 4.1.2.

It is important to note the confidence in the study might vary depending on the specific analysis presented (i.e., greater confidence could be placed on the results of an exposure-response analysis with an internal comparison group than on a summary standardized mortality ratio in an occupational exposure study); thus, the confidence characterization could apply only to one outcome or one analysis of a study. With a few exceptions, the evaluation does not incorporate information about the study results (i.e., do the results provide evidence of an association?); this information is addressed in the synthesis phase described in Chapter 6. Review of some of the results might be needed to complete some evaluations. For example, within the context of the evaluation of confounding, the results are considered because confounding depends on the strength of various relationships (i.e., between the exposure and the confounder and between the confounder and the outcome).

Lastly, critically deficient and uninformative ratings are uncommon; these ratings are reserved for critical flaws where the study findings are truly uninterpretable due to identified biases. The most frequent situation where they are used for epidemiological studies is when potential confounding has not been considered using any method (e.g., adjustment, stratification, restriction), including unadjusted correlation coefficients or means in cases/controls in a heterogeneous population where confounding is likely.

4.3. EVALUATION OF CONTROLLED HUMAN EXPOSURE STUDIES

Controlled human exposure studies involve intentionally exposing volunteer human subjects to an agent over short periods to test specific hypotheses about short-term exposures and biological responses. For these studies to be ethically conducted, the effects being studied must be temporary and reversible, unless there is an expectation of possible benefit (e.g., vanadium supplementation). Reviewers should confirm the authors included an explicit declaration that the study protocol was approved by an institutional review board. For study evaluation, a process incorporating aspects of the evaluation approaches used for epidemiological studies and experimental animal studies, such as the Cochrane RoB tools for randomized trials (ROB2) ([Sterne et al., 2019](#)) and the ROBINS-I tool discussed in Section 4.2 ([Sterne et al., 2016](#)), should be used to

evaluate controlled exposure studies in humans.⁹ Generally, controlled human exposure studies should be evaluated for important attributes of experimental studies, including randomization of exposure assignments, blinding of subjects and investigators, exposure generation and characterization, appropriateness of control exposures or comparisons, outcome ascertainment, missing data, deviations from the intended exposure (when relevant), study sensitivity, and other aspects of the exposure protocol.

4.4. EVALUATION OF EXPERIMENTAL ANIMAL TOXICOLOGICAL STUDIES

Using the principles described in Section 4.1, the animal studies of health effects are evaluated for RoB and sensitivity using the following domains: allocation, observational bias/blinding, confounding, attrition, chemical administration and characterization, endpoint measurement, results presentation, selective reporting, and sensitivity (see Table 4-8).

The IRIS RoB evaluation is influenced by several other existing approaches used in environmental health or preclinical research to evaluate animal studies, including the Office of Health Assessment and Translation [OHAT ([NTP, 2019](#); [NIEHS, 2015](#))], the Office of Report on Carcinogens ([NIEHS, 2015](#)), Navigation Guide ([Woodruff and Sutton, 2014](#)), Systematic Review Centre for Laboratory Animal Experimentation ([Hooijmans et al., 2014](#)), and Science in Risk Assessment and Policy [SciRAP ([Molander et al., 2015](#))]. The IRIS approach includes a sensitivity domain to capture certain aspects of study design that do not strictly fall under RoB defined as “a systematic error, or deviation from the truth, in results or inferences” ([Cooper et al., 2016](#)). Briefly, evaluation of the sensitivity of experimental animal toxicity studies seeks to establish the level of confidence in an effect being truly detected and the potential for false negative results. For example, a study could have been conducted in way that is bias-free but looked at an inappropriate period of exposure for the outcome of interest. The IRIS approach considers sensitivity separately to distinguish these considerations more clearly from RoB.

Table 4-8 provides core and prompting questions for each evaluation domain and general considerations to guide the reviewers during study evaluations. For some domains, the general considerations described below might need to be refined by assessment teams to meet the specific needs of the assessment (e.g., considerations specific to the test chemical) or the evidence base (e.g., developing assay specific considerations). In addition to the general considerations, example ratings and rationales have been developed and are available in the HAWC project “SEM Template Figures and Resources” (see “Example answers to the animal study evaluation domain” attachment).

⁹The Cochrane ROB2 and ROBINS-I tools are valuable resources for identifying considerations in evaluating these studies but in most cases cannot be used “off the shelf” for these studies due to differences in the typical study design. Controlled human exposure studies of chemical exposures are most commonly performed without randomization and a control group. Participants act as their own controls with measurement of outcomes at baseline and post-exposure, which these tools are not designed to evaluate.

Table 4-8. Domains, questions, and general considerations to guide the evaluation of animal studies

Domain and core question	Prompting questions	General considerations
<p>Allocation</p> <p>Were animals assigned to experimental groups using a method that minimizes selection bias?</p>	<p>For each study:</p> <p>Did each animal or litter have an equal chance of being assigned to any experimental group (i.e., random allocation)?^a</p> <p>Is the allocation method described?</p> <p>Aside from randomization, were any steps taken to balance variables across experimental groups during allocation?</p>	<p>These considerations typically do not need to be refined by assessment teams.</p> <p>A judgment and rationale for this domain should be given for each cohort or experiment in the study.</p> <p>Good: Experimental groups were randomized, and any specific randomization procedure was described or inferable (e.g., computer-generated scheme. Note that normalization is not the same as randomization [see response for <i>adequate</i>]).</p> <p>Adequate: Authors report that groups were randomized but do not describe the specific procedure used (e.g., “animals were randomized”). Alternatively, authors used a nonrandom method to control for important modifying factors across experimental groups (e.g., body-weight normalization).</p> <p>Not reported (interpreted as <i>deficient</i>): No indication of randomization of groups or other methods (e.g., normalization) to control for important modifying factors across experimental groups.</p> <p>Deficient: Bias in the animal allocations was reported or inferable but is not expected to be severe.</p> <p>Critically deficient: Severe bias in the animal allocations was reported or inferable.</p>
<p>Observational bias/blinding</p> <p>Did the study implement measures to reduce observational bias?</p>	<p>For each endpoint/outcome or grouping of endpoints/outcomes in a study:</p> <p>Does the study report blinding or other procedures for reducing observational bias?</p> <p>If not, did the study use a design or approach for which such procedures can be inferred?</p> <p>What is the expected impact of failure to implement (or report implementation) of these procedures on results?</p>	<p>These considerations typically need not be refined by the assessment teams. (Note that it can be useful for teams to identify highly subjective measures of endpoints/outcomes where observational bias might strongly influence results prior to performing evaluations.)</p> <p>A judgment and rationale for this domain should be given for each endpoint/outcome or group of endpoints/outcomes investigated in the study.</p> <p>Good: Measures to reduce observational bias were described (e.g., blinding to conceal treatment groups during endpoint evaluation; consensus-based evaluations of histopathology—lesions).^b</p> <p>Adequate: Methods for reducing observational bias (e.g., blinding) is not explicitly stated but can be inferred.</p>

ORD Staff Handbook for Developing IRIS Assessments

Domain and core question	Prompting questions	General considerations
		<p>Not reported: Measures to reduce observational bias were not described.</p> <p>(Interpreted as <i>adequate</i>): The potential concern for bias was mitigated on the basis of using automated/computer driven systems, standard laboratory kits, relatively simple, objective measures (e.g., body or tissue weight), or screening-level evaluations of histopathology.</p> <p>(Interpreted as <i>deficient</i>): The potential impact on the results is major (e.g., outcome measures are highly subjective).</p> <p>Critically deficient: Strong evidence for observational bias that impacted the results.</p>
<p>Confounding</p> <p>Are variables with the potential to confound or modify results controlled for and consistent across experimental groups?</p> <p><i>Note:</i></p> <p>Consideration of overt toxicity (possibly masking more specific effects) is addressed under endpoint measurement reliability.</p>	<p>For each study:</p> <p>Are there differences across the treatment groups, considering both differences related to the exposure (e.g., coexposures, vehicle, diet, palatability) and other aspects of the study design or animal groups (e.g., animal source, husbandry, or health status), that could bias the results?</p> <p>If differences are identified, to what extent are they expected, based on a specific scientific understanding, to impact the results?</p>	<p>These considerations might need to be refined by assessment teams, as the specific variables of concern can vary by experiment or chemical.</p> <p>A judgment and rationale for this domain should be given for each cohort or experiment in the study, noting when the potential for confounding is restricted to specific endpoints/outcomes.</p> <p>Good: Outside of the exposure of interest, variables likely to confound or modify results appear to be controlled for and consistent across experimental groups.</p> <p>Adequate: Some concern that variables likely to confound or modify results were uncontrolled or inconsistent across groups but are expected to have a minimal impact on the results.</p> <p>Deficient: Notable concern that potentially confounding variables were uncontrolled or inconsistent across groups and are expected to substantially impact the results.</p> <p>Critically deficient: Confounding variables were presumed to be uncontrolled or inconsistent across groups and are expected to be a primary driver of the results.</p>
<p>Attrition</p> <p>Did the study report results for all tested animals?</p>	<p>For each study:</p> <p>Are all animals accounted for in the results?</p> <p>If there is attrition, do authors provide an explanation (e.g., death or unscheduled sacrifice during the study)?</p>	<p>These considerations typically do not need to be refined by assessment teams.</p> <p>A judgment and rationale for this domain should be given for each cohort or experiment in the study.</p> <p>Good: Results were reported for all animals. If animal attrition is identified, the authors provide an explanation, and these are not expected to impact the interpretation of the results.</p>

ORD Staff Handbook for Developing IRIS Assessments

Domain and core question	Prompting questions	General considerations
	<p>If unexplained attrition of animals for outcome assessment is identified, what is the expected impact on the interpretation of the results?</p>	<p>Adequate: Results are reported for most animals. Attrition is not explained but this is not expected to significantly impact the interpretation of the results.</p> <p>Deficient: Moderate to high level of animal attrition that is not explained and could significantly impact the interpretation of the results.</p> <p>Critically deficient: Extensive animal attrition that prevents comparisons of results across treatment groups.</p>
<p>Chemical administration and characterization</p> <p>Did the study adequately characterize exposure to the chemical of interest and the exposure administration methods?</p> <p><i>Note:</i></p> <p><i>Consideration of the appropriateness of the route of exposure (not the administration method) is not a risk of bias consideration. Relevance and utility of the routes of exposure are considered in the PECO criteria for study inclusion and during evidence synthesis.</i></p> <p><i>Relatedly, consideration of exposure level selection (e.g., were levels sufficiently high to elicit effects) is addressed during evidence synthesis and is not a risk of bias consideration.</i></p>	<p>For each study:</p> <p>Are there concerns [specific to this chemical] regarding the source and purity or composition (e.g., identity and percent distribution of different isomers) of the chemical?</p> <p>Was independent analytical verification of the test article (e.g., composition, homogeneity, and purity) performed?</p> <p>Were nominal exposure levels verified analytically? Are there concerns about the methods used to administer the chemical (e.g., inhalation chamber type, gavage volume)?</p>	<p>It is essential these considerations are considered, and potentially refined, by assessment teams, as the specific variables of concern can vary by chemical (e.g., stability might be an issue for one chemical but not another).</p> <p>A judgment and rationale for this domain should be given for each cohort or experiment in the study.</p> <p>Good: Chemical administration and characterization is complete (i.e., source and purity are provided or can be obtained from the supplier and test article is analytically verified). There are no notable concerns about the composition, stability, or purity of the administered chemical, or the specific methods of administration. Exposure levels are verified using reliable analytical methods.</p> <p>Adequate: Some uncertainties in the chemical administration and characterization are identified, but these are expected to have minimal impact on interpretation of the results (e.g., purity of the test article is suboptimal but interpreted as unlikely to have a significant impact; analytical verification of exposure levels is not reported or verified with nonpreferred methods).</p> <p>Deficient: Uncertainties in the exposure characterization are identified and expected to substantially impact the results (e.g., source of the test article is not reported, and composition is not independently verified; impurities are substantial or concerning; administration methods are considered likely to introduce confounders, such as use of static inhalation chambers or a gavage volume considered too large for the species or lifestage at exposure).</p> <p>Critically deficient: Uncertainties in the exposure characterization are identified and there is reasonable certainty that the study results are largely attributable to factors other than exposure to the chemical of interest (e.g., identified impurities are expected to be a primary driver of the results).</p>

Domain and core question	Prompting questions	General considerations
<p>Endpoint measurement</p> <p>Are the selected procedures, protocols, and animal models adequately described and appropriate for the endpoint(s)/outcome(s) of interest?</p> <p><i>Notes:</i></p> <p><i>Considerations related to the sensitivity of the animal model and timing of endpoint measurement are evaluated under sensitivity</i></p> <p><i>considerations related to adjustments/corrections to endpoint measurements (e.g., organ weight corrected for body weight) are addressed under results presentation.</i></p>	<p>For each endpoint/outcome or grouping of endpoints/outcomes in a study:</p> <p>Are the evaluation methods and animal model adequately described and appropriate?</p> <p>Are there concerns regarding the methodology selected for endpoint evaluation?</p> <p>Are there concerns about the specificity of the experimental design?</p> <p>Are there serious concerns regarding the sample size or how endpoints were sampled?</p> <p>Are appropriate control groups for the study/assay type included?</p>	<p>Considerations for this domain are highly variable depending on the endpoint(s)/outcome(s) of interest and typically must be refined by assessment teams.</p> <p>A judgment and rationale for this domain should be given for each endpoint/outcome or group of endpoints/outcomes investigated in the study.</p> <p>Some considerations include the following:</p> <p>Good:</p> <ul style="list-style-type: none"> • Adequate description of methods and animal models. • Use of generally accepted and reliable endpoint methods. • Sample sizes are generally considered adequate for the assay or protocol of interest and there are no notable concerns about sampling in the context of the endpoint protocol (e.g., sampling procedures for histological analysis). • Includes appropriate control groups and any use of nonconcurrent or historical control data (e.g., for evaluation of rare tumors) is justified (e.g., authors or evaluators considered the similarity between current experimental animals and laboratory conditions to historical controls). <p>Ratings of Adequate, Deficient, and Critically Deficient are generally defined as follows:</p> <p>Adequate: Issues are identified that could affect endpoint measurement but are considered unlikely to substantially impact the overall findings or the ability to reliably interpret those findings.</p> <p>Deficient: Concerns are raised that are expected to notably affect endpoint measurement and reduce the reliability of the study findings.</p> <p>Critically deficient: Severe concerns are raised about endpoint measurement and any findings are likely to be largely explained by these limitations.</p> <p>The following specific examples of relevant concerns are typically associated with a Deficient rating, but Adequate or Critically Deficient might be applied depending on the expected impact of limitations on the reliability and interpretation of the results:</p> <ul style="list-style-type: none"> • Study report lacks important details necessary to evaluate the appropriateness of the study design (e.g., description of the assays or protocols; information on the strain, sex, or lifestage of the animals)

ORD Staff Handbook for Developing IRIS Assessments

Domain and core question	Prompting questions	General considerations
		<ul style="list-style-type: none"> • Selection of protocols that are nonpreferred or lack specificity for investigating the endpoint of interest. This includes omission of additional experimental criteria (e.g., inclusion of a positive control or dosing up to levels causing minimal toxicity) when required by specific testing guidelines/protocols.* • Overt toxicity (e.g., mortality, extreme weight loss) is observed or expected on the basis of findings from similarly designed studies and might mask interpretation of outcome(s) of interest. • Sample sizes are smaller than is generally considered adequate for the assay or protocol of interest. Inadequate sampling can also be raised within the context of the endpoint protocol (e.g., in a pathology study, bias that is introduced by only sampling a single tissue depth or an inadequate number of slides per animal).** • Control groups are not included, considered inappropriate, or comparisons to nonconcurrent or historical controls are not adequately justified. <p>*These limitations typically also raise a concern for insensitivity.</p> <p>**Sample size alone is not a reason to conclude an individual study is critically deficient.</p>
<p>Results presentation</p> <p>Are the results presented and compared in a way that is appropriate and transparent?</p>	<p>For each endpoint/outcome or grouping of endpoints/outcomes in a study:</p> <p>Does the level of detail allow for an informed interpretation of the results?</p> <p>Are the data compared, or presented, in a way that is inappropriate or misleading?</p>	<p>Considerations for this domain are highly variable depending on the outcomes of interest and typically must be refined by assessment teams.</p> <p>A judgment and rationale for this domain should be given for each endpoint/outcome or group of endpoints/outcomes investigated in the study.</p> <p>Some considerations include the following:</p> <p>Good:</p> <ul style="list-style-type: none"> • No concerns with how the data are presented. • Results are quantified or otherwise presented in a manner that allows for an independent consideration of the data (assessments do not rely on author interpretations). • No concerns with completeness of the results reporting.*

Domain and core question	Prompting questions	General considerations
		<p>Ratings of Adequate, Deficient, and Critically Deficient are generally defined as follows:</p> <p>Adequate: Concerns are identified that could affect results presentation but are considered unlikely to substantially impact the overall findings or the ability to reliably interpret those findings.</p> <p>Deficient: Concerns with results presentation are identified and expected to substantially impact results interpretation and reduce the reliability of the study findings.</p> <p>Critically deficient: Severe concerns about results presentation were identified and study findings are likely to be largely explained by these limitations or failure to report any results (qualitative or quantitative) for a prespecified outcome.*</p> <p>The following specific examples of relevant concerns are typically associated with a Deficient rating, but Adequate or Critically Deficient might be applied depending on expected impact of limitations on the reliability and interpretation of the results:</p> <ul style="list-style-type: none"> • Nonpreferred presentation of data (e.g., developmental toxicity data averaged across pups in a treatment group, when litter responses are more appropriate; presentation of only absolute organ weight data when relative weights are more appropriate). • Pooling data when responses are known or expected to differ substantially (e.g., across sexes or ages). • Incomplete presentation of the data* (e.g., presentation of mean without variance data; concurrent control data are not presented; dichotomizing or truncating continuous data). <p>*Failure to describe <u>any</u> findings for assessed outcomes (i.e., report lacks any qualitative or quantitative description of the results in tables, figures, or text) results in a critically deficient rating for the outcome(s) of interest for results presentation; overall completeness of reporting at the study level is addressed under selective reporting.</p>
<p>Selective reporting</p> <p>Did the study report result for all prespecified outcomes?</p> <p><i>Note:</i></p> <p><i>This domain does not consider the appropriateness of the</i></p>	<p>For each study:</p> <p>Are results presented for all endpoints/outcomes described in the methods (see note)?</p>	<p>These considerations typically need not be refined by assessment teams.</p> <p>A judgment and rationale for this domain should be given for each cohort or experiment in the study.</p> <p>Good: Quantitative or qualitative results were reported for all prespecified outcomes (explicitly stated or inferred), exposure groups and evaluation time points. Data not reported in the primary article are available from supplemental material. If results</p>

Domain and core question	Prompting questions	General considerations
<p><i>analysis/results presentation. This aspect of study quality is evaluated in another domain.</i></p>	<p>If unexplained results omissions are identified, what is the expected impact on the interpretation of the results?</p>	<p>omissions are identified, the authors provide an explanation, and these are not expected to impact the interpretation of the results.</p> <p>Adequate: Quantitative or qualitative results are reported for most prespecified outcomes (explicitly stated or inferred) and evaluation time points. Omissions are not explained but are not expected to significantly impact the interpretation of the results.</p> <p>Deficient: Quantitative or qualitative results are missing for many prespecified outcomes (explicitly stated or inferred), omissions are not explained and could significantly impact the interpretation of the results.</p> <p>Critically deficient: Extensive results omission is identified and prevents comparisons of results across treatment groups.</p>
<p>Sensitivity</p> <p>Are there concerns sensitivity in the study is not adequate to detect an effect?</p> <p><i>Note:</i></p> <p><i>Consideration of exposure level selection (e.g., were levels sufficiently high to elicit effects) is addressed during evidence synthesis and is not a study sensitivity consideration.</i></p>	<p>Was the exposure period, timing (e.g., lifestage), frequency, and duration sensitive for the outcome(s) of interest?</p> <p>Based on knowledge of the health hazard of concern, did the selection of species, strain, or sex of the animal model reduce study sensitivity?</p> <p>Are there concerns regarding the timing (e.g., lifestage) of the outcome evaluation?</p> <p>Are there aspects related to risk of bias domains that raise concerns about insensitivity (e.g., selection of protocols that are known to be insensitive or nonspecific for the outcome(s) of interest)?</p>	<p>These considerations might require customization to the specific exposure and outcomes. Some study design features that affect study sensitivity might have already been included in the other evaluation domains; these should be noted in this domain, along with any features that have not been addressed elsewhere. Some considerations include:</p> <p>Good</p> <ul style="list-style-type: none"> • The experimental design (considering exposure period, timing, frequency, and duration) is appropriate and sensitive for evaluating the outcome(s) of interest. • The selected animal model (considering species, strain, sex, or lifestage) is known or assumed to be appropriate and sensitive for evaluating the outcome(s) of interest. • No significant concerns with the ability of the experimental design to detect the specific outcome(s) of interest (e.g., outcomes evaluated at the appropriate lifestage; study designed to address known endpoint variability that is unrelated to treatment, such as estrous cyclicity or time of day). • Timing of endpoint measurement in relation to the chemical exposure is appropriate and sensitive (e.g., behavioral testing is not performed during a transient period of test chemical-induced depressant or irritant effects; endpoint testing does not occur only after a prolonged period, such as weeks or months, of nonexposure). • Potential sources of bias toward the null are not a substantial concern. <p>Adequate</p> <p>Same considerations as <i>Good</i>, except:</p>

Domain and core question	Prompting questions	General considerations
		<ul style="list-style-type: none"> The duration and frequency of the exposure was appropriate, and the exposure covered most of the critical window (if known) for the outcome(s) of interest. Potential issues are identified that could reduce sensitivity, but they are unlikely to impact the overall findings of the study. <p>Deficient</p> <ul style="list-style-type: none"> Concerns were raised about the considerations described for <i>Good</i> or <i>Adequate</i> that are expected to notably decrease the sensitivity of the study to detect a response in the exposed group(s). <p>Critically deficient</p> <ul style="list-style-type: none"> Severe concerns were raised about the sensitivity of the study and experimental design such that any observed associations are likely explained by bias. The rationale should indicate the specific concern(s).
<p>Overall confidence</p> <p>Considering the identified strengths and limitations, what is the overall confidence rating for the endpoint(s)/outcome(s) of interest?</p>	<p>For each endpoint/outcome or grouping of endpoints/outcomes in a study:</p> <p>Were concerns (i.e., limitations or uncertainties) related to the risk of bias or sensitivity identified?</p> <p>If yes, what is their expected impact on the overall interpretation of the reliability and validity of the study results, including (when possible) interpretations of impacts on the magnitude or direction of the reported effects?</p>	<p>The overall confidence rating considers the likely impact of the noted concerns (i.e., limitations or uncertainties) in reporting, bias, and sensitivity on the results.</p> <p>Reviewers should mark studies that are rated lower than high confidence due only to low sensitivity (i.e., bias toward the null) for additional consideration during evidence synthesis. If the study is otherwise well conducted and an effect is observed, it might increase the strength of evidence judgment.</p> <p>A confidence rating and rationale should be given for each endpoint/outcome or group of endpoints/outcomes investigated in the study. Confidence ratings are described above (see Section 4.1.1).</p>

PECO = population, exposure, comparators, outcome.

^aSeveral studies have characterized the relevance of randomization, allocation concealment, and blind outcome assessment in experimental studies ([Hirst et al., 2014](#); [Krauth et al., 2013](#); [Macleod, 2013](#); [Higgins and Green, 2011b](#)).

^bFor nontargeted or screening-level histopathology outcomes often used in guideline studies, blinding during the initial evaluation of tissues is generally not recommended as masked evaluation can make “the task of separating treatment-related changes from normal variation more difficult” and “there is concern that masked review during the initial evaluation might result in missing subtle lesions.” Generally, blinded evaluations are recommended for targeted secondary review of specific tissues or in instances when there is a predefined set of outcomes that is known or predicted to occur ([Crissman et al., 2004](#)).

4.5. PRIORITIZATION AND EVALUATION OF NON-POPULATIONS, EXPOSURES, COMPARATORS, AND OUTCOMES (PECO) STUDIES

4.5.1. Prioritization of Non-Populations, Exposures, Comparators, and Outcomes (PECO) Studies

Although potentially informative, mechanistic studies initially tagged as “potentially relevant supplemental material” do not meet PECO criteria (discussed in Section 2.2) and are not routinely evaluated for RoB and sensitivity. This is because the process of evaluating mechanistic information differs from evaluations of the other evidence streams, as it focuses on the analysis of individual mechanistic “events” or sets of related events, typically with less focus on individual studies. In addition, an intensive analysis may not be warranted for mechanistic events not expected to meaningfully impact assessment approaches or conclusions or for those already well accepted scientifically. For many chemicals, the sheer number of mechanistic studies warrants the use of pragmatic approaches to help narrow the scope of studies that might require detailed summarization and evaluation at the individual study level. Therefore, mechanistic studies are prioritized to identify the most relevant evidence, including studies that could be evaluated for RoB and sensitivity. This prioritization process is also applicable to other types of non-PECO studies prioritized as influential to a key judgment (e.g., a subset of zebrafish studies included within the unit of analysis for an assessment of effects on neurobehavior; non-PECO exposure routes when pharmacokinetic [PK] related to an outcome is well understood).

The prioritization of mechanistic and other non-PECO studies is a stepwise process that begins with the selection of health effects and associated units of analysis, exposure levels, and lifestage(s) are to be included in the hazard synthesis. The next step of prioritization is to identify the most mechanistically relevant studies on the basis of the extent to which the reported endpoints, and the experimental models, assays, and study designs used to experimentally evaluate these endpoints, inform the identified hazard questions of interest. This is facilitated by the formation of the supplemental content inventory and subcategories for screening and tagging mechanistic studies (see Section 2.5.2). Depending on the available human and animal evidence base and the needs indicated by the human and animal evidence syntheses (see Chapter 6), a subset of the most relevant mechanistic studies will be prioritized for inventory and evaluation. For example, a detailed analysis of mechanistic information might be prioritized when (1) little or no evidence is available from epidemiological studies or animal bioassays, (2) the reported findings on a critical mechanistic event are conflicting, or (3) the available mechanistic evidence addresses a complex and influential aspect of the assessment, particularly those expected to significantly impact hazard conclusions or assumptions about dose-response analysis.

There is no one-size-fits-all approach; therefore, the method used to synthesize and integrate the prioritized mechanistic evidence will be customized depending on what uncertainty(ies) the evidence addresses. Development of this stepwise approach is based on staff

experience in conducting assessments and input from several EPA-sponsored workshops organized to discuss pragmatic approaches for considering mechanistic information when conducting a systematic review [e.g., [NAS \(2018\)](#) and [NAS \(2019\)](#)]. A comprehensive mechanistic evaluation (which might include a mode of action (MOA) analysis) is not necessarily conducted for every potential hazard discussed in the assessment. The analysis of mechanistic evidence can range from a high-level summary of potential mechanisms of action to answering specific, focused questions needed to address key uncertainties. The scope, complexity, and depth of the mechanistic analyses will vary on the basis of the key science issues and confidence of the studies used to inform the evidence synthesis judgments. For example, effort spent on an in-depth analysis of mechanisms associated with a health effect supported by exposure-dependent findings from multiple *medium* and *high* confidence human studies may have relatively little impact on hazard characterization conclusions; in this case, it may make more sense to focus the mechanistic analyses on identifying information on potentially susceptible populations and lifestages or data that could inform the shape of the dose-response curve (i.e., if the available human data have substantial quantitative uncertainties). The same could be true for animal and human outcomes with well-accepted mechanistic associations, where a broad overview can provide the appropriate context.

The approach to prioritizing non-PECO studies for RoB and sensitivity evaluation is intentionally flexible to accommodate varied evidence bases and mechanistically based predictive approaches to testing and assessment. The decision to evaluate non-PECO studies will also consider if the evidence is applicable to multiple health outcomes (e.g., chemical-molecular or molecular-molecular interactions that are shared or interact between biological systems), the abundance of information, and the assessment resources available. The decision to conduct evaluations of non-PECO studies additionally includes a tailored approach that considers the evidence in the context of one or more factors (i.e., canonical biological pathway knowledge, chemical toxicodynamic knowledge, human relevance, experimental design and methodology, potential to inform susceptibility, or certainty in the outcome assessment to inform human relevance). Regardless of the approach, the steps taken for the selective evaluation of non-PECO studies should be transparently documented during assessment development. Additional evaluation of the available mechanistic information can strengthen the justification for or against an endpoint of concern or related unit(s) of analysis (see Chapter 3).

4.5.2. Evaluation of Non-Populations, Exposures, Comparators, and Outcomes (PECO) Studies

Study evaluation tools for supplemental evidence not meeting PECO criteria are developed on the basis of the design of the assessed studies and not necessarily where those studies fit within an evidence synthesis and integration narrative. Sources of evidence for the potential mechanism(s) of toxicity for a health effect can span a wide range of evidence types, including mechanistic endpoint evaluations across in vivo human and animal studies (see Sections 4.2–4.4), in vitro studies (see Section 4.5.3), and other types of non-PECO supplemental studies, such as

alternative animal models (e.g., zebrafish; *C. elegans*). Although validated tools designed for evaluating all potential alternative animal models do not exist, these studies can be evaluated using a combination of considerations from the in vivo animal and in vitro approaches.

Similar to the evaluation of apical outcomes reported in epidemiological and animal evidence, study evaluation considerations for individual mechanistic studies will differ depending on multiple factors, including the type of endpoints, study and experimental designs, model systems, and population(s) evaluated. Thus, any of the evaluation methods described in Sections 4.2 through 4.5 might be applicable to the mechanistic evidence synthesis, although some of the general considerations must be refined (i.e., in the protocol or assessment) to address the utility of a study to assess mechanistic endpoints. As mechanistic methods are rapidly evolving and often no “standard practice” exists, it should be determined early in assessment development whether the assessment team has the knowledge and familiarity with the available mechanistic study designs, methodologies, and endpoints to perform the study evaluation approach(es).

4.5.3. Evaluation of In Vitro Studies

The development of methods for the evaluation of in vitro studies lags behind that of human and animal studies, although it is an active area of development in the field of systematic review. Historically, most in vitro study tools focused on reporting quality and RoB (internal validity) ([NASEM, 2018](#); [NTP, 2015](#)). Current trends are to expand the assessment of mechanistic data to include methodological quality with consideration of potential bias ([U.S. EPA, 2015a](#)). The approach taken by the IRIS Program is based on the domains described for animal study evaluations (see Section 4.4), namely RoB and sensitivity, with modifications (see Table 4-9). The IRIS Program is aware of other tools and approaches for evaluating in vitro studies ([Beronius et al., 2018](#); [NASEM, 2018](#); [OECD, 2018](#); [U.S. EPA, 2018c](#)) and will continue to monitor developments through collaborative engagement with communities with similar motivation. Existing tools to evaluate in vitro studies tend to be general and designed for application to all in vitro studies. However, it should be acknowledged that to be truly useful in evaluating the RoB and sensitivity of in vitro studies, additional assay-specific considerations will need to be developed and applied to these domains. Variations in application will include more complex in vitro cell culture test systems (e.g., co-cultures of two or more cell types; pluripotent cells from human volunteers; intact, functional tissues grown in a dish). In addition, some elements of this approach might be useful when evaluating studies in alternative animal model systems (e.g., *C. elegans*). This increases the challenge of operationalizing a one-size-fits-all approach. Therefore, pilot testing across assessments with different in vitro evidence bases will be key for refining these considerations to be useful and practical for all in vitro studies that require evaluation. Adaptations in the use of this approach will be documented within assessment-specific protocols as necessary.

Table 4-9. Domains, questions, and general considerations to guide the evaluation of in vitro studies

Domain and core question	Prompting questions	General considerations
<p>Observational bias/blinding</p> <p>Did the study implement measures, where possible, to reduce observational bias?</p> <p>Considerations will vary depending on the specific assay/model system used and may not be applicable to some analyses.</p>	<p>For each assay or endpoint in a study:</p> <p>Did the study report steps taken to minimize observational bias during analysis (e.g., blinding/coding of slides or plates for analysis, collection of data from randomly selected fields, positive controls that are not immediately identifiable)?</p> <p>If not, did the study use a design or approach for which such procedures can be inferred, or which would not be possible to implement?</p> <p>Were the assays evaluated using automated approaches (e.g., microplate readers) that reduce concern for observational bias?</p> <p>What is the expected impact of failure to implement (or report implementation) of these methods/procedures on results?</p>	<p>These considerations typically do not need to be refined by the assessment teams. Prior to performing evaluations, teams should consider the specific assay to identify highly subjective measures of endpoints where observational bias may strongly influence results.</p> <p>A judgment and rationale for this domain should be given for each assay or endpoint or group of endpoints investigated in the study.</p> <p>Good: Measures to reduce observational bias were described (e.g., specific mention of blinding or coding of slides for analysis) or observational bias is not a concern because of use of automated/computer driven systems or standard laboratory kits.</p> <p>Not reported, interpreted as adequate: Measures to reduce observational bias were not described, but the potential concern for bias was mitigated because protocol cited includes a description of requirements for blinding/coding, or the impact on results is expected to be minor because the specific measurement is more objective.</p> <p>Not reported, interpreted as deficient: No protocol cited; the potential impact on the results is major because the endpoint measures are highly subjective (e.g., counting plaques or live vs. dead cells).</p> <p>Critically deficient: Strong evidence for observational bias that could have impacted the results.</p>
<p>Variable Control</p> <p>Are all introduced variables with the potential to affect the results of interest controlled for and</p>	<p>For each study:</p> <p>Are there any known or presumed differences across treatment groups (e.g., coexposures, culture conditions, cell passages, variations in reagent production lots, mycoplasma infections)</p>	<p>These considerations will need to be refined by assessment teams as the specific variables of concern can vary by the experimental test system and chemical.</p>

Domain and core question	Prompting questions	General considerations
<p>consistent across experimental groups?</p>	<p>that could bias the results? If differences are identified, to what extent are they expected to impact the results?</p> <p>Did the study address features inherent to the physicochemical properties of the test substance(s) that have the potential to bias the results away from the null? For example, could the test article interfere with a given assay (e.g., auto-fluoresces or inhibits enzymatic processes necessary for assay signals), potentially leading to an erroneous positive signal? <i>(Note that concerns related to dose are addressed in chemical administration and characterization.)</i></p> <p>Are there known variations in cellular signaling unique to the model system that could influence the possibility of detecting the effect(s) of interest?</p> <p>Are there concerns regarding the negative (untreated or vehicle) controls used? Were negative controls run concurrently?</p>	<p>A judgment and rationale for this domain should be given for each experiment in the study, noting when the potential to affect results is restricted to specific assays or endpoints.</p> <p>Good: Outside of the exposure of interest, variables or features of the test system or chemical properties likely to impact results appear to be controlled for and consistent across experimental groups.</p> <p>Adequate: Some concern that variables or features of the test system or chemical properties likely to modify or interfere with results were uncontrolled or inconsistent across groups but are expected to have a minimal impact on the results.</p> <p>Deficient: Notable concern that important study variables or features of the test system lacked specificity or were uncontrolled or inconsistent across groups and are expected to substantially impact the results.</p> <p>Critically deficient: Features of the test system are known to be nonspecific for this endpoint or influential study variables were presumed to be uncontrolled or inconsistent across groups and are expected to be a primary driver of the results.</p>
<p>Selective Reporting</p> <p>Did the study present results, quantitatively or qualitatively, for all prespecified assays or endpoints and replicates described in the methods?</p> <p><i>Note: The appropriateness of the analysis or results presentation is considered under results presentation.</i></p>	<p>For each study:</p> <p>Are results presented for all endpoints/outcomes described in the methods?</p> <p>Did the study clearly indicate the number of replicate experiments performed? Were the replicates technical (from the same sample) or independent (from separate, distinct exposures)?</p> <p>If unexplained results omissions are identified, what is the expected impact on the interpretation of the results?</p>	<p>These considerations typically do not need to be refined by assessment teams.</p> <p>A judgment and rationale for this domain should be given for each assay or endpoint in the study.</p> <p>Good: Quantitative or qualitative results were reported for all prespecified assays or endpoints (explicitly stated or inferred), exposure groups and evaluation timepoints. Data not reported in the primary article are available from supplemental material. If results omissions are identified, the authors provide an explanation, and these are not expected to impact the interpretation of the results.</p>

Domain and core question	Prompting questions	General considerations
		<p>Adequate: Quantitative or qualitative results are reported for most prespecified assays or endpoints (explicitly stated or inferred), exposure groups, and evaluation timepoints. Omissions are not explained but are not expected to significantly impact the interpretation of the results.</p> <p>Deficient: Quantitative or qualitative results are missing for many prespecified assays or endpoints (explicitly stated or inferred), exposure groups, and evaluation timepoints; omissions are not explained and may significantly impact the interpretation of the results.</p> <p>Critically Deficient: Extensive results omissions are identified, preventing comparisons of results across treatment groups.</p>
<p>Chemical administration and characterization</p> <p>Did the study adequately characterize exposure to the chemical of interest and the exposure administration methods?</p>	<p>For each study:</p> <p>Are there concerns regarding the purity or composition (e.g., identity and percent distribution of different isomers) of the test material/chemical? If so, can the purity or composition be obtained from the supplier (e.g., as reported on the website)?</p> <p>Was independent analytical verification of the test article purity and composition performed? If not, is this a significant concern for this substance?</p> <p>Are there concerns about the stability of the test chemical in the vehicle or culture media (e.g., pH, solubility, volatility, adhesion to plastics) that were not corrected for, leading to potential bias away from the null (e.g., observed precipitate formation at high concentrations) or toward the null (e.g., enclosed chambers not used for testing volatile chemicals)?</p> <p>Are there concerns about the preparation or storage conditions of the test substance?</p> <p>Are there concerns about the methods used to administer the chemical?</p>	<p>It is essential that these criteria are considered, and potentially refined, by assessment teams, as the specific variables of concern can vary by chemical (e.g., stability could be an issue for one chemical but not another).</p> <p>A judgment and rationale for this domain should be given for each experiment in the study.</p> <p>Good: Chemical administration and characterization is complete (i.e., source, purity, and analytical verification of the test article are provided). There are no concerns about the composition, stability, or purity of the administered chemical, or the specific methods of administration.</p> <p>Adequate: Some uncertainties in the chemical administration and characterization are identified, but these are expected to have minimal impact on interpretation of the results (e.g., source and vendor-reported purity are presented but not independently verified, purity of the test article is suboptimal but not concerning).</p> <p>Deficient: Uncertainties in the exposure characterization are identified and expected to substantially impact the</p>

Domain and core question	Prompting questions	General considerations
		<p>results (e.g., the source and purity of the test article are not reported, and no independent verification of the test article was conducted; levels of impurities are substantial or concerning; deficient administration methods were used).</p> <p>Critically deficient: Uncertainties in the exposure characterization are identified and there is reasonable certainty that the results are largely attributable to factors other than exposure to the chemical of interest (e.g., identified impurities are expected to be a primary driver of the results).</p>
<p>Endpoint measurement</p> <p>Are the selected protocols, procedures, and test systems adequately described and appropriate for evaluating the endpoint(s) of interest?</p> <p><i>Notes:</i></p> <p><i>Considerations related to adjustments or corrections to endpoint measurements are addressed under results presentation.</i></p> <p><i>Considerations related to the sensitivity of the animal model and timing of endpoint measurement are evaluated under sensitivity.</i></p>	<p>For each endpoint or grouping of endpoints in a study:</p> <p>Are the evaluation methods and test systems adequately described and appropriate?</p> <p>Are there concerns regarding the methodology selected (e.g., accepted guidelines, established criteria) for endpoint evaluation?</p> <p>Are there concerns about the specificity of the experimental design? Did the study address features inherent to the test system or experiment that have the potential to lead to bias away from the null?</p> <p>Are there serious concerns about the number of replicates or sample size in the study?</p> <p>Are appropriate control groups for the study/assay type included? Was there a need for the assay to include specific controls to reduce potential sources of underlying bias?</p> <p>Did the test compound induce cytotoxicity (known, or expected based on other studies of similar design) to a degree expected to affect interpretation of results?</p>	<p>Considerations for this domain are highly variable depending on the assay or endpoint(s) of interest and must be refined by assessment teams.</p> <p>A judgment and rationale for this domain should be given for each assay or endpoint or group of endpoints investigated in the study.</p> <p>Some considerations include the following:</p> <p>Good:</p> <ul style="list-style-type: none"> • Adequate description of methods and test system. • Use of generally accepted and reliable endpoint methods that are consistent with accepted guidelines or established criteria for the assay(s)/endpoint(s) of interest. • Sample sizes are generally considered adequate for the assay or protocol of interest and there are no notable concerns about sampling in the context of the endpoint protocol. • Includes appropriate control groups (e.g., use of loading controls) and any use of nonconcurrent or historical control data (e.g., for comparison to background levels in negative controls) is justified (e.g., authors or evaluators considered the similarity between current cell cultures and

Domain and core question	Prompting questions	General considerations
		<p>laboratory conditions to historical controls).</p> <p>Ratings of Adequate, Deficient, and Critically Deficient are generally defined as follows:</p> <p>Adequate: Issues are identified that could affect endpoint measurement but are considered unlikely to substantially impact the overall findings or the ability to reliably interpret those findings.</p> <p>Deficient: Concerns are raised that are expected to notably affect endpoint measurement and reduce the reliability of the study findings.</p> <p>Critically deficient: Severe concerns are raised about endpoint measurement and any findings are likely to be largely explained by these limitations.</p> <p>The following specific examples of relevant concerns are typically associated with a Deficient rating, but Adequate or Critically Deficient might be applied depending on the expected impact of limitations on the reliability and interpretation of the results:</p> <ul style="list-style-type: none"> • Study report lacks important details necessary to evaluate the appropriateness of the study design (e.g., description of the assays or protocols; information on the cell line, passage number). • Selection of protocols that are nonpreferred or lack specificity for investigating the endpoint of interest. This includes omission of additional experimental criteria (e.g., inclusion of a positive control or dosing up to levels causing minimal toxicity) when required by specific testing guidelines/protocols.* • Cytotoxicity is observed or expected on the basis of findings from similarly designed studies and may mask interpretation of outcome(s) of interest.

Domain and core question	Prompting questions	General considerations
		<ul style="list-style-type: none"> • Sample sizes are smaller than is generally considered adequate for the assay or protocol of interest. Inadequate sampling can also be raised within the context of the endpoint protocol (e.g., in a pathology study, bias that is introduced by only sampling a single tissue depth or an inadequate number of slides per animal)** • Controls are not included or considered inappropriate. <p>*These limitations typically also raise a concern for insensitivity.</p> <p>**Sample size alone is not a reason to conclude an individual study is critically deficient.</p>
<p>Results presentation</p> <p>Are the results presented and compared in a way that is appropriate and transparent and makes the data usable?</p>	<p>For each assay/endpoint or grouping of endpoints in a study:</p> <p>Does the level of detail allow for an informed interpretation of the results?</p> <p>If applicable, was the assay signal normalized to account for nonbiological differences across replicates and exposure groups?</p> <p>Are the data compared or presented in a way that is inappropriate or misleading (e.g., presenting western blot images without including numerical values for densitometry analysis, or vice versa)?</p> <p>Flag potentially inappropriate statistical comparisons for further review.</p>	<p>Considerations for this domain are highly variable depending on the endpoints of interest and must be refined by assessment teams.</p> <p>A judgment and rationale for this domain should be given for each assay or endpoint or group of endpoints investigated in the study.</p> <p>Some considerations include the following:</p> <p>Good:</p> <ul style="list-style-type: none"> • No concerns with how the data are presented. • Results are quantified or otherwise presented in a manner that allows for an independent consideration of the data (assessments do not rely on author interpretations). • No concerns with completeness of the results reporting.*

Domain and core question	Prompting questions	General considerations
		<p>Ratings of Adequate, Deficient, and Critically Deficient are generally defined as follows:</p> <p>Adequate: Concerns are identified that might affect results presentation but are considered unlikely to substantially impact the overall findings or the ability to reliably interpret those findings.</p> <p>Deficient: Concerns with results presentation are identified and expected to substantially impact results interpretation and reduce the reliability of the study findings.</p> <p>Critically deficient: Severe concerns about results presentation were identified and study findings are likely to be largely explained by these limitations.</p> <p>The following specific examples of relevant concerns are typically associated with a Deficient rating but Adequate or Critically Deficient might be applied depending on expected impact of limitations on the reliability and interpretation of the results:</p> <ul style="list-style-type: none"> • Nonpreferred presentation of data (e.g., averaging technical replicates rather than independent replicates). • Failure to present quantitative results. • Pooling data when responses are known or expected to differ substantially (e.g., across cell types or passage number). • Incomplete presentation of the data* (e.g., presentation of mean without variance data, concurrent control data are not presented, failure to report or address overt cytotoxicity). <p>*Failure to describe <u>any</u> findings for assessed outcomes (i.e., report lacks any qualitative or quantitative description of the results in tables, figures, or text) will result in a critically deficient rating for the outcome(s) of interest for results presentation; overall completeness of reporting at the study level is addressed under selective reporting.</p>

Domain and core question	Prompting questions	General considerations
<p>Sensitivity</p> <p>Are there concerns that sensitivity in the study is not adequate to detect an effect?</p>	<p>Was the exposure period, timing (i.e., cell passage number, insufficient culture maturity for the adequate expression of mature cell markers; insufficient treatment or measurement duration for the production of protein above the level of detection), frequency, and duration of exposure sensitive for the assay/model system of interest, particularly in the absence of a positive control?</p> <p>Assay-specific considerations regarding sensitivity, specificity, and validity of the selection of the test methods will be described here (e.g., metabolic competency, antibody specificity) (some of these external considerations might have been applied during prioritization of studies for evaluation). Are there aspects related to risk of bias domains that raise concerns about insensitivity (e.g., selection of protocols or methods that are known to be insensitive or nonspecific for the outcome(s) of interest)?</p> <p>Are there concerns regarding the need for positive controls (e.g., concerns that the effects of interest may be inhibited or otherwise poorly manifest in the test system, for example due to differences from in vivo biology)? If used, was the selected positive test substance (and dose) reasonable and appropriate and was the intended positive response induced?</p>	<p>Are there concerns regarding the need for positive controls (e.g., concerns that the effects of interest might be inhibited or otherwise poorly manifest in the test system, e.g., due to differences from in vivo biology)? If used, was the selected positive test substance (and dose) reasonable and appropriate and was the intended positive response induced?</p> <p>Considerations for this domain are highly variable depending on the specific assay/model system used or endpoint(s) of interest and must be refined by assessment teams. Some study design features that affect study sensitivity might have already been included in the other evaluation domains; these should be noted in this domain, along with any features that have not been addressed elsewhere.</p> <p>Some considerations include:</p> <p>Good</p> <ul style="list-style-type: none"> • The experimental design (considering exposure period, timing, frequency, and duration) is appropriate and sensitive for evaluating the outcome(s) of interest. • The selected test system is appropriate and sensitive for evaluating the outcome(s) of interest (e.g., cell line/cell type is appropriate and routinely used for the selected assay). • No significant concerns with the ability of the experimental design to detect the specific outcome(s) of interest. (e.g., study designed to address known endpoint variability that is unrelated to treatment, such as doubling time or confluency). • Timing of endpoint measurement in relation to the chemical exposure is appropriate and sensitive (e.g., cultures adequately express mature cell markers).

Domain and core question	Prompting questions	General considerations
		<ul style="list-style-type: none"> Potential sources of bias toward the null are not a substantial concern. <p>Adequate</p> <ul style="list-style-type: none"> Potential issues are identified and related to the considerations described for <i>Good</i> that could reduce sensitivity, but they are unlikely to impact the overall findings of the study. <p>Deficient</p> <ul style="list-style-type: none"> Concerns were raised about the considerations described for <i>Good</i> that are expected to notably decrease the sensitivity of the study to detect a response in the exposed group(s). <p>Critically deficient</p> <ul style="list-style-type: none"> Severe concerns were raised about the sensitivity of the study and experimental design such that any observed associations are likely explained by bias. The rationale should indicate the specific concern(s).
<p>Overall confidence</p> <p>Considering the identified strengths and limitations, what is the overall confidence rating for the assay(s) or endpoint(s) of interest?</p> <p><i>Note:</i></p> <p><i>Reviewers should mark studies for additional consideration during evidence synthesis if, due to low sensitivity only (i.e., bias toward the null), these studies are rated as lower than high confidence. If the study is otherwise well conducted and an effect is observed, the confidence may be increased.</i></p>	<p>For each assay or endpoint or grouping of endpoints in a study:</p> <ul style="list-style-type: none"> Were concerns (i.e., limitations or uncertainties) related to the risk of bias or sensitivity identified? If yes, what is their expected impact on the overall interpretation of the reliability and validity of the study results, including (when possible) interpretations of impacts on the magnitude or direction of the reported effects? 	<p>The overall confidence rating considers the likely impact of the noted concerns (i.e., limitations or uncertainties) in reporting, bias, and sensitivity on the results.</p> <p>A confidence rating and rationale should be given for each assay or endpoint, or group of endpoints investigated in the study. Confidence rating definitions are described above (see Section 4.1).</p>

4.6. EVALUATION OF EXISTING COMPUTATIONAL PHYSIOLOGICALLY BASED PHARMACOKINETIC MODELS

For a specific target organ/tissue, it might be possible to employ or adapt an existing PBPK model, develop a new PBPK model, or develop an alternative quantitative approach to use instead of a PBPK model (e.g., a classical PK model or other empirical use of dosimetry data). A useful source of information is EPA's *Approaches for the Application of Physiologically Based Pharmacokinetic (PBPK) Models and Supporting Data in Risk Assessment* ([U.S. EPA, 2006a](#)). Here, the identification and evaluation of PK data will be necessary. These data could come from studies with animals or humans and might be in vitro or in vivo in design. It should be recognized that chemicals produce multiple toxicities, through different MOAs, which could vary by lifestage ([U.S. EPA, 2006b](#)), and with different dose-response functions. If data are available from studies evaluating susceptible lifestages (e.g., in utero/pregnant women, lactating women, growing child, adolescent), it should be considered as part of a PBPK model that reflects the ADME differences that could affect dose. It is recommended that ADME information be interpreted in the context of single effects first, then evaluated as a body of information when applicable (e.g., in instances where dose-response functions for multiple and apparently independent adverse effects are similar in the low-dose region).

When a quantitative understanding of ADME leads to the development of PBPK models or other quantitative approaches for animals and humans (e.g., classical PK model), summaries of ADME studies will require a slightly higher level of detail than when these approaches are not used. Important points about computational models from EPA's *A Review of the Reference Dose and Reference Concentration Processes* ([U.S. EPA, 2002b](#)) for noncancer assessment apply equally to PBPK model use for cancer assessments, including

- The use of a PBPK model provides the optimal approach for extrapolating from one exposure-duration response situation to another.
- A chemical -specific PBPK model parameterized for the species and regions (e.g., respiratory tract) involved in the toxicity is the preferred option for calculating a human equivalent exposure (oral dose or human equivalent dose [HED] or inhalation concentration or human equivalent concentration [HEC]).

Given these preferences, it follows that sound justification should be provided for *not* using a PBPK (or classical PK) model when an applicable one exists and no equal or better alternative for dosimetric extrapolation is available. It should also be noted, however, that these preferences only apply to models that faithfully represent current scientific knowledge and accurately translate the science into computational code in a reproducible, transparent manner. In practice, it has been found that many published models have errors of varying degrees of impact on their predictions; hence, an evaluation of a model is required before it can be accepted for use in an assessment. Typically, the review process includes contacting the authors of the model for the source code to

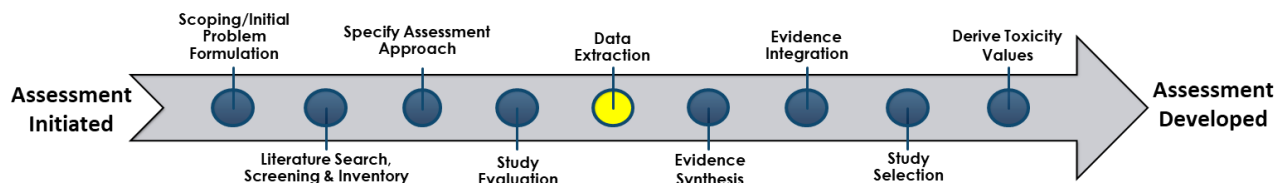
review and modifying the model to correct any errors ([U.S. EPA, 2018e](#)). There are also cases where one must choose among several different models, which a formal evaluation can facilitate.

Considerations for judging the suitability of a model are separated into two categories: scientific and technical. In summary, the scientific criteria focus on whether the biology, chemistry, and other information available for chemical MOA(s) are appropriately represented by the model structure and equations. Significant to the overall efficiency of this process, the scientific criteria can be judged by reading the publication or report that describes the model and do not require evaluation of the computer code. Preliminary technical criteria include availability of the computer code and apparent completeness of parameter listing and documentation. The in-depth technical and scientific criteria focus on the accurate implementation of the conceptual model in the computational code, use of correct or biologically consistent parameters in the model, and reproducibility of model results reported in journal publications and other documents. Additional details are provided in *An Umbrella Quality Assurance Project Plan for PBPK Models* ([U.S. EPA, 2018e](#)) and in the protocol template.

If no PBPK model exists or the existing PBPK models are determined technically or scientifically inadequate, EPA will evaluate the cost and effort of developing or significantly revising a PBPK model against the potential value of such a model, compared to standard methods of extrapolation [e.g., body-weight scaling to the $3/4$ power ($BW^{3/4}$) scaling ([U.S. EPA, 2011a](#))]. For example, PBPK models have a high potential to impact an assessment where there are significant nonlinearities in the exposure-dose relationship in the range of interest, animal and human metabolic data significantly differ from $BW^{3/4}$ scaling, or data exist to quantify human variability via PBPK modeling. These cases all depend on availability of the data necessary to support model development or revision. These are not exclusive or strict criteria because they are highly dependent on chemical-specific scientific and technical factors and resource considerations.

This approach stresses: (1) clarity in the documentation of model purpose, structure, and biological characterization; (2) validation of mathematical descriptions, parameter values, and computer implementation; and (3) evaluation of each plausible dose metric. Such transparency and documentation are important for compliance with the Agency's information quality guidelines ([U.S. EPA, 2002a](#)). The critical points and model evaluation criteria characterized by the World Health Organization (WHO)/International Programme on Chemical Safety (IPCS) ([IPCS, 2010](#)) are largely mirrored in the present EPA draft criteria as described in [U.S. EPA \(2018e\)](#). In addition to providing transparency through documentation, the process will confirm objectivity and scientific rigor.

5. EXTRACTION AND DISPLAY OF STUDY RESULTS FROM EPIDEMIOLOGICAL AND TOXICOLOGICAL STUDIES



Purpose

- To extract data to support evidence synthesis and integration and normalize the data in digital formats that support data visual presentation, digital dissemination, and reuse.

This chapter is intended to provide a general framework for extraction of data to support evidence synthesis and integration through development of visual representations of the evidence base. At this stage in assessment development, health outcomes have generally been prioritized as explained in the systematic review protocol, and studies might have also undergone evaluation. The next task is to extract study information and results, then organize this information into tables, graphs, and integrative constructs, which facilitates evaluation of results and relevant features (e.g., exposure range, study duration and design) across studies and the collective interpretation of those results. Although examples are provided, this chapter is not intended to establish strict “rules” for developing tables or graphical summaries of information, recognizing that a single presentation format will not work well for all sets of studies. Additionally, data visualization formats are evolving rapidly in the Integrated Risk Information System (IRIS) Program with increasing use of interactive web-based interfaces in HAWC. The most effective graphical depictions of the data being used in the assessment should be utilized, regardless of whether that has been explicitly mentioned in this document.

The decision to extract a study is made when developing the systematic review protocol (see Chapter 3). Not all studies that meet the assessment populations, exposures, comparators, and outcomes (PECO) criteria will necessarily undergo data extraction. Studies considered *uninformative* during study evaluation might be summarized in the synthesis section to highlight data gaps but would not undergo data extraction. These studies are relatively rare and have critical flaws that make the findings uninterpretable. In large evidence bases with an abundance of *medium* and *high* confidence studies, assessment teams might decide not to extract data from *low*

confidence studies. The decision whether to extract a study is considered in the context of the available information and (e.g., if there is a large number of *high* confidence studies assessing the long-term effects of exposure, the most informative studies will be prioritized for data extraction; whereas, if the literature inventory is limited to a few studies, all available studies might be extracted). Outcomes or study designs determined to be less informative for dose-response and toxicity value derivation might not go through full detailed data extraction (e.g., acute or short-term studies, single dose studies, studies with confounding exposures that were not adequately controlled) but are considered in the overall assessment of evidence. The direction of effect does not influence the decision on whether to extract the study (i.e., null results should be extracted). Supplemental materials considered important to cite in the assessment typically do not undergo the same level of extraction as studies that meet the PECO criteria; most commonly these studies are described in narrative or tabular format.

The steps of study evaluation and data extraction might not always be strictly sequential and co-occur, especially for animal toxicological studies. When it becomes clear during study evaluation that a study is likely to be extracted (e.g., it is sufficiently well reported, and no major issues are identified from an initial scan of the methods), it might be more efficient to extract the data during study evaluation because much of the extracted data extraction directly informs the study evaluation judgments.

5.1. DATA EXTRACTION

Data extraction is one of the most time-intensive stages of conducting a systematic review and should be approached strategically. Assessment teams should plan on one person for data extraction and another person for quality assurance of the extracted information. This typically requires 1–4 hours per study, depending on the complexity of the study, experimental design, and endpoints evaluated. Further, extraction time increases substantially if information is not numerical (presented in figures such as bar/line graphs that must be digitized) compared with tables. Presentation of results should be designed to be inclusive of all informative study results regardless of the direction or magnitude of individual effect estimates; however, the level of data extraction might vary across endpoints. For example, data extraction decisions include consideration of whether information for dose-response needs to be extracted versus a summary level description of key dose levels (e.g., doses associated with specific magnitudes of effect) versus a narrative summary. Detailed extraction of information at the level of effect size is generally pursued for key study findings. Assessment teams should consider the utility of extracting other contextual findings (e.g., null biochemical findings in an animal study with apical results), repeated measures designs, or health outcomes where findings across studies are mostly null and, therefore, not likely to be a primary focus for developing toxicity values. Efficient data extraction could require some knowledge of what and how information is presented in the set of studies to help make decisions on the extent of data extraction appropriate for a given health outcome/endpoint. In some cases,

attempts will be made to obtain missing information from human and animal health effect studies (e.g., if the missing information is considered influential during study evaluations or is required to conduct an additional analysis).

The IRIS Program commonly uses the U.S. Environmental Protection Agency (EPA) version of Health Assessment Workspace Collaborative (HAWC) (<https://hawcprd.epa.gov/portal/>) for structured data extraction of epidemiological and animal toxicological studies. Extracting into HAWC allows users access to the data (which are fully downloadable as an MS Excel spreadsheet) and the IRIS Program to create visuals and tables from the extracted data. The structured extraction entities are consistent across assessments and facilitate reusing data across assessments and conducting updates. The visuals that can be created in HAWC are interactive, and users can hover and “click to see more” information presented and linked to content made available in the visual displays. Note that the visual functionality within HAWC is intended to facilitate data aggregation and dissemination of the information (computationally). In addition, files created outside of HAWC for data extraction purposes can be imported into HAWC to create visuals, but these visuals will not be interactive, that is, they will not have the “click to see more” functionality that requires direct extraction into HAWC. The visualization features of HAWC (including the data set overview dashboards) also make it easier to identify and present patterns of findings that support evidence synthesis, the integration of findings, and overall conclusions (see Chapter 6). Examples of the types of standard visualizations that can be generated in HAWC are provided in Section 5.5.

Currently, HAWC is best suited for graphical displays of health outcome data. Tabular or narrative presentations or summarization of nonhealth outcome content (e.g., absorption, distribution, metabolism, and excretion [ADME]/pharmacokinetic) is best pursued using other approaches, including Microsoft Word, Excel, or customized DistillerSR forms. Although in vitro studies can be extracted into HAWC, in many cases a detailed extraction in HAWC is a greater level of effort than needed to summarize in vitro or other types of mechanistic evidence, and other approaches should be considered (i.e., narrative, tabular, or graphical presentation based on Word, Excel, or DistillerSR customized forms, and Tableau visualization software). In addition to HAWC, R-based graphical scripts developed for use with other software tools (e.g., GraphPad Prism) also might be useful.

5.1.1. Health Assessment Workspace Collaborative (HAWC)

Instructions for summarizing specific data extraction elements are described within the HAWC extraction modules. A list of data extraction fields for animal bioassay, epidemiological, and in vitro studies in Excel is available at the public HAWC website [see “[IRIS PPRTV SEM Template Figures and Resources](#) (2021),” then “Downloads”]. In addition to fields used for collecting information on study design and results, extraction fields are available to gather other information such as funding source, conflicts of interest, details on any author correspondence, and documentation on use of digitization tools (used to extract information from figures/graphs). A frequently asked questions document “HAWC FAQs for assessment readers” is also available in

“Downloads” in the “IRIS PPRTV SEM Template Figures and Resources (2021)” project to help readers of an assessment learn how to access HAWC content. This document can be referenced as an assessment appendix or directly through use of this publicly accessible HAWC URL (<https://hawc.epa.gov/assessment/100000039/>).

Certain aspects of data extraction can be done independently by support staff who are familiar with HAWC (e.g., contractors, student interns). Further, assessment managers can access the HAWC management dashboard to assign tasks and manage quality assurance (QA)/quality control (QC). Activities most amenable to delegation include uploading studies into HAWC and extracting summary-level study design information and methods for animal toxicological studies. However, extraction of results and creation of graphics by support staff should be done under close supervision by the assessment team. Typically, epidemiological studies are more difficult to summarize and extract than animal toxicological studies because of greater heterogeneity in study designs and reporting. Any delegation of extraction for epidemiological studies should be done under close supervision by epidemiologists on the assessment team.

The selection and level of detail of individual findings to be extracted from each study are dependent on author reporting and the needs of the assessment. When large amounts of quantitative analyses are presented in a published study, decisions to select the most informative effect estimates might be needed. Considerable heterogeneity in study designs and presentation of results can be expected among the studies included in the review. Some types of analysis common across studies (e.g., “ever” exposed compared with “never” exposed) might not be as informative as a more comprehensive analysis (e.g., analyses considering level of exposure) developed in only one or two studies.

5.1.2. Quality Control during Data Extraction

Data extraction is a laborious process even when conducted using specialized software such as HAWC. The following approaches can be used to promote high quality and consistent data extraction.

- Plan for a training period to orient new staff to the extraction process. Ideally, new staff should do a pilot extraction of one study with review by someone experienced in data extraction/HAWC, followed by extraction of another two or three studies with an additional round of review.
- Create tables and visualizations early in the process to help QA/QC the extraction and aid the evidence synthesis process.
- Ensure the extraction of study information into HAWC or other applications is complete and accurate at initial entry because it can be used as a template for adding additional experiments and results for a given study. Any errors or incompleteness in the initial extraction can proliferate and be time intensive to adjust.

- For consistent outcome/endpoint extraction, use the suggested terminology in the “Environmental Health Vocabulary” database available in HAWC (<https://hawc.epa.gov/vocab/ehv/>). The suggested terminology can be applied directly from within HAWC or used as a reference material for other extraction tools. This terminology not only promotes consistency across assessments but also interoperability with other databases (e.g., ToxRefDB, Chemical Effects in Biological Systems [CEBS], Organisation for Economic Co-operation and Development [OECD] Harmonised Templates, and other ontologies) and coded using the Unified Medical Language System (UMLS; <https://www.nlm.nih.gov/research/umls/>).
- Terminology for other data fields in HAWC are typically controlled using picklists and the picklist term or abbreviation should be used to control for ambiguity and redundancy. Note that the data clean-up feature can be used to quickly check for consistency across terminology extracted into HAWC.
- Use digitizing software applications to estimate numbers from graphs, such as Grab It! (<http://www.datatrendsoftware.com/instructions.html>), WebPlotDigitizer (<https://automeris.io/WebPlotDigitizer/>), or Universal Desktop Ruler (<https://avpsoft.com/products/udruler/>). Document when values are estimated (i.e., in HAWC, check the box “values estimated” in the results extraction module).
- Have at least one member of the assessment team review the entire extraction. Following verification, the assessment should be “locked” to prevent inadvertent modifications.
- Frequently monitor the consistency of extraction across studies, including consistency of the extracted data in visuals, bulk data extraction clean-up, etc.
- Use the management dashboard to track QA/QC.

5.1.3. Best Practices for Data Extraction in Health Assessment Workspace Collaborative (HAWC) and Tabular Presentation

Although instructions for summarizing specific data extraction elements are described within the HAWC extraction modules, some general best practices for data extraction and presentation might be useful when using HAWC and other tools. These tips are summarized below.

General Tips for Data Extraction in Health Assessment Workspace Collaborative (HAWC)

- In HAWC, the extraction comment box in the “Study Details” module can be used to summarize endpoint extraction decisions. For example, “Extraction” focused on fertility and malformation findings might result in general observations for dams (bodyweight gain, feed consumption, liver weight) not being fully extracted. Findings for these outcomes from an existing data extraction are shown below as examples (quoted text indicates the text was taken from the published report):
 - “During the first few days of exposure, a slight decrease in body weight gain was observed among the dams exposed to chloroform from Days 6–15 of gestation. Body weight gain was significantly reduced among the mice in the Days 1–7 or 8–15 groups. Slightly less food and water were consumed by each experimental group as compared during the first few days of exposure by controls.” As no other details were provided

and these observations were not being considered for dose-response analysis, no attempt was made to fully extract these data.

- “The absolute and relative liver weights were significantly increased among the pregnant mice exposed to chloroform from Days 6 through 15 or from Days 8 through 15 of gestation. A similar effect was not discerned among the dams exposed from Days 1 through 7 of gestation. This pattern of liver weight changes also was observed among bred mice that were not pregnant at sacrifice.” As these results were not deemed exposure-related, the data for these observations were not extracted.
- In the event dose-response data are not fully extracted, a user can “dummy code” the endpoint to generate exposure-response array figures that display the direction of effect. This can be especially useful when authors desire to relay a treatment-related effect. Dummy coding is not a significant resource saving step when effect size information is presented in tabular format. To develop figures for animal studies in HAWC, coding can be used to generate graphs with symbols that indicate direction of effect (control and no effect findings can be coded as “0” to graph a ●; treatment-related increases coded as “1” to graph a ▲; and treatment-related decreases coded as “-1” to graph a ▼). When this approach is used, it should be indicated as a caption in the HAWC figure and annotated as a result note in the “Endpoint Module.”
- The assessment team should consider contacting authors when effect size and variability information in a study are presented extensively in figures. The request need not be for the underlying individual participant/animal data; even obtaining the summary information presented in the figure can make the data extraction process less time intensive and more accurate.

Time course measurements can be difficult to extract, especially when the information is presented in figures and values must be estimated. Several strategies can be considered depending on the content being presented and whether the result is a primary endpoint of interest or a peripheral finding. In some cases, presenting the difference between the initial and final time point might be reasonable. Animal studies of learning can be especially challenging to summarize because they often include repeated measurements, and judgments need to be made on whether a difference score or other measure, such as number of trials to achieve the learning goal, represents the best summary. In other cases, a representative value might be summarized for effect size purposes and a figure note used to indicate that a similar response was observed at the other time points measured. Alternatively, the time point with a significant finding might be summarized and a figure note used to indicate that no significant findings were observed at the other time points. A digital measurement approach can also be used to extract the information as area under the curve, although this process can be laborious and can transform the unit of measure in a manner that is confusing compared to how the information was presented in the study. When complete extraction is required for time course information, use of a tabular presentation or seeking copywrite permission to reproduce the original figure might be more appropriate.

- Many of the data extraction fields in HAWC include a controlled picklist of terms or abbreviations for extracting author-reported terms into HAWC. The data extractor should use these picklists to control for data ambiguity and redundancy.

Epidemiological Evidence

- When available, adjusted statistical estimates should be extracted rather than unadjusted or raw estimates. In some cases, it might be desirable to extract both the adjusted and unadjusted estimates (e.g., to show the difference that adjustment for confounding made on the estimates).
- When several different group numbers are reported in studies (e.g., total participants, numbers included in a specific analysis), study size should reflect the number of participants in the primary analysis of interest.
- Description of the population could include demographic characteristics and important potential confounders relevant to the endpoint of concern (e.g., percentage of males, mean age, percentage of smokers), as relevant for interpretation of the results.
- Exposure estimate format will vary according to the study; where applicable, it is helpful to have some measure of both the average [such as median (preferred) or mean] and range [such as interquartile range (preferred) or standard deviation]. If this information is not available, whatever information is available on exposure levels should be extracted.
- Include a summary of the study evaluation and the overall study confidence conclusion (see Chapter 4).
- For studies and outcomes prioritized for data extraction, results should be extracted regardless of statistical significance. When available, there should be some indication of the uncertainty in the result (e.g., 95% confidence interval [CI]), and it might be informative to include the number of individuals (e.g., cases by exposure level, exposure level by case status) that contributed to each displayed effect estimate. In some cases, multiple results of the same exposure and outcome might be reported (e.g., sensitivity analyses such as a different exposure categorization or exclusion of specific participants). Not every similar result needs to be extracted; the extractor should use their judgment to avoid extracting duplicative results that are unlikely to add to the interpretation of the findings.
- If multiple exposure measures are provided (e.g., cumulative and peak exposure), all could be presented in the table or selected metric(s) might be presented with a note that multiple metrics were considered, as well as a summary of similarities and differences between them. At a minimum, extracting the most relevant/highest quality exposure measure should be done, along with others that might be informative.
- If few or no quantitative results are reported, a qualitative description of results could be provided using brief sentences or phrases. Also note instances where quantitative results were not reported (e.g., “Authors state no differences between groups; quantitative results not reported”).

Animal Evidence

- When present, the organization of the information in the “Reference and study design” column is flexible but should include the key information about the study design (e.g., study confidence, species, duration, age/lifestage, route) but should be consistent as possible both within and across tables.
- Include a summary of the study evaluation and the overall confidence conclusion (see Chapter 4). These summary level tables can be prepared in HAWC from study evaluation summaries and included in tabular format in HAWC for import into Microsoft Word.
- Exposure levels should be extracted as common units depending on exposure route (e.g., mg/kg-day oral or mg/m³ inhalation) and be reported in the results column in line with the row of results corresponding to that animal group. If it was necessary to convert the reported exposures to a common metric, the converted numbers should be provided in parentheses or a footnote with sufficient information to replicate the conversion (including references). When available, study specific information will be used to make the conversions; however, EPA defaults can also be used ([U.S. EPA, 1988](#)). Assumptions used in performing dose conversions will be documented.
- Results presented in the table should be those reported by the study authors (e.g., mean, and standard deviation [SD] or standard error [SE], or incidence and number at risk), including all exposed groups and the control. In addition, outcome measures should be transformed to a common metric to help assess related outcomes measured with different scales (discussed in Section 5.2). The evidence tables should specify how the data were transformed (e.g., absolute difference in means, normalized mean difference [NMD], percentage of change from control) including the formula that was used as a footnote. Qualitative results should be included as a brief sentence or phrase; note also that quantitative results were not reported. For example: “Treatment-related histopathological changes were reported to be absent; quantitative results were not reported.”

Mechanistic Evidence

- The mechanistic studies tagged as supplemental material are typically categorized by the biological focus of the available information. Importantly, this categorization is typically done based on title/abstract (TIAB) content only. Full-text retrieval for supplemental material is not typically done unless chapter leads determine the available evidence could impact assessment conclusions. Extraction of the supplemental mechanistic evidence (tagged during TIAB screening) should initially consist of high-level tagging (categorization) by health outcome (for example) using screening software and two screeners. After health outcome tagging, the supplemental mechanistic studies can be assigned for full text review by health outcome to chapter leads. Note that the review of the available supplemental mechanistic studies should be informed by expert input. At a minimum, extraction includes core information (i.e., model system, endpoint evaluated, experimental design, other expert tailored content) describing evaluated endpoint, measurement method, and any dose-response information (see Section 2.5.2).
- Outside of HAWC, mechanistic information can be captured using the Adverse Outcome Pathway (AOP) integrative construct. Data extraction is performed using the AOPWiki where data are extracted using structured data extraction pages and fields. These AOP data

are stored in the AOP knowledge base (AOPKB). Data from the AOPKB can be accessed and analyzed using various third-party tools. Additional information on AOP development and documentation is available at the AOPWiki [website](#).

5.2. STANDARDIZING REPORTING OF OUTCOME MEASURES

Designations of treatment-related findings could differ from study to study, thereby contributing to inconsistent bases for comparing and integrating evidence. For example, no-observed-adverse-effect levels (NOAELs) and lowest-observed-adverse-effect levels (LOAELs) are commonly used by study authors to summarize, interpret, and make conclusions from study findings. However, the NOAEL/LOAEL approach is less suitable than presenting effect size-based measures to evaluate consistency across studies, summarize findings, and interpret those findings.¹⁰ Further, the treatment-related findings extracted from studies and presented in IRIS assessment text, tables, graphs, etc. represent all information identified during assessment development and all information is made available (digitally and without interpretation). This standardization process facilitates transparency and tracking of the interpretation of that information supporting EPA conclusions during assessment development.

In addition to providing quantitative outcomes in their original units for all study groups, results from outcome measures are transformed, when possible, to a common metric to help compare distinct but related outcomes measured with different measurement scales. These standardized effect size estimates facilitate systematic evaluation and evidence integration for hazard identification and whether meta-analysis is feasible for an assessment (see Section 7.2.1). The following summary of effect size metrics by data type outlines issues in selecting the most appropriate common metric for a collection of related endpoints ([Vesterinen et al., 2014](#)). Note that it is important to consider the variability associated with effect size estimates, with stronger studies generally showing more precise estimates. Effect size estimation can be affected, however, by such factors as variances that differ substantially across treatment groups or by a lack of information to characterize variance, especially for animal studies in biomedical research ([Vesterinen et al., 2014](#)).

Common metrics for continuous outcomes:

- *Absolute difference in means.* This metric is the difference between the means in the control and treatment groups, expressed in the units in which the outcome is measured. When the outcome measure and its scale are the same across all studies, this approach is the simplest to implement and analyze.

¹⁰EPA's reference dose/reference concentration (RfD/RfC) review ([U.S. EPA, 2002b](#)) emphasizes balancing statistical and biological significance in identifying NOAELs and LOAELs. Inconsistency in published NOAEL and LOAEL values results largely from reliance only on statistical significance, which varies with different statistical tests between study authors and with different study designs and sizes. See EPA's *Benchmark Dose Technical Guidance* ([U.S. EPA, 2012b](#)) for other limitations of NOAELs and LOAELs.

- *Percentage of control response or NMD.* This metric is the difference between control and treatment means divided by the control mean, expressed as a percentage. Note that some outcomes reported as percentages, such as mean percentage of affected offspring per litter, can lead to distorted effect sizes when further characterized as a percentage of change from control. Such measures are better expressed as absolute difference in means or are preferably transformed to incidences using approaches for event or incidence data (see below).
- *Standardized mean difference.* This metric is the difference between control and treatment means divided by the estimated standard deviation among individual experimental units. The standard deviation is often based on the pooled variance for controls and treated units. Pooling variances can be problematic if variances differ substantially, in which case it might be preferable to standardize using the standard deviation of controls. This metric converts all outcome measures to a standardized scale with units of standard deviations. This approach can also be applied to data using different units of measurement (e.g., different measures of lesion size such as infarct volume and infarct area).

Common metrics for event or incidence data:

- *Absolute difference in proportions or percentages.* This metric can be used to estimate a population-wide increase, assuming the study population was similar to the population for which the extrapolation is made.
- *Percentage of change from control.* This metric is analogous to the NMD approach described for continuous data above. Note the warning for the NMD approach above; this metric might be inappropriate for summary measures expressed in terms of percentages. For example, a 50% decrease (halving) from control might be viewed differently when the control percentage is 2% versus 20%. Also note that a control percentage of zero leads to an undefined percentage change; 0% can readily occur when the control incidence probability is small relative to sample size.
- *Extra risk.* Often used for defining toxicity values, this metric is the difference between control and treatment proportions or percentages responding, divided by the control level not responding.
- *Odds ratio.* For binary outcomes, such as the number of individuals that developed a disease or died, and with only one treatment evaluated, data can be represented in a 2×2 table. Note that when the value in any cell is zero, 0.5 is added to each cell to avoid problems with the computation of the standard error. For each comparison, the odds ratio (OR) and its standard error should be calculated. Odds ratios are normally combined on a logarithmic scale. Some outcome measures are polytomous, having $k > 2$ outcomes (usually ordinal, such as severity ranks), leading to a $2 \times k$ table at each dose. The metrics above can be applied to each control-treated comparison in a $2 \times k$ table, resulting in k 2×2 metrics at each dose. One simplifying approach is to reduce the $2 \times k$ table to a 2×2 table (e.g., severity rank ≤ 3 and > 3). Statisticians and subject matter experts might suggest other approaches for reducing a $2 \times k$ table to a single metric.

5.3. STANDARDIZING ADMINISTERED DOSE LEVELS/CONCENTRATIONS

Exposures are standardized to common units when appropriate. Exposure levels in oral studies are expressed in units of mg/kg-day. When study authors provide exposure levels in concentrations in the diet or drinking water, dose conversions can be made by EPA using study-specific food or water consumption rates and body weights when available. Otherwise, EPA defaults will be used ([U.S. EPA, 1988](#)) when addressing age and study duration as relevant for the species/strain and sex of the animal of interest. Exposure levels in inhalation studies will be expressed in units of mg/m³. Assumptions used in performing dose conversions will be documented. Administered doses for animal studies can be presented in multiple dose metrics in HAWC by adding new dose representations, although the calculations are not automatic. Instead, the conversions are done outside of HAWC and are manually entered. For metals and other chemicals (e.g., salts such as potassium nitrate or sodium fluoride) that exist in various chemical forms, exposure levels will typically be converted to chemical equivalents.

5.4. GENERAL PRINCIPLES FOR PRESENTING EVIDENCE

Each type of data presentation should be constructed in a manner that clearly conveys the key information to the reader. Tabular or graphical formats should be used to present study summaries, and narrative text should focus on evidence synthesis observations. Although the specific organization and level of detail can vary, as much consistency as possible should be maintained across tables and graphics with similar purposes. This includes nomenclature (e.g., abbreviations, units, grouping, sorting criteria) and structural choices (e.g., types of information in columns and rows, axes, symbols). Contextual information provided by peripheral analysis in a study or from supplemental material is often not extracted and might be described only in narrative form or table/graph notes.

There might be some results for an outcome that are more commonly reported across multiple studies, which could be presented graphically to evaluate consistency within the unit of analysis. Additional analyses (e.g., summary measures, trend tests) could add value to the analysis when deciding the set of effect estimates and results to present in tables and text. The ordering of information should be used to tell the story of the evidence, as opposed to being organized alphabetically. For example, depending on the nature of the evidence, the tables might be organized by study confidence, study design/exposure duration, species/population, or lowest tested exposure level. Sort orders often involve nested schemes (e.g., sorting by outcome such as motor activity, then by endpoint such as horizontal activity or rearing). Regardless of how the information is organized in the tables and graphics, a thorough QA check to ensure all the relevant details are either included in the table/figure or are properly cross-referenced elsewhere in the document (preferably with hyperlinks).

5.5. GRAPHICAL AND TABULAR DISPLAY

The use of arrays and other types of graphical representations (of raw data and analyses of those data) is a foundation of hazard identification and is also used in dose-response analysis. Several graphical formats are routinely used in assessments, notably exposure-response arrays that show direction of effects at a given dose level in animal toxicological studies, exposure- or dose-response graphs, and forest plots for epidemiological studies. The display of data facilitates identification of patterns of response associated with chemical exposure and can aid in those evaluations and help identify data gaps ([Woodall and Goldberg, 2008](#)). To the extent possible, the presentations should incorporate study evaluation judgments and information that facilitates consistent judging of the biological significance of the effects seen across studies, including effect sizes (e.g., magnitude of effect relative to a control level) or benchmark doses (BMDs) corresponding to 10% responses.

The following sections discuss and provide examples for both graphical and tabular display (see Figures 5-1 through 5-5). HAWC figures can be downloaded as PowerPoint, PDF, or Scalable Vector Graphics (SVG) files. HAWC images can be exported as SVG files for further editing using applications such as Inkscape (<https://inkscape.org/en/>), a commonly used free application.

An additional aspect important to consider in the development of visualizations is the presentation of outcome-specific confidence in a study based on study evaluation. There are multiple ways to present this information, including sorting studies by confidence level or using color-coding and a legend.

5.5.1. GRAPHICAL DISPLAY

Dose-Response Graphs

One of the most basic concepts in toxicology is the principle of dose-response. A commonly used graphic demonstrating this principle is the dose-response curve. Most simply, a dose-response curve is plotted on an x-y graph showing the level of the causative agent (drug, chemical, radiation, temperature, etc.) on the x-axis versus the response level plotted on the y-axis. Responses can be measured as counts of an effect in a population or test group (e.g., incidence), categories of the severity of an effect (e.g., pathological gradations of a lesion), or continuous measurements (e.g., blood pressure). The direction of a response might be an increase (e.g., higher incidence) or a decrease (e.g., decrease in body-weight gain compared to a control group). The scale of the axes can distort the shape of the dose-response curve, however, and should be considered carefully ([Lutz et al., 2005](#)). When either doses or responses range across two or more orders of magnitude and a large amount of those data is in the lower range, the differences between values will tend to get lost in a linear arithmetic graphic. Use of a log scale on one or both axes of the graph (semi-log or log-log, respectively) is useful in response to skewness, i.e., when one or a few data points are much larger than a bulk of the data, or to show percent change or multiplicative factors. Log scaling spreads out the graphical presentation of those data sets (so that the values are not all clustered

around the low end of the scale) and allows a more critical visual inspection for any trends. As mentioned in [Lutz et al. \(2005\)](#), care needs to be taken to ensure data are not misrepresented by misapplication of log scaling.

In the example shown in Figure 5-1, in which the information being displayed is for a single study, a notation of the study confidence should be included in the caption for the figure. In examples for which data are displayed for multiple studies (as in Figure 5-2), data for higher confidence studies should be emphasized in the graphic. Examples for doing so are to add an indicator line as a demarcation of where study confidence changes [Figure 5-2(a) and (b)] or add the line to the legend to indicate the quality of the studies as a parenthetical [Figure 5-2(c)]. When confidence ratings within a study vary by outcome, those indicators of confidence should be outcome specific. Another potential consideration in results display is the biological significance of the measure, which might be relevant in addition to an indicator of statistical significance. Biological significance is loosely interpreted to reflect the judgment that the observed level of effect is likely to impair the organism’s function or ability to respond to additional challenge (or is consistent with steps in an established mode-of-action [MOA]). Thus, a consideration related to this interpretation is the historical range of effect responses established across a large number of animals of the same species, strain, and sex. As an example, when the “historical range” of a response is not similar to the control group response, the “historical range” for the measure can be added as a band overlaid with the range of the responses observed in the study.

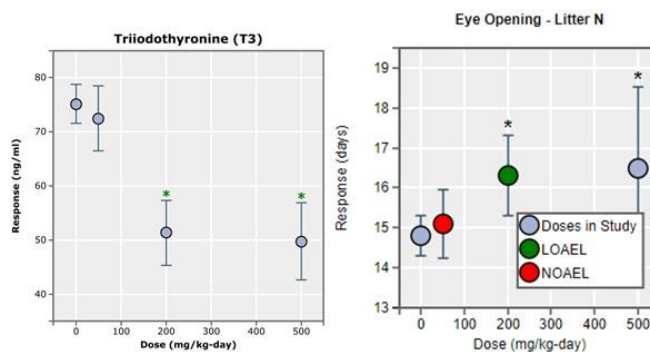


Figure 5-1. Examples of dose-response graphical displays for single endpoint created in Health Assessment Workspace Collaborative (HAWC) (for illustrative purposes only).

BMDS = Benchmark Dose Software; LOAEL = lowest-observed-adverse-effect level; NOAEL = no-observed-adverse-effect level.

The above visualizations are automatically created in HAWC for animal data when effect size information is added in the results extraction module. Within HAWC, the scale can be adjusted (linear, logarithmic) and the image downloaded. Dose-response displays can also be created using software applications such as BMDS, Excel, GraphPad Prism, or SAS.

The examples are available at: <https://hawcprd.epa.gov/ani/endpoint/100002336/> and <https://hawcprd.epa.gov/ani/endpoint/99902179/>. The standard figure in HAWC includes a LOAEL/NOAEL legend. The legend can be removed, data point color(s) adjusted, and further edited by downloading the image as an SVG file. Inkscape (<https://inkscape.org/en/>) is a commonly used free application for editing SVG files.

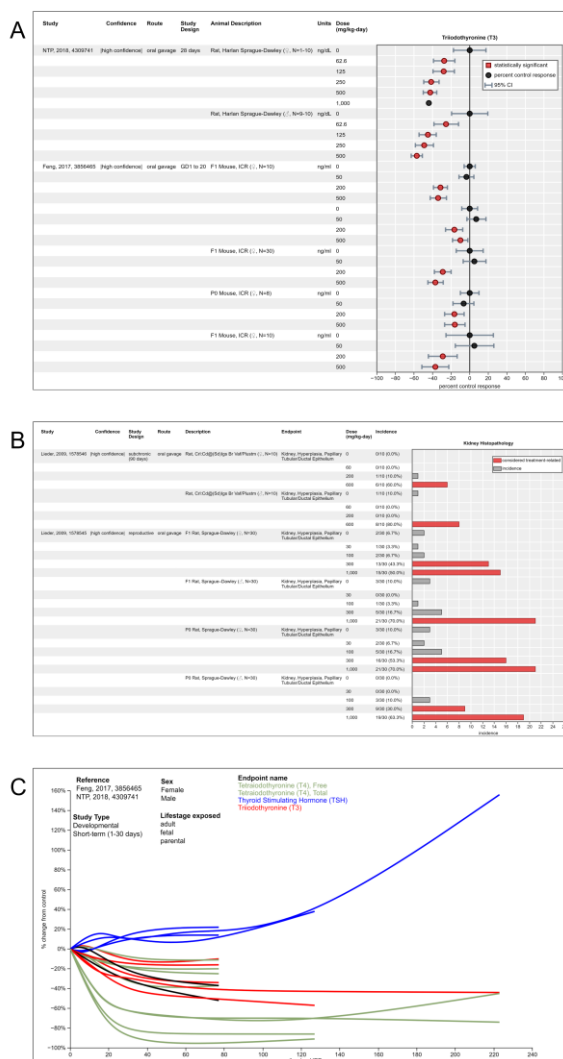


Figure 5-2. Examples of dose-response graphical displays across endpoints and studies created in Health Assessment Workspace Collaborative (HAWC) (for illustrative purposes only). (a) Data pivot (continuous variable); (b) data pivot (dichotomous variable); (c) animal bioassay endpoint cross view with detailed pop-out of a single study.

T3 = triiodothyronine.

These images can be created in HAWC for animal data using the “data pivot” visualization option when effect size information has been extracted. Within HAWC, many options are available for customizing the content (e.g., column text content, sort order, selection of endpoints, use of color and shapes). Instructions for creating visuals in HAWC are available in the training videos (see “About”). The HAWC Crossview plot can also be used to show dose-response relationships across endpoints with options to select specific studies, e.g., based on study evaluation judgments, sex, species, lifestage. In addition, new figures can be created by selecting the “copy from existing” option and adjusting the endpoint content as needed.

HAWC currently does not have meta-analysis capabilities; if meta-analysis is needed, the extracted data should be imported into other software, such as programs in R or CatReg, for analysis and visualization.

The examples are available at: <https://hawcprd.epa.gov/summary/data-pivot/assessment/10000037/pfbs-estrous-cyclicty-effect-size-animal/>; <https://hawcprd.epa.gov/summary/data-pivot/assessment/10000039/pfbs-kidney-histopathology-effect-size-animal/>; and <https://hawcprd.epa.gov/summary/visual/10000087/>.

Forest Plots

Forest plots can be used to present summary results from meta-analyses or to display results from a set of citations evaluating the same endpoints. The latter application is commonly used during hazard identification to facilitate analyses of patterns of associations across citations. In these scenarios, citations in the same plot might not have used the same exposure metrics or outcome measures, and therefore quantitative comparisons of the magnitude of the associations between studies are not appropriate.

Increasingly, forest plot displays are applied to animal studies to present effect size information for each studied dose level, rather than just those with statistical or biological significance, e.g., NOAEL or LOAEL dose levels. A forest plot can be a useful display of consistency (or heterogeneity) of results and can be used to examine sources of heterogeneity [i.e., differences in populations, exposure measures, ranges of exposures, or potential biases ([White et al., 2013](#))].

When applied to epidemiological data, forest plots typically array multiple point estimates of the effects of a specific exposure with a specific health endpoint (e.g., relative risks, odds ratios, hazard ratios) and their associated CIs (e.g., 95% CI) represented by lines from the lower bound of the CI to the upper bound with the point estimate clearly identified (see Figure 5-3). Additional details (e.g., design, numbers of cases, specific exposure metric, study confidence evaluation) can be annotated as needed to describe the available data transparently. A reference line is typically plotted at the value consistent with the null hypothesis (i.e., no association; for relative effect measures, the reference line is at unity, e.g., relative risk = 1). The natural log or logarithmic scale is used for ratio measures to retain symmetry between the ratio and its inverse. In cases for which additive effect measures or linear regression coefficients are being compared, the reference line is plotted at zero (0) and the standard linear scale is used for the effect measure. If the forest plot was generated to display the results of a meta-analysis and calculation of a summary effect measure across multiple studies, the size of the symbols for each study will vary according to the weight (often determined by the variance of the effect estimate) contributed to the summary estimate by each study.

For animal evidence, outcome measures presented in forest plot displays should be transformed to a common metric to help assess related outcomes that are measured with different scales. The graph should specify how the data were transformed (e.g., percentage change from control, absolute difference in means, normalized mean difference).

The ability to incorporate confidence ratings (if needed) is more limited for some types of forest plots than others. When results are primarily organized by the outcomes and secondarily by the study [Figure 5-3(a)], a column can be added with study confidence rating or a notation can be added to another column [e.g., as a part of the study identifier as to the confidence rating for that study (e.g., L, M, or H), with inclusion of a definition for those indicators in the caption]. When a

figure is organized by study first [Figure 5-3(b)], ordering the studies from top to bottom by study confidence with labeled lines as demarcations of where confidence changes is a possibility.

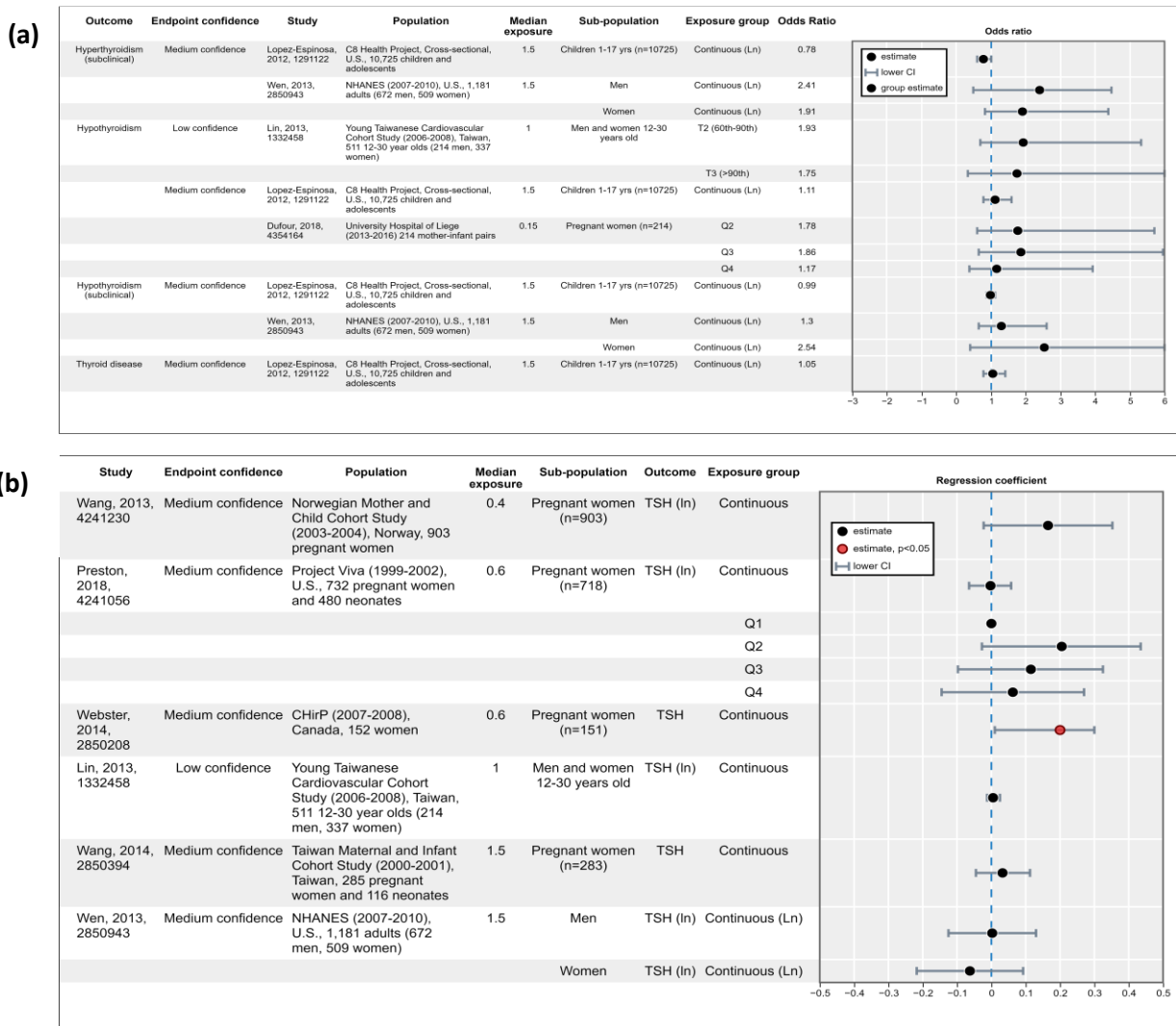


Figure 5-3. Examples of forest plots used for epidemiological evidence (for illustrative purposes only). (a) Health Assessment Workspace Collaborative (HAWC) forest plot (odds ratio, null of 1), all medium confidence studies; (b) HAWC forest plot (regression coefficient, null of 0), all medium confidence studies.

CI = confidence interval; T3 = triiodothyronine; TSH = thyroid stimulating hormone; Q = quartile; T = tertile.

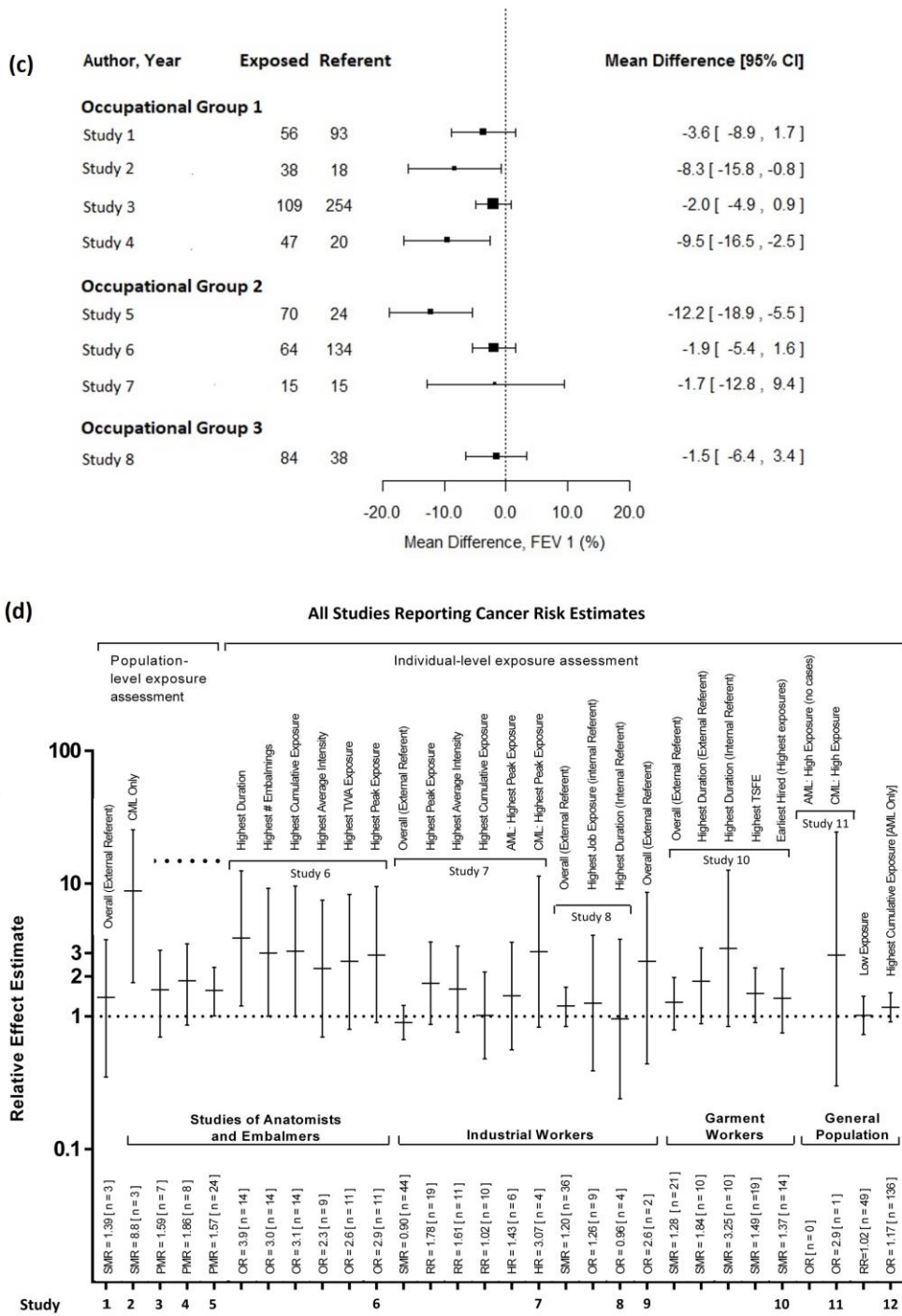


Figure 5-4. Examples of forest plots used for epidemiological evidence (for illustrative purpose only). (c) R forest plot (mean difference, null of 0); (d) GraphPad Prism Forest plot (null of 1).

CI = confidence interval; FEV = forced expiratory volume.

Forest plots for individual results are automatically created in HAWC when effect size information is added in the results extraction module. (a) and (b) can be created in HAWC using the data pivot visualization option to display

multiple findings in one study or across studies. In HAWC, forest plots can be developed using the data pivot visualization option for results presented on a null of 1 (e.g., odds ratio) or null of 0 (e.g., regression coefficients) but studies with different null lines cannot be combined in the same graphic. HAWC currently does not have meta-analysis capabilities; if meta-analysis is needed, the extracted data should be imported into other software, such as R, for analysis and visualization, as shown in (c).

The examples in HAWC are available at: <https://hawcprd.epa.gov/summary/data-pivot/assessment/100000026/pfna-and-thyroid-disease/> and <https://hawcprd.epa.gov/summary/data-pivot/assessment/100000026/example-forest-plot/> (currently limited to IRIS staff).

Exposure-Response Arrays

Exposure-response arrays are visual representations of health effect data most often derived from experimental or clinical observations. In an array, each line represents the exposure range for a single study-endpoint combination. Study information represented on each line can include the following.

- All exposures to which the test subjects were exposed.
- Indications of judgments on statistical/biological significance.

Exposure-response arrays differ from dose-response graphs in allowing comparisons across multiple studies, several types of effects, and other characteristics of the health effect data. The principal limitation of arrays is that they do not effectively convey the magnitude of the response at any given exposure.

Information in an array can be organized to illustrate patterns or differences in response associated with exposure duration, toxicity endpoint (including those of different severity), species, sex, or lifestage ([Woodall and Goldberg, 2008](#)). Study confidence should be incorporated using the same techniques as described for other graphic formats discussed in this section. Figure 5-5(a) includes confidence ratings as a part of the figure. Several stylistic and formatting conventions have been adopted in the development of exposure-response arrays and are described in [Woodall \(2014\)](#); these are also likely to be applicable to other types of graphical depictions of data.



Figure 5-5. Examples of exposure-response arrays. (a) Health Assessment Workspace Collaborative (HAWC) exposure- response array for animal studies. (b) HAWC exposure-response array for in vitro studies.

GD = gestation day.

Exposure-response arrays can be created in HAWC for animal data using the “data pivot” visualization option. When effect size information has been extracted, directional symbols (e.g., up and down triangles) can be used to show direction of the effect using conditional formatting options. Within HAWC, many options are available for customizing the content (e.g., column text content, sort order, selection of endpoints, use of color and shapes). Instructions for creating visuals in HAWC are available in the training videos (see “About”). In addition, new figures can be created by selecting the “copy from existing” option and adjusting the endpoint content as needed. Conditional formatting in the data pivot is used to apply the colors and shapes. To implement, make sure the “AND” button is checked under settings > data filtering and ordering tab. If conditional formatting is set to “base,” it will be applied if the condition is true. If conditional formatting is set to “---,” no changes will be applied. Generally, “---” should be used.

The examples in HAWC are available at: <https://hawcprd.epa.gov/summary/data-pivot/assessment/10000039/pfbs-thyroid-effects/> and <https://hawcprd.epa.gov/summary/data-pivot/assessment/10000039/estrogen-receptor-reporter-gene-assays/>.

5.5.2. TABULAR DISPLAY

Although graphical displays (e.g., exposure-response arrays) provide a visual snapshot of available data in a form easily digested by readers, inclusion of all clarifying or explanatory details in the graphic might not be possible and would unnecessarily clutter the display. Tables can be used as standalone depictions of evidence or can accompany an array to provide critical ancillary information, such as additional description of the studies and endpoints. In addition, in some cases,

data are less amenable to graphical illustrations. For example, when there is inconsistency in the effect estimates, units, or other factors across studies being reviewed, a tabular summary might be the most appropriate way to present the data. Tables 5-1 and 5-2 show example summary tables for epidemiological and animal studies, respectively.

General Tips for Tabular Presentation

- Callouts to footnotes move from left to right, top to bottom. Use the scheme (a, b, c) for general footnote callouts. Occasionally, a table style imported from applications other than Microsoft Word can get corrupted when imported into Microsoft Word and cause issues with in-text citations using Health and Environmental Research Online (HERO). In this case, the corrupted style definition should be replaced. Contact the information management team for document support if needed.
- Tables can be formatted using either portrait or landscape orientation. In general, portrait orientation is easier to read, but landscape orientation could be needed if additional columns (e.g., more detailed study design or results information) are presented.
- Examples of tables for epidemiological and animal toxicological studies are shown below. Space constraints, and the most effective communication of key aspects of the data being presented, will affect the ultimate format and content of the table. The amount of detail and information presented should be customized to the assessment needs.

Table 5-1. Example epidemiological summary table of selected data on exposure antibody response to vaccines in children

Reference, N, confidence	Exposure timing and concentration in serum ^a	Outcome measure timing	Diphtheria vaccine β (95%) ^b	Tetanus vaccine β (95%) ^b
Grandjean et al. (2012) , N = 380–537, Medium	Maternal; mean (IQR): 0.6 (0.5–0.8) ng/mL	Children (age 5), prebooster	-14.8 (-31.2, 5.5)	11.2 (-8.6, 35.1)
		Children (age 5), postbooster	-12.9 (-26.7, 3.5)	-3.7 (-23.1, 20.7)
		Children (age 7)	-5.1 (-24.4, 19.2)	22.1 (-4.2, 55.5)
	Children (age 5); mean (IQR): 1.0 (0.8–1.2) ng/mL	Children (age 5), prebooster	-17.7 (-33.0, 1.1)	-5.9 (-21.8, 13.4)
		Children (age 5), postbooster	-16.1 (-28.8, -1.0)	-18.2 (-34.0, 1.4)
		Children (age 7)	-17.1 (-32.8, 2.2)	-17.4 (-34.1, 3.6)
	Children (age 7); mean (IQR): 1.1 (0.9–1.5) ng/mL	Children (age 13)	-11.3 (-27.4, 8.5)	31.0 (-2.7, 76.4)
Children (age 13); mean (IQR): 0.7 (0.6–0.9) ng/mL	Children (age 13)	-4.5 (-24.2, 20.2)	15.2 (-16.9, 59.7)	
Grandjean et al. (2017) , ^c N = 349, Medium	At birth, not reported	Children (age 5), prebooster	4.79 (-18.21 to 34.27)	-7.11 (-26.59, 17.53)
	Infant (18 m); median (IQR): 1.0 (0.6–1.5) ng/mL	Children (age 5), prebooster	2007–2009 cohort 24.43 (5.72, 46.45)	2007–2009 cohort -6.98 (-21.10, 9.67)

ORD Staff Handbook for Developing IRIS Assessments

Reference, N, confidence	Exposure timing and concentration in serum ^a	Outcome measure timing	Diphtheria vaccine β (95%) ^b	Tetanus vaccine β (95%) ^b
			1997–2000 cohort -35.28 (-64.95, 19.48)	1997–2000 cohort -33.79 (-64.36, 23.01)
	Children (age 5); median (IQR): 1.1 (0.8–1.6) ng/mL	Children (age 5), prebooster	-8.85 (-23.95, 9.25)	-10.31 (-24.39, 6.40)
Granum et al. (2013) , N = 49 Medium	Maternal 0–3 d post-delivery; median (IQR): 0.3 (0.2–0.4) ng/mL	Children (age 3)	n/a	-0.01 (-0.41, 0.39)
			Measles vaccine β (95%)^a	Rubella vaccine β (95%)^a
Granum et al. (2013) , N = 50 Medium	Maternal 0–3 d post-delivery; median (IQR): 0.3 (0.2–0.4) ng/mL	Children (age 3)	-0.55 (-1.51 to 0.41)	-1.38 (-2.35 to -0.40)
Stein et al. (2016) N = 1101–1190, Medium	Children (age 12–19); mean: 0.8 ng/mL	Children (age 12–19)	1.1 (-11.8 to 15.9) (seropositive)	0.6 (-6.7 to 8.5) (seropositive)
			Hib vaccine β (95%)^a	Mumps vaccine β (95%)^a
Granum et al. (2013) , N = 50, Medium	Maternal 0–3 d post-delivery; median: 0.3 ng/mL	Children (age 3)	4.9 (-10.7 to 20.5)	n/a
Stein et al. (2016) N = 1101–1190, Medium	Children (age 12–19); mean: 2.5 ng/mL	Children (age 12–19)	n/a	-2.7 (-8.4 to 3.4) (seropositive)

IQR = interquartile range; N = number.

Bold font indicates $p < 0.05$.

^aExposure timing is organized into groups based on maternal exposure, childhood exposure (including from birth through age 13), and adult exposure. Linear regression (β or % change in antibody per 2-fold increase of PFNA).

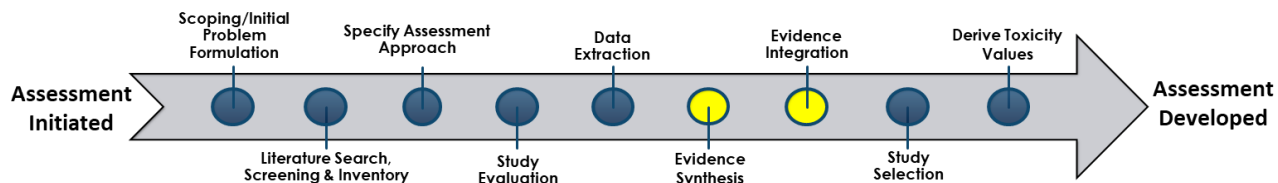
^bNumbers in parentheses are 95% confidence intervals.

^cResults for Faroe Islands Cohort 5 (2007–2009) unless otherwise stated.

Table 5-2. Example animal summary table showing percent change of liver weight

Study Design and Reference	Dose (mg/kg-d)															
	2.5	5	10	15	20	30	35	50	62.5	100	125	175	200	250	500	1,000
28-d female rat Study 1									1		2			7	15*	47*
28-d male rat Study 1									8		7			14*	32*	64*
90-d female rat Study 2			4					6					5			
90-d male rat Study 2			1					1					22*			
90-d female rat Study 3					-1					5						37*
90-d male rat Study 3					0					11						63*

6. EVIDENCE SYNTHESIS AND INTEGRATION



Purpose

- To interpret the evidence and make judgments on the overall potential that a substance can be hazardous to humans: first, through within-stream evidence synthesis judgments, and second, through evidence integration judgments across evidence streams.

Evidence synthesis¹¹ is a within-stream analysis, conducted separately for human, animal, and mechanistic evidence. Findings from human and animal evidence for each unit of analysis are separately judged to reach an expression of certainty in the evidence for a hazard (*robust, moderate, slight, indeterminate, or compelling evidence of no effect*). Within-stream evidence synthesis conclusions directly inform the integration across the evidence streams to draw overall conclusions for each of the assessed health effect categories (*evidence demonstrates, evidence indicates (likely), evidence suggests, evidence inadequate, or strong evidence supports no effect*). A structured framework approach is used to guide both evidence synthesis and integration. Although there are circumstances where specific mechanistic evidence (typically biological precursors) is included in the unit of analysis for human or animal evidence synthesis (see Chapter 3), in most cases mechanistic findings are presented separately from the human and animal evidence and used to inform conclusions on (1) the coherence, directness of outcome measures, and biological significance of findings within the animal or human evidence streams during evidence synthesis and, (2) evidence integration judgments on the human relevance of findings in animals, coherence across evidence streams (“cross-stream coherence”), information on susceptible populations or lifestages, understanding of biological plausibility and mode-of-action (MOA), and possibly other critical inferences (e.g., read-across analyses). The structured framework also accommodates consideration of other supplemental information [e.g., ADME (absorption, distribution, metabolism, and excretion), non-PECO (populations, exposures, comparators, and outcomes) routes of exposure] that can inform evidence synthesis and integration judgments.

¹¹The phrases “evidence synthesis” and “evidence integration” used here are analogous to the phrases “strength of evidence” and “weight of evidence,” respectively, used in some other assessment processes ([EFSA, 2017](#); [U.S. EPA, 2017a](#); [NRC, 2014](#); [U.S. EPA, 2005a](#)).

- *Evidence synthesis*: A summary of findings and judgment(s) regarding the certainty in the evidence for hazard for each unit of analysis from the human and animal studies are made in parallel, but separately. A unit of analysis is an outcome or group of related outcomes within a health effect category that are considered together during evidence synthesis. These judgments can incorporate mechanistic and other supplemental evidence when the unit of analysis is defined as such (see Chapter 3). The units of analysis can also include or be framed to focus on precursor events (e.g., biomarkers). In addition, this can include an evaluation of coherence across units of analysis within an evidence stream. At this stage, the animal evidence judgment(s) does not yet consider the human relevance of that evidence.
- *Evidence integration*: The animal and human evidence judgments are combined to draw an overall evidence integration judgment(s) that incorporates inferences drawn on the basis of information on the human relevance of the animal evidence, coherence across evidence streams, potential susceptibility, understanding of biological plausibility and MOA and other critical inferences informed by mechanistic, ADME, or other supplemental data.

Evidence synthesis and integration judgments are expressed both narratively in the assessment and summarized in tabular format in evidence profile tables (see Table 6-1). Key findings and analyses of mechanistic and other supplemental content are also summarized in narrative and tabular format to inform evidence synthesis and integration judgments (see Table 6-2). In brief, after synthesis a certainty in the evidence judgment is drawn for each unit of analysis summarized as *robust*, *moderate*, *slight*, *indeterminate*, or *compelling evidence of no effect* (see Section 6.1). Next, these judgments are used to inform evidence integration judgments summarized as ***evidence demonstrates***, ***evidence indicates (likely)***, ***evidence suggests***, ***evidence inadequate***, or ***strong evidence supports no effect*** (see Section 6.2). These summary judgments are included as part of the evidence synthesis and integration narratives. When multiple units of analysis are synthesized, the main evidence integration judgments typically focus on the unit of analysis with the strongest evidence synthesis judgments, although exceptions can occur.¹² Health outcomes or endpoints for which the unit of analysis is considered to present *slight*, *indeterminate*, or *compelling evidence of no effect* can inform the evidence integration hazard judgment but would typically not be used as the basis for deriving a toxicity value. Structured evidence profile tables are

¹²In some cases, as discussed in Section 6.2, it might be appropriate to draw multiple evidence integration judgments within a given health effect category. This is generally dependent on data availability (i.e., more narrowly defined categories may be possible with more evidence) and the ability to integrate the different evidence streams at the level of these more granular categories. More granular categories will generally be organized by predefined manifestations of potential toxicity. For example, within the health effect category of immune effects, separate and different evidence integration judgments might be appropriate for immunosuppression, immunostimulation, and sensitization and allergic response [i.e., the three types of immunotoxicity described in the WHO guidance (2012)]. Likewise, within the category of developmental effects, it may be appropriate to draw separate judgments for potential effects on fetal death, structural abnormality, altered growth, and functional deficits (i.e., the four manifestations of developmental toxicity described in EPA guidelines (U.S. EPA, 1991b)). These separate judgments are particularly important when the evidence supports that the different manifestations might be based on different toxicological mechanisms. As described for the evidence synthesis judgments, the strongest evidence integration judgment will typically be used to reflect certainty in the broader health effect category.

used to summarize these analyses and foster consistency within and across assessments. Instructions for using HAWC to create these tables are available at the HAWC project "[IRIS PPRTV SEM Template Figures and Resources](#) (2021)" (see "Attachments," then select the "Creating Evidence Profile Tables in HAWC"). A repository of examples is also available (see "Example Evidence Synthesis and Integration Scenarios" attachment).

Table 6-1. Generalized evidence profile table showing the relationship between evidence synthesis and evidence integration to reach a judgment of certainty in the evidence for hazard

Evidence Synthesis Judgments (note that many factors and judgments require elaboration or evidence-based justification)					Evidence integration (weight of evidence) judgment(s)
Studies	Summary of key findings	Factors that increase certainty (Applied to each unit of analysis)	Factors that decrease certainty (Applied to each unit of analysis)	Evidence synthesis judgment(s)	Describe overall evidence integration judgment(s): ⊕⊕⊕ <i>Evidence demonstrates</i> ⊕⊕⊖ <i>Evidence indicates</i> (likely) ⊕⊖⊖ <i>Evidence suggests</i> ⊖⊖⊖ <i>Evidence inadequate</i> --- <i>Strong evidence supports no effect</i> Highlight the primary supporting evidence for each integration judgment ^a Present inferences and conclusions on:
Evidence from human studies					
Unit of analysis #1 Studies considered and study confidence	Description of the primary results	<ul style="list-style-type: none"> All/Mostly <i>medium</i> or <i>high</i> confidence studies Consistency Dose-response gradient Large or concerning magnitude of effect Coherence* 	<ul style="list-style-type: none"> All/Mostly <i>low</i> confidence studies Unexplained inconsistency Imprecision Concerns about biological significance^a Indirect outcome measures^a Lack of expected coherence^a 	Judgment reached for each unit of analysis^a ⊕⊕⊕ <i>Robust</i> ⊕⊕⊖ <i>Moderate</i> ⊕⊖⊖ <i>Slight</i> ⊖⊖⊖ <i>Indeterminate</i> --- <i>Compelling evidence of no effect</i>	
Evidence from animal studies					
Unit of analysis #1 Studies considered and study confidence	Description of the primary results	<ul style="list-style-type: none"> All/Mostly <i>medium</i> or <i>high</i> confidence studies Consistency Dose-response gradient Large or concerning magnitude of effect Coherence^a 	<ul style="list-style-type: none"> All/Mostly <i>low</i> confidence studies Unexplained inconsistency Imprecision Concerns about biological significance^a Indirect outcome measures^a Lack of expected coherence^a 	Judgment reached for each unit of analysis ⊕⊕⊕ <i>Robust</i> ⊕⊕⊖ <i>Moderate</i> ⊕⊖⊖ <i>Slight</i> ⊖⊖⊖ <i>Indeterminate</i> --- <i>Compelling evidence of no effect</i>	

MOA = mode of action.

^aCan be informed by key findings from the mechanistic analyses (see Table 6-2).

Table 6-2. Generalized evidence profile table to show the key findings and supporting rationale from mechanistic analyses

Mechanistic analyses		
Biological events or pathways (or other relevant evidence grouping)	Summary of key findings and interpretation	Judgment(s) and rationale
<p><u>Different analyses can be presented separately, e.g., by exposure route or key uncertainty addressed</u></p> <p><u>Each analysis can include multiple rows separated by biological events or other feature of the approach used for the analysis</u></p> <ul style="list-style-type: none"> • Generally, will cite mechanistic synthesis (e.g., for references, for detailed analysis) • Does not have to be chemical-specific (e.g., read-across) 	<p><u>Can include separate summaries, for example by study type (e.g., new approach methods vs. in vivo biomarkers), dose, or design</u></p> <p><i>Interpretation:</i> Summary of expert interpretation for the body of evidence and supporting rationale</p> <p><i>Key findings:</i> Summary of findings across the body of evidence (can focus on or emphasize highly informative designs or findings), including key sources of uncertainty or identified limitations of the study designs tested (e.g., regarding the biological event or pathway being examined)</p>	<p>Overall summary of expert interpretation across the assessed set of biological events, potential mechanisms of toxicity, or other analysis approach (e.g., AOP)</p> <ul style="list-style-type: none"> • Includes the primary evidence supporting the interpretation(s) • Describes and informs the extent to which the evidence influences inferences across evidence streams • Characterizes the limitations of the evaluation and highlights existing data gaps • May have overlap with factors summarized for other streams

AOP = adverse outcome pathway.

6.1. EVIDENCE SYNTHESIS

Integrated Risk Information System (IRIS) assessments synthesize the evidence separately for each unit of analysis by focusing on factors that increase or decrease certainty in the reported findings as evidence for hazard. These factors are adapted from considerations for causality introduced by Austin Bradford Hill ([Hill, 1965](#)) with some expansion and adaptation of how they are applied to facilitate transparent application to chemical assessments that consider multiple streams of evidence. Specifically, the factors considered are confidence in study findings (risk of bias [RoB] and sensitivity), consistency across studies or experiments, dose/exposure-response gradient, strength (effect magnitude) of the association, directness of outcome or endpoint measures, and coherence [Table 6-3; see additional discussion in [U.S. EPA \(2005a\)](#), [U.S. EPA \(1994\)](#), and [U.S. EPA \(2020b\)](#)]. These factors are similar to the domains considered in the GRADE (Grading of Recommendations Assessment, Development, and Evaluation) Quality of Evidence framework ([Schünemann et al., 2013](#)). Each of the considered factors and the certainty of evidence judgments requires elaboration or evidence-based justification in the synthesis narrative. Analysis of evidence synthesis considerations is qualitative (i.e., numerical scores are not developed, summed, or subtracted).

Biological understanding (e.g., knowledge of how an effect manifests or progresses) or mechanistic inference (e.g., dependency on a conserved key event across outcomes) can be used to define which related outcomes are considered as a unit of analysis. The units of analysis can also include predefined categories of mechanistic evidence (typically precursor events). When mechanistic evidence is included in the units of analysis, it is evaluated against all evidence synthesis factors. Mechanistic and other supplemental evidence not included in the units of analysis can be analyzed to inform select evidence synthesis factors (i.e., coherence, directness of outcome measures, or biological significance) within the animal and human evidence synthesis. Additional mechanistic evaluations (e.g., biological plausibility) are considered as part of across-stream evidence integration (see Section 6.2).

Five levels of certainty in the evidence for a hazard are used to summarize evidence synthesis judgments: *robust* ($\oplus\oplus\oplus$, very little uncertainty exists); *moderate* ($\oplus\oplus\ominus$, some uncertainty exists); *slight* ($\oplus\ominus\ominus$, large uncertainty exists); *indeterminate* ($\ominus\ominus\ominus$); or *compelling evidence of no effect* (- - -, little to no uncertainty exists for lack of hazard) (see Tables 6-4 and 6-5 for descriptions). Conceptually, before the evidence synthesis framework is applied, certainty in the evidence is neutral (i.e., functionally equivalent to *indeterminate*). Next, the level of certainty regarding the evidence for (or against) hazard is increased or decreased depending on interpretations using the factors described in Table 6.3, noting that these analyses are conducted for each unit of analysis within an evidence stream. Evidence factors that increase certainty include having an evidence base that exhibits a signal of an effect on the health outcome based on evaluation of consistency across studies or experiments, the presence of a dose or exposure-

response gradient, observing a large or concerning magnitude of effect, and coherent findings for closely related endpoints (can include mechanistic endpoints). These patterns are more compelling when observed among *high* or *medium* confidence studies. Evidence factors that decrease certainty include having an evidence base of mostly *low* confidence studies, unexplained inconsistency, imprecision, concerns about biological significance, indirect measures of outcomes, and lack of expected coherence. Study sensitivity considerations can be expressed as a factor that can either increase or decrease certainty in the evidence, depending on whether an association is observed. An evidence base of mostly null findings for which insensitivity is a serious concern decreases certainty that the evidence is sufficient to support a lack of health effect or association. Conversely, there may be an increase in the evidence certainty in cases for which an association is observed, although the expected impact of study sensitivity is toward the null.

Table 6-3. Considerations that inform evidence synthesis judgments of the certainty in the animal or human evidence for hazard for each unit of analysis

Consideration	Increased evidence certainty (of the human or animal evidence for hazard ^a)	Decreased evidence certainty (of the human or animal evidence for hazard ^a)
Risk of bias and sensitivity (across studies)	<ul style="list-style-type: none"> • An evidence base of mostly (or all) <i>high</i> or <i>medium</i> confidence studies is interpreted as being only minimally affected by bias and insensitivity. • This factor should not be used if no other factors would increase or decrease the confidence for a given unit of analysis. • In addition, consideration of risk of bias and sensitivity should inform how other factors are evaluated, i.e., can inconsistency be potentially explained by variation in confidence judgments? 	<ul style="list-style-type: none"> • An evidence base of mostly (or all) <i>low</i> confidence studies decreases certainty. An exception to this is an evidence base of studies in which the issues resulting in <i>low</i> confidence are related to insensitivity. This might increase evidence certainty in cases where an association is identified because the expected impact of study insensitivity is toward the null. • An evidence base of mostly null findings where insensitivity is a serious concern decreases certainty that the evidence is sufficient to support a lack of health effect or association. • Decisions to increase certainty for other considerations in this table should generally not be made if there are serious concerns for risk of bias.
Consistency	<ul style="list-style-type: none"> • Similarity of findings for a given outcome (e.g., of a similar direction) across independent studies or experiments, especially when <i>medium</i> or <i>high</i> confidence, increases certainty. The increase in certainty is larger when consistency is observed across populations (e.g., geographical location) or exposure scenarios in human studies, and across laboratories, species, or exposure scenarios (e.g., route; timing) in animal studies. When seemingly inconsistent findings are identified, patterns should be further analyzed to discern if the inconsistencies can potentially be explained based on study confidence, dose or exposure levels, population, or experimental model differences, etc. This factor is typically given the most attention during evidence synthesis. 	<ul style="list-style-type: none"> • Unexplained inconsistency [i.e., conflicting evidence; see U.S. EPA (2005a)] decreases certainty. Generally, certainty should not be decreased if discrepant findings can be reasonably explained by considerations such as study confidence conclusions (including sensitivity); variation in population or species, sex, or lifestage (including understanding of differences in pharmacokinetics); or exposure patterns (e.g., intermittent versus continuous), levels (<i>low</i> versus <i>high</i>), or duration. Similar to current recommendations in the Cochrane Handbook [Higgins et al. (2022b), see Section 7.8.6], clear conflicts of interest related to funding source can be considered as a factor to explain apparent inconsistency. For small evidence bases, it might be hard to assess consistency. An evidence base of a single or a few studies where consistency cannot be accurately assessed does not, alone, increase or decrease evidence certainty. Similarly, a reasonable explanation for inconsistency does not necessarily result in an increase in evidence certainty.

Consideration	Increased evidence certainty (of the human or animal evidence for hazard ^a)	Decreased evidence certainty (of the human or animal evidence for hazard ^a)
Effect magnitude and imprecision	<ul style="list-style-type: none"> Evidence of a large or concerning magnitude of effect can increase certainty (generally only when observed in <i>medium</i> or <i>high</i> confidence studies). Judgments on effect magnitude and imprecision consider the rarity and severity of the effect. 	<ul style="list-style-type: none"> Certainty could be decreased if the findings are considered not likely to be biologically significant. Effects that are small in magnitude might not be considered to be biologically significant (adverse^b) based on information such as historical responses and variability. However, effects that appear to be of small magnitude could be meaningful at the population level (e.g., IQ shifts); in such cases, certainty would not be decreased. Certainty might also be decreased for imprecision, particularly if there are only a few studies available to evaluate consistency in effect magnitude across studies.
Dose-response	<ul style="list-style-type: none"> Evidence of dose-response or exposure-response in <i>high</i> or <i>medium</i> confidence studies increases certainty. Dose-response can be demonstrated across studies or within studies, and it can be dose- or duration-dependent. It could also not be a monotonic dose-response (monotonicity should not necessarily be expected as different outcomes might be expected at low vs. high doses or long vs. short durations due to factors such as activation of different mechanistic pathways, systemic toxicity at high doses, or tolerance/acclimation). Sometimes, grouping studies by level of exposure is helpful to identify the dose-response pattern. Decreases in a response (e.g., symptoms of current asthma) after a documented cessation of exposure also might increase certainty in a relationship between exposure and outcome (this is primarily applicable to epidemiological studies because of their observational nature). 	<ul style="list-style-type: none"> A lack of dose-response when expected on the basis of biological understanding can decrease certainty in the evidence. If the data are not adequate to evaluate a dose-response pattern, however, certainty is neither increased nor decreased. In some cases, duration-dependent patterns in the dose-response can decrease evidence certainty. Such patterns are generally only observable in experimental studies. Specifically, the magnitude of effects at a given exposure level might decrease with longer exposures (e.g., due to tolerance or acclimation). Or, effects might rapidly resolve under certain experimental conditions (e.g., reversibility after removal of exposure). As many reversible and short-lived effects can be of high concern, decisions about whether such patterns decrease evidence certainty depend on considering the pharmacokinetics of the chemical and the conditions of exposure [see U.S. EPA (1998)], endpoint severity, judgments regarding the potential for delayed or secondary effects, the underlying mechanism(s) involved, and the exposure context focus of the assessment (e.g., addressing intermittent or short-term exposures).
Directness of outcome/ endpoint measures	<ul style="list-style-type: none"> Not applicable 	<ul style="list-style-type: none"> If the evidence base primarily includes outcomes or endpoints that are indirect measures (e.g., biomarkers) of the unit of analysis, certainty (for that unit of analysis) is typically decreased. Judgments to decrease certainty based on indirectness should focus on findings for measures

Consideration	Increased evidence certainty (of the human or animal evidence for hazard ^a)	Decreased evidence certainty (of the human or animal evidence for hazard ^a)
		<p>that have an unclear linkage to an apical or clinical (adverse^b) outcome. Scenarios where the magnitude of the response is not considered to reflect a biologically meaningful level of change (i.e., biological significance; see “effect magnitude and imprecision” row, above) are not considered under indirectness of outcome measures.</p> <ul style="list-style-type: none"> • Related to indirectness, certainty in the evidence can be decreased when the findings are determined to be nonspecific to the hazard under evaluation. This consideration is generally only applicable to animal evidence and the most common example is effects only with exposures (level, duration) shown to cause excessive toxicity in that species and lifestage (including consideration of maternal toxicity in developmental evaluations). This does not apply when an effect is viewed as secondary to other changes (e.g., effects on pulmonary function because of disrupted immune responses).

Consideration	Increased evidence certainty (of the human or animal evidence for hazard ^a)	Decreased evidence certainty (of the human or animal evidence for hazard ^a)
Coherence	<ul style="list-style-type: none"> Biologically related findings within or across studies, within an organ system or across populations (e.g., sex), increase certainty (generally only when observed in <i>medium</i> or <i>high</i> confidence studies). Certainty is further increased when a temporal or dose-dependent progression of related effects is observed within or across studies, or when related findings of increasing severity are observed with increasing exposure. Coherence across findings within a unit of analysis (e.g., consistent changes in disease markers and biological precursors in exposed humans) can increase certainty in the evidence for an effect. Coherence within or across biologically related units of analysis can also increase certainty for a given (or multiple) unit(s) of analysis. This considers certainty in the biological relationships between the endpoints being compared, and the sensitivity and specificity of the measures used. Mechanistic support for, or biological understanding of, the relatedness between different endpoints within (or across different) units of analysis, can inform an understanding of coherence. 	<ul style="list-style-type: none"> An observed lack of expected coherent changes (e.g., in well-established biological relationships) within or across biologically related units of analysis typically decrease evidence certainty. This includes mechanistic changes when included in the unit of analysis. However, as described for decisions to increase certainty, confidence in the understanding of the biological relationships between the endpoints being compared, and the sensitivity and specificity of the measures used, need to be carefully examined. The decision to decrease certainty depends on the availability of evidence across multiple related endpoints for which changes would be anticipated, and it considers factors (e.g., dose and duration of exposure, strength of expected relationship) across the studies of related changes.
Other factors	<ul style="list-style-type: none"> Unusual scenarios that cannot be addressed by the considerations above, e.g., read-across inferences supporting the adversity of observed changes. 	<ul style="list-style-type: none"> Unusual scenarios that cannot be addressed by the considerations above, e.g., strong evidence of publication bias.^c

^aAlthough the focus is on identifying potential adverse human health effects (hazards) of exposure, these factors can also be used to increase or decrease certainty in the evidence supporting lack of an effect (e.g., leading to a judgment of compelling evidence of no effect). The latter application is not explicitly outlined here.

^bWithin this framework, evidence synthesis judgments reflect an interpretation of the evidence for a hazard; thus, consideration of the adversity of the findings is an explicit aspect of the analyses. To better define how adversity is evaluated, the consideration of adversity is broken into the two, sometimes related, considerations of the indirectness of the outcome measures and the interpreted biological significance of the effect magnitude.

^cPublication bias involves the influence of the direction, magnitude, or statistical significance of the results on the likelihood of a paper being published; it can result from decisions made, consciously or unconsciously, by study authors, journal reviewers, and journal editors ([Dickersin, 1990](#)). This could make the available evidence base unrepresentative. However, publication bias can be difficult to evaluate ([NTP, 2019](#)) and should not be used as a factor that decreases certainty unless there is strong evidence.

A structured framework approach is used to draw evidence synthesis judgments for human and animal evidence. Tables 6-3 and 6-4 (for human and animal evidence, respectively) provide the example-based criteria that guide how to draw the evidence synthesis judgments for each unit of analysis within a health effect category and the terms used to summarize those judgments. These terms are applied to human and animal evidence separately. The terms *robust* and *moderate* are characterizations for judgments that the evidence (across studies) supports a conclusion that the effect(s) results from the exposure being assessed. These two terms are differentiated by the quality and amount of information available to rule out alternative explanations for the results. For example, repeated observations of effects by independent studies or experiments examining various aspects of exposure or response (e.g., different exposure settings, dose levels or patterns, populations or species, biologically related endpoints) result in increased certainty in the evidence for hazard. The term *slight* indicates situations in which there is some evidence supporting an association within the evidence stream, but substantial uncertainties in the data exist to prevent judgments that the effect(s) can be reliably attributed to the exposure being assessed. *Indeterminate* reflects judgments for a wide variety of evidence scenarios, including when no studies are available or when the evidence from studies of similar confidence has a high degree of unexplained inconsistency. *Compelling evidence of no effect* represents a rare situation in which extensive evidence across a range of populations and exposures has demonstrated that no effects are likely attributable to the exposure being assessed. This category is applied at the health effect level (e.g., hepatic effects) rather than more granular units of analysis level to avoid giving the impression of confidence in lack of a health effect when aspects of potential toxicity have not been adequately examined. Reaching this judgment is infrequent because it requires both a high degree of confidence in the conduct of individual studies, including consideration of study sensitivity, as well as comprehensive assessments of outcomes and lifestages of exposure that adequately address concern for the hazard under evaluation.

Table 6-4. Framework for evidence synthesis judgments from studies in humans

Evidence synthesis judgment	Description
<p>Robust (⊕⊕⊕) ...evidence in human studies <i>(strong signal of effect with very little uncertainty)</i></p>	<p>A set of <i>high</i> or <i>medium</i> confidence independent studies (e.g., in different populations) reporting an association between the exposure and the health outcome(s), with reasonable confidence that alternative explanations, including chance, bias, and confounding, can be ruled out across studies. The set of studies is primarily consistent, with reasonable explanations when results differ; the findings are considered adverse (i.e., biologically significant and without notable concern for indirectness); and an exposure-response gradient is demonstrated. Additional supporting evidence, such as associations with biologically related endpoints in human studies (coherence) or large estimates of risk or severity of the response, can increase certainty but are not required. Supplemental evidence included in the unit of analysis (e.g., mechanistic studies in exposed humans or human cells) could raise the certainty in the evidence to <i>robust</i> for a set of studies that otherwise would be described as <i>moderate</i>. Such evidence not included in the unit of analysis can also inform evaluations of the coherence of the human evidence, the directness of the outcome measures, and the biological significance of the findings. Causality is inferred for a human evidence base of <i>robust</i>.</p>
<p>Moderate (⊕⊕○) ...evidence in human studies <i>(signal of effect with some uncertainty)</i></p>	<p>A set of evidence that does not reach the degree of certainty required for <i>robust</i>, but which includes at least one <i>high</i> or <i>medium</i> confidence study reporting an association and additional information increasing certainty in the evidence. For multiple studies, there is primarily consistent evidence of an association with reasonable support for adversity, but there might be some uncertainty due to potential chance, bias, or confounding or because of the indirectness of some measures. When only a single study is available in the unit of analysis, there is a large magnitude or severity of the effect, or a dose-response gradient, or other supporting evidence, and there are no serious residual methodological uncertainties. Supplemental evidence included in the unit of analysis might address the above factors and raise certainty in the evidence to <i>moderate</i> for a set of studies that otherwise would be described as <i>slight</i> or, in exceptional cases, could support raising to <i>moderate</i> evidence that would otherwise be described as <i>indeterminate</i>. Mechanistic evidence not included in the unit of analysis can also inform evaluations of the coherence of the human evidence, the directness of the outcome measures, and the biological significance of the findings.</p>
<p>Slight (⊕○○) ...evidence in human studies <i>(signal of effect with large amount of uncertainty)</i></p>	<p>One or more studies reporting an association between exposure and the health outcome, but considerable uncertainty exists and supporting coherent evidence is sparse. In general, the evidence is limited to a set of consistent <i>low</i> confidence studies, or higher confidence studies with significant unexplained heterogeneity or other serious residual uncertainties. It also applies when one <i>medium</i> or <i>high</i> confidence study is available within the unit of analysis without additional information strengthening the likelihood of a causal association (e.g., coherent findings within the same study or from other studies). This category serves primarily to encourage additional study where evidence does exist that might provide some support for an association, but for which the evidence does not reach the degree of confidence required for <i>moderate</i>.</p>

Evidence synthesis judgment	Description
Indeterminate (⊖⊖⊖) ...evidence in human studies (signal cannot be determined for or against an effect)	No studies available in humans or situations when the evidence is inconsistent and primarily of <i>low</i> confidence. In addition, this might include situations where higher confidence studies exist, but there are major concerns with the evidence base such as unexplained inconsistency, a lack of expected coherence from a stronger set of studies, very small effect magnitude (i.e., major concerns about biological significance), or uncertainties or methodological limitations that result in an inability to discern effects from exposure. It also applies for a single <i>low</i> confidence study in the absence of factors that increase certainty. A set of largely null studies could be concluded to be <i>indeterminate</i> if the evidence does not reach the level required for <i>compelling evidence of no effect</i> .
Compelling evidence of no effect (- - -) ...in human studies (strong signal for lack of an effect with little uncertainty)	A set of <i>high</i> confidence studies examining a reasonable spectrum of endpoints showing null results (e.g., an odds ratio of 1.0), ruling out alternative explanations including chance, bias, and confounding with reasonable confidence. Each of the studies should have used an optimal outcome and exposure assessment and adequate sample size (specifically for higher exposure groups and for susceptible populations). The set as a whole should include diverse sampling (across sexes [if applicable] and different populations) and include the full range of levels of exposures that human beings are known to encounter, an evaluation of an exposure-response gradient, and an examination of at-risk populations and lifestages. Supplemental evidence can help to address the above considerations or, when included in the unit of analysis, provide additional support for this judgment.

Table 6-5. Framework for evidence synthesis judgments from studies in animals

Evidence synthesis judgment	Description
Robust (⊕⊕⊕) ...evidence in animal studies (strong signal of effect with very little uncertainty)	The set of <i>high</i> or <i>medium</i> confidence, independent experiments (i.e., across laboratories, exposure routes, experimental designs [for example, a subchronic study and a multigenerational study], or species) reporting effects of exposure on the health outcome(s). The set of studies is primarily consistent, with reasonable explanations when results differ (i.e., due to differences in study design, exposure level, animal model, or study confidence), and the findings are considered adverse (i.e., biologically significant and without notable concern for indirectness). At least two of the following additional factors in the set of experiments increase certainty in the evidence: coherent effects across multiple related endpoints (within or across biologically related units of analysis); an unusual magnitude of effect, rarity, age at onset, or severity; a strong dose-response relationship; or consistent observations across animal lifestages, sexes, or strains. Supplemental evidence included in the unit of analysis (e.g., mechanistic studies in exposed animals or animal cells) might raise the certainty of evidence to <i>robust</i> for a set of studies that otherwise would be described as <i>moderate</i> . Such evidence not included in the unit of analysis can also inform evaluations of the coherence of the animal evidence, the directness of the outcome measures, and the biological significance of the findings.

Evidence synthesis judgment	Description
<p><i>Moderate</i> (⊕⊕⊖) ...evidence in animal studies <i>(signal of effect with some uncertainty)</i></p>	<p>A set of evidence that does not reach the degree of certainty required for <i>robust</i>, but which includes at least one <i>high</i> or <i>medium</i> confidence study and additional information increasing certainty in the evidence. For multiple studies or a single study, the evidence is primarily consistent or coherent with reasonable support for adversity, but there are notable remaining uncertainties (e.g., difficulty interpreting the findings due to concerns for indirectness of some measures); however, these uncertainties are not sufficient to reduce or discount the level of concern regarding the positive findings and any conflicting findings are from a set of experiments of lower confidence. The set of experiments supporting the effect provide additional information increasing certainty in the evidence, such as consistent effects across laboratories or species; coherent effects across multiple related endpoints (can include mechanistic endpoints within the unit of analysis); an unusual magnitude of effect, rarity, age at onset, or severity; a strong dose-response relationship; or consistent observations across exposure scenarios (e.g., route, timing, duration), sexes, or animal strains. Supplemental evidence included in the unit of analysis could address the above factors and raise certainty in the evidence to <i>moderate</i> for a set of studies that otherwise would be described as <i>slight</i> or, in exceptional cases, might support raising to <i>moderate</i> evidence that would otherwise be described as <i>indeterminate</i>. Mechanistic evidence not included in the unit of analysis can also inform evaluations of the coherence of the animal evidence, the directness of the outcome measures, and the biological significance of the findings.</p>
<p><i>Slight</i> (⊕⊖⊖) ...evidence in animal studies <i>(signal of effect with large amount of uncertainty)</i></p>	<p>One or more studies reporting an effect on an exposure on the health outcome, but considerable uncertainty exists and supporting coherent evidence is sparse. In general, the evidence is limited to a set of consistent <i>low</i> confidence studies, or higher confidence studies with significant unexplained heterogeneity or other serious uncertainties (e.g., concerns about adversity) across studies. It also applies when one <i>medium</i> or <i>high</i> confidence experiment is available within the unit of analysis without additional information increasing certainty in the evidence (e.g., coherent findings within the same study or from other studies). Biological evidence from mechanistic studies could also be independently interpreted as <i>slight</i>. This category serves primarily to encourage additional study where evidence does exist that might provide some support for an association, but for which the evidence does not reach the degree of confidence required for <i>moderate</i>.</p>
<p><i>Indeterminate</i> (⊖⊖⊖) ...evidence in animal studies <i>(signal cannot be determined for or against an effect)</i></p>	<p>No studies available in animals or situations when the evidence is inconsistent and primarily of <i>low</i> confidence. In addition, this might include situations where higher confidence studies exist, but there are major concerns with the evidence base such as unexplained inconsistency, a lack of expected coherence from a stronger set of studies, very small effect magnitude (i.e., major concerns about biological significance), or uncertainties or methodological limitations that result in an inability to discern effects from exposure. It also applies for a single <i>low</i> confidence study in the absence of factors that increase certainty. A set of largely null studies could be concluded to be <i>indeterminate</i> if the evidence does not reach the level required for <i>compelling evidence of no effect</i>.</p>

Evidence synthesis judgment	Description
<p><i>Compelling evidence of no effect</i> (- - -) ...in animal studies <i>(strong signal for lack of an effect with little uncertainty)</i></p>	<p>A set of <i>high</i> confidence experiments examining a reasonable spectrum of endpoints that demonstrate a lack of biologically significant effects across multiple species, both sexes, and a broad range of exposure levels. The data are compelling in that the experiments have examined the range of scenarios across which health effects in animals could be observed, and an alternative explanation (e.g., inadequately controlled features of the studies' experimental designs; inadequate sample sizes) for the observed lack of effects is not available. Each of the studies should have used an optimal endpoint and exposure assessment and adequate sample size. The evidence base should represent both sexes and address potentially susceptible populations and lifestages. Supplemental evidence can help to address the above considerations or, when included in the unit of analysis, provide additional support for this judgment.</p>

6.1.1. Considerations for Developing the Human and Animal Evidence Syntheses

Several considerations specific to evaluating the evidence from human and animal studies to draw synthesis judgments are worth elaborating upon, including evaluating the consistency of the available studies within each unit of analysis and the use and interpretation of statistical testing.

Considerations for Evidence Synthesis of Human (Primarily Epidemiological) Studies

The complexity of the analysis of the evidence in a synthesis will be determined by the breadth of the evidence base, confidence in study results, and the differences encompassed by the studies. As previously noted, given the often-heterogeneous nature of studies, evaluating the consistency of the available results across studies is typically one of the most time-consuming and consequential pieces of the human evidence synthesis.

Grouping studies by the level and variation or range of exposure experienced by the study populations might explain a set of seemingly inconsistent results or provide evidence of a biological gradient or exposure-response relationship. Associations among populations exposed to lower levels could be null or highly variable with wide confidence intervals (CIs), while associations from studies at higher levels might be stronger. Sometimes, a comparison across exposure levels also will involve comparisons by exposure setting (e.g., occupational vs. residential, or between industry types). An example of how grouping studies on the basis of exposure level can inform the synthesis of evidence is seen in the IRIS evaluation of evidence on carcinogenicity of trichloroethylene [TCE ([U.S. EPA, 2011b](#))]. Figure 6-1 illustrates how forest plots can be used to present effect estimates in relation to levels of exposure. The shape of the exposure-response relationship observed in a given study can depend on various factors, including population characteristics, dose-response model used, range of exposure, sample size, and others [e.g., exposure measurement error ([Park and Stayner, 2006](#); [Brauer et al., 2002](#))]. In some cases, these analyses can also be done quantitatively (e.g., by combining results across studies using meta-analysis as described below and stratifying by factors of interest).

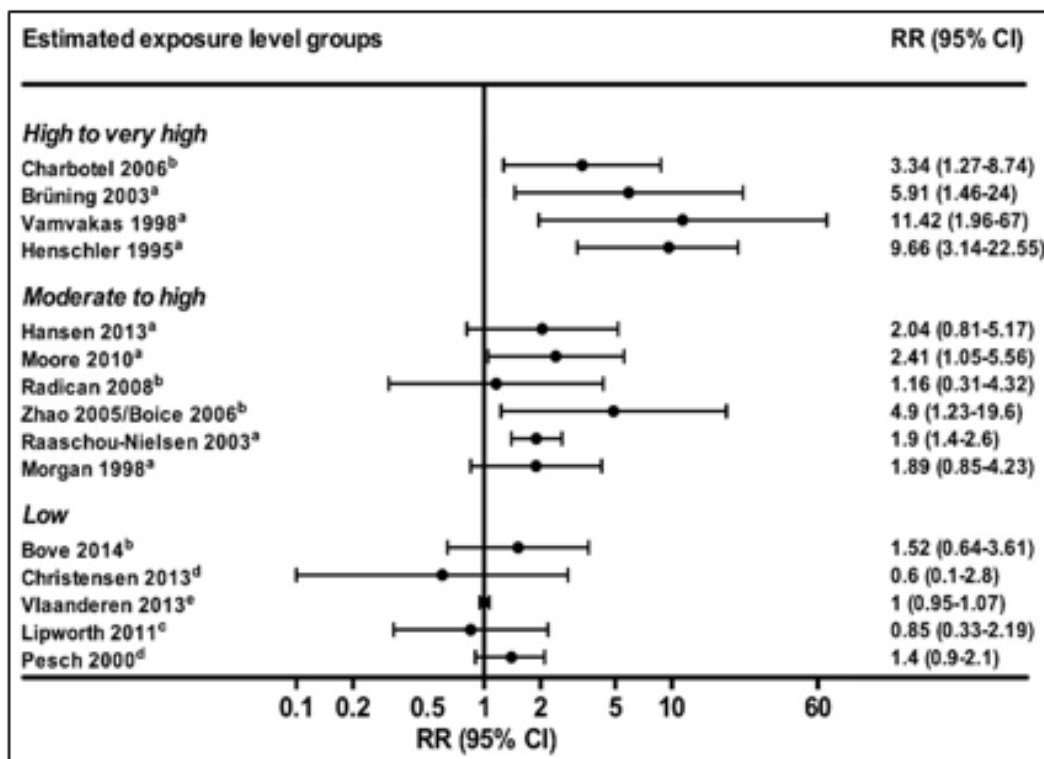


Figure 6-1. Trichloroethylene (TCE) and kidney cancer: stratification by exposure level (U.S. EPA, 2011b).

CI = confidence interval; RR = relative risk.

All figures comparing study results by potentially explanatory factors should include information about each study's confidence.

Some evidence synthesis considerations, including the effect magnitude and precision of an association, also can be used to assess the impact of limitations identified in individual studies to increase confidence that the association is not due to chance or bias. Higher precision, as reflected by narrow confidence bounds or smaller standard errors (SEs), adds confidence in the observed association; as described previously, however, precision of individual studies might not be as important to consider as the pattern that is seen across studies, or the precision of a combined effect estimate.

The evaluation of findings across studies also can facilitate assessments of confounding when an important characteristic or coexposure was not considered by all studies or could not be ruled out in individual studies. Similar observations in different populations (e.g., different types of industries, or different geographical areas) reduce the likelihood that confounding is a reasonable explanation for the findings. An example of an analysis of confounding in the synthesis of results across studies is found in the IRIS Toxicological Review of TCE and kidney cancer (U.S. EPA, 2011b). Several cohort and case-control studies that met defined standards for design and analysis were included in the systematic review. Although the case-control studies adjusted for potential

confounding by smoking (a known risk factor for kidney cancer) most of the cohort studies did not. The Toxicological Review concluded that the expected impact was minimal because smoking was not expected to be associated with TCE exposure in the study populations. In addition, lung cancer was not associated with TCE exposure in most of the studies. If smoking were a strong confounder of the observed association with kidney cancer, a stronger association with TCE would have been expected for lung cancer, as the smoking-related relative risk for lung cancer is more than threefold higher than the risk for kidney cancer ([IARC, 2004](#)). Confounding by smoking also was evaluated using the results of a meta-analysis by comparing the common estimates of relative risk for kidney cancer and lung cancer.

Considerations for Evidence Synthesis of Animal Studies

Paralleling the human evidence synthesis considerations, the syntheses of the available animal evidence often emphasize the evaluation of the consistency (and, in some cases, coherence) of the available results across studies over the other considerations summarized in Table 6-3. In addition to study evaluation judgments, some examples of considerations especially pertinent to evaluating the consistency of animal evidence synthesis include:

- *Exposure ranges:* Did a null study use an exposure range or periodicity that might be too low or infrequent (e.g., were the highest exposure levels in the null study similar to, or lower than levels tested in the other available studies observing effects)? Conversely, if only excessively high exposure levels were tested, is there reason (e.g., an experimentally validated, substantial difference in pharmacokinetics at different exposure levels; observed or inferable nonspecific toxicity) to believe that the observed responses might be dissimilar to responses that might occur at lower exposure levels?
- *Pharmacokinetics:* Can differences in response be explained by differences in pharmacokinetics (e.g., metabolism) across different animal species?¹³ (This factor can also be considered within the context of differences in response seen by route of exposure.)
- *Endpoint comparisons:* Are there notable differences in the specific endpoints evaluated across studies, or in the way those endpoints were assessed? For some effects, the seemingly similar evaluations across animal studies can be highly heterogeneous and might be better considered as coherence across biologically related endpoints rather than as consistency.

Coherence of results is another important consideration in the synthesis of the animal evidence. Correlated toxicity measures in individual studies or across studies strengthen the evidence for a hazard. An example is related effects in a target organ (e.g., changes in serum enzymes that are markers of liver damage, increased liver weight, and liver histopathology), particularly when the coherent effects are observed within the same cohort of exposed animals. Within the context of coherence, it is often useful to examine the concordance between the

¹³Although pharmacokinetics can also differ due to differences in age, sex, or strain, chemical-specific data describing such differences are rarely available.

sequence of observed effects and the timing, duration, and level of exposure (e.g., do mild effects occur prior to, or at lower exposure levels than, more severe changes?). If an expected coherence between findings is not observed, possible explanations should be explored including the biology of the effects and the sensitivity and specificity of the measures used. Typically, the synthesis should consider and discuss the relative sensitivity and severity of the different endpoints and emphasize those most informative to the health effect in question (e.g., endpoints indicating impaired or loss of function in an organ are generally prioritized over change in its weight).

Consideration of Statistical Tests and Meta-Analysis

Statistical significance testing

Statistical significance testing is an important tool for supporting a decision that there is a demonstrable effect, especially when the biological significance ([U.S. EPA, 2002b](#)) of an outcome is uncertain or unclear (e.g., based on historical responses and variability). A consistent pattern of statistically significant results for an effect (or related effects), of similar size, across comparable, well-designed studies increases confidence that the effect results from the exposure of interest. However, consideration of the consistency in patterns of results does not require that all findings are statistically significant. A presence of a change that lacks statistical significance can be used to support conclusions of consistency. Nor should all statistically significant results be interpreted as evidence of an effect. The limitations of sole reliance on statistical significance for reaching conclusions are well recognized ([Ziliak, 2011](#); [Rothman, 2010](#); [Newman, 2008](#); [Hoenig and Heisey, 2001](#); [Sterne et al., 2001](#); [Savitz, 1993](#)). In particular, the American Statistical Association “Statement on Statistical Significance and *P*-Values” ([Wasserstein and Lazar, 2016](#)) has clarified widely agreed upon statistical principles in support of the validity, reproducibility, and replicability of scientific conclusions. Overall, a careful analysis of results across a set of comparable studies should include those that are statistically significant and those that are not.

The following summarizes several principles relevant for interpreting reported statistical significance testing for hazard evaluation.

- The use of $p = 0.05$ as a decision point for statistical significance is a conventionally used but arbitrary criterion, with no a priori connection to biological significance [e.g., [Rothman \(2010\)](#)].
- *P* values alone provide no information about effect size or inform risk assessors about the biological significance of reported results.
 - Lack of statistical significance should not automatically be interpreted as evidence of no effect. Because statistical significance is a function of sample size, an effect’s prevalence, and strength of the association with an exposure, the lack of statistical significance in the presence of an elevated effect estimate often means that chance cannot be ruled out with confidence. For example, if a particular exposure level leads to an adverse effect, studies with low statistical power might not show statistical significance for this effect.

- Not all statistically significant results (“ $p < 0.05$ ”) should be interpreted as evidence of an effect. Several situations can lead to spuriously low p -values, such as unusually low variability in control or treated groups. One concern is that the greater the number of statistical tests performed, the greater the chance that some negligible effects will be recognized as statistically significant, a consequence of the statistical testing paradigm (i.e., “false positives”). These instances of statistically significant results could also be reconciled by examining patterns in effect estimates across similar studies and evaluating coherence with related evidence (see previous bullet).
- Consistency of results across studies is a question of the direction and magnitude of the effect sizes rather than the magnitude of the p -value, especially whether $p < 0.05$. Challenges in interpreting p -values reported by different investigators—due to, for example, variation in study designs and sizes, and the variety of statistical significance tests that can be used¹⁴—are also important to address when distinguishing between “conflicting” and “differing” evidence.

These points are raised to clarify the overall role of statistical significance testing and its interpretation in the systematic evaluation of hazard evidence. In some cases, statistical analysis of individual studies beyond that reported (e.g., use of a consistent statistical method to evaluate several similar studies) or across a related set of studies in a meta-analysis can increase confidence in findings for an outcome.

Additional statistical analyses of individual studies (see “Trend testing” below) or across a set of studies (see “Meta-analysis” below) might increase precision in estimating the magnitude of the association, help determine whether an association exist or does not exist, and identify a dose-response pattern.

Trend testing of individual studies

One relatively common application at the individual study level is trend testing to evaluate response patterns across treated groups. Detection of a dose-response trend across all treated groups can inform evidence synthesis judgments on dose-gradient. Identifying trends can be missed in analyses based on one-at-a-time, multiple pairwise comparisons between each dose group and the control group. When trend tests are not presented in published studies (or details of the trend test used are not provided), U.S. Environmental Protection Agency (EPA) can conduct trend tests using summary statistics in published studies, such as means and variance estimates.

Meta-analysis or other quantitative analysis across studies

With respect to statistical analyses across studies, some data sets can support calculating a summary effect estimate using a common measure reported by some or all the studies and provide a more precise estimate and a better understanding of the overall magnitude of the effect than could be achieved by estimate(s) from individual studies. The preferred statistical method for

¹⁴Sometimes it might be possible to obtain additional results that are comparable by requesting analyses or results from the authors of the studies, or if appropriate data are available, to conduct additional analyses.

synthesizing evidence within a such a set of studies is some type of statistical meta-analysis. This might use a measure of effect (e.g., extra risk, percentage of difference from the control, risk ratio, odds ratio, trend statistics, slopes) with their variances.¹⁵ Meta-regression, examining the influence of various factors on results across studies, can be used in some circumstances (e.g., with sufficient numbers of studies). Meta-analysis is more typically considered for human evidence data sets, although it can be considered for animal evidence. However, environmental health evidence often is too heterogeneous to provide a compelling reason to conduct a meta-analysis. As noted in Section 5.5.2, forest plots can be used to help assess patterns across human or animal studies even if a quantitative summary estimate is not developed. Below are general considerations for conducting a meta-analysis or other quantitative analysis across studies.

- What could the analysis contribute to the synthesis of the evidence?
- The criteria used to select studies, weights, and validity of the assumption that the studies are examining a common effect estimate must be carefully considered. The question of the suitability of a set of studies for meta-analysis requires more than a statistical test of heterogeneity ([Vesterinen et al., 2014](#); [Fu et al., 2011](#)). Statistical significance or other criteria based on the study results should not be used for selecting studies for the meta-analysis (i.e., studies with null findings should not be excluded from the meta-analysis).
- What factors, if any, should be used to stratify a meta-analysis? Study confidence, exposure levels, exposure route, species, lifestage, and numerous other considerations could contribute to the observed results and to heterogeneity among studies.
- What study results can be combined? If studies cannot be included in the meta-analysis (e.g., because of different measures or forms of the results), they should be discussed in the synthesis.

6.1.2. Approaches to Facilitate Evidence Synthesis of Mechanistic Studies

As described previously, the mechanistic evidence synthesis section on a health effect has multiple roles, as it can inform the human and animal evidence synthesis conclusions (the focus of this section) and the integration judgments across evidence streams (described in Section 6.2). When mechanistic information is included in a predefined unit of analysis (as specified in the protocol, see Chapter 3), the approach for synthesizing the available mechanistic evidence to inform the within-stream judgments parallels the approach to the human and animal evidence syntheses, and the same factors included in Table 6-3 are considered during the mechanistic evidence synthesis. The synthesis of mechanistic information included in a unit of analysis is typically summarized in narrative form by unit of analysis, although in some cases, and particularly for robust evidence bases, it might make sense to use a different organization (e.g., describing an analysis of an MOA pertinent to multiple units of analysis). Similar to the human and animal

¹⁵A meta-analysis is most often conducted on effect estimates but can also be conducted using *p*-values.

synthesis sections, the takeaways from the mechanistic evidence synthesis are tracked in evidence profile table entries (accompanying the evidence integration narrative, see Table 6-1).

In addition to the considerations described for evaluating the factors in Table 6-3, it is important to consider whether the evidence for a mechanistic endpoint is consistent, and if not, to determine whether there is a plausible explanation for the heterogeneous findings. Some examples of considerations that are especially pertinent to evaluating the consistency of the mechanistic evidence include:

Endpoint comparisons: Are there notable differences in the specific endpoints evaluated across studies or experiments, or in the way those endpoints were assessed (i.e., measurement methods, assay conditions, model system)? For some effects, seemingly similar evaluations across in vitro studies can be highly heterogeneous and may be better considered as coherence across biologically related endpoints rather than as consistency.

Exposure range: Did a study or experiment use an exposure range or periodicity that might be too low or too high? For example, were the highest exposure levels in the null study similar to, or lower than, levels tested in the other available studies observing effects? Did the positive and negative control samples perform as expected? Conversely, if only high exposure levels were tested, is there reason to believe that the observed responses might be dissimilar to responses that might occur at lower exposure levels? For in vitro studies, was cytotoxicity measured?

Adversity: As described previously, the evaluation of adversity typically encompasses an interpretation of the biological significance of the effect magnitude and the directness of the outcome measures (i.e., in relation to outcomes accepted as adverse or clinically relevant). Mechanistic evidence, on its own, is generally not interpreted as adverse because it is defined as representing indirect measures of apical effects. Instead, mechanistic data are typically used to inform potential linkages between indirect measures and adverse responses (see “Directness of Outcome Measures” below).

Mechanistic and other supplemental information not included in a predefined unit of analysis are also considered for inclusion in the mechanistic evidence synthesis depending on the characteristics and uncertainties of the other available evidence. Thus, in many cases, understanding which additional reviews of mechanistic information are warranted can only occur after the human and animal evidence syntheses have been completed and once the uncertainties in those data become apparent. The mechanistic narrative emphasizes coherence, biological significance, and directness as factors that could increase or decrease the overall certainty in the evidence considered for each unit of analysis.

The resulting mechanistic inferences inform, when possible, the interpretation of human and animal evidence for each unit of analysis. In less common scenarios, the mechanistic evidence might be the only evidence considered within a unit of analysis. Considerations (as applied to using mechanistic evidence to interpret the available human and animal studies) include the following:

Coherence

Mechanistic evidence informing the biological relatedness of outcomes within or across related units of analysis within an evidence stream can increase or decrease certainty in the evidence for an individual unit of analysis. Outcome measurements from mechanistic studies (e.g., biomarkers, molecular changes) can inform an understanding of coherence across studies and across biologically related endpoints within a study. A more complete understanding of the biological interactions associated with the observed effects, based on established biological and medical knowledge, can increase certainty in the evidence when changes are related. Mechanistic findings that strengthen the linkage between different outcomes potentially associated with the chemical exposure, or which are coherent with the changes observed within the unit(s) of analysis, can also inform certainty in the evidence, which is particularly consequential when the findings are from the same exposed population or experimental model. For example, deiodinase activity within tissues affects thyroid signaling, even when blood thyroid hormone (TH) concentrations remain constant. Therefore, changes in tissue deiodinase activity (e.g., in the brain) can facilitate interpretations of coherence between blood TH level changes and tissue-specific effects of endocrine disruption (e.g., brain histopathology potentially related to neuroendocrine effects). Often, the mechanistic findings are outside the units of analysis, but could provide coherence that supports certainty in the evidence within a predefined unit of analysis. The interpretation of the pattern of changes across the outcomes should consider the underlying biology (e.g., one outcome may be expected to precede the other, or be more sensitive), and compare the available data against that understanding. If the mechanistic evidence is insufficient to provide a mechanistic or biological understanding of coherence (or lack thereof), this will not change the interpretation of the results from the human or animal studies (i.e., there is no increase or decrease in certainty).

Biological Significance

The biological significance of an effect reflects an interpretation regarding whether an effect magnitude is biologically meaningful or adverse. In most cases, interpretations of biological significance are based on the established clinical relevance of the effect magnitude (at the individual or population level), understanding of the natural variability in the response, or assumptions based either on historical linkages to more overt manifestations of toxicity (e.g., leveraging studies in patients with various types of hepatic injury to interpret the importance and time-course of serum biomarker changes) or health-protective defaults (e.g., a 10% change in organ weight being interpreted as biologically significant). Typically, when chemical-nonspecific mechanistic understanding of the underlying biology is used for this purpose, it makes sense to discuss and provide this justification as part of the mechanistic evidence synthesis. However, chemical-specific mechanistic evidence might also sometimes (rarely) be available to inform these interpretations. For example, establishing the progression of key events in the biological pathway from exposure to outcome (e.g., key event relationships) can aid in identifying potential thresholds

for changes in one event that can be reasonably expected to lead to cascading changes toward more overt phenotypes. Although the level of mechanistic evidence required to establish such an understanding will be high, this inference can substantially increase or decrease certainty that the observed effect magnitude supports identifying a hazard.

Directness of Outcome Measures

The directness of the outcome measures can also inform interpretations regarding whether an effect is biologically meaningful or adverse. Units of analysis are typically defined as an outcome(s) that signals potential toxicity (e.g., disease incidence, change in tissue structure or function). Thus, direct measures of the units of analysis are typically straightforward to use as evidence informing a hazard judgment. Indirect measures (e.g., serum biomarkers of an organ-level effect) decrease evidence certainty when they are determined to be nonspecific or not linked to the unit of analysis under evaluation, or when the unit of analysis is itself defined as an indirect measure. Mechanistic evidence (or, as described in the context of biological significance) chemical-nonspecific mechanistic understanding can inform the potential linkage between an indirect measure and more direct indicators of potential toxicity. Even sparse amounts of mechanistic evidence can help to inform this interpretation, so the potential for assessments to use chemical-specific mechanistic evidence in this way is far more common than for an interpretation of biological significance.

The mechanistic evidence synthesized for use in this way is often not included within the units of analysis (although some key pieces of information may be). Primarily, this is because an analysis of such linkages typically spans multiple mechanistic events in a pathway and may consider findings from markedly heterogeneous study designs, including from both mechanistic and apical evaluations, for each event. The interpretation of the pattern of changes across the assessed events or the outcomes themselves should consider the underlying biology (e.g., one event may be expected to precede the other, or be more sensitive). Note that if the mechanistic evidence (or biological knowledge) is conflicting or is otherwise considered insufficient to provide support for an association with more overtly adverse or direct measures of toxicity, this will not change the interpretation of the results from the human and animal syntheses (i.e., certainty would be decreased due to the indirectness of the outcome measures).

6.2. EVIDENCE INTEGRATION

The phase of evidence integration combines animal and human evidence synthesis judgments while also considering information on the human relevance of findings in animal evidence, coherence across evidence streams (“cross-stream coherence”), information on susceptible populations or lifestages, understanding of biological plausibility and MOA, and potentially other critical inferences (e.g., read-across analyses) that can draw on mechanistic and other supplemental evidence (see Table 6-6). This analysis culminates in an evidence integration

judgment and narrative for each potential health effect category (i.e., each noncancer health effect and specific type of cancer, or broader grouping of related outcomes as defined during problem formulation; see Chapter 3). To the extent it can be characterized prior to conducting dose-response analyses, exposure context is also provided.

6.2.1. Considerations That Inform Evidence Integration

To inform the overall judgment on certainty in the evidence for a hazard (see Section 6.2.2), IRIS assessments integrate the evidence synthesis judgments drawn for each unit of analysis while considering additional factors assessed across evidence streams. These additional factors addressed during evidence integration are summarized in Table 6-6, with some elaboration on considering the human relevance of findings, cross-stream coherence, susceptible populations, and biological plausibility provided in the following subsections. Similar to evidence synthesis, each of the considered factors in Table 6-6 require elaboration or evidence-based justification in the evidence integration narrative. Also similar to the consideration of factors during evidence synthesis, the analysis of these considerations across streams during evidence integration is qualitative (i.e., numerical scores are not developed, summed, or subtracted).

During evidence integration (or the later stages of evidence synthesis), it may be determined that there are additional analyses that have not been conducted but that are important for drawing a more reliable evidence integration judgment (e.g., to address a critical uncertainty identified during draft development). If the literature inventory (see Section 2.5) does not include studies on the topic, it might be determined that an additional focused search for information is worthwhile. Depending on the extent of any such additional analyses and whether they ultimately are able to inform the judgment, it might be necessary to refine the problem formulation (see Chapter 3) and the steps that follow. In most cases, however, it is unlikely that such data will be available, and the uncertainty should be highlighted in the narrative.

Table 6-6. Considerations that inform the evidence integration judgment

Judgment	Description
Human relevance of findings	Used to describe and justify the interpreted relevance of the data from experimental animals (or other model systems) to humans. In the absence of chemical-specific evidence informing human relevance, the evidence integration narrative will briefly describe the interpreted underlying biological similarity across species. As noted in EPA guidelines (U.S. EPA, 2005a), there needs to be evidence or a biological explanation to support an interpreted lack of human relevance for findings in animals, and site concordance is neither expected nor required. Thus, in the absence of specific evidence or cross-species understanding of the underlying biology, it is appropriate to use a statement such as, “without evidence to the contrary, [health effect] responses in animals are presumed relevant to humans.”
Cross-stream coherence	Used to address the concordance of biologically related findings across human, animal, and mechanistic studies, considering features of the available evidence such as exposure timing and levels. Notably, for many health effects (e.g., some nervous system and reproductive effects;

Judgment	Description
	cancer), it is not necessary or expected that effects manifest in humans are identical to those observed in animals (e.g., tumors in animals can be predictive of carcinogenic potential in humans, but not necessarily at the same site), although this typically provides stronger evidence. Biological understanding of the manner in which the outcomes are manifest in different species can inform cross-stream coherence. Evidence supporting a biologically plausible mechanistic pathway across species adds coherence (see below).
Susceptible populations and lifestages	Used to summarize analyses relating to individual and social factors that may increase susceptibility to exposure-related health effects in certain populations or lifestages, or to highlight the lack of such information. These analyses are based on knowledge about the health outcome or organ system affected and focus on the influence of intrinsic biological factors but can also include consideration of mechanistic and ADME evidence.
Biological plausibility and MOA considerations	Used to summarize the interpreted biological plausibility of an association between exposure and the health effect, based primarily on the extent to which the available evidence comports with the known development and characteristics of the health effect (and thus dependent on sufficient information being available to draw such an interpretation). Importantly, because this interpretation is dependent on canonical scientific knowledge about the health effect, the lack of such understanding does not provide a rationale to decrease certainty in the evidence for an effect (NTP, 2015 ; NRC, 2014). These analyses can be detailed (e.g., when attempting to establish MOA understanding) and, if so, are typically conducted separately (e.g., as part of the mechanistic evidence synthesis) and then referenced in the evidence integration narrative.
Other critical inferences (optional)	Can be used to describe the consideration of other evidence or non-chemical-specific information that informs evidence integration judgments (e.g., use of read across analyses or ADME understanding used to inform the other considerations described below; judgments on other health effects expected to be linked to the health effect under evaluation).

ADME = absorption, distribution, metabolism, and excretion; MOA = mode of action.

Human Relevance of Findings

Although overlapping with the analysis of cross-stream coherence (see below), observations of changes in exposed humans that are coherent with mechanistic or toxicological endpoints observed in experimental studies (and that are interpreted to be associated with the outcome and/or unit of analysis) strengthen the human relevance of the findings from experimental animals (or other model systems). Evidence of biological precursors that link the exposure to the observed outcome in humans and animals strengthens human relevance and inferences that the effect is relevant between species. If evidence establishes that the mechanism underlying the animal response does not operate in humans, or that animal models do not suitably inform a specific human health outcome or unit of analysis, this can support the view that the animal response is not relevant to humans. Mechanistic explanations for differing responses across populations (e.g., by species, sex, strain) informs judgments on relevance or a lack of relevance to humans. [Note that in the absence of sufficient information to the contrary, effects in animal models are assumed to be relevant to humans ([U.S. EPA, 2005a, 1998, 1991a](#)).]

When considering the human relevance of animal evidence, some questions to focus this analysis include the following:

- What is known about the biology underlying the development of the outcomes observed in animals and are there important species differences in the functions or responses of the organs or systems involved? At this stage, this does not refer to whether the studies employed typical human exposure levels, but rather focuses on critical differences in biology between animals and humans, e.g., knowledge that humans lack a critical enzyme.
- When human evidence is lacking or has results that differ from animals, is there evidence that the mechanisms underlying the effects in animals do not operate in humans? Analyses of the mechanisms underlying the animal response in relation to those presumed to operate in humans, or the suitability of the animal models to a specific human health outcome, can inform the extent to which the animal response is likely to be relevant to humans.
- The analysis of human relevance focuses on evaluations of the following issues. The extent of the analysis varies depending on the anticipated impact of the animal evidence to the overall evidence integration judgment.
 - ADME comparisons across species, primarily relating to distribution (e.g., to the likely target tissue) and metabolism (particularly if a metabolite is known to be more/less toxic).
 - Coherence of changes observed in exposed humans with animal evidence of mechanistic or toxicological changes (see also cross-stream coherence below).
 - Understanding of similarities (or differences) in the underlying biology of the organ or system that is the target of the health effect (e.g., thyroid signaling processes are well conserved across rodents and humans).
 - Evidence for a plausible biological pathway or MOA, within which the endpoints/outcomes/units of analysis and relationships are evaluated regarding the likelihood of similarities (e.g., in presence or function) across species.

Cross-stream Coherence

Consideration of cross-stream coherence during evidence integration addresses the coherence of findings when compared across evidence streams. Judgments on the coherence of findings integrates the available information on biologically related units of analysis across human and animal lines of evidence, considering one or more units of analysis from each evidence stream. Similar to the discussion of the analysis of coherence within an evidence stream (see Section 6.1.2), the interpretation of the coherence of changes across evidence streams during evidence integration considers the underlying biology within each species (and the manner in which the related effects are manifest in different species) and compares the available data against that understanding.

During evidence integration, both chemical-specific mechanistic evidence and established biological understanding can inform the relatedness of findings across evidence streams, possibly

increasing or decreasing certainty in the overall evidence for an effect. EPA guidelines and other resources (e.g., OECD guidance) are consulted when drawing these inferences.

Susceptible Subpopulations and Lifestages

A description of the information on potential susceptibility to the health effects caused by exposure to the agent can help to not only identify those that might be most at risk of developing the health effects following exposure but also identify factors that could improve dose-response estimates, as discussed below. In addition to assessment-specific health effects evidence, an understanding of biological mechanisms and chemical-specific pharmacokinetics, as well as biochemical and physiological differences among species, lifestages and sexes, can be used. At a minimum, particular consideration is given to infants and children, pregnant women, and women of childbearing age. Many of the foundational analyses for summarizing susceptibility in the evidence integration narrative are undertaken during evidence synthesis as patterns across studies are evaluated with respect to consistency, coherence, and the magnitude and direction of effect measures. Relevant factors for exploring patterns include intrinsic biological factors such as race/ethnicity, genetic variability, sex, age or lifestage, and pre-existing health conditions (which can also have an extrinsic basis), as well as certain extrinsic factors (e.g., socioeconomic status, access to health care), although information on the latter is rarely available in human health studies of environmental chemicals. Information on extrinsic factors potentially influencing susceptibility (e.g., proximity to exposure; certain lifestyle factors including subsistence living) are not considered in IRIS assessments as part of characterizing potential susceptibility. These and other exposure-focused factors are considered by risk managers as part of exposure assessment and after the human health assessment is complete (<https://www.epa.gov/iris/basic-information-about-integrated-risk-information-system>, see “What’s the Role of IRIS Assessments in Risk Assessment?”

A summary of all potential susceptibility factors should be included in the evidence integration narrative. For more information, see U.S. EPA (2005b). When characterizing the potential susceptible populations or lifestages to inform evidence integration, important considerations include whether the results appear to differ by categories that indicate the apparent presence of susceptible populations (e.g., across demographics, species, strains, sexes, or lifestages). In data-rich scenarios, it may be possible to conduct focused analyses of supplemental evidence to better characterize the sources and impact of potential susceptibilities; for example, those that can be explained by mechanistic understanding (e.g., due to genetic polymorphisms or metabolic deficiencies). Information on susceptible populations and lifestages is also considered for use in dose-response analyses. As described in Chapters 7 and 8, this information can inform the selection of studies advanced for quantification (e.g., selecting those studies that stratify results for populations or lifestages identified as more susceptible) and the uncertainty factors applied (e.g., the intraspecies uncertainty factor [UF_H]). In addition, if the mechanistic evidence base, including an understanding of pharmacokinetic differences, allows for an understanding of which populations or lifestages might be particularly susceptible to the MOA, this information should be

flagged for consideration during dose response assessment. A mechanistic understanding of how a health outcome develops, even without a full MOA analysis, can clarify characteristics of important events (e.g., their presence or sensitivity across lifestages or across genetic variations) and help to identify susceptible populations, informing the dose-response analysis decisions described above.

Biological Plausibility and Mode of Action (MOA) Considerations

Mechanistic information can strengthen the evidence for an association (or the lack of an association) between exposure and the health effect on the basis of existing biological knowledge of how the health effect develops (biological plausibility). A more complete understanding of the biological interactions associated with the observed effects, based on established biological and medical knowledge, paired with the mechanistic support (e.g., a shared key event) for linkages across outcomes, increases the certainty of the evidence when changes are related. The interpretation of the pattern of changes across the outcomes should consider the underlying biology (e.g., one outcome may be expected to precede the other, or to be more sensitive). The plausibility of an association observed in human or animal studies could be diminished if expected findings are not apparent in mechanistic evidence, or an expected pattern among biologically linked health effects is not observed. If there is unexplained inconsistency in the available mechanistic evidence or if it is otherwise insufficient to provide a mechanistic explanation for an association (or lack thereof), this will not change the interpretation of the results from the human and animal syntheses (i.e., consideration of biological plausibility will not influence the evidence integration judgment).

Mechanistic evidence informing biological plausibility and MOA analyses can be provided by data from experimental studies of mechanistic pathways; such evidence can be particularly meaningful to evidence integration judgments when strong support is provided for key events or multiple components of a potential pathway. Mechanisms or biological changes with broad scientific acceptance for their relevance to chemical toxicity or the health effect (e.g., key characteristics, hallmarks of cancer) are typically used to organize the chemical-specific evidence and identify key events leading from exposure to the health effect (see Section 2.5.2). For each key event and key event relationship, the evidence is considered regarding the consistency of experimental data and the generalizability, or likelihood of similarities (e.g., in presence or function) across species, as well as the strength of the support for the biological mechanism.

Mechanistic evidence from well-conducted studies that demonstrates that the health effect is unlikely to occur (i.e., species-specific effects, non-relevant exposure conditions) can support a judgment that the effects from animal or human studies are not biologically relevant, which weakens the summary evidence integration judgment. Such a decision depends on an evaluation of the strength of the information supporting vs. opposing biological plausibility, as well as the strength of the health effect specific findings (e.g., stronger health effect data require more certainty in mechanistic evidence opposing plausibility). If sufficiently supported, MOA understanding can

serve to strengthen (e.g., strong support for mutagenicity) or weaken (e.g., critical dependence on a key event not likely to be operant in humans) evidence integration judgments.

Interpreting the mechanistic evidence

The mechanistic analysis often focuses on precursors, biomarkers, or other molecular or cellular changes known to be closely related to the health effect(s) of interest. If available, such information can often inform the likelihood of whether the observed effects result from exposure. In many cases, foundational literature might not be available for the chemical under evaluation, and it will be necessary to evaluate mechanistic events and pathways only tangentially related to the apical effect(s). Thus, this analysis might not be limited to supplemental evidence relevant to the assessment PECO criteria but could also include evaluations of biological pathways; for example, those evaluations based on chemical-nonspecific information of the health effect, or those that have been established for other, possibly related chemicals. When interpreting the mechanistic evidence, some considerations include:

- Are the hypothesized MOA(s) biologically plausible, considering the chemical's pharmacokinetic processes, the biological processes known to contribute to the health effect, and the biological or experimental support for connections between mechanistic events? Consider consistency with established MOAs for related agents.
- Are there mechanistic key events potentially related to the health effects of interest? Is there dose-response information supporting linkages between identified key events in the biological pathway leading to the adverse health effect? Key events, if sufficiently supported by the available evidence, may be considered for use in dose-response analyses.
- Do independent studies and different experimental hypothesis-testing approaches identify key events in the MOA(s) that have been demonstrated to be associated with the health effect in question? What is the directness of this association (e.g., if blocking a key event supported by strong chemical-specific evidence reduces or prevents the appearance of the health effect, this provides a very high level of certainty)? MOA hypotheses or key events that have been shown to be reproducible in different species, populations, model systems, or laboratories strengthen confidence in the validity of an MOA.
- Are there key events in the biological pathway (or known consequences of mechanistic events that have been clearly demonstrated to occur after exposure) that were not observed despite well-designed, appropriate studies? This can reduce confidence in an MOA and certainty in the evidence integration judgment.
- How well do key events in the MOA correlate with the health effect, in terms of temporality and dose-response concordance? For example, do key events precede the appearance of the health effect (e.g., with shorter exposure durations or lower exposure levels)? If not, is this explainable (e.g., consider detection sensitivity or susceptibility)?
- How well does the MOA explain demonstrated differences across health effect studies (e.g., by sex, timing of exposure)? If there are major unexplainable differences, this could indicate that the agent produces effects other than those hypothesized, or that other

pathways are being activated. This might warrant separate evaluations. Is the appearance of some effects inconsistent with the proposed MOA (e.g., the appearance of treatment-related kidney tumors in female rats and/or mice of either sex would be inconsistent with an $\alpha 2u$ globulin MOA being solely operative in rodent tumorigenesis)?

Application of organizing constructs to evidence integration

Several existing organizing constructs developed by EPA and others for the assessment of mechanistic events and the MOA of an agent—including consideration of biological plausibility, human relevance, and the identification of susceptible subpopulations—are available.¹⁶ The type of organizing construct will be assessment specific and likely dependent on the pathway or outcome evaluated (e.g., cancer vs. noncancer outcomes). Below, several existing frameworks for assessing mechanistic data are presented for consideration. Scoping and refinement of the evaluation plan will indicate an MOA for some health hazards (i.e., cancer), whereas for other outcomes, informative mechanistic information might not be available or there might not be evidence to indicate that the human and animal findings warrant mechanistic evaluation (i.e., no significant findings were observed or there is evidence of null effects). The latter type of mechanistic information can be included in the synthesis; however, the available human and animal data may be limited to the extent that an MOA is not feasible or practical for the purpose of an overall assessment of causality between exposure to outcome. A synthesis of mechanistic events is part of an MOA analysis and overall cross-stream evidence integration that includes a variety of factors (see Tables 6-6 and 6-7) to determine whether the available data for a chemical's effects can support a proposed MOA(s) for the toxic effect(s) of an agent.

The analysis of assessment-specific health effects evidence culminates in an evidence integration judgment and narrative for each potential health effect category (i.e., each noncancer health effect and specific type of cancer, or broader grouping of related outcomes as defined in the evaluation plan. Organizing the health effects evidence narrative and overall judgment using an integrative construct may be pertinent to the assessment of the overall mechanistic evidence. The selection of organizing construct can depend on the hazard (e.g., cancer vs. noncancer); several example constructs are included below.

- 1) The 2005 EPA Cancer Guidelines: The cancer guidelines were developed in conjunction with efforts by the World Health Organization (WHO) International Programme on Chemical Safety (IPCS) to harmonize the approaches used to assess the risk of cancer ([IPCS, 2007a](#)) and noncancer ([IPCS, 2007b](#)) outcomes from chemical exposures by establishing an MOA framework based on modified Bradford Hill considerations for causality. Consideration of the evidence strength, consistency, specificity of association, dose-response concordance, temporal relationship, biological plausibility, and coherence are described in [U.S. EPA \(2005a\)](#) and can be very useful for constructing an effective narrative of evidence linking exposure to toxic effects. These considerations are not a checklist; no one aspect is either

¹⁶*Guidelines for Carcinogenic Risk Assessment* ([U.S. EPA, 2005a](#)); Integrated Science Assessment (ISA) for particulate matter ([U.S. EPA, 2019b](#)); Preamble to the Integrated Science Assessments ([U.S. EPA, 2015c](#)); Assessing Health Risk of Environmental Exposures to Children ([U.S. EPA, 2006b](#)).

necessary or sufficient for drawing inferences of causality ([U.S. EPA, 2005a](#)). Rather, these considerations should be used to emphasize strength (or the lack thereof) in the mechanistic evidence.

- 2) Other EPA frameworks for organizing biological plausibility: Although much emphasis in this chapter is placed on the cancer guidelines ([U.S. EPA, 2005a](#)), the same concept can be applied to noncancer health effects. Similar frameworks in EPA guideline documents ([U.S. EPA, 1998, 1996, 1991a](#)) can be specified and used as an effective framework for a narrative of evidence linking exposure to toxic effects. These frameworks and syntheses should also include consideration of the evidence strength, consistency, specificity of association, dose-response concordance, temporal relationship, biological plausibility, and coherence.
- 3) Adverse outcome pathways (AOPs): AOPs have become functional and versatile tools for use in the risk assessment workflow. AOPs organize the sequential connections of empirically measured key events between a single molecular initiating event and an adverse outcome. AOPs are focused on the molecular initiating events rather than exposure to a specific chemical, although they establish, when possible, a quantitative understanding of the key event relationships that describe the progression from one key event to the next ([Villeneuve et al., 2014a, b](#)). AOPs provide a clear visual representation and integrative construct for organizing the more complex relationships and associations described in an MOA. Thus, the outcomes from other exposures with similar molecular initiating events or key events can be predicted from measurable upstream events.

6.2.2. Evidence Integration Judgment

Using a structured framework approach, one of five phrases is used to summarize the evidence integration judgment based on the integration of the evidence synthesis judgments (see Section 6.1), taking into account the additional considerations assessed across evidence streams (see Section 6.2.1): ***evidence demonstrates***, ***evidence indicates (likely)***, ***evidence suggests***, ***evidence is inadequate***, or ***strong evidence supports no effect*** (see Table 6-7). The five evidence integration judgment levels reflect the differences in the amount and quality of the data that inform the evaluation of whether exposure is interpreted as capable of causing the health effect(s). As it is assumed that any identified health hazards will only be manifest given exposures of a certain type and amount (e.g., a specific route; a minimal duration, periodicity, and level), the evidence integration narrative and summary judgment levels include the generic phrase, “given sufficient exposure conditions.” This highlights that, for those assessment-specific health effects identified as potential hazards, the exposure conditions associated with those health effects will be defined (as will the uncertainties in the ability to define those conditions) during dose-response analysis (see Chapter 8). More than one evidence integration judgment level can be used when the evidence base is able to support that a chemical’s effects differ by exposure level or route ([U.S. EPA, 2005a](#)). The analyses and judgments are summarized in the evidence profile table (see Tables 6-1 and 6-2).

For evaluations of carcinogenicity, consistent with EPA’s Cancer Guidelines ([U.S. EPA, 2005a](#)), one of EPA’s standardized cancer descriptors is used to describe the overall potential for carcinogenicity within the evidence integration narrative for carcinogenicity. These descriptors are: (1) ***carcinogenic to humans***, (2) ***likely to be carcinogenic to humans***, (3) ***suggestive evidence of***

carcinogenic potential, (4) inadequate information to assess carcinogenic potential, or (5) not likely to be carcinogenic to humans. The standardized cancer descriptors will often align with the evidence integration judgments (i.e., “**evidence demonstrates**” aligns with “carcinogenic to humans”) but not in all cases. For example, the evidence integration judgments are generally used for individual tumor or cancer types and the standardized EPA descriptors are used to characterize overall cancer hazard.

For each type of cancer evaluated (e.g., lung cancer; renal cancer) or sets of related cancer types, an evidence integration narrative and summary judgment level are provided as described above for noncancer health effects. When considering evidence on carcinogenicity across human and animal evidence, site concordance is not required (U.S. EPA, 2005a). If a systematic review of more than one cancer type was conducted, then the strongest evidence integration judgment(s) is used as the basis for selecting the standardized cancer descriptor in accordance with the EPA cancer guidelines (U.S. EPA, 2005a), including application of the MOA framework (incorporating an evaluation of evidence relevant to potential mutagenicity).

Similar to the description for summarizing noncancer judgments above, the cancer descriptor and evidence integration narrative for carcinogenicity also consider the conditions of carcinogenicity, including exposure (e.g., route; level) and susceptibility (e.g., genetics; lifestage), as the data allow (Farland, 2005; U.S. EPA, 2005a, b). As with noncancer effects, the specific exposure conditions necessary for carcinogenicity are further defined during dose-response analysis (see Chapter 8).

Table 6-7. Framework for summary evidence integration judgments in the evidence integration narrative

Summary evidence integration judgment ^a in narrative	Evidence integration judgment level	Explanation and example scenarios ^b
The currently available evidence demonstrates that [chemical] causes [health effect] in humans ^c given sufficient exposure conditions. This conclusion is based on studies of [humans or animals] that assessed [exposure or dose] levels of [range of	Evidence demonstrates	A strong evidence base demonstrating that [chemical] exposure causes [health effect] in humans. <ul style="list-style-type: none"> • This conclusion level <u>is</u> used if there is <i>robust</i> human evidence supporting an effect. • This conclusion level <u>could also be</u> used with <i>moderate</i> human evidence and <i>robust</i> animal evidence if there is strong mechanistic evidence that MOAs and key precursors identified in animals are anticipated to occur and progress in humans.

Summary evidence integration judgment ^a in narrative	Evidence integration judgment level	Explanation and example scenarios ^b
concentrations or specific cutoff level concentration ^d].		
The currently available evidence indicates that [chemical] likely causes [health effect] in humans given sufficient exposure conditions. This conclusion is based on studies of [humans or animals] that assessed [exposure or dose] levels of [range of concentrations or specific cutoff level concentration].	Evidence indicates (likely^e)	<p>An evidence base that indicates that [chemical] exposure likely causes [health effect] in humans, although there might be outstanding questions or limitations that remain, and the evidence is insufficient for the higher conclusion level.</p> <ul style="list-style-type: none"> • This conclusion level <u>is</u> used if there is <i>robust</i> animal evidence supporting an effect and <i>slight-to-indeterminate</i> human evidence, or with <i>moderate</i> human evidence when strong mechanistic evidence is lacking. • This conclusion level <u>could also be</u> used with <i>moderate</i> human evidence supporting an effect and <i>moderate-to-indeterminate</i> animal evidence, or with <i>moderate</i> animal evidence supporting an effect and <i>moderate-to-indeterminate</i> human evidence. In these scenarios, any uncertainties in the <i>moderate</i> evidence are not sufficient to substantially reduce confidence in the reliability of the evidence, or mechanistic evidence in the <i>slight</i> or <i>indeterminate</i> evidence base (e.g., precursors) exists to increase confidence in the reliability of the <i>moderate</i> evidence.
The currently available evidence suggests that [chemical] might cause [health effect] in humans. This conclusion is based on studies of [humans or animals] that assessed [exposure or dose] levels of [range of concentrations or specific cutoff level concentration].	Evidence suggests	<p>An evidence base that suggests that [chemical] exposure might cause [health effect] in humans, but there are very few studies that contributed to the evaluation, the evidence is very weak or conflicting, or the methodological conduct of the studies is poor.</p> <ul style="list-style-type: none"> • This conclusion level <u>is</u> used if there is <i>slight</i> human evidence and <i>indeterminate-to-slight</i> animal evidence. • This conclusion level <u>is</u> also used with <i>slight</i> animal evidence and <i>indeterminate-to-slight</i> human evidence. • This conclusion level <u>could also be</u> used with <i>moderate</i> human evidence and <i>slight</i> or <i>indeterminate</i> animal evidence, or with <i>moderate</i> animal evidence and <i>slight</i> or <i>indeterminate</i> human evidence. In these scenarios, there are outstanding issues or uncertainties regarding the <i>moderate</i> evidence (i.e., the synthesis judgment was borderline with <i>slight</i>), or mechanistic evidence in the <i>slight</i> or <i>indeterminate</i> evidence base (e.g., null results in well-conducted evaluations of precursors) exists to decrease confidence in the reliability of the <i>moderate</i> evidence. • Exceptionally, when there is general scientific understanding of mechanistic events that result in a health effect, this conclusion level <u>could also be</u> used if there is strong mechanistic evidence that is sufficient to highlight potential human toxicity^f—in the

Summary evidence integration judgment ^a in narrative	Evidence integration judgment level	Explanation and example scenarios ^b
		absence of informative conventional studies in humans or in animals (i.e., <i>indeterminate</i> evidence in both).
The currently available <i>evidence is inadequate</i> to assess whether [chemical] might cause [health effect] in humans.	<i>Evidence inadequate</i>	<p>This conveys either a lack of information or an inability to interpret the available evidence for [health effect]. On an assessment-specific basis, a single use of this “inadequate” conclusion level might be used to characterize the evidence for multiple health effect categories (i.e., all health effects that were examined and did not support other conclusion levels).^g</p> <ul style="list-style-type: none"> • This conclusion level <u>is</u> used if there is <i>indeterminate</i> human and animal evidence. • This conclusion level <u>could also be</u> used with <i>slight-to-robust</i> animal evidence and <i>indeterminate</i> human evidence if strong mechanistic information indicates that the animal evidence is unlikely to be relevant to humans. • This conclusion level <u>could also be</u> used with <i>compelling evidence of no effect</i> in human studies and <i>moderate-to-robust</i> animal evidence if there is not strong mechanistic information that the animal evidence is unlikely to be relevant to humans. • A conclusion of inadequate is not a determination that the agent does not cause the indicated health effect(s). It simply indicates that the available evidence is insufficient to reach conclusions.
<i>Strong evidence supports no effect</i> in humans. This conclusion is based on studies of [humans or animals] that assessed [exposure or dose] levels of [range of concentrations].	<i>Strong evidence supports no effect</i>	<p>This represents a situation in which extensive evidence across a range of populations and exposure levels has identified no effects/associations. This scenario requires a <i>high</i> degree of confidence in the conduct of individual studies, including consideration of study sensitivity, and comprehensive assessments of the endpoints and lifestages of exposure relevant to the health effect of interest.</p> <ul style="list-style-type: none"> • This conclusion level <u>is</u> used if there is <i>compelling evidence of no effect</i> in human studies and <i>compelling evidence of no effect</i> in animal studies-to-<i>slight</i> animal evidence. • This conclusion level <u>is</u> also used if there is <i>indeterminate</i> human evidence and <i>compelling evidence of no effect</i> in animal models concluded to be relevant to humans. • This conclusion level <u>could also be</u> used with <i>compelling evidence of no effect</i> in human studies and <i>moderate-to-robust</i> animal evidence if strong mechanistic information indicates that the animal evidence is unlikely to be relevant to humans.

MOA = mode of action.

^gEvidence integration judgments are typically developed at the level of the health effect when there are sufficient studies on the topic to evaluate the evidence at that level; this should always be the case for “***evidence demonstrates***” and “***strong***”

evidence supports no effect,” and typically for *“evidence indicates (likely).”* However, some databases only allow for evaluations at the category of health effects examined; this will more frequently be the case for conclusion levels of *“evidence suggests”* and *“evidence inadequate.”* A judgment of *“strong evidence supports no effect”* is drawn at the health effect level.

^bTerminology of “is” refers to the default option; terminology of “could also be” refers to situational options dependent on mechanistic understanding. Scenarios with “could also be” typically reflect situational decisions dependent on the results of evaluating the additional evidence integration considerations outlined in Section 6.2.1 (e.g., human relevance of findings).

^cIn some assessments, these conclusions might be based on data specific to a particular lifestage of exposure, sex, or population (or another specific group). In such cases, this would be specified in the narrative conclusion, with additional detail provided in the narrative text. This applies to all conclusion levels.

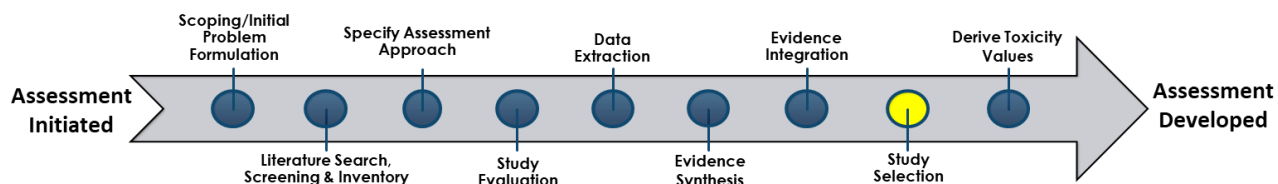
^dIf concentrations cannot be estimated, an alternative expression of exposure level such as “occupational exposure levels,” is provided. This applies to all conclusion levels.

^eFor some applications, such as benefit-cost analysis, to better differentiate the categories of *“evidence demonstrates”* and *“evidence indicates (likely),”* the latter category should be interpreted as evidence that supports an exposure-effect linkage that is likely to be causal.

^fScientific understanding of adverse outcome pathway (AOP) and of the human implications of new toxicity testing methods (e.g., from high-throughput screening, from short-term in vivo testing of alternative species or from new in vitro testing) will continue to increase. This might make possible the development of hazard conclusions when there are mechanistic or other relevant data that can be interpreted with a similar level of confidence to positive animal results in the absence of conventional studies in humans or in animals.

^gSpecific narratives for each of these health effects could also be deemed unnecessary.

7. HAZARD CONSIDERATIONS AND STUDY SELECTION FOR DERIVING TOXICITY VALUES



Purpose

- Summarize and apply the hazard identification judgments to prioritize outcomes and select studies, among those that characterize each health hazard, for use in deriving human toxicity values.

The previous chapters outline principles that support the transparent identification of health outcomes for which human toxicity values are needed and identification of the most important studies from which to derive these toxicity values. The derivation of reference values and cancer risk estimates depends on the nature of the health hazard conclusions drawn during evidence integration (see Chapter 6). When suitable data are available, as described in this chapter, toxicity values should always be developed for evidence integration conclusions of *evidence demonstrates* and *evidence indicates (likely)* and for carcinogenicity descriptors of carcinogenic to humans or likely carcinogenic to humans.

In general, toxicity values would not be developed for noncancer or cancer hazards with *evidence suggests* or suggestive evidence of carcinogenicity conclusions, respectively. However, for these scenarios a value might be useful for some purposes when the evidence includes a well-conducted study (particularly when that study could also demonstrate a credible concern for greater toxicity in a susceptible population or lifestage). For example, *evidence suggests* could be based on either a single high or medium confidence study or multiple low confidence studies. In the former case, a value could be developed. The U.S. Environmental Protection Agency (EPA) Cancer Guidelines ([U.S. EPA, 2005a](#)) discuss such evidence scenarios and the potential use of toxicity values derived in these scenarios: “When there is suggestive evidence [of carcinogenicity], the Agency generally would not attempt a dose-response assessment, as the nature of the data generally would not support one; however, when the evidence includes a well-conducted study, quantitative analyses might be useful for some purposes, for example, providing a sense of the magnitude and uncertainty of potential risks, ranking potential hazards, or setting research priorities. In each case, the rationale for the quantitative analysis is explained, considering the uncertainty in the data and the suggestive nature of the weight of evidence. These analyses

generally would not be considered Agency consensus estimates ([U.S. EPA, 2005a](#)).” Toxicity values should not be developed for other evidence integration judgments (i.e., ***evidence suggests, strong evidence supports no effect, inadequate information to assess carcinogenic potential, or not likely to be carcinogenic to humans***).

As discussed in Section 7.1, selection of specific endpoints for toxicity value derivation is primarily a result of the hazard characterization. Ideally, the hazard synthesis and integration has clarified any important considerations, including mechanistic understanding, that would indicate the use of particular dose-response models, including chemical-specific or biologically based models, over more generic models (see Chapter 8). These considerations also include whether linked health effects within and between organ systems should be characterized together, as well as whether there is suitable mechanistic information to support combining related outcomes or to identify internal dose measures that could differ among outcomes (generally for animal studies). Section 7.2 builds on these considerations, as well as general principles of dose-response analysis, to prioritize the studies most appropriate for use in deriving toxicity values.

7.1. HAZARD CONSIDERATIONS FOR DOSE-RESPONSE

The section of the assessment between the hazard identification and dose-response chapters provides a transition from hazard identification to dose-response analysis, highlighting information that (1) informs the selection of outcomes or broader health effect categories for which toxicity values will be derived; (2) helps determine whether toxicity values can be derived to protect specific populations or lifestages; (3) describes how dose-response modeling will be informed by pharmacokinetic data; and (4) aids the identification of biologically based benchmark response (BMR) levels. The pool of informative outcomes and study-specific findings (e.g., summarized in evidence profile tables) is used to identify which categories of effects and study designs are considered the strongest and most appropriate for quantitative dose-response assessment of a given health effect. Health effects from studies that utilized exposure levels within or closer to the range of exposures encountered in the environment are particularly informative. When there are multiple endpoints for an organ/system, considerations for characterizing the overall impact on this organ/system should be discussed. For example, if there are multiple histopathological alterations relevant to liver function changes, liver necrosis can be selected as the most representative endpoint to consider for dose-response analysis. This section can review or clarify which endpoints or combination of endpoints in each organ/system characterize the overall effect for dose-response analysis. For cancer types, consideration is given to deciding whether and how to develop quantitative estimate(s) across multiple types of cancer. Similarly, multiple tumor types (if applicable) will be discussed, and a rationale given for any grouping.

Biological considerations important for dose-response analysis (e.g., that could help with selection of a BMR) should also be discussed. The impact of route of exposure on toxicity to different organs/systems will be examined, if appropriate and as possible. The existence and

validity of physiologically based pharmacokinetic (PBPK) models or pharmacokinetic information that could allow the estimation of internal dose for route-to-route extrapolation should be presented and used if appropriate (see Chapter 8 for more details). In addition, mechanistic evidence influential to the dose-response analyses should be highlighted, for example evidence related to susceptibility or potential shape of the dose-response curve (i.e., linear, nonlinear, or threshold model).

The hazard considerations for dose-response section also summarizes the evidence (i.e., human, animal, mechanistic) regarding populations and lifestages that appear to be susceptible to the health hazards identified and factors that increase risk of developing (or exacerbating) these health effects, depending on the available evidence. This section should include a discussion of the populations that might be susceptible to the health effects identified to be hazards of exposure to the assessed chemical, even if there are no specific data on effects of exposure to that chemical in the potentially susceptible population. In addition, if there is evidence or an expectation that susceptibility could be conferred by lifestage, this should be explicitly discussed. Differences in absorption, distribution, metabolism, and excretion (ADME) can be conferred by lifestage, sex, or genetic variability, which can result in differences in key metabolic pathways, and the form or amount of the toxic moiety that interacts with target molecules and tissues. Background information about biological mechanisms or ADME, as well as biochemical and physiological differences among lifestages, can be used to guide the selection of populations and lifestages to consider. At a minimum, particular consideration should be given to infants and children, pregnant women, and women of childbearing age. Evidence on factors that might confer susceptibility (see below) is typically summarized and evaluated with respect to patterns across studies pertinent to consistency, coherence, and the magnitude and direction of effect measures. Relevant factors could include intrinsic factors (e.g., age, sex, genetics, health or nutritional status, behaviors), extrinsic factors (e.g., socioeconomic status, access to health care), and differential exposure levels or frequency (e.g., occupation-related exposure, residential proximity to locations with greater exposure intensity). If studies directly addressing identified susceptibilities are unusable for quantitative analyses, susceptibility data might still support refined human variability uncertainty factors or probabilistic uncertainty analyses. Table 7-1 provides a partial list of examples that could define a susceptible population or lifestage.

There could be a variety of logical approaches to the organization of the analysis of susceptibility. The evidence is drawn from discussions in the hazard sections for specific outcomes, although some additional details from the studies might need to be highlighted in this section. The section should explicitly consider options for using data related to susceptible populations to impact dose-response analysis. An attempt should be made to highlight where it might be possible to use identified data to develop separate risk estimates for a specific population or lifestage, or if evidence is available to select a data-derived uncertainty factor.

Table 7-1. Factors that can increase susceptibility to exposure-related health effects

Factor	Examples
Lifestage	In utero, childhood, puberty, pregnancy, women of child-bearing age, and old age
Demographics	Gender, race/ethnicity, education, income level, occupation, and geography
Social determinants	Socioeconomic status, neighborhood factors, health care access, and social, economic, or political inequality
Behaviors or practices	Diet, mouthing, smoking, alcohol consumption, pica, and subsistence or recreational hunting and fishing
Health status	Preexisting conditions or disease such as psychosocial stress, elevated body mass index, frailty, nutritional status, and chronic disease
Genetic variability	Polymorphisms in genes regulating cell cycle, DNA repair, cell division, cell signaling, cell structure, gene expression, apoptosis, and metabolism

DNA = deoxyribonucleic acid.

7.2. SELECTION OF STUDIES

As previously discussed, for both cancer and noncancer hazards, preference is given to health effects (or outcomes) and cancer types with stronger evidence integration conclusions. When more evidence is available, this strength of evidence characterization can also be used to narrow the focus of the dose-response assessment for a given hazard to a particular endpoint(s) or study design(s). In general, all studies identified as influential to drawing the aforementioned judgments are considered for deriving toxicity values (see Chapter 6 for discussion on how different studies can influence the overall judgments); thus, focus should be almost exclusively on *high* or *medium* confidence studies. However, there are additional considerations specific to their use in quantitative analyses, as discussed in Section 7.2.1. It is critical that the decisions and the supporting rationale for the health effects, studies, and endpoints considered (and ultimately selected) for candidate toxicity value derivation are transparently documented in the assessment, typically in summary tables.

7.2.1. SYSTEMATIC ASSESSMENT OF STUDY ATTRIBUTES TO SUPPORT DERIVATION OF TOXICITY VALUES

In addition to the evidence integration considerations described above and the study confidence determinations of the hazard identification, attributes of the studies identified for each hazard are reviewed for additional factors such as relevance of the test species, relevance of the studied exposure to human environmental exposures, quality of measurements of exposure and outcomes, and other aspects of study design (including specific reconsideration of the potential for

bias in the reported association between exposure and outcomes). See Table 7-2 for a general summary of these considerations, which can be further refined based on the specific details of the exposure and hazard under review. Higher confidence studies demonstrating more of the preferred considerations, and those which demonstrate the considerations to a greater extent, are expected to provide more accurate human equivalent toxicity values. Often, studies in an endpoint-specific database (i.e., the body of evidence identified for an endpoint) demonstrate many of the preferred considerations, but in different combinations, so that it is not clear that one data set (i.e., the quantitative data from a single dose-response relationship for a single endpoint from a single study) is the optimal choice; therefore, all data sets should be considered for toxicity value derivation. Further, even studies showing less of the preferred considerations still can be important for toxicity value derivation, depending on the biological significance of the endpoint relative to others, and in light of extrapolations (e.g., interspecies) or uncertainty factors (UFs) that might be relevant (see Section 8.3).

Table 7-2. Attributes used to evaluate studies for derivation of toxicity values

Study attributes		Considerations	
		Human studies	Animal studies
Study confidence		<i>High or medium</i> confidence studies (see Chapter 6) are highly preferred over <i>low</i> confidence studies. The selection of low confidence studies should include an additional explanatory justification (e.g., only low confidence studies had adequate data for toxicity value derivation). The available <i>high</i> and <i>medium</i> confidence studies are further differentiated on the basis of the study attributes below, as well as a reconsideration of the specific limitations identified and their potential impact on dose-response analyses.	
Rationale for choice of species		Human data are preferred over animal data to eliminate interspecies extrapolation uncertainties (e.g., in pharmacodynamics, dose-response pattern in relevant dose range, relevance of specific health outcomes to humans).	Animal studies provide supporting evidence when adequate human studies are available, and they are considered the studies of primary interest when adequate human studies are not available. For some hazards, studies of particular animal species known to respond similarly to humans would be preferred over studies of other species.
Relevance of exposure paradigm	Exposure route	Studies involving human environmental exposures (oral, inhalation).	Studies by a route of administration relevant to human environmental exposure are preferred. A validated pharmacokinetic model can also be used to extrapolate across exposure routes.
	Exposure durations	When developing a chronic toxicity value, chronic or subchronic studies are preferred over studies of acute exposure durations. Exceptions exist, such as when a susceptible population or lifestage is more sensitive in a particular time window (e.g., developmental exposure).	
	Exposure levels	Exposures near the range of typical environmental human exposures are preferred. Studies with a broad exposure range and multiple exposure levels are preferred to the extent that they can provide information about the shape of the exposure-response relationship (see the EPA <i>Benchmark Dose Technical Guidance</i> , §2.1.1) and facilitate extrapolation to more relevant (generally lower) exposures.	
Subject selection		Studies that provide risk estimates in the most susceptible groups are preferred.	
Controls for possible confounding ^a		Studies with a design (e.g., matching procedures, blocking) or analysis (e.g., covariates or other procedures for statistical adjustment) that adequately address the relevant sources of potential critical confounding for a given outcome are preferred.	

Study attributes	Considerations	
	Human studies	Animal studies
Measurement of exposure	Studies that can reliably distinguish between levels of exposure in a time window considered most relevant for development of a causal effect are preferred. Exposure assessment methods that provide measurements at the level of the individual and that reduce measurement error are preferred. Measurements of exposure should not be influenced by knowledge of health outcome status.	Studies providing actual measurements of exposure (e.g., analytical inhalation concentrations vs. target concentrations) are preferred. Relevant internal dose measures might facilitate extrapolation to humans, as would availability of a suitable animal PBPK model in conjunction with an animal study reported in terms of administered exposure.
Health outcome(s)	Studies that can reliably distinguish the presence or absence (or degree of severity) of the outcome are preferred. Outcome ascertainment methods using generally accepted or standardized approaches are preferred.	
	Studies with individual data are preferred in general. For example, individual data allow you to characterize experimental variability more realistically and to characterize overall incidence of individuals affected by related outcomes (e.g., phthalate syndrome).	
	Among several relevant health outcomes, preference is generally given to those outcomes with less concerns for indirectness or with greater biological significance.	
Study size and design	Preference is given to studies using designs reasonably expected to have power to detect responses of suitable magnitude. ^b This does not mean that studies with substantial responses but low power would be ignored, but that they should be interpreted in light of a confidence interval or variance for the response. Studies that address changes in the number at risk (through decreased survival, loss to follow-up) are preferred.	

PBPK = physiologically based pharmacokinetic.

^aIn epidemiological studies, this is an exposure or other variable that is associated with both exposure and outcome but is not an intermediary between the two. Although the potential for confounding is considered during evaluations of study confidence (see Chapter 6), some aspects (e.g., covariate-adjusted effect estimates) are important to reconsider for developing more informative quantitative estimates.

^bPower is an attribute of the design and population parameters, based on a concept of repeatedly sampling a population; it cannot be inferred post hoc using data from one experiment ([Hoening and Heisey, 2001](#)).

Typically, candidate toxicity values are derived from each data set selected, and the specific attributes for each chemical and health endpoint as evaluated here are balanced in selecting final toxicity values (see Section 8.5). In some cases, if there are many data sets in an endpoint-specific database, the number of studies considered for toxicity value derivation can (and should) be reduced to a specified subset of suitable studies—e.g., only studies involving exposures near environmental exposure levels as opposed to those using only very high exposures, or only studies demonstrating the most sensitive effects among those of most concern for humans.¹⁷ The rationale for focusing on the particular subset, and distinguishing between studies included and excluded in the subset, is generally articulated in a study selection summary table.

In some cases, a common effect measure reported by some or all studies in a database can be used in a meta-analysis to provide a more precise estimate, and better understanding of the magnitude of effect, than could be achieved by estimates from individual studies. It might also be possible to derive a toxicity value by combining suitable studies in an endpoint-specific database in a meta-regression or dose-response meta-analysis [e.g., combining male and female responses for the same outcome from the same study, or combining several similar experiments conducted in the same laboratory; §2.1.6 ([U.S. EPA, 2012b](#))], as described further in Section 7.2.2.

In addition to the more general considerations described above, specific statistical issues could impact the feasibility of dose-response modeling for individual data sets, such as the lack of variability measures for continuous data; these issues are described in more detail in the *Benchmark Dose Technical Guidance*, §2.1.4, ([U.S. EPA, 2012b](#)). Several important considerations from the Benchmark Dose Technical Guidance concerning the levels and patterns of response observed across treatment groups are highlighted below.

- Data sets that are most useful for dose-response analysis generally have at least one exposure level in the region of the dose-response curve near the BMR (the response level to be used for estimating a point of departure [POD] to derive a toxicity value), to minimize low-dose extrapolation, and more exposure levels and larger sample sizes overall ([U.S. EPA, 2012b](#)). These attributes support a more complete characterization of the shape of the exposure-response curve and decrease the uncertainty in the associated exposure-response metric (e.g., inhalation unit risk or reference concentration [RfC]) by reducing statistical uncertainty in the POD and minimizing the need for low-dose extrapolation.
- The minimum data set to be used for estimating the benchmark dose (BMD) and benchmark dose lower confidence limit (BMDL) should show a biologically or statistically significant dose-related trend in response for the selected endpoint(s) [see §2.1.5 and Figure 2A ([U.S. EPA, 2012b](#))]. Within an endpoint-specific evidence stream, studies showing no or very weak responses, but judged to be consistent or coherent with studies showing stronger

¹⁷Note that no-observed-adverse-effect levels/lowest-observed-adverse-effect levels (NOAELs/LOAELs) are generally not useful for choosing between studies for dose-response assessment. The apparent relative sensitivities of endpoints based on NOAELs/LOAELs generally do not correspond to the same relative sensitivities based on benchmark doses (BMDs) or benchmark dose lower confidence levels (BMDLs), because NOAELs/LOAELs do not correspond to similar response levels across studies of the same endpoints ([U.S. EPA, 2012b](#)).

responses (e.g., because of differences in study design such as exposure levels or sensitivity), generally would not support their own toxicity value derivations in an assessment that generates study-by-study values. However, such studies could be included in any meta-regressions or meta-analyses, with appropriate incorporation of the noted differences in study confidence evaluation or other relevant attributes (see Section 7.2.2).

- In cases where the biological significance of a response is not well understood, statistical significance often supports identifying an endpoint suitable for dose-response assessment. In cases of elevated responses without a statistically significant trend (monotonic trends in rare endpoints, adverse endpoints in studies with low power), biological significance could be inferred from other data on the same chemical and endpoint [see §2.1.5 and Figure 2A ([U.S. EPA, 2012b](#))].
- Dose-response analysis might not be supported if only the highest treatment group shows a response different from controls (the major concern in situations like this is that there is a lack of data between the high dose and next tested dose to inform the shape of the dose-response models, and this leads to model uncertainty) [see §2.1.5 and Figures 2A and 2B ([U.S. EPA, 2012b](#))]. If the one elevated response is near the BMR, however, adequate BMD and BMDL computation might result ([Kavlock et al., 1996](#)). Also, fitting multiple models to the data set can help evaluate the magnitude of uncertainty regarding BMD and BMDL estimates.
- Data sets in which *all* the exposure levels show significantly (see previous bullets) elevated responses compared with controls (i.e., a no-observed-adverse-effect level [NOAEL] is not identified) are generally useable in dose-response analyses, with the possible exception of those with a relatively high response at the lowest exposure [see §2.1.5 ([U.S. EPA, 2012b](#))]. In this situation, depending on the needs of the assessment, low-dose extrapolation might be too uncertain, and a lowest-observed-adverse-effect level (LOAEL) would likely need to be identified.
- Responses exhibiting nonmonotonic exposure-response relationships should not necessarily be excluded from the analysis. For example, a diminished response at higher exposure levels, suggesting a nonmonotonic relationship, might be satisfactorily explained by factors such as competing toxicity, saturation of absorption or metabolism, exposure misclassification, or selection bias [see §2.3.6 ([U.S. EPA, 2012b](#))].

In cases where dose-response modeling is not feasible or involves substantial uncertainty (see points discussed above), the NOAEL/LOAEL approach might still be applicable in selection of PODs [see §2.1.5 and Figure 2A ([U.S. EPA, 2012b](#))]. In addition to providing a thorough rationale for the data sets selected for dose-response analysis or NOAEL/LOAEL identification, reasons for not analyzing particular studies or data sets quantitatively should be documented with discussion of the impact on the overall toxicity value derivation of excluding any data sets judged not suitable for dose-response analysis.

7.2.2. COMBINING DATA FOR DOSE-RESPONSE MODELING

This section discusses general considerations for combining dose-response data for the same endpoint across more than one study (or across multiple subgroups within a study, e.g., males

and females) into one overall analysis. The evaluation of study strengths and similarities described above (see Section 7.2.1) is essential for supporting such a combined analysis and would ideally be considered at the start of the dose-response modeling phase of an assessment. This type of analysis can be conducted with group-level data, or when available, with individual-level data. One situation in which combining data is often reasonable occurs when responses in different subgroups of one study—such as males and females—do not differ materially for the same outcome. If the dose-response data are very similar, it might be desirable to combine the data to obtain more precise estimates of PODs [see the Integrated Risk Information System (IRIS) assessment of tetrachloroethylene (U.S. EPA, 2012c) for example (Swartout, 2009; Allen et al., 1996; Stiteler et al., 1993; Vater et al., 1993)]. Alternatively, a covariate might be included in the combined analysis to account for any group differences.

When there are multiple studies deemed adequate for the same outcome, candidate PODs typically will be derived individually based on data from each study. The magnitude of an effect might differ among these data sets based on biological or study design differences. Sources of potential heterogeneity across studies include laboratory procedures used (e.g., type of assay), population, animal species or strain studied, sex, and route of exposure. It might be possible, however, to conduct dose-response modeling that combines data from multiple studies, accounting for study-specific characteristics (e.g., by inclusion of covariates or statistical weights), resulting in a single POD based on multiple data sets (i.e., meta-regression). This might increase the precision of the estimated POD and could be useful for quantifying the impact of specific sources of heterogeneity. Considerations for judging whether studies are potentially suitable to derive a POD based on combining multiple data sets include the following.

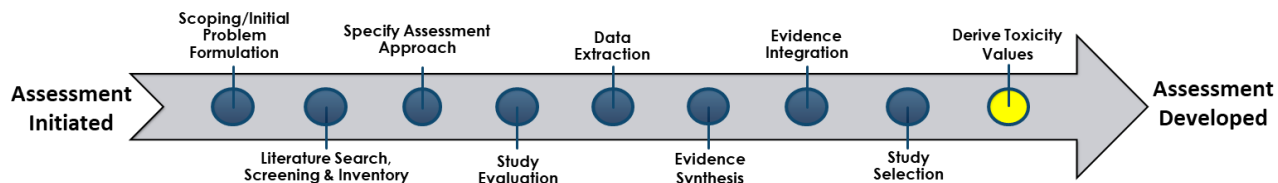
- *In addition to the established study confidence, does the study support POD derivation (see Section 7.2.1)?* Note that statistical precision (e.g., study size or number of treatment groups) for any one study should not be a consideration for this question, as it can be automatically accounted for by statistical weighting. Indeed, one of the reasons for considering combining data sets might be to increase the overall precision in the POD.
- *Is a common endpoint of concern reported?* Note that “common endpoint” in this case refers to the same specific outcome measurement, not just any endpoint in a common target organ. An exception might be, for example, a categorical regression analysis of endpoints within a target system that are amenable to severity categorization, particularly for (but not necessarily limited to) endpoints that represent progressive effects in the same adverse outcome pathway (AOP).
- *Is a common measure of exposure available?* In the absence of a common measure of exposure, a validated PBPK model might be useful for estimating a common (internal) dose measure, particularly across routes of exposure.
- *Is there evidence of homogeneous responses to exposure?* Species, sexes, and lifestages often differ in dose-response, so convincing evidence of similar responses would be needed to consider combining the data from these groups. For example, a hypothesis test of no

difference across groups can be performed to evaluate possible heterogeneity, based on the dose-response model that best fits the pooled data. A likelihood ratio test that compares the fit of the pooled data to the fits of the individual groups can be used [e.g., [Stiteler et al. \(1993\)](#)].

- Other aspects of the studies, including study duration and confidence level, should also be considered, and incorporated into the analysis as warranted. Statistical significance or other criteria based on the study results should not be used for selecting studies (i.e., studies with null findings should not be excluded).

If potentially suitable data sets are available, statistical and relevant subject area experts (e.g., in epidemiology or toxicology) should confer to evaluate support for combining data sets, and if data sets are combined, what modeling approaches to employ. Specific criteria for such evaluations will depend on the design of the underlying studies and the sources of potential heterogeneity. Statistical testing results could be considered among inclusion criteria, but a lack of statistical significance might be less important than any biological differences that should be addressed in the analysis. Also, all higher confidence studies with either null results or potentially supporting a lack of effect are essential to include. Additional evidence, especially mode-of-action (MOA) data, is useful for supporting a decision whether to combine subgroups in a combined analysis. PBPK models can provide estimates of a common dose measure, further increasing the number of studies that might be combined and leading to greater precision in the POD. Methods in common use for combined data include models that fit a common potency parameter while allowing background response levels to vary (e.g., multiple regression, multivariate analysis, categorical regression).

8. DERIVATION OF TOXICITY VALUES



Purpose

- Derive toxicity values (e.g., reference doses [RfDs], reference concentrations [RfCs], cancer slope factors, or unit risk values) from chemical and endpoint specific studies using statistical approaches (e.g., dose-response modeling) that support quantitative risk assessment.

This chapter describes the process involved in deriving toxicity values, particularly statistical considerations specific to dose-response analysis. A number of U.S. Environmental Protection Agency (EPA) guidance and support documents provide background for the development of these toxicity values, especially EPA's reference dose (RfD)/reference concentration (RfC) review ([U.S. EPA, 2002b](#)), the EPA *Guidelines for Carcinogen Risk Assessment* ([U.S. EPA, 2005a](#)), and the EPA *Supplemental Guidance for Assessing Susceptibility from Early-Life Exposure to Carcinogens* ([U.S. EPA, 2005b](#)). Some familiarity with the development and use of these toxicity values is presumed. As discussed in detail in Section 1.2 of EPA's *Benchmark Dose Technical Guidance* ([U.S. EPA, 2012b](#)), dose-response modeling (i.e., benchmark dose modeling) is the preferred approach for deriving points of departure given several limitations in the no-observed-adverse-effect level/ lowest-observed-adverse-effect level (NOAEL/LOAEL) approach. However, there are situations where benchmark dose (BMD) modeling might not be feasible due to data constraints (see Section 7.2.1) or attempts to model data fail to produce actionable results (see Section 8.2.2). In these cases, the NOAEL/LOAEL approach could be considered on a case-by-case basis for suitability in identifying points of departure.

This chapter highlights topics and principles underlying making thorough use of an environmental agent's database for deriving toxicity values. Specific topics are presented in the order they typically occur in this process and include selecting benchmark response (BMR) values (see Section 8.1), dose characterization and dose-response modeling (see Section 8.2), developing candidate toxicity values (see Section 8.3), characterizing uncertainty and confidence (see Section 8.4), and selecting final toxicity values (see Section 8.5). These topics build from the selection of hazards, studies, and outcomes for dose-response analyses, as discussed in Chapter 7.

8.1. SELECTING BENCHMARK RESPONSE VALUES FOR DOSE-RESPONSE MODELING

When dose-response modeling is feasible and appropriate (see Section 7.2), the BMR that determines the point of departure (POD) for each toxicity value is selected prior to modeling, irrespective of the particular dose-response models under consideration (e.g., multistage). However, BMR selection generally takes into account the type of low-dose extrapolation to be used, linear or nonlinear [see EPA *Guidelines for Carcinogen Risk Assessment* ([U.S. EPA, 2005a](#)) p 1–11, Footnote 3], as discussed further below.

When linear low-dose extrapolation is used (see Section 8.3.1), the result is typically a slope, such as an oral slope factor or an inhalation unit risk, from a point near the low end of the data range to the background response. In this case, the BMR selected does not highly influence the result, so standard BMR values near the low end of the observable range of the data are generally used, such as 10% extra risk for cancer bioassay data and 1% for epidemiological cancer data ([U.S. EPA, 2012b, 2005a](#)). Lower BMR values might be selected in either case to reduce low-dose extrapolation uncertainty if supported by the data.

For nonlinear low-dose extrapolation, the result typically is a reference dose or reference concentration, and both statistical and biological considerations are taken into account when selecting the BMR. For deriving an RfD or RfC, the objective is to determine an exposure level “likely to be without an appreciable risk of deleterious effects during a lifetime,” and the BMR selected should correspond to a low or minimal level of response in a population for the outcome under consideration.¹⁸ The following recommendations for BMR selection for nonlinear low-dose extrapolation (for both human and animal effects) focus on biological considerations, and are for data sets that either contain the response level of interest or involve minimal extrapolation below the observed data.

- For dichotomous data (e.g., presence or absence), a BMR of 10% extra risk is generally used for minimally adverse effects. Lower BMRs (5% or lower) can be selected for severe or frank effects. For example, developmental effects are relatively serious effects, and BMDs derived for these effects could use a 5% extra risk BMR. Developmental malformations considered severe enough to lead to early mortality could use an even lower BMR [see [U.S. EPA \(2012b\)](#), §2.2.1].
- For continuous data, a BMR is ideally based on an established definition of biological significance in the effect of interest. In the absence of such a definition, a difference of one standard deviation (SD) from the mean response of the control mean is often used and one-half the standard deviation is used for more severe effects. Note that the standard deviation used should reflect underlying variability in the outcome to the extent possible,

¹⁸The BMR for an outcome would generally be the same across assessments, reflecting understanding of the outcome rather than the sensitivity of varying study designs. The BMR could change over time, however, based on new data or scientific developments that update the understanding of population response.

separate from variability attributable to laboratory procedures, etc. [see [U.S. EPA \(2012b\)](#), §2.2.2].

- In the case of a nonlinear carcinogen, the outcome of interest would be a key precursor leading to cancer, generally with low severity relative to the ultimate cancer. The points above would apply in selecting a BMR for the precursor.

With respect to statistical considerations, when data sets available for dose-response modeling exhibit response ranges that do not include the BMR, some degree of extrapolation to the BMR is often feasible but must be evaluated on a case-by-case basis. For the most severe effects, such as frank toxicity leading to death, the BMR would ideally be <1% extra risk (i.e., 10^{-6} – 10^{-5}), generally not close enough to observable data for humans or animals to support extrapolation. When extrapolation to the desired BMR is not supported and a more suitable data set is not available (e.g., a precursor effect to the more extreme outcome), the only option is to identify an exposure level that corresponds to a higher response level—either a BMD at a higher BMR, or a LOAEL. In either case, an adjustment for extrapolating to a lower exposure, such as a LOAEL-to-NOAEL uncertainty factor (UF_L), also typically should be used.

In addition to the BMRs outlined above, BMRs of 10% extra risk for dichotomous data and 1 SD difference in the mean response from the control mean for continuous data are recommended for standard reporting purposes across all effects, to facilitate POD comparisons across chemicals or endpoints. A justification should always be provided for each BMR selected. These approaches for selecting BMRs for dichotomous and continuous data are discussed further in the Agency's *Benchmark Dose Technical Guidance* [[U.S. EPA \(2012b\)](#), §2.2].

8.2. CONDUCTING DOSE-RESPONSE MODELING

EPA uses a two-step approach that distinguishes analysis of the observed dose-response data from any inferences about lower exposure levels generally needed to develop toxicity values [[U.S. EPA \(2012b, 2005a\)](#), §3].

- 1) Within the observed range, the preferred approach is to use dose-response modeling to incorporate as much of the data set as possible into the analysis. This modeling yields a POD, an exposure level near the lower end of the observed range of the data, without significant extrapolation to lower exposure levels. Selecting the BMR was discussed in Section 8.1.
- 2) To derive toxicity values, extrapolation below the POD is typically necessary. This step is described further in Section 8.3, "Developing Candidate Toxicity Values."

When both laboratory animal data and human data with sufficient information to perform exposure-response modeling are available, human data are generally preferred for the derivation of toxicity values (see Chapter 7). Key practices are described in Section 8.2.1 for modeling human data and in Section 8.2.2 for modeling animal data.

8.2.1. Exposure-Response Modeling of Human Data

Observational epidemiological studies require evaluation of several attributes, as described in Sections 6.1 and 7.1, before conducting exposure-response modeling. If multiple human studies are suitable for exposure-response modeling and if no single study is judged appreciably better than the others for the purposes of deriving toxicity values, data or results from multiple studies could be combined where justified, or toxicity values might be developed from different studies for comparison.

Cancer Data

Cumulative exposure (or a dose metric that can be converted to cumulative exposure) is generally the preferred exposure metric for cancer responses; exposure estimates can include a lag period, if warranted. Additionally, data on incident cases are generally preferred over mortality data ([U.S. EPA, 2005a](#)), as toxicity values are intended to reflect effect incidences. Adjustments can be made to derive incidence estimates from mortality data, and for some cancers, mortality is a reasonable estimation of incidence. Further discussion of modeling human data can be found in Section 3.2.1 of EPA's *Guidelines for Carcinogen Risk Assessment* ([U.S. EPA, 2005a](#)).

The modeling of cancer epidemiological data typically involves relative risk models. For grouped or categorical exposure data, results might not be sufficiently precise to discern the shape of the exposure-response relationship, and a linear model is often used ([U.S. EPA, 2005a](#)). For individual continuous exposure data, a model such as the Cox proportional hazards model is frequently used because it can easily account for time-dependent and time-independent covariates.

Once an exposure-response model is obtained, the result is applied within a life-table analysis to derive a POD. As noted in Section 8.1, a BMR of 1% extra risk is typically used for relatively common cancers; a lower BMR, for example for less common cancers, might be more suitable for establishing a POD near the lower end of the observed range [[U.S. EPA \(2005a\)](#), §3.2]. Cancer unit risk estimates are derived for individual chemical-associated cancer types that are then generally combined to obtain an overall cancer unit risk estimate [[U.S. EPA \(2005a\)](#); see §2.2.1.1, §3.2.1, §3.3.5; also see Section 8.2.3].

Noncancer Data

Grouped epidemiological data for noncancer effects can be modeled by Benchmark Dose Software (BMDS) models, in the same way as grouped laboratory animal data (see Section 8.2.2). Some situations, such as the need to account for covariates, might call for specialized methods and software. Individual continuous exposure data might similarly involve more specialized models. As with laboratory animal data, BMRs for noncancer effects depend on the effect severity and characteristics of the data set (see Section 8.1 for general recommendations).

In some circumstances with adequate human epidemiological data for noncancer effects, the output of the dose-response analysis might be dose-response functions and associated risk-specific doses, in addition to BMDs and reference values ([NRC, 2013](#)).

8.2.2. Exposure-Response Modeling of Animal Data

Characterization of Exposure for Extrapolation to Humans

This section outlines considerations for characterizing human equivalent exposure levels when deriving risk values from animal data, depending on the extent and complexity of the available data. One useful principle to keep in mind when dose correspondence between animals and humans follows linear relationships is that it is often adequate for this interspecies extrapolation to occur following the estimation of the POD.

The preferred approach for *dose estimation* for dose-response modeling is physiologically based pharmacokinetic (PBPK) modeling because it can incorporate a wide range of relevant chemical-specific information, describe the active agent more accurately, and provide a better basis for extrapolation to human equivalent exposures. To support dose-response modeling for development of toxicity values, optimal absorption, distribution, metabolism, and excretion (ADME) studies underlying PBPK models are those that have been peer reviewed, have been conducted in humans or in the species/strain of animal used in the toxicity study(ies) advanced for dose-response analysis, and have employed a range of doses surrounding the POD. The preferred dose metric would refer to the active agent at the site of its biological effect or to a reliable surrogate measure. The active agent might be the administered chemical or one of its metabolites. Confidence in the use of a PBPK model depends on the robustness of its validation process and the results of sensitivity analyses [[U.S. EPA \(2006a\)](#); [U.S. EPA \(2005a\)](#), §3.1; [U.S. EPA \(1994\)](#), §4.3]. See Section 4.6 for more information.

Use of physiologically based pharmacokinetic (PBPK) models

When a PBPK model supports dose-response modeling, whether using a biologically based model or an empirical curve-fitting model, the most rigorous approach for characterizing dose-response relationships is to use the animal PBPK model to estimate internal doses for each external (applied) exposure, simulating the exposure profile of the bioassay, then use the internal doses in a dose-response analysis to estimate an internal dose metric POD for the animal data. The human PBPK model is then applied to estimate human equivalent concentration (HEC) or human equivalent dose (HED) levels, in terms of external exposure, which result in the same internal dose POD, thereby completing the interspecies extrapolation. This approach might be preferred if the data being modeled are in a nonlinear PBPK range, as it could provide dose-response data that are more amenable to modeling using available dose-response models.

The relationship between internal dose and external exposure is often linear within the range of exposures being modeled. In these cases, it is adequate and simpler to derive the POD using the administered exposure as the dose metric first, obtaining a POD in terms of environmental exposure for the animal results. The animal PBPK model, simulating the exposure profile of the bioassay, is then used to estimate the internal dose metric corresponding to the POD

for the animal, followed by application of the human PBPK model as above to complete interspecies extrapolation.

Also note that if the human PBPK model is nonlinear in the range of the POD, the correspondence of exposure ranges underlying each PBPK model could impact confidence in the human extrapolation; these situations need to be considered on a case-by-case basis. For example, if the human PBPK model can only be calibrated at exposure levels far below the range of exposures needed for the extrapolation, the PBPK results might not reliably support deriving a reference value. One approach to increase confidence in the PBPK predictions is to consider applying the relevant components of the uncertainty factor (UF) for human variation (see Section 8.3.2) to the animal-based POD prior to application of the human PBPK model (doing some prior to PBPK-based dosimetric adjustments might allow the PBPK model to do those adjustments in a dose range that it is calibrated for, although this is not always the case).

Approaches when a physiologically based pharmacokinetic (PBPK) model is not available

When a PBPK model or comparable data are not available, EPA has developed standard approaches that can be applied to typical data sets. These standard approaches also facilitate comparison across exposure patterns and species.

- Intermittent study exposures (e.g., exposure only on weekdays) are standardized to a daily average over the duration of exposure. Exposures during a critical period, such as gestation, however, are not averaged over a longer duration [[U.S. EPA \(2005a\)](#), §3.1.1; [U.S. EPA \(1991a\)](#), §3.2].
- Exposures are standardized to equivalent human terms to facilitate comparison of results from different species, and to estimate final risk values.
- Oral doses are scaled allometrically using $\text{mg}/\text{kg}^{3/4}\text{day}$ as the equivalent dose metric across species. Allometric scaling pertains to equivalence across species, not across lifestages, and is not used to scale doses from adult humans or mature animals to infants or children [[U.S. EPA \(2011a\)](#) and [U.S. EPA \(2005a\)](#), §3.1.3].
- Inhalation exposures are scaled using dosimetry models that apply species-specific physiological and anatomical factors and consider whether the effect occurs at the site of first contact or after systemic circulation [[U.S. EPA \(2012a\)](#) and [U.S. EPA \(1994\)](#), §3].

In the absence of study-specific data for physical parameters (e.g., intake rates or body weight), standard values are recommended for use in dose-response analysis ([U.S. EPA, 1988](#)).

Route-to-Route Extrapolation

PBPK models can be used to estimate human equivalent values for routes of exposure that differ from those administered to test animals. To be used for route-to-route extrapolations, a PBPK model would need to be appropriately structured and parameterized to account for differences in uptake and distribution that occur between inhalation, oral, dermal, and other routes of exposure

for which it is intended and must pass a quality review (metabolism and excretion are not expected to vary with route of exposure, but otherwise need to be described appropriately). The same standards apply for use of PBPK model for animal-to-human extrapolation within a given route. The model should appropriately account for the timing and relative rate of distribution to various tissues and be able to predict a dose metric appropriate for the endpoint being evaluated (e.g., parent chemical concentration, rate of metabolism, metabolite concentration). In-short, there are no new or additional uncertainties introduced by route-to-route extrapolation compared to animal-to-human extrapolation when using a valid PBPK model and an appropriate endpoint dose metric, with regard to the model's ability to predict the metric. However, there remains the possibility that unknown pharmacodynamic differences, including those closely related to pharmacokinetics are not accounted for by the model, and failure to account for these residual pharmacodynamic differences could lead to a significant underprediction of response or risk when extrapolating across routes of exposure. Therefore, in the case of noncancer assessments when a PBPK model is required and used for route-to-route extrapolation, the potential added uncertainties from this application might be considered within the context of the database deficiency uncertainty factor, if warranted (other UFs would typically remain the same unless specific data are available to identify different UFs).

When route-to-route extrapolation of study results can be reasonably accomplished without PBPK models, the assessment needs to describe the underlying data, algorithms, and assumptions [[U.S. EPA \(2005a\)](#), §3.1.4]. For example, doses in human ADME studies in the range of the POD are ideal for informing animal-to-human extrapolation. In many circumstances, however, simple route-to-route extrapolation might not be supported [e.g., [U.S. EPA \(1994\)](#), §4.1.2; [U.S. EPA \(2006a\)](#)].

Modeling Response in the Range of Observation to Obtain a Point of Departure (POD)

When evaluating animal data, EPA first considers pharmacodynamic, or biologically based, models if any relevant to the assessment are available. Pharmacodynamic modeling that incorporates data on biological processes leading to an effect can be used to establish a POD and might reduce the extent of low-dose extrapolation needed for toxicity value derivation. Such models require sufficient data to ascertain the mode-of-action (MOA) and to support model parameters associated with its key events. Because different models could provide equivalent fits to the observed data but diverge substantially at lower exposure levels, critical biological parameters should be measured from laboratory studies, not by model fitting. Confidence in the use of a pharmacodynamic model depends on the robustness of its validation process and on the results of sensitivity analyses. Peer review of the scientific basis and performance of a model is essential [[U.S. EPA \(2005a\)](#), §3.2.2].

Because pharmacodynamic models are frequently not available, EPA has developed a standard set of dose-response models consistent with biological processes (<http://www.epa.gov/bmds/>) that can be applied to typical data sets. Refer to Appendix C of the

EPA *Benchmark Dose Technical Guidance* ([U.S. EPA, 2012b](#)) and the “Model Descriptions” section of the BMDS User Manual for more information on these models

(https://www.epa.gov/sites/production/files/2018-09/documents/bmds_3.0_user_guide.pdf).

Currently, there is no recommended hierarchy of models that would expedite model selection, in part because of the many different types of data sets and study designs affecting dose-response patterns. As more flexible models are developed, hierarchies for some categories of endpoints will likely be more feasible. See the EPA *Benchmark Dose Technical Guidance* ([U.S. EPA, 2012b](#)) for more information on model fitting, model selection, and reporting of decisions and results.

If dose-response modeling does not provide an estimate of the BMD and benchmark dose lower confidence limit (BMDL) at the desired BMR without undue extrapolation (i.e., the response at the lowest exposure substantially exceeds the desired BMR), sensitivity of the BMD and BMDL to model choices could be evaluated by fitting a variety of parametric and nonparametric dose-response models and then by applying a model-averaging procedure (see Section 8.4.1). Based on an explicit, case-specific evaluation of the uncertainties, a POD might be selected, or a decision could be reached that the data do not support a reasonable POD inference.

If data are not amenable to dose-response modeling, e.g., due to substantial low-dose extrapolation or no models provide appropriate fit to the observed data, the NOAEL (or absent that, the LOAEL) could then be used as the POD. Given that the hazard synthesis (see Chapter 7) supports the importance of the data considered for developing a toxicity value, identification of a NOAEL or LOAEL focuses on the *biological significance* of the degree of effect at the candidate exposure level [see also [U.S. EPA \(2012b\)](#), §1.2 and [U.S. EPA \(2002b\)](#), §4.3.1.1, §4.4.4 for more information].

8.2.3. Composite Risk

If there are multiple tumor types in a study population (human or animal), it is important to consider composite or overall risk to characterize the risk of developing a tumor in at least one site. The risk of experiencing tumors across several sites was termed “composite risk” by [Bogen \(1990\)](#) and “aggregate risk” by the [NRC \(1994\)](#). The EPA *Guidelines for Carcinogen Risk Assessment* ([U.S. EPA, 2005a](#)) suggest several approaches for characterizing total risk for multiple tumor sites, including estimating cancer risk from all tumor-bearing animals. EPA traditionally used the tumor-bearing animal approach until *Science and Judgment in Risk Assessment* ([NRC, 1994](#)) concluded that this would tend to underestimate composite risk when tumor types occur in a statistically independent manner; that is, the occurrence of a hemangiosarcoma, for example, would not depend on whether there was a hepatocellular tumor. [NRC \(1994\)](#) argued that a general assumption of statistical independence of tumor-type occurrences within animals would not always be verifiable but was not likely to introduce substantial error in assessing carcinogenic potency from rodent bioassay data. See the Integrated Risk Information System (IRIS) assessment of 1,3-butadiene ([U.S. EPA, 2002c](#)) for an example.

Several additional methods are available for estimating composite tumor risk, depending on considerations of MOA(s) and independence of tumors, and relevant dose metrics. For combinations of tumors with independent MOAs, but using a common dose metric, and with dose-response data for individuals that can be adequately modeled by the multistage model, EPA's BMDs includes specific software (MS-Combo) for estimating a POD for the overall tumor risk. When different dose metrics are relevant for some tumor types in a data set, facilitated usually using a PBPK model, the use of Markov Chain Monte Carlo methods (e.g., via WinBUGS) to derive a distribution of BMDs for the multistage model facilitates estimation of overall risk ([Kopylev et al., 2007](#)).

8.2.4. Tools and Documentation to Support Dose-Response Modeling

The decisions and processes used for derivation of toxicity values should be documented clearly enough to permit independent verification. There should be explicit documentation of methods and decisions regarding:

- Selection of the studies and endpoints
- Exact identification and source of the data used
- Exposure level
- Conversions and other calculations
- Endpoint transformations (if any)
- A generally accepted level of detail documenting PBPK modeling
- A generally accepted level of detail documenting biologically based modeling
- Choices of response metrics (e.g., BMR types and numerical values)
- Dose-response modeling methods and assumptions
- Model selection
- For model-derived PODs, both the BMD and the BMDL to support central and lower bound estimates of risk values
- For NOAELs or LOAELs used as PODs when dose-response modeling is not feasible, response level relative to control, and a 95% confidence interval (CI) if feasible, to clarify comparability of responses across studies
- Methods of combining or weighting studies, data, or PODs, if applicable
- Selection of a single toxicity value to represent each type of health effect

The dose-response modeling template of each chemical assessment documents BMDS-based modeling assumptions and conditions (including parameter constraints and parameters at boundaries) as well as model selection.

8.3. DEVELOPING CANDIDATE TOXICITY VALUES

This section provides an overview of linear and nonlinear low-dose extrapolation approaches to yield candidate toxicity values for each identified hazard, building on recommendations provided by EPA's RfD/RfC review ([U.S. EPA, 2002b](#)) and Cancer Guidelines ([U.S. EPA, 2005a](#)).

8.3.1. Linear Low-Dose Extrapolation

A linear approach is most commonly used for cancer endpoints. In such cases, linear extrapolation is used if the dose-response curve is expected to have a linear component below the POD. This includes agents or their metabolites that are deoxyribonucleic acid (DNA) reactive and have direct mutagenic activity. Linear extrapolation is also used when data are insufficient to establish the MOA and when scientifically plausible ([U.S. EPA, 2005a](#)). The result of linear extrapolation is described by the slope of the line from the response at the POD to the background or control response, such as an oral slope factor or an inhalation unit risk.

Not all carcinogens are consistent with low-dose linearity, and in some cases both linear and nonlinear approaches can be used if there are multiple MOAs identified for the agent's carcinogenicity ([U.S. EPA, 2005a](#)). For example, modeling to a low response level can be useful for estimating the response where a high-exposure MOA would be less important. Also, comparing linear and nonlinear models can provide insights into uncertainties related to model choice and mechanisms. In this context, note that "... it is impossible to determine the correct functional form of a population dose-response curve solely from mechanistic information derived from animal studies and in vitro systems" [[NRC \(2014\)](#), p.111].

Derivation of Cancer Risk Values

If linear extrapolation is used for cancer risk estimation, the assessment develops a candidate slope factor or unit risk for each suitable data set. These results are arrayed, using common dose metrics, to show the distribution of relative potency across various effects and experimental systems. Cancer risk values are predictive risk estimates, derived for low-dose linear extrapolation, by inferring the slope of a line drawn from the POD (e.g., BMDL) to the background response for the function relating risk (e.g., extra risk) to exposure.

- An inhalation unit risk is a plausible upper bound lifetime risk of cancer from chronic inhalation of the agent per unit of air concentration (expressed as ppm or $\mu\text{g}/\text{m}^3$).
- An oral slope factor can be derived based on food intake, gavage dosing, or drinking water concentration. When derived from food intake or gavage, it is defined per unit of mass

consumed per unit body weight, per day (mg/kg-day). When derived from drinking water, it is defined per unit of concentration in drinking water (expressed as µg/L).

- Additionally, if there are data that support a mutagenic MOA for a suspected carcinogen, age-dependent adjustment factors (ADAF) should be applied to account for the fact that early life exposures to mutagens increase the risk for cancer. Supplemental Cancer Guidelines ([U.S. EPA, 2005b](#)) provide more guidance on how and when to apply these ADAFs.

8.3.2. Nonlinear Low-Dose Extrapolation

Reference value derivation is EPA's most frequently used type of nonlinear extrapolation method and is most commonly used for noncancer effects (see Derivation of Reference Values below). This approach is also used for cancer effects if there are sufficient data to ascertain the MOA and conclude that it is not linear at low doses, but without enough data to support chemical-specific modeling at low doses. For these cases, reference values for each relevant route of exposure are developed following EPA's established practices [[U.S. EPA \(2005a\)](#), §3.3.4]; in general, the reference value is based not on tumor incidence, but on a key precursor event in the MOA that is necessary for tumor formation.

Derivation of Reference Values

An oral RfD or an inhalation RfC is an estimate of an exposure to the human population (including in susceptible groups) that is likely to be without an appreciable risk of deleterious health effects over a lifetime [[U.S. EPA \(2002b\)](#), §4.2]. These health effects are either effects other than cancer or related to cancer if a well-characterized MOA indicates that a necessary key precursor event does not occur below a specific exposure level. Reference values are not predictive risk values; they provide no information about risks at higher exposure levels.

For each data set analyzed for dose-response (see Section 7.2), reference values are estimated by applying relevant adjustments, i.e., uncertainty factors, to the PODs to account for five possible areas of uncertainty and variability: extrapolation from animals to humans, human variation, extrapolation to chronic exposure duration, the type of POD being used for reference value derivation, and extrapolation to a minimal level of risk (if not observed in the data set). The particular value for these adjustments depends on the quality of the studies and data, the breadth of the chemical-specific database, and scientific judgment. The default uncertainty factor values typically cover a single order of magnitude (10^1) and usually take the values of 10, 3, or 1. By convention, the half-power ($10^{0.5}$) value is rounded to 3 when considered individually but is considered a log-value when considered with other half-power factors. Thus, combination of two uncertainty factors of 10 and 3 would result in a composite value of 30, whereas combination of two uncertainty factors of 3 each would result in a composite value of 10. Ultimately, although default values are recommended for the individual uncertainty factors, the final selected values should rely on a careful consideration of all available chemical-specific data and the scientific rationale for their selection must be justified in the dose-response section. Uncertainty factor values

other than the standard values can be used but must be based on chemical-specific information if sufficient information exists in the chemical database.

- *Animal-to-human extrapolation:* If animal results are used to make inferences about humans, the toxicity value incorporates cross-species differences, which could arise from differences in pharmacokinetics or pharmacodynamics. Typically, the pharmacokinetic and pharmacodynamic portions are considered to address an equivalent (i.e., $10^{0.5}$) amount of the total uncertainty factor. If the POD is standardized to equivalent human terms or is based on pharmacokinetic or dosimetry modeling ([U.S. EPA, 2014b](#), [2011a](#)), a factor of $10^{0.5}$ (rounded to 3) is applied to account for the remaining uncertainty involving pharmacokinetic and pharmacodynamic differences. If a biologically based model adjusts fully for pharmacokinetic and pharmacodynamic differences across species, this factor is not used. Similarly, although this is not a common scenario, if chemical-specific information is available sufficient to reasonably conclude that the experimental animal species is less or equally sensitive as humans, the pharmacodynamic portion of this uncertainty factor can be reduced to 1.
- *Human variation:* The assessment accounts for variation in susceptibility across the human population and the possibility that the available data might not be representative of individuals who are most susceptible to the effect. If population-based data for the effect or for characterizing the internal dose are available, the potential for data-based adjustments for pharmacodynamics or pharmacokinetics is considered ([U.S. EPA, 2014b](#)).¹⁹ Further, “when sufficient data are available, an intraspecies UF either less than or greater than $10\times$ may be justified ([U.S. EPA, 2002b](#)). However, a reduction from the default (10) is only considered in cases when there is dose-response data for the most susceptible population” ([U.S. EPA, 2002b](#)). This factor is reduced only if the POD is derived or adjusted specifically for susceptible individuals [not for a general population that includes both susceptible and nonsusceptible individuals; [U.S. EPA \(2002b\)](#), §4.4.5; [U.S. EPA \(1998\)](#), §4.2; [U.S. EPA \(1996\)](#), §4; [U.S. EPA \(1994\)](#), §4.3.9.1; [U.S. EPA \(1991a\)](#), §3.4]. Otherwise, a factor of 10 is generally used to account for this variation. Note that when a PBPK model is available for relating human internal dose to environmental exposure, relevant portions of this UF might be more usefully applied prior to animal-to-human extrapolation, depending on the correspondence of any nonlinearities (e.g., saturation levels) between species (also see Section 8.2.2).
- *LOAEL to NOAEL:* If a POD is based on a LOAEL, the assessment must infer an exposure level where such effects are not expected. This can be a matter of great uncertainty if there is no evidence available at lower exposures. The ratio of the doses at the LOAEL and NOAEL are expected to vary considerably across studies, and consideration of cross-study information might not be informative. A factor of up to 10 is generally applied to extrapolate to a lower exposure expected to be without appreciable effects. A factor other than 10 can be used depending on the magnitude and nature of the response and the shape of the dose-response curve ([U.S. EPA, 2002b](#), [1998](#), [1996](#), [1994](#), [1991a](#)). For example, LOAELs associated with lower response levels or less adverse effects might warrant smaller uncertainty factors, whereas higher response levels likely warrant the default value of 10, or in rare instances,

¹⁹Examples of adjusting the pharmacokinetic portion of interhuman variability include the IRIS boron assessment’s use of nonchemical-specific kinetic data [glomerular filtration rate in pregnant humans as a surrogate for boron clearance ([U.S. EPA, 2004](#))]; and the IRIS trichloroethylene assessment’s use of population variability in trichloroethylene metabolism via a PBPK model to estimate the lower 1st percentile of the dose metric distribution for each POD ([U.S. EPA, 2011b](#)).

values higher than 10. Regardless, the available data should be carefully evaluated and any decision to apply a nondefault value requires adequate discussion in the dose-response section.

- *Subchronic-to-chronic exposure:* If a chronic reference value is being developed, a POD is based on subchronic evidence, the assessment considers whether lifetime exposure could have effects at lower levels of exposure. A factor of up to 10 is applied (after adjustment of intermittent exposures to continuous) when using subchronic studies to make inferences about lifetime exposure. A factor other than 10 can be used, depending on the duration or timing of the studies and the nature of the response ([U.S. EPA, 2002b, 1998, 1994](#)). For example, studies that occur during a sensitive lifestage might not warrant application of this uncertainty factor. A prime example of this is developmental toxicity studies and effects observed in offspring. Typically, developmental toxicity studies use exposure durations either encompassing a specific portion of gestation (e.g., organogenesis) or the entirety of gestation as these are expected to be the critical windows of susceptibility for developmental effects. Thus, there is no concern that a longer duration exposure would result in more severe effects and an uncertainty factor would not be applied. This factor could be applied, albeit rarely, for developmental or reproductive effects if exposure covered less than the full critical period. A value different from 10 can be applied if there exists sufficient information from the chemical database. For example, if a chemical database contains subchronic and short-term studies and there is no evidence of an exacerbation of effect when moving from short-term to subchronic exposure durations, an uncertainty factor lower than 10 might be warranted.
- In addition to the adjustments above, if database deficiencies raise concern that further studies might identify a more sensitive effect, organ system, or lifestage, the assessment can apply a database UF ([U.S. EPA, 2002b, 1998, 1996, 1994, 1991a](#)). The size of the factor depends on the nature of the database deficiency. For example, EPA typically follows the suggestion that a factor of 10 be applied if a prenatal toxicity study and a two-generation reproduction study are both missing, and a factor of $10^{0.5}$ (rounded to 3) if either one or the other is missing [[U.S. EPA \(2002b\)](#), §4.4.5]. A database UF would still be applied if this type of study were available but considered to be a low confidence study based on the evaluation process described in Chapters 4 and 7. However, when deciding to apply this uncertainty factor and the value of the factor, risk assessors should consider the data missing and available for specific organ systems or lifestages, meaning this uncertainty factor can still be applied in scenarios when both developmental and two-generation reproduction studies are available if sufficient *evidence suggests* that effects could occur in other organ systems at lower doses. This uncertainty factor should still be applied even if the POD being adjusted comes from human data, and information from both human and animal studies should be considered when selecting the value of this factor. Information on structurally related chemicals could be used to select the value of this factor if it suggests effects in organ systems for which chemical-specific data are missing.

The POD for a particular reference value (RfV) is divided by the product of these factors. The RfD/RfC review recommends that any composite factor that exceeds 3,000 represents excessive uncertainty and recommends against relying on the associated RfV. A tabular display of deriving candidate toxicity values (for an RfD) is shown in Figure 8-1.

EPA will continue to seek improvements in uncertainty characterization. Increasingly, data-based adjustments ([U.S. EPA, 2014b](#)) and Bayesian methods for characterizing population variability ([NRC, 2014](#)) are feasible [e.g., [Simon et al. \(2016\)](#)] and can be distinguished from the UF considerations outlined above.

In addition, uncertainty in deriving toxicity values can be accounted for probabilistically. As an example of this, the World Health Organization (WHO) International Programme on Chemical Safety (IPCS) developed a unified probabilistic approach based on the concept of the HDMI, which is defined as the human dose at which a fraction I of the population shows an effect of magnitude M or greater for the critical endpoint considered ([IOMC ED, 2017](#)). Under this approach, the HDMI is treated as a random variable with its own probability distribution. From here, one can estimate the distribution of a “risk-specific dose,” defined as the dose at which a prespecified risk occurs. For the appropriate selection of values M and I, a probabilistic toxicity value can be set to some low percentile of the HDMI distribution estimate. WHO/IPCS also developed the Excel-based spreadsheet tool Approximate Probabilistic Analysis (APROBA) to estimate the HDMI distribution and probabilistic toxicity values. In APROBA, the HDMI is assumed to be lognormally distributed. An example of APROBA being applied in an IRIS assessment context is [Blessinger et al. \(2020\)](#).

Endpoint and reference	POD _{HED} ^a	POD type	UF _A	UF _H	UF _L	UF _S	UF _D	Composite UF	Candidate value (mg/kg-day)
Nervous system (rat)									
Convulsions Crouse et al. (2006)	0.27	BMDL ₀₁	3	10	1	3	3	300	8.8×10^{-4}
Convulsions Cholakis et al. (1980)	0.06	BMDL ₀₁	3	10	1	3	3	300	2.0×10^{-4}
Kidney/urogenital system (rat)									
Prostate suppurative inflammation Levine et al. (1983)	0.23	BMDL ₁₀	3	10	1	1	3	100	2.3×10^{-3}
Male reproductive system (mouse)									
Testicular degeneration Lish et al. (1984)	2.4	BMDL ₁₀	3	10	1	1	3	100	2.5×10^{-2}

Figure 8-1. Example summary of candidate toxicity values (for reference dose [RfD] derivation). Candidate values for three effects (nervous system, kidney/urogenital system, and male reproductive system).

BMDL = benchmark dose lower confidence limit; HED = human equivalent dose; POD = point of departure; UF_A = interspecies uncertainty factor; UF_D = database uncertainty factor; UF_H = intraspecies uncertainty factor; UF_L = LOAEL-to-NOAEL uncertainty factor; UF_S = subchronic-to-chronic uncertainty factor.

8.4. CHARACTERIZING UNCERTAINTY AND CONFIDENCE IN TOXICITY VALUES

8.4.1. Uncertainty in Toxicity Values

In addition to the UFs discussed in the preceding section, which are applied to derived reference values through prescribed extrapolations if agent-specific data are not available, the assessment should address, at least qualitatively, other principal sources of uncertainty. Common issues relevant to both reference values and cancer risk values include:

- *Consistency of the overall database for estimating toxicity values associated with important adverse outcomes:* For each toxicity value derivation, the variability among candidate values for the same outcome is evaluated, taking into account potential explanations for differences (e.g., different durations, different species/strains).
- *Dose metric(s) used for dose-response modeling, route-to-route extrapolation, or extrapolation to humans:* Relevant issues include the strength of evidence associating a dose metric with the critical effects, strength of evidence for human relevance of the dose metric (if based on an animal study), and whether extrapolation to humans relies on chemical-specific evidence or default allometric relationships (whether or not a PBPK model is used).
- *Model uncertainty underlying POD selections:* If there is no biologically based model on which to base human estimates of toxicity values, uncertainties attributable to the use of empirical models should be evaluated. While PODs generally do not vary significantly across dose-response models if they are within the observed data ranges, PODs can vary considerably across models if extrapolation outside the observed data is needed.
- *Statistical uncertainty in the POD:* Statistical uncertainty, as characterized by the model-estimated CI, generally represents the experimental variability associated with the data set. It might also increase with increasing extrapolation outside a data range, overlapping with model uncertainty. The degree of statistical uncertainty associated with each POD, and its sources, should be discussed and compared among PODs. For each toxicity value relying on dose-response modeling, the central tendency value (BMD) is reported in addition to the POD [(lower bound, or BMDL) also see [U.S. EPA \(2005a\)](#), Sections 3.2 and 3.6]. For toxicity values relying on NOAELs or LOAELs, the observed response level at that exposure is reported.

In addition to the uncertainties listed above, there is currently no accommodation in cancer risk values for addressing susceptible populations and lifestages. There might be data available to qualify the estimated potential risk either qualitatively or quantitatively. To account for the fact that early life exposures to mutagens increase the risk for cancer, ADAFs are applied when estimating cancer risk associated with specific exposure levels. The Supplemental Cancer Guidelines ([U.S. EPA, 2005b](#)) provide more guidance on how and when to apply these ADAFs.

Depending on the availability of suitable information and the needs of individual assessments, the qualitative discussion and synthesis of uncertainty in values could be enhanced by quantitative analyses, including sensitivity analyses for decisions made in selecting study

populations, dose metrics, and PBPK model parameters. Modeling uncertainty using ranges or probability distributions might also be useful in cases where the data are adequate. Whether it is quantitative or qualitative, characterization of uncertainty is communicated clearly and transparently to facilitate decision-making.

EPA will continue to seek improvements in its dose-response methods, including improved methods for characterizing model uncertainty. To rely less on selecting a single best-fitting model from among a limited set of parametric models, EPA is evaluating more model-robust approaches such as model averaging ([Wheeler et al., 2022](#); [Wheeler et al., 2020](#); [Shao and Gift, 2013](#); [Shao, 2012](#); [Wheeler and Bailer, 2009](#)). Model averaging is a technique for inference over multiple models that accounts for model uncertainty by estimating a predictor-response relationship as a weighted sum of individual model estimates. Model averaging has been shown to be statistically superior to single model selection methods ([West et al., 2012](#); [Wheeler and Bailer, 2009](#)). A Bayesian model averaging method has recently been developed for dichotomous and continuous endpoints that approximates the posterior density using maximum a posteriori estimation and constructs model weights based on a Laplace approximation ([Wheeler et al., 2022](#); [Wheeler et al., 2020](#)). Other approaches to addressing model uncertainty include application of nonparametric dose-response modeling ([Guha et al., 2013](#); [Bhattacharya and Lin, 2011](#); [Wheeler and Bailer, 2009](#)) and flexible model forms that are validated with historical data ([Slob and Setzer, 2014](#)).

8.4.2. Characterizing Confidence

In assessments for which an RfD or RfC is derived, the level of confidence in the primary studies, the health effect database associated with that reference value, the quantification of the POD, and the overall reference value (based on the three aforementioned confidence judgments) are provided. Details on characterizing confidence in the derived toxicity values are provided in *Methods for Derivation of Inhalation Reference Concentrations and Application of Inhalation Dosimetry* ([U.S. EPA, 1994](#)). Briefly, the confidence ranking of the derived toxicity value (i.e., low, medium, or high) reflects the degree of belief that the reference value (RfD or RfC) will change (in either direction) with the acquisition of new data; it is not a statement about confidence in the degree of health protection provided by the reference value. In addition, the confidence ranking is intended to reflect considerations not already covered by the UFs and is not linked directly to the UF values. The confidence ranking for each of these parameters is accompanied with a narrative describing strengths, limitations, and data gaps. The overall determination of the confidence in the derived toxicity values can consider multiple topics, including confidence in the study(ies) used to derive the POD, the evidence base supporting the hazard, and the methods of quantitation used to derive the POD. It is important to recognize that characterizing confidence requires a narrative description and does not solely entail the designation of a confidence ranking. Confidence rankings are not discrete entities and for any given parameter, the level of confidence could fall along the continuum between low to high. There is no algorithm that links the designated level of confidence applied to the study/studies used in dose-response analysis, the database, the quantification of the

POD, or overall risk estimate. For example, a designation of high confidence in the study(ies) used in dose-response analysis might not translate to the assessment reporting a high level of confidence in the database of available studies or the overall confidence in the derived risk estimate.

Additionally, different components of the overall confidence in the derived risk estimate could factor more heavily in that final determination given assessment- or endpoint-specific situations. In other words, confidence in the database might be the predominating factor in the overall confidence in one risk estimate, whereas the quantification of the POD might be the most important factor in the confidence for another risk estimate.

8.5. SELECTING FINAL TOXICITY VALUES

8.5.1. Organ/System-specific Toxicity Values

The next step is to select an organ/system-specific toxicity value for each hazard identified in the assessment. This selection can be based on the study confidence considerations, the most sensitive outcome, a clustering of values, or a combination of such factors; the rationale for the selection is presented in the assessment. By providing these organ/system-specific toxicity values, IRIS assessments facilitate subsequent cumulative risk assessments that consider the combined effect of multiple agents acting at a common site or through common mechanisms ([NRC, 2009](#)).

Given multiple candidate toxicity values for an organ or system, each candidate value should be evaluated with respect to multiple considerations. The following key considerations should be included, but are not presented in a hierarchy.

- *Weight of evidence of hazard for the specific health effect or endpoint within the broader hazard category:* In general, effects and endpoints with stronger evidence of a causal relationship are preferred.
- *Attributes evaluated when selecting studies for deriving candidate toxicity values:* These include the study population/species, exposure paradigm, and quality of exposure and outcome measurement (see Chapter 7). Studies of higher confidence, when evaluated according to these attributes, are preferred.
- *Sensitivity of POD:* Concerning the identification of the most sensitive outcome or toxicity value, note that BMDs (not BMDLs) should be the starting point for evaluating relative sensitivity. Similarities of the BMDs between candidate outcomes suggest very little difference between candidate toxicity values. BMDLs characterize associated statistical uncertainty and should be examined in determining which data sets provide more reliable PODs. Note: this is not the driver of the selection of a final RfV, rather one of several considerations that prioritize preferences for a relatively stronger, more confident foundation for a particular POD and BMD/BMDL (see other five bullets in this section).
- *Basis of the POD:* A modeled BMDL is preferred over a NOAEL, which is in turn preferred over a LOAEL. Additionally, when there is sufficient knowledge of pharmacokinetics and the active toxic agent for the effect, a POD based on an internal dose metric would be preferred over one based on administered exposure.

- *Other uncertainties in dose-response modeling:* These include the uncertainty in the BMD (e.g., reflected in the relative proximity of the BMD and BMDL) and model uncertainty due to less optimal model fit or to extrapolation below the range of observation.
- *Uncertainties due to other extrapolations:* Toxicity values for which other extrapolations are less uncertain are preferred. For example, a reference value relying on a data-derived adjustment factor for interspecies extrapolation would be less uncertain than a reference value relying on an interspecies extrapolation UF of 10. Note that the size of the composite UF (see Section 8.3) might not be a good indication of the remaining uncertainty because all UFs but the database UF address needed extrapolations (adjustments) or variability, rather than uncertainty ([NRC, 2009](#)). Therefore, to avoid “double counting” or otherwise mischaracterizing uncertainty, the remaining uncertainties that are discussed should be explicitly identified.

Because of this evaluation, the organ/system-specific toxicity value could be:

- Based on selecting a single candidate value considered to be most appropriate for protecting against toxicity in the given organ or system, or
- Based on deriving a “composite” value, supported by multiple candidate toxicity values, which protects against toxicity in the given organ or system. The designation of the supporting candidate toxicity values and the derivation of the composite value are documented in the assessment. (Note that this composite value approach is distinct from a combined analysis approach described in Section 7.2; the composite approach could be practical in situations in which a combined data set approach cannot be carried out [e.g., because of differences in exposure metrics or other measures].)

8.5.2. Overall Toxicity Values

The selection of overall toxicity values for noncancer and cancer effects involves the study preferences discussed in Chapter 7, consideration of overall toxicity, study confidence, and confidence in each value, including the strength of various dose-response analyses and the possibility of basing a more robust result on multiple data sets. In addition to the information described above, the direct graphical comparison of PODs and toxicity values can inform selection of a final value (i.e., before and after application of UFs to PODs).

When the bulk of toxicity values exhibit a relatively small range of variation, it is questionable whether formal quantitative methods will add much value or change the risk assessment conclusions and final toxicity value(s). In such cases, simple graphical methods [[NRC \(2014\)](#), see Figure 7-6; [NRC \(2011\)](#)] might be sufficient for both communicating uncertainty and selecting a final toxicity value.

REFERENCES

- [ACGIH](#) (American Conference of Governmental Industrial Hygienists). (2007). 2007 TLVs and BEIs: Based on the documentation of the threshold limit values for chemical substances and physical agents and biological exposure indices [TLV/BEI]. Cincinnati, OH.
- [AIHA](#) (American Industrial Hygiene Association). (2016). Current ERPG Values. In 2016 ERPG/WEEL Handbook. Fairfax, VA: American Industrial Hygiene Association Guideline Foundation.
- [Allen, BC; Strong, PL; Price, CJ; Hubbard, SA; Daston, GP.](#) (1996). Benchmark dose analysis of developmental toxicity in rats exposed to boric acid. *Fundam Appl Toxicol* 32: 194-204. <http://dx.doi.org/10.1093/toxsci/32.2.194>
- [Arzuaga, X; Smith, MT; Gibbons, CF; Skakkebaek, NE; Yost, EE; Beverly, BEJ; Hotchkiss, AK; Hauser, R; Pagani, RL; Schrader, SM; Zeise, L; Prins, GS.](#) (2019). Proposed key characteristics of male reproductive toxicants as an approach for organizing and evaluating mechanistic evidence in human health hazard assessments. *Environ Health Perspect* 127: 1-12. <http://dx.doi.org/10.1289/EHP5045>
- [ATSDR](#) (Agency for Toxic Substances and Disease Registry). (2021). Toxic substances portal: Toxicological profiles [Database]. Atlanta, GA: Centers for Disease Control and Prevention. Retrieved from <https://www.atsdr.cdc.gov/toxprofiledocs/index.html>
- [Beronius, A; Molander, L; Rudén, C; Hanberg, A.](#) (2014). Facilitating the use of non-standard in vivo studies in health risk assessment of chemicals: A proposal to improve evaluation criteria and reporting. *J Appl Toxicol* 34: 607-617. <http://dx.doi.org/10.1002/jat.2991>
- [Beronius, A; Molander, L; Zilliacus, J; Rudén, C; Hanberg, A.](#) (2018). Testing and refining the Science in Risk Assessment and Policy (SciRAP) web-based platform for evaluating the reliability and relevance of in vivo toxicity studies. *J Appl Toxicol* 38: 1460-1470. <http://dx.doi.org/10.1002/jat.3648>
- [Bhattacharya, R; Lin, L.](#) (2011). Nonparametric benchmark analysis in risk assessment: A comparative study by simulation and data analysis. *Sankhya Ser B* 73: 144-163. <http://dx.doi.org/10.1007/s13571-011-0019-7>
- [Blessinger, T; Davis, A; Chiu, WA; Stanek, J; Woodall, GM; Gift, J; Thayer, KA; Bussard, D.](#) (2020). Application of a unified probabilistic framework to the dose-response assessment of acrolein. *Environ Int* 143: 105953. <http://dx.doi.org/10.1016/j.envint.2020.105953>
- [Bogen, KT.](#) (1990). Uncertainty in environmental health risk assessment (Environment - Problems and solutions). New York, NY: Garland Publishing.
- [Bragge, P; Clavisi, O; Turner, T; Tavender, E; Collie, A; Gruen, RL.](#) (2011). The Global evidence mapping initiative: Scoping research in broad topic areas. *BMC Med Res Methodol* 11: 92. <http://dx.doi.org/10.1186/1471-2288-11-92>
- [Brauer, M; Brumm, J; Vedal, S; Petkau, AJ.](#) (2002). Exposure misclassification and threshold concentrations in time series analyses of air pollution health effects. *Risk Anal* 22: 1183-1193.

- [CalEPA](#) (California Environmental Protection Agency). (2016). OEHHA toxicity criteria database. Sacramento, CA: Office of Environmental Health Hazard Assessment. Retrieved from <http://www.oehha.ca.gov/tcdb/index.asp>
- [Cooper, GS; Lunn, RM; Ågerstrand, M; Glenn, BS; Kraft, AD; Luke, AM; Ratcliffe, JM](#). (2016). Study sensitivity: Evaluating the ability to detect effects in systematic reviews of chemical exposures. *Environ Int* 92-93: 605-610. <http://dx.doi.org/10.1016/j.envint.2016.03.017>
- [CRD](#) (Centre for Reviews and Dissemination). (2013). Systematic reviews: CRD's guidance for undertaking reviews in health care. In J Akers (Ed.), (3rd ed.). York, UK: Centre for Reviews and Dissemination, University of York.
- [Crissman, JW; Goodman, DG; Hildebrandt, PK; Maronpot, RR; Prater, DA; Riley, JH; Seaman, WJ; Thake, DC](#). (2004). Best practices guideline: Toxicologic histopathology. *Toxicol Pathol* 32: 126-131. <http://dx.doi.org/10.1080/01926230490268756>
- [CT DEEP](#) (Connecticut Department of Energy and Environmental Protection). (2015). Hazardous air pollutants (pp. 1-31). <https://eregulations.ct.gov/eRegsPortal/Browse/getDocument?guid={00D6A654-0300-CC47-9B95-397D2AD21304}>
- [CT DEEP](#) (Connecticut Department of Energy and Environmental Protection). (2018). Technical support document: Recommended numeric criteria for common additional polluting substances and certain alternative criteria. Hartford, CT. https://portal.ct.gov/-/media/DEEP/site_clean_up/remediation_regulations/TechnicalSupportDocumentAPSAltCriteriapdf.pdf
- [DFG](#) (German Research Foundation). (2020). List of MAK and BAT values 2020: Permanent senate commission for the investigation of health hazards of chemical compounds in the work area. (Report 56). Bonn, Germany. http://dx.doi.org/10.34865/mbwl_2020_eng
- [Dickersin, K](#). (1990). The existence of publication bias and risk factors for its occurrence. *JAMA* 263: 1385-1389.
- [DOE](#) (Department of Energy). (2018). Table 2: Protective Action Criteria (PAC) Rev. 29a based on applicable 60-minute AEGs, ERPGs, or TEELs. https://edms3.energy.gov/pac/docs/Revision_29A_Table2.pdf
- [Dusseldorp, A; van Bruggen, M; Douwes, J; Janssen, P; Kelfkens, G](#). (2011). Health-based guideline values for the indoor environment. (RIVM report 609021044/2007). Bilthoven, The Netherlands: National Institute for Public Health and the Environment (RIVM). <https://www.rivm.nl/bibliotheek/rapporten/609021044.pdf>
- [EFSA](#) (European Food Safety Authority). (2017). Guidance on the use of the weight of evidence approach in scientific assessments. *EFSA J* 15: 1-69. <http://dx.doi.org/10.2903/j.efsa.2017.4971>
- [Elsevier](#). (2017). Guidance notes for authors of systematic reviews, systematic maps and other related manuscripts. Available online at <https://www.elsevier.com/journals/environment-international/0160-4120/guidance-notes>
- [Emerson, JD; Burdick, E; Hoaglin, DC; Mosteller, F; Chalmers, TC](#). (1990). An empirical study of the possible relation of treatment differences to quality scores in controlled randomized clinical trials. *Contemp Clin Trials* 11: 339-352.

- [Farland, WH.](#) (2005). [Memo to Science Policy council regarding implementation of the cancer guidelines and accompanying supplemental guidance - Science Policy Council Cancer Guidelines. Implementation Workgroup communication I: Application of the mode of action framework in mutagenicity determinations for carcinogenicity]. Available online at https://www.epa.gov/sites/production/files/2015-01/documents/cgiwgcommuniatio_n_i.pdf
- [Fu, R; Gartlehner, G; Grant, M; Shamliyan, T; Sedrakyan, A; Wilt, TJ; Griffith, L; Oremus, M; Raina, P; Ismaila, A; Santaguida, P; Lau, J; Trikalinos, TA.](#) (2011). Conducting quantitative synthesis when comparing medical interventions: AHRQ and the Effective Health Care Program. *J Clin Epidemiol* 64: 1187-1197. <http://dx.doi.org/10.1016/j.jclinepi.2010.08.010>
- [Germolec, DR; Lebrec, H; Anderson, SE; Burleson, GR; Cardenas, A; Corsini, E; Elmore, SE; Kaplan, BLF; Lawrence, BP; Lehmann, GM; Maier, CC; McHale, CM; Myers, LP; Pallardy, M; Rooney, AA; Zeise, L; Zhang, L; Smith, MT.](#) (2022). Consensus on the key characteristics of immunotoxic agents as a basis for hazard identification. *Environ Health Perspect* 130: 105001. <http://dx.doi.org/10.1289/EHP10800>
- [Government of Canada.](#) (2021). Publications: Healthy living. Available online at <https://www.canada.ca/en/services/health/publications/healthy-living.html> (accessed March 30, 2021).
- [Grandjean, P; Andersen, EW; Budtz-Jørgensen, E; Nielsen, F; Mølbak, K; Weihe, P; Heilmann, C.](#) (2012). Serum vaccine antibody concentrations in children exposed to perfluorinated compounds. *JAMA* 307: 391-397. <http://dx.doi.org/10.1001/jama.2011.2034>
- [Grandjean, P; Heilmann, C; Weihe, P; Nielsen, F; Mogensen, UB; Timmermann, A; Budtz-Jørgensen, E.](#) (2017). Estimated exposures to perfluorinated compounds in infancy predict attenuated vaccine antibody concentrations at age 5-years. *J Immunotoxicol* 14: 188-195. <http://dx.doi.org/10.1080/1547691X.2017.1360968>
- [Granum, B; Haug, LS; Namork, E; Stølevik, SB; Thomsen, C; Aaberge, IS; van Loveren, H; Løvik, M; Nygaard, UC.](#) (2013). Pre-natal exposure to perfluoroalkyl substances may be associated with altered vaccine antibody levels and immune-related health outcomes in early childhood. *J Immunotoxicol* 10: 373-379. <http://dx.doi.org/10.3109/1547691X.2012.755580>
- [Guha, N; Roy, A; Kopylev, L; Fox, J; Spassova, M; White, P.](#) (2013). Nonparametric Bayesian methods for benchmark dose estimation. *Risk Anal* 33: 1608-1619. <http://dx.doi.org/10.1111/risa.12004>
- [Haddaway, NR; Macura, B; Whaley, P; Pullin, AS.](#) (2018). ROSES RepOrting standards for Systematic Evidence Syntheses: Pro forma, flow-diagram and descriptive summary of the plan and conduct of environmental systematic reviews and systematic maps. *Environ Evid* 7. <http://dx.doi.org/10.1186/s13750-018-0121-7>
- [Health Canada.](#) (1996). Health-based tolerable daily intakes/concentrations and tumorigenic doses/concentrations for priority substances. (96-EHD-194). Ottawa, Canada: Environmental Health Directorate, Health Protection Branch. <http://publications.gc.ca/collections/Collection/H46-2-96-194E.pdf>

- [Health Canada](https://www.canada.ca/content/dam/hc-sc/migration/hc-sc/ewh-semt/alt_formats/pdf/pubs/water-eau/sum_guide-res_recom/summary-table-EN-2020-02-11.pdf). (2020). Guidelines for Canadian drinking water quality: Summary table. Ottawa, Ontario: Water and Air Quality Bureau, Healthy Environments and Consumer Safety Branch, Health Canada. https://www.canada.ca/content/dam/hc-sc/migration/hc-sc/ewh-semt/alt_formats/pdf/pubs/water-eau/sum_guide-res_recom/summary-table-EN-2020-02-11.pdf
- [Higgins, J; Green, S](http://handbook.cochrane.org). (2011a). Cochrane handbook for systematic reviews of interventions. Version 5.1.0: The Cochrane Collaboration, 2011. <http://handbook.cochrane.org>
- [Higgins, J; Morgan, R; Rooney, A; Taylor, K; Thayer, K; Silva, R; Lemeris, C; Akl, A; Arroyave, W; Bateson, T; Berkman, N; Demers, P; Forastiere, F; Glenn, B; Hróbjartsson, A; Kirrane, E; LaKind, J; Luben, T; Lunn, R; ... Sterne, J](https://www.riskofbias.info/welcome/robins-e-tool). (2022a). Risk Of Bias In Non-randomized Studies - of Exposure (ROBINS-E). Launch version. Available online at <https://www.riskofbias.info/welcome/robins-e-tool>
- [Higgins, JPT; Green, S](http://handbook.cochrane.org/). (2011b). Cochrane handbook for systematic reviews of interventions. Version 5.1.0 (Updated March 2011). London, UK: The Cochrane Collaboration. <http://handbook.cochrane.org/>
- [Higgins, JPT; Thomas, J; Chandler, J; Cumpston, M; Li, T; Page, MJ; Welch, VA](http://www.training.cochrane.org/handbook). (2022b). Cochrane handbook for systematic reviews of interventions version 6.3. Cochrane Reviews. <http://www.training.cochrane.org/handbook>
- [Hill, AB](#). (1965). The environment and disease: Association or causation? Proc R Soc Med 58: 295-300.
- [Hirst, JA; Howick, J; Aronson, JK; Roberts, N; Perera, R; Koshiaris, C; Heneghan, C](http://dx.doi.org/10.1371/journal.pone.0098856). (2014). The need for randomization in animal trials: An overview of systematic reviews [Review]. PLoS ONE 9: e98856. <http://dx.doi.org/10.1371/journal.pone.0098856>
- [Hoenig, JM; Heisey, DM](#). (2001). The abuse of power: The pervasive fallacy of power calculations for data analysis. Am Stat 55: 19-24.
- [Hooijmans, CR; Rovers, MM; De Vries, RB; Leenaars, M; Ritskes-Hoitinga, M; Langendam, MW](http://dx.doi.org/10.1186/1471-2288-14-43). (2014). SYRCLE's risk of bias tool for animal studies. BMC Med Res Methodol 14: 43. <http://dx.doi.org/10.1186/1471-2288-14-43>
- [Howard, BE; Phillips, J; Miller, K; Tandon, A; Mav, D; Shah, MR; Holmgren, S; Pelch, KE; Walker, V; Rooney, AA; Macleod, M; Shah, RR; Thayer, K](http://dx.doi.org/10.1186/s13643-016-0263-z). (2016). SWIFT-Review: A text-mining workbench for systematic review. Syst Rev 5: 87. <http://dx.doi.org/10.1186/s13643-016-0263-z>
- [Howard, BE; Phillips, J; Tandon, A; Maharana, A; Elmore, R; Mav, D; Sedykh, A; Thayer, K; Merrick, BA; Walker, V; Rooney, A; Shah, RR](http://dx.doi.org/10.1016/j.envint.2020.105623). (2020). SWIFT-Active Screener: Accelerated document screening through active learning and integrated recall estimation. Environ Int 138: 105623. <http://dx.doi.org/10.1016/j.envint.2020.105623>
- [HSA](https://www.hsa.ie/eng/publications_and_forms/publications/codes_of_practice/chemical_agents_cop_2020.pdf) (Health & Safety Authority). (2020). Chemical agents and carcinogens: Code of practice 2020. Dublin, Ireland. https://www.hsa.ie/eng/publications_and_forms/publications/codes_of_practice/chemical_agents_cop_2020.pdf
- [HSL](http://www.hse.gov.uk/research/hsl_pdf/2002/hsl02-23.pdf) (Health & Safety Laboratory). (2002). Draft 2nd indicative occupational exposure limit value (IOELV) list: Workplace measurement method summaries. (HSL/2002/23). Bootle, England: Health and Safety Executive. http://www.hse.gov.uk/research/hsl_pdf/2002/hsl02-23.pdf

- [IARC](#) (International Agency for Research on Cancer). (2004). IARC Monographs on the evaluation of carcinogenic risks to humans. Volume 83: Tobacco smoke and involuntary smoking. Lyon, France: World Health Organization, IARC. <https://monographs.iarc.fr/wp-content/uploads/2018/06/mono83.pdf>
- [Idaho DEQ](#). (2019). Rules for the control of air pollution in Idaho. Boise, ID: Department of Environmental Quality, Air Quality Division. <https://adminrules.idaho.gov/rules/current/58/580101.pdf>
- [IDEM](#) (Indiana Department of Environmental Management). (2019). Air toxics toxicity information. Available online at <https://web.archive.org/web/20200922041251/https://www.in.gov/idem/toxic/2343.htm> (accessed July 18, 2019).
- [IFA](#) (Institute for Occupational Safety and Health of the German Social Accident Insurance). (2020). GESTIS international limit values database (Version April 2020) [Database]. Retrieved from <https://limitvalue.ifa.dguv.de/>
- [IOM](#) (Institute of Medicine). (2011). Introduction. In Finding what works in health care: Standards for systematic reviews. Washington, DC: The National Academies Press. <http://dx.doi.org/10.17226/13059>
- [IOMC ED](#) (Inter-organization Programme for the Sound Management of Chemicals). (2017). Guidance document on evaluating and expressing uncertainty in hazard characterization. Harmonization Project Document 11 – 2nd edition (2nd ed.). Geneva, Switzerland: World Health Organization. http://www.who.int/ipcs/methods/harmonization/areas/hazard_assessment/en/
- [IPCS](#) (International Programme on Chemical Safety). (2007a). Harmonization project document no. 4: Part 1: IPCS framework for analysing the relevance of a cancer mode of action for humans and case-studies: Part 2: IPCS framework for analysing the relevance of a non-cancer mode of action for humans. Geneva, Switzerland: World Health Organization. http://www.who.int/ipcs/methods/harmonization/areas/cancer_mode.pdf?ua=1
- [IPCS](#) (International Programme on Chemical Safety). (2007b). Harmonization project document no. 4: Part 2: IPCS framework for analysing the relevance of a non-cancer mode of action for humans. Geneva, Switzerland: World Health Organization. http://www.who.int/ipcs/methods/harmonization/areas/cancer_mode.pdf?ua=1
- [IPCS](#) (International Programme on Chemical Safety). (2010). Characterization and application of physiologically based pharmacokinetic models in risk assessment. (Harmonization Project Document No 9). Geneva, Switzerland: World Health Organization. <http://www.inchem.org/documents/harmproj/harmproj/harmproj9.pdf>
- [IPCS](#) (International Programme of Chemical Safety). (2012). Harmonization project document no. 10: Guidance for immunotoxicity risk assessment for chemicals. (Harmonization Project Document No. 10). Geneva, Switzerland: World Health Organization. <http://www.inchem.org/documents/harmproj/harmproj/harmproj10.pdf>
- [ISOH](#) (The Japan Society for Occupational Health). (2017). Recommendation of occupational exposure limits (2017-2018). J Occup Health 59: 436-469. <http://dx.doi.org/10.1539/joh.ROEL2017>
- [Juni, P; Witschi, A; Bloch, R; Egger, M](#). (1999). The hazards of scoring the quality of clinical trials for meta-analysis. JAMA 282: 1054-1060. <http://dx.doi.org/10.1001/jama.282.11.1054>

- [Kase, R; Korkaric, M; Werner, I; Ågerstrand, M.](#) (2016). Criteria for Reporting and evaluating Ecotoxicity Data (CRED): Comparison and perception of the Klimisch and CRED methods for evaluating reliability and relevance of ecotoxicity studies. *Environ Sci Eur* 28: 7. <http://dx.doi.org/10.1186/s12302-016-0073-x>
- [Kavlock, RJ; Schmid, JE; Setzer, RW, Jr.](#) (1996). A simulation study of the influence of study design on the estimation of benchmark doses for developmental toxicity. *Risk Anal* 16: 399-410. <http://dx.doi.org/10.1111/j.1539-6924.1996.tb01474.x>
- [Kopylev, L; Chen, C; White, P.](#) (2007). Towards quantitative uncertainty assessment for cancer risks: Central estimates and probability distributions of risk in dose-response modeling [Review]. *Regul Toxicol Pharmacol* 49: 203-207. <http://dx.doi.org/10.1016/j.yrtph.2007.08.002>
- [Krauth, D; Woodruff, TJ; Bero, L.](#) (2013). Instruments for assessing risk of bias and other methodological criteria of published animal studies: a systematic review [Review]. *Environ Health Perspect* 121: 985-992. <http://dx.doi.org/10.1289/ehp.1206389>
- [La Merrill, MA; Vandenberg, LN; Smith, MT; Goodson, W; Browne, P; Patisaul, HB; Guyton, KZ; Kortenkamp, A; Cogliano, VJ; Woodruff, TJ; Rieswijk, L; Sone, H; Korach, KS; Gore, AC; Zeise, L; Zoeller, RT.](#) (2020). Consensus on the key characteristics of endocrine-disrupting chemicals as a basis for hazard identification [Review]. *Nat Rev Endocrinol* 16: 45-57. <http://dx.doi.org/10.1038/s41574-019-0273-8>
- [Légis Québec.](#) (2020). Regulation respecting occupational health and safety. http://legisquebec.gouv.qc.ca/en/showdoc/cr/S-2.1,%20r.%2013?csi_scan_9222d36c6a354dc6=B09xyrMZ+270UP3j0MGuOD0kZigFAAAAXrM3HA==&bcsi_scan_filename=S-2.1,%20r.%2013&bcsi_scan_9222d36c6a354dc6=KXzmpPueuN0L1AjnJOB1Zerr85YMAAAyhrPTg==&bcsi_scan_filename=S-2.1,%20r.%2013
- [Lind, L; Araujo, JA; Barchowsky, A; Belcher, S; Berridge, BR; Chiamvimonvat, N; Chiu, WA; Cogliano, VJ; Elmore, S; Farraj, AK; Gomes, AV; Mchale, CM; Meyer-Tamaki, KB; Posnack, NG; Vargas, HM; Yang, X; Zeise, L; Zhou, C; Smith, MT.](#) (2021). Key characteristics of cardiovascular toxicants. *Environ Health Perspect* 129: 95001. <http://dx.doi.org/10.1289/EHP9321>
- [Luderer, U; Eskenazi, B; Hauser, R; Korach, KS; Mchale, CM; Moran, F; Rieswijk, L; Solomon, G; Udagawa, O; Zhang, L; Zlatnik, M; Zeise, L; Smith, MT.](#) (2019). Proposed key characteristics of female reproductive toxicants as an approach for organizing and evaluating mechanistic data in hazard assessment. *Environ Health Perspect* 127: 75001. <http://dx.doi.org/10.1289/EHP4971>
- [Lutz, WK; Gaylor, DW; Conolly, RB; Lutz, RW.](#) (2005). Nonlinearity and thresholds in dose-response relationships for carcinogenicity due to sampling variation, logarithmic dose scaling, or small differences in individual susceptibility [Review]. *Toxicol Appl Pharmacol* 207: S565-S569. <http://dx.doi.org/10.1016/j.taap.2005.01.038>
- [Macleod, MR.](#) (2013). Systematic reviews of experimental animal studies. Presentation presented at Workshop on weight of evidence; US National Research Council Committee to review the Integrated Risk Information System (IRIS) process, March 27-28, 2013, Washington, DC.
- [MassDEP](#) (Massachusetts Department of Environmental Protection). (2019). MassDep Ambient Air Toxics Guidelines. Available online at <https://www.mass.gov/service-details/massdep-ambient-air-toxics-guidelines>

- [MDH](https://www.health.state.mn.us/communities/environment/risk/guidance/air/table.html) (Minnesota Department of Health). (2019). Air guidance values. Available online at <https://www.health.state.mn.us/communities/environment/risk/guidance/air/table.html>
- [Miake-Lye, JM; Hempel, S; Shanman, R; Shekelle, PG.](#) (2016). What is an evidence map? A systematic review of published evidence maps and their definitions, methods, and products [Review]. *Syst Rev* 5: 28. <http://dx.doi.org/10.1186/s13643-016-0204-x>
- [Michigan DEQ.](#) (2016). Cleanup criteria and screening levels development and application: Remediation and redevelopment division resource materials (draft). Lansing, MI. https://www.michigan.gov/documents/deq/deq-rrd-chem-CleanupCriteriaTSD_527410_7.pdf
- [Moher, D; Jadad, AR; Tugwell, P.](#) (1996). Assessing the quality of randomized controlled trials. Current issues and future directions [Review]. *Int J Technol Assess Health Care* 12: 195-208.
- [Moher, D; Liberati, A; Tetzlaff, J; Altman, DG.](#) (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Med* 6. <http://dx.doi.org/doi.org/10.1136/bmj.b2535>
- [Molander, L; Ågerstrand, M; Beronius, A; Hanberg, A; Rudén, C.](#) (2015). Science in Risk Assessment and Policy (SciRAP): An online resource for evaluating and reporting in vivo (eco) toxicity studies. *Hum Ecol Risk Assess* 21: 753-762. <http://dx.doi.org/10.1080/10807039.2014.928104>
- [NAS](#) (National Academy of Science). (2018). Workshop to discuss systematic review with the NAS (Dec 2018). Workshop to discuss systematic review with the NAS, December 10-11, 2018, Washington, DC.
- [NAS](#) (National Academy of Science). (2019). Evidence Integration in Chemical Assessments Workshop (Jun 2019). Evidence integration in chemical assessments workshop, June 3-4, 2019, Washington, DC.
- [NASEM](#) (National Academies of Sciences, Engineering, and Medicine). (2018). Progress toward transforming the Integrated Risk Information System (IRIS) program: A 2018 evaluation. Washington, DC: National Academies Press. <http://dx.doi.org/10.17226/25086>
- [NC DEQ](#) (North Carolina Department of Environmental Quality). (2014). Toxic air pollutant guidelines. (15A NCAC 02D .1104). Raleigh, NC. <https://files.nc.gov/ncdeq/Air%20Quality/rules/rules/D1104.pdf>
- [NDEP](#) (Nevada Department of Environmental Protection). (2017). Basic comparison levels. Carson City, NV. <https://ndep.nv.gov/uploads/documents/july-2017-ndep-bcls.pdf>
- [Newman, MC.](#) (2008). "What exactly are you inferring?" A closer look at hypothesis testing. *Environ Toxicol Chem* 27: 1013-1019. <http://dx.doi.org/10.1897/07-373.1>
- [NIEHS](#) (National Institute of Environmental Health Sciences). (2015). Handbook for preparing report on carcinogens monographs. U.S. Department of Health and Human Services, Office of the Report on Carcinogens. https://ntp.niehs.nih.gov/ntp/roc/handbook/roc_handbook_508.pdf
- [NIOSH](#) (National Institute for Occupational Safety and Health). (2021). NIOSH pocket guide to chemical hazards: Index of chemical abstracts service registry numbers (CAS No.) [Website]. Atlanta, GA: Center for Disease Control and Prevention, U.S. Department of Health, Education and Welfare. <http://www.cdc.gov/niosh/npg/npgdcas.html>

- [NJ DEP](https://www.state.nj.us/dep/aqpp/downloads/risk/ToxAll2020.pdf) (New Jersey Department of Environmental Protection). (2020). Toxicity values for inhalation exposure. (NJDEP/DAQ/AQEv). Trenton, NJ.
<https://www.state.nj.us/dep/aqpp/downloads/risk/ToxAll2020.pdf>
- [NRC](http://dx.doi.org/10.17226/366) (National Research Council). (1983). Risk assessment in the federal government: Managing the process. Washington, DC: National Academy Press. <http://dx.doi.org/10.17226/366>
- [NRC](http://dx.doi.org/10.17226/2125) (National Research Council). (1994). Science and judgment in risk assessment. Washington, DC: The National Academies Press. <http://dx.doi.org/10.17226/2125>
- [NRC](http://dx.doi.org/10.17226/12209) (National Research Council). (2009). Science and decisions: Advancing risk assessment. Washington, DC: National Academies Press. <http://dx.doi.org/10.17226/12209>
- [NRC](http://dx.doi.org/10.17226/13142) (National Research Council). (2011). Review of the Environmental Protection Agency's draft IRIS assessment of formaldehyde (pp. 1-194). Washington, DC: The National Academies Press. <http://dx.doi.org/10.17226/13142>
- [NRC](https://www.nap.edu/catalog/18594/critical-aspects-of-epas-iris-assessment-of-inorganic-arsenic-interim) (National Research Council). (2013). Critical aspects of EPA's IRIS assessment of inorganic arsenic: Interim report. Washington, DC: The National Academies Press.
<https://www.nap.edu/catalog/18594/critical-aspects-of-epas-iris-assessment-of-inorganic-arsenic-interim>
- [NRC](http://dx.doi.org/10.17226/18764) (National Research Council). (2014). Review of EPA's Integrated Risk Information System (IRIS) process. Washington, DC: The National Academies Press.
<http://dx.doi.org/10.17226/18764>
- [NTP](https://ntp.niehs.nih.gov/ntp/ohat/pubs/handbookjan2015_508.pdf) (National Toxicology Program). (2015). Handbook for conducting a literature-based health assessment using OHAT approach for systematic review and evidence integration. Research Triangle Park, NC: U.S. Department of Health and Human Services, National Toxicology Program, Office of Health Assessment and Translation.
https://ntp.niehs.nih.gov/ntp/ohat/pubs/handbookjan2015_508.pdf
- [NTP](https://ntp.niehs.nih.gov/ntp/ohat/pubs/handbookmarch2019_508.pdf) (National Toxicology Program). (2019). Handbook for conducting a literature-based health assessment using OHAT approach for systematic review and evidence integration. Research Triangle, NC: U.S. Department of Health and Human Services, National Toxicology Program, Office of Health Assessment and Translation.
https://ntp.niehs.nih.gov/ntp/ohat/pubs/handbookmarch2019_508.pdf
- [NYSDEC](https://www.dec.ny.gov/docs/remediation_hudson_pdf/techsuppdoc.pdf) (New York State Brownfield Cleanup Program). (2006). Development of soil cleanup objectives: Technical support document. Albany, NY.
https://www.dec.ny.gov/docs/remediation_hudson_pdf/techsuppdoc.pdf
- [OECD](http://dx.doi.org/10.1787/9789264304796-en) (Organisation for Economic Co-operation and Development). (2018). Guidance document on good in vitro method practices (GIVIMP). Paris, France.
<http://dx.doi.org/10.1787/9789264304796-en>
- [Ontario Ministry of Labour](https://www.labour.gov.on.ca/english/hs/pubs/oel_table.php). (2020). Current occupational exposure limits for Ontario workplaces required under regulation 833. Ontario, Canada.
https://www.labour.gov.on.ca/english/hs/pubs/oel_table.php
- [Oregon DEQ](https://www.oregon.gov/deq/FilterDocs/airtox-abc.pdf). (2018). Ambient benchmark concentrations (ABCs): Based on 2014-2017 ATSAC review. Portland, OR. <https://www.oregon.gov/deq/FilterDocs/airtox-abc.pdf>
- [OSHA](https://www.osha.gov/dsg/annotated-pels/tablez-1.html) (Occupational Safety & Health Administration). (2019). Permissible exposure limits: OSHA annotated table Z-1 [Website]. Washington, DC: United States Department of Labor, Occupational Safety & Health Administration. <https://www.osha.gov/dsg/annotated-pels/tablez-1.html>

- [OSHA](#) (Occupational Safety & Health Administration). (2020a). Air contaminants: Occupational safety and health standards for shipyard employment, subpart Z, toxic and hazardous substances. (OSHA Standard 1915.1000). Washington, DC.
https://www.osha.gov/pls/oshaweb/owadisp.show_document?p_table=STANDARDS&p_id=10286
- [OSHA](#) (Occupational Safety and Health Administration). (2020b). Safety and health regulations for construction: Occupational health and environmental controls: Gases, vapors, fumes, dusts, and mists: Appendix A. Available online at
http://www.osha.gov/pls/oshaweb/owadisp.show_document?p_table=STANDARDS&p_id=10629
- [Park, RM; Stayner, LT.](#) (2006). A search for thresholds and other nonlinearities in the relationship between hexavalent chromium and lung cancer. *Risk Anal* 26: 79-88.
<http://dx.doi.org/10.1111/j.1539-6924.2006.00709.x>
- [Radke, EG; Braun, JM; Meeker, JD; Cooper, GS.](#) (2018). Phthalate exposure and male reproductive outcomes: A systematic review of the human epidemiological evidence [Review]. *Environ Int* 121: 764-793. <http://dx.doi.org/10.1016/j.envint.2018.07.029>
- [RI DEM](#) (Rhode Island Department of Environmental Management). (2008). Rhode island air toxics: Guidelines. Providence, RI.
<http://www.dem.ri.gov/programs/benviron/air/pdf/airtoxgl.pdf>
- [RIVM](#) (National Institute for Public Health and the Environment (Netherlands)). (2001). Re-evaluation of human-toxicological maximum permissible risk levels. (RIVM report 711701025). Bilthoven, Netherlands: National Institute for Public Health and the Environment (RIVM). <https://www.rivm.nl/bibliotheek/rapporten/711701025.pdf>
- [Rothman, K.](#) (2010). Curbing type I and type II errors. *Eur J Epidemiol* 25: 223-224.
<http://dx.doi.org/10.1007/s10654-010-9437-5>
- [Rusyn, I; Arzuaga, X; Cattley, RC; Corton, JC; Ferguson, SS; Godoy, P; Guyton, KZ; Kaplowitz, N; Khetani, SR; Roberts, RA; Roth, RA; Smith, MT.](#) (2021). Key characteristics of human hepatotoxicants as a basis for identification and characterization of the causes of liver toxicity [Review]. *Hepatology* 74: 3486-3496. <http://dx.doi.org/10.1002/hep.31999>
- [Safe Work Australia.](#) (2019). Workplace exposure standards for airborne contaminants. Canberra, Australia.
<https://www.safeworkaustralia.gov.au/system/files/documents/1912/workplace-exposure-standards-airborne-contaminants.pdf>
- [Savitz, DA.](#) (1993). Is statistical significance testing useful in interpreting data? [Review]. *Reprod Toxicol* 7: 95-100.
- [Schulz, KF; Chalmers, I; Hayes, RJ; Altman, DG.](#) (1995). Empirical evidence of bias: Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA* 273: 408-412.
- [Schünemann, H; Brożek, J; Guyatt, G; Oxman, A.](#) (2013). GRADE handbook. Available online at <https://gdt.gradepro.org/app/handbook/handbook.html> (accessed April 22, 2022).
- [Segal, D; Makris, SL; Kraft, AD; Bale, AS; Fox, J; Gilbert, M; Bergfelt, DR; Raffaele, KC; Blain, RB; Fedak, KM; Selgrade, MK; Crofton, KM.](#) (2015). Evaluation of the ToxRTool's ability to rate the reliability of toxicological data for human health hazard assessments. *Regul Toxicol Pharmacol* 72: 94-101. <http://dx.doi.org/10.1016/j.yrtph.2015.03.005>

- [Shao, K.](#) (2012). A comparison of three methods for integrating historical information for Bayesian model averaged benchmark dose estimation. *Environ Toxicol Pharmacol* 34: 288-296.
<http://dx.doi.org/10.1016/j.etap.2012.05.002>
- [Shao, K; Gift, JS.](#) (2013). Model uncertainty and Bayesian model averaged benchmark dose estimation for continuous data. *Risk Anal* 34: 101-120.
<http://dx.doi.org/10.1111/risa.12078>
- [Shapiro, AJ; Antoni, S; Guyton, KZ; Lunn, RM; Loomis, D; Rusyn, I; Jahnke, GD; Schwingl, PJ; Mehta, SS; Addington, J; Guha, N.](#) (2018). Software tools to facilitate systematic review used for cancer hazard identification. *Environ Health Perspect* 126: 104501.
<http://dx.doi.org/10.1289/EHP4224>
- [Simon, TW; Zhu, Y; Dourson, ML; Beck, NB.](#) (2016). Bayesian methods for uncertainty factor application for derivation of reference values. *Regul Toxicol Pharmacol* 80: 9-24.
<http://dx.doi.org/10.1016/j.yrtph.2016.05.018>
- [Slob, W; Setzer, RW.](#) (2014). Shape and steepness of toxicological dose-response relationships of continuous endpoints [Review]. *Crit Rev Toxicol* 44: 270-297.
<http://dx.doi.org/10.3109/10408444.2013.853726>
- [Smith, MT; Guyton, KZ; Gibbons, CF; Fritz, JM; Portier, CJ; Rusyn, I; DeMarini, DM; Caldwell, JC; Kavlock, RJ; Lambert, PF; Hecht, SS; Bucher, JR; Stewart, BW; Baan, RA; Coglian, VJ; Straif, K.](#) (2016). Key characteristics of carcinogens as a basis for organizing data on mechanisms of carcinogenesis [Review]. *Environ Health Perspect* 124: 713-721.
<http://dx.doi.org/10.1289/ehp.1509912>
- [Stein, CR; MCGovern, KJ; Pajak, AM; Maglione, PJ; Wolff, M.](#) (2016). Perfluoroalkyl and polyfluoroalkyl substances and indicators of immune function in children aged 12-19 y: National Health and Nutrition Examination Survey. *Pediatr Res* 79: 348-357.
<http://dx.doi.org/10.1038/pr.2015.213>
- [Sterne, JAC; Hernán, MA; Reeves, BC; Savović, J; Berkman, ND; Viswanathan, M; Henry, D; Altman, DG; Ansari, MT; Boutron, I; Carpenter, JR; Chan, AW; Churchill, R; Deeks, JJ; Hróbjartsson, A; Kirkham, J; Juni, P; Loke, YK; Pigott, TD; Ramsay, CR; Regidor, D; Rothstein, HR; Sandhu, L; Santaguida, PL; Schünemann, HJ; Shea, B; Shrier, I; Tugwell, P; Turner, L; Valentine, JC; Waddington, H; Waters, E; Wells, GA; Whiting, PF; Higgins, JPT.](#) (2016). ROBINS-I: A tool for assessing risk of bias in non-randomised studies of interventions. *BMJ* 355: i4919.
<http://dx.doi.org/10.1136/bmj.i4919>
- [Sterne, JAC; Savovic, J; Page, MJ; Elbers, RG; Blencowe, NS; Boutron, I; Cates, CJ; Cheng, HY; Corbett, MS; Eldridge, SM; Emberson, JR; Hernan, MA; Hopewell, S; Hrobjartsson, A; Junqueira, DR; Juni, P; Kirkham, JJ; Lasserson, T; Li, T; McAleenan, A; Reeves, BC; Shepperd, S; Shrier, I; Stewart, LA; Tilling, K; White, IR; Whiting, PF; Higgins, JPT.](#) (2019). RoB 2: A revised tool for assessing risk of bias in randomised trials. *BMJ* 366: l4898.
<http://dx.doi.org/10.1136/bmj.l4898>
- [Sterne, JAC; Smith, GD; Cox, DR.](#) (2001). Sifting the evidence -- What's wrong with significance tests? *Br Med J* 322: 226-231.
- [Stitel, WM; Knauf, LA; Hertzberg, RC; Schoeny, RS.](#) (1993). A statistical test of compatibility of data sets to a common dose-response model. *Regul Toxicol Pharmacol* 18: 392-402.
<http://dx.doi.org/10.1006/rtph.1993.1065>

- [Swartout, J.](#) (2009). Analysis of dose-response uncertainty using benchmark dose modeling. Chapter 1. In RM Cooke (Ed.), *Uncertainty modeling in dose response: Bench testing environmental toxicity*. New York, NY: Wiley, John & Sons, Inc.
- [SWCAA](#) (Southwest Clean Air Agency). (2021). Pollutant search [Database]. Retrieved from <http://www.swcleanair.gov/epages/polsrch.asp>
- [TCEQ](#) (Texas Commission on Environmental Quality). (2018). TRRP protective concentration levels: April 2018 PCL and supporting tables [Database]. Retrieved from <https://www.tceq.texas.gov/remediation/trrp/trrppcls.html>
- [TCEQ](#) (Texas Commission on Environmental Quality). (2021). Final development support documents (DSDs). Available online at <https://www.tceq.texas.gov/toxicology/dsd/final>
- [TERA](#) (Toxicology Excellence For Risk Assessment). (2021). ITER database [Database]. Cincinnati, OH. Retrieved from <https://iter.tera.org/>
- [Thayer, KA; Angrish, M; Arzuaga, X; Carlson, LM; Davis, A; Dishaw, L; Druwe, I; Gibbons, C; Glenn, B; Jones, R; Kaiser, JP; Keshava, C; Keshava, N; Kraft, A; Lizarraga, L; Persad, A; Radke, EG; Rice, G; Schulz, B; ... Vetter, N.](#) (2022a). Systematic evidence map (SEM) template: Report format and methods used for the US EPA Integrated Risk Information System (IRIS) program, Provisional Peer Reviewed Toxicity Value (PPRTV) program, and other "fit for purpose" literature-based human health analyses. *Environ Int* 169: 107468. <http://dx.doi.org/10.1016/j.envint.2022.107468>
- [Thayer, KA; Shaffer, RM; Angrish, M; Arzuaga, X; Carlson, LM; Davis, A; Dishaw, L; Druwe, I; Gibbons, C; Glenn, B; Jones, R; Kaiser, JP; Keshava, C; Keshava, N; Kraft, A; Lizarraga, L; Persad, A; Radke, EG; Rice, G; Schulz, B; Shannon, T; Shapiro, A; Thacker, S; Vulimiri, S; Woodall, G; Yost, E.](#) (2022b). Use of systematic evidence maps within the US Environmental Protection Agency (EPA) Integrated Risk Information System (IRIS) program: Advancements to date and looking ahead. *Environ Int* 169: 107363.
- [Tiesjema, B; Baars, AJ.](#) (2009). Re-evaluation of some human-toxicological Maximum Permissible Risk levels earlier evaluated in the period 1991-2001. (RIVM Report 711701092). Bilthoven, the Netherlands: National Institute for Public Health and the Environment (Netherlands). <http://www.rivm.nl/bibliotheek/rapporten/711701092.pdf>
- [Tsafnat, G; Glasziou, P; Choong, MK; Dunn, A; Galgani, F; Coiera, E.](#) (2014). Systematic review automation technologies [Editorial]. *Syst Rev* 3: 74. <http://dx.doi.org/10.1186/2046-4053-3-74>
- [U.S. APHC](#) (U.S. Army Public Health Center). (2013). Environmental health risk assessment and chemical exposure guidelines for deployed military personnel. (Technical guide 230, 2013 revision). Aberdeen Proving Ground, MD. <https://phc.amedd.army.mil/PHC%20Resource%20Library/TG230-DeploymentEHRA-and-MEGs-2013-Revision.pdf>
- [U.S. EPA](#) (U.S. Environmental Protection Agency). (1988). Recommendations for and documentation of biological values for use in risk assessment [EPA Report]. (EPA/600/6-87/008). Washington, DC: U.S. Environmental Protection Agency. <http://cfpub.epa.gov/ncea/cfm/recordisplay.cfm?deid=34855>
- [U.S. EPA](#) (U.S. Environmental Protection Agency). (1991a). Guidelines for developmental toxicity risk assessment [EPA Report] (pp. 1-71). (EPA/600/FR-91/001). Washington, DC: U.S. Environmental Protection Agency, Risk Assessment Forum. <https://www.epa.gov/risk/guidelines-developmental-toxicity-risk-assessment>

- [U.S. EPA](#) (U.S. Environmental Protection Agency). (1991b). Guidelines for developmental toxicity risk assessment. Federal Register 56(234):63798-63826 [EPA Report].
<http://www.epa.gov/iris/backgr-d.htm>
- [U.S. EPA](#) (U.S. Environmental Protection Agency). (1993). NATICH: Data base report on state, local and EPA air toxics activities [EPA Report]. (EPA-453/R-93-041). Research Triangle Park, NC: U.S. Environmental Protection Agency, Office of Air Quality Planning and Standards.
<http://nepis.epa.gov/exe/ZyPURL.cgi?Dockey=2000NS7S.txt>
- [U.S. EPA](#) (U.S. Environmental Protection Agency). (1994). Methods for derivation of inhalation reference concentrations and application of inhalation dosimetry [EPA Report]. (EPA/600/8-90/066F). Research Triangle Park, NC: U.S. Environmental Protection Agency, Office of Air Quality Planning and Standards.
<https://cfpub.epa.gov/ncea/risk/recordisplay.cfm?deid=71993&CFID=51174829&CFTOKEN=25006317>
- [U.S. EPA](#) (U.S. Environmental Protection Agency). (1996). Guidelines for reproductive toxicity risk assessment [EPA Report] (pp. 1-143). (EPA/630/R-96/009). Washington, DC: U.S. Environmental Protection Agency, Risk Assessment Forum.
https://www.epa.gov/sites/production/files/2014-11/documents/guidelines_repro_toxicity.pdf
- [U.S. EPA](#) (U.S. Environmental Protection Agency). (1998). Guidelines for neurotoxicity risk assessment [EPA Report] (pp. 1-89). (EPA/630/R-95/001F). Washington, DC: U.S. Environmental Protection Agency, Risk Assessment Forum.
<http://www.epa.gov/risk/guidelines-neurotoxicity-risk-assessment>
- [U.S. EPA](#) (U.S. Environmental Protection Agency). (2002a). Guidelines for ensuring and maximizing the quality, objectivity, utility, and integrity of information disseminated by the Environmental Protection Agency [EPA Report]. (EPA/260/R-02/008). Washington, DC: U.S. Environmental Protection Agency, Office of Environmental Information.
<https://www.epa.gov/sites/production/files/2017-03/documents/epa-info-quality-guidelines.pdf>
- [U.S. EPA](#) (U.S. Environmental Protection Agency). (2002b). A review of the reference dose and reference concentration processes [EPA Report]. (EPA/630/P-02/002F). Washington, DC.
<https://www.epa.gov/sites/production/files/2014-12/documents/rfd-final.pdf>
- [U.S. EPA](#) (U.S. Environmental Protection Agency). (2002c). Toxicological review and IRIS summary of 1,3-butadiene [EPA Report]. Washington, DC: U.S. Environmental Protection Agency, National Center for Environmental Assessment.
http://ofmpub.epa.gov/eims/eimscomm.getfile?p_download_id=530289
- [U.S. EPA](#) (U.S. Environmental Protection Agency). (2004). Toxicological review of boron and compounds. In support of summary information on the Integrated Risk Information System (IRIS) [EPA Report]. (EPA/635/04/052). Washington, DC: U.S. Environmental Protection Agency, Integrated Risk Information System.
<http://nepis.epa.gov/exe/ZyPURL.cgi?Dockey=P1006CK9.txt>
- [U.S. EPA](#) (U.S. Environmental Protection Agency). (2005a). Guidelines for carcinogen risk assessment [EPA Report]. (EPA/630/P-03/001F). Washington, DC: U.S. Environmental Protection Agency, Risk Assessment Forum.
https://www.epa.gov/sites/production/files/2013-09/documents/cancer_guidelines_final_3-25-05.pdf

- [U.S. EPA](#) (U.S. Environmental Protection Agency). (2005b). Supplemental guidance for assessing susceptibility from early-life exposure to carcinogens [EPA Report]. (EPA/630/R-03/003F). Washington, DC: U.S. Environmental Protection Agency, Risk Assessment Forum. <https://www.epa.gov/risk/supplemental-guidance-assessing-susceptibility-early-life-exposure-carcinogens>
- [U.S. EPA](#) (U.S. Environmental Protection Agency). (2006a). Approaches for the application of physiologically based pharmacokinetic (PBPK) models and supporting data in risk assessment (Final Report) [EPA Report] (pp. 1-123). (EPA/600/R-05/043F). Washington, DC: U.S. Environmental Protection Agency, Office of Research and Development, National Center for Environmental Assessment. <http://cfpub.epa.gov/ncea/cfm/recordisplay.cfm?deid=157668>
- [U.S. EPA](#) (U.S. Environmental Protection Agency). (2006b). A framework for assessing health risk of environmental exposures to children [EPA Report] (pp. 1-145). (EPA/600/R-05/093F). Washington, DC: U.S. Environmental Protection Agency, Office of Research and Development, National Center for Environmental Assessment. <http://cfpub.epa.gov/ncea/cfm/recordisplay.cfm?deid=158363>
- [U.S. EPA](#) (U.S. Environmental Protection Agency). (2009). Graphical arrays of chemical-specific health effect reference values for inhalation exposures [EPA Report]. (EPA/600/R-09/061). Washington, DC: U.S. Environmental Protection Agency. <http://cfpub.epa.gov/ncea/cfm/recordisplay.cfm?deid=211003>
- [U.S. EPA](#) (U.S. Environmental Protection Agency). (2010). Toxicological review of chloroprene (CASRN 126-99-8) in support of summary information on the Integrated Risk Information System (IRIS) [EPA Report]. (EPA/635/R-09/010F). Washington, DC: U.S. Environmental Protection Agency. <https://nepis.epa.gov/Exe/ZyPURL.cgi?Dockey=P1008LW6.txt>
- [U.S. EPA](#) (U.S. Environmental Protection Agency). (2011a). Recommended use of body weight 3/4 as the default method in derivation of the oral reference dose. (EPA/100/R11/0001). Washington, DC: U.S. Environmental Protection Agency, Risk Assessment Forum. <https://www.epa.gov/sites/production/files/2013-09/documents/recommended-use-of-bw34.pdf>
- [U.S. EPA](#) (U.S. Environmental Protection Agency). (2011b). Toxicological review of trichloroethylene (CAS No. 79-01-6) in support of summary information on the Integrated Risk Information System (IRIS) [EPA Report]. (EPA/635/R-090/11F). Washington, DC: U.S. Environmental Protection Agency, National Center for Environmental Assessment. <https://nepis.epa.gov/Exe/ZyPURL.cgi?Dockey=P100CB6V.txt>
- [U.S. EPA](#) (U.S. Environmental Protection Agency). (2012a). Advances in inhalation gas dosimetry for derivation of a reference concentration (RfC) and use in risk assessment [EPA Report] (pp. 1-140). (EPA/600/R-12/044). Washington, DC: U.S. Environmental Protection Agency. <https://cfpub.epa.gov/ncea/risk/recordisplay.cfm?deid=244650&CFID=50524762&CFTOKEN=17139189>
- [U.S. EPA](#) (U.S. Environmental Protection Agency). (2012b). Benchmark dose technical guidance [EPA Report]. (EPA/100/R-12/001). Washington, DC: U.S. Environmental Protection Agency, Risk Assessment Forum. <https://www.epa.gov/risk/benchmark-dose-technical-guidance>

- [U.S. EPA](#) (U.S. Environmental Protection Agency). (2012c). Toxicological review of tetrachloroethylene (perchloroethylene) (CASRN 127-18-4) in support of summary information on the Integrated Risk Information System (IRIS) [EPA Report]. (EPA/635/R-080/11F). Washington, DC: U.S. Environmental Protection Agency, National Center for Environmental Assessment.
https://cfpub.epa.gov/ncea/iris/iris_documents/documents/toxreviews/0106tr.pdf
- [U.S. EPA](#) (U.S. Environmental Protection Agency). (2013). Mn and BTEX reference value arrays (final reports) [EPA Report]. (EPA/600/R-12/047F). Washington, DC: U.S. Environmental Protection Agency. <https://cfpub.epa.gov/ncea/risk/recordisplay.cfm?deid=250571>
- [U.S. EPA](#) (U.S. Environmental Protection Agency). (2014a). Framework for human health risk assessment to inform decision making. Final [EPA Report]. (EPA/100/R-14/001). Washington, DC: U.S. Environmental Protection, Risk Assessment Forum.
<https://www.epa.gov/risk/framework-human-health-risk-assessment-inform-decision-making>
- [U.S. EPA](#) (U.S. Environmental Protection Agency). (2014b). Guidance for applying quantitative data to develop data-derived extrapolation factors for interspecies and intraspecies extrapolation [EPA Report]. (EPA/100/R-14/002F). Washington, DC: Risk Assessment Forum, Office of the Science Advisor. <https://www.epa.gov/sites/production/files/2015-01/documents/ddef-final.pdf>
- [U.S. EPA](#) (U.S. Environmental Protection Agency). (2015a). Advancing systematic review for chemical risk assessment: Agenda. Advancing Systematic Review Workshop, December 16-17, 2015, Arlington, VA.
- [U.S. EPA](#) (U.S. Environmental Protection Agency). (2015b). Peer review handbook [EPA Report] (4th ed.). (EPA/100/B-15/001). Washington, DC: U.S. Environmental Protection Agency, Science Policy Council. <https://www.epa.gov/osa/peer-review-handbook-4th-edition-2015>
- [U.S. EPA](#). (2015c). Preamble to the integrated science assessments [EPA Report]. (EPA/600/R-15/067). Research Triangle Park, NC: U.S. Environmental Protection Agency, Office of Research and Development, National Center for Environmental Assessment, RTP Division.
<https://cfpub.epa.gov/ncea/isa/recordisplay.cfm?deid=310244>
- [U.S. EPA](#) (U.S. Environmental Protection Agency). (2017a). Guidance to assist interested persons in developing and submitting draft risk evaluations under the Toxic Substances Control Act [EPA Report]. (EPA/740/R17/001). Washington, DC: U.S. Environmental Protection Agency, Office of Chemical Safety and Pollution Prevention.
https://www.epa.gov/sites/production/files/2017-06/documents/tsca_ra_guidance_final.pdf
- [U.S. EPA](#) (U.S. Environmental Protection Agency). (2017b). IRIS assessment plan for ethylbenzene [CASRN 100-41-4]. Washington, DC: U.S. Environmental Protection Agency, National Center for Environmental Assessment.
- [U.S. EPA](#) (U.S. Environmental Protection Agency). (2018a). 2018 Edition of the drinking water standards and health advisories tables [EPA Report]. (EPA/822/F-18/001). Washington, DC: Office of Water, U.S. Environmental Protection Agency.
<https://www.epa.gov/sites/production/files/2018-03/documents/dwtable2018.pdf>

- [U.S. EPA](https://www.epa.gov/aegl/access-acute-exposure-guideline-levels-aegls-values#chemicals) (U.S. Environmental Protection Agency). (2018b). Access Acute Exposure Guideline Levels (AEGs) values database [Database]. Washington, DC: National Research Council, National Academy of Sciences. Retrieved from <https://www.epa.gov/aegl/access-acute-exposure-guideline-levels-aegls-values#chemicals>
- [U.S. EPA](https://www.epa.gov/sites/production/files/2018-06/documents/final_application_of_sr_in_tsca_05-31-18.pdf) (U.S. Environmental Protection Agency). (2018c). Application of systematic review in TSCA risk evaluations [EPA Report]. (EPA/740/P1/8001). Washington, DC: U.S. Environmental Protection Agency, Office of Chemical Safety and Pollution Prevention. https://www.epa.gov/sites/production/files/2018-06/documents/final_application_of_sr_in_tsca_05-31-18.pdf
- [U.S. EPA](https://cfpub.epa.gov/ncea/iris_drafts/recordisplay.cfm?deid=340791) (U.S. Environmental Protection Agency). (2018d). IRIS assessment plan for naphthalene (scoping and problem formulation materials) [CASRN 91-20-3]. Problem formulation draft [EPA Report]. (EPA/635/R-18/007). U.S. Environmental Protection Agency, Integrated Risk Information System. https://cfpub.epa.gov/ncea/iris_drafts/recordisplay.cfm?deid=340791
- [U.S. EPA](https://www.epa.gov) (U.S. Environmental Protection Agency). (2018e). An umbrella Quality Assurance Project Plan (QAPP) for PBPK models [EPA Report]. (ORD QAPP ID No: B-0030740-QP-1-1). Research Triangle Park, NC: U.S. Environmental Protection Agency.
- [U.S. EPA](https://chemview.epa.gov/chemview) (U.S. Environmental Protection Agency). (2019a). ChemView [Database]. Retrieved from <https://chemview.epa.gov/chemview>
- [U.S. EPA](https://cfpub.epa.gov/ncea/isa/recordisplay.cfm?deid=347534) (U.S. Environmental Protection Agency). (2019b). Integrated Science Assessment (ISA) for particulate matter (final report, Dec 2019) [EPA Report]. (EPA/600/R-19/188). Washington, DC: U.S. Environmental Protection Agency. <https://cfpub.epa.gov/ncea/isa/recordisplay.cfm?deid=347534>
- [U.S. EPA](https://www.epa.gov/fera/dose-response-assessment-assessing-health-risks-associated-exposure-hazardous-air-pollutants) (U.S. Environmental Protection Agency). (2020a). Dose-response assessment for assessing health risks associated with exposure to hazardous air pollutants [Website]. Washington, DC: U.S. Environmental Protection Agency. <https://www.epa.gov/fera/dose-response-assessment-assessing-health-risks-associated-exposure-hazardous-air-pollutants>
- [U.S. EPA](https://cfpub.epa.gov/ncea/iris_drafts/recordisplay.cfm?deid=350086) (U.S. Environmental Protection Agency). (2020b). ORD staff handbook for developing IRIS assessments (public comment draft) [EPA Report]. (EPA/600/R-20/137). Washington, DC: U.S. Environmental Protection Agency, Office of Research and Development, Center for Public Health and Environmental Assessment. https://cfpub.epa.gov/ncea/iris_drafts/recordisplay.cfm?deid=350086
- [U.S. EPA](https://cfpub.epa.gov/ncea/iris_drafts/recordisplay.cfm?deid=345309) (U.S. Environmental Protection Agency). (2020c). Systematic review protocol for the methylmercury (MeHg) IRIS assessment (preliminary assessment materials). (EPA/635/R-19/243). Washington, DC: U.S. Environmental Protection Agency. https://cfpub.epa.gov/ncea/iris_drafts/recordisplay.cfm?deid=345309
- [U.S. EPA](https://nepis.epa.gov/Exe/ZyPURL.cgi?Dockey=P1010Y0L.txt) (U.S. Environmental Protection Agency). (2020d). Systematic review protocol for the PFBA, PFHxA, PFHxS, PFNA, and PFDA (anionic and acid forms) IRIS assessments: Supplemental information appendix A [EPA Report]. (EPA/635/R-20/131). Washington, DC: U.S. Environmental Protection Agency, Office of Research and Development, Integrated Risk Information System. <https://nepis.epa.gov/Exe/ZyPURL.cgi?Dockey=P1010Y0L.txt>
- [U.S. EPA](https://comptox.epa.gov/dashboard) (U.S. Environmental Protection Agency). (2021a). CompTox chemicals dashboard. Washington, DC. Retrieved from <https://comptox.epa.gov/dashboard>
- [U.S. EPA](http://www.epa.gov/iris/) (U.S. Environmental Protection Agency). (2021b). Integrated Risk Information System (IRIS) database [Database]. Washington, DC. Retrieved from <http://www.epa.gov/iris/>

- U.S. EPA (U.S. Environmental Protection Agency). (2021c). Pesticide chemical search [Database]. Retrieved from <https://iaspub.epa.gov/apex/pesticides/?p=chemicalsearch:1>
- U.S. EPA (U.S. Environmental Protection Agency). (2021d). Systematic review protocol for the inorganic mercury salts IRIS assessment (preliminary assessment materials). (EPA/635/R-20/239). Washington, DC: U.S. Environmental Protection Agency. https://cfpub.epa.gov/ncea/iris_drafts/recordisplay.cfm?deid=349284
- U.S. EPA (U.S. Environmental Protection Agency). (2021e). Systematic review protocol for the vanadium and compounds (oral exposure) IRIS assessment (preliminary assessment materials). (EPA/635/R-21/047). Washington, DC: U.S. Environmental Protection Agency.
- Vater, ST; McGinnis, PM; Schoeny, RS; Velazquez, SF. (1993). Biological considerations for combining carcinogenicity data for quantitative risk assessment. *Regul Toxicol Pharmacol* 18: 403-418. <http://dx.doi.org/10.1006/rtp.1993.1066>
- Vesterinen, HM; Sena, ES; Egan, KJ; Hirst, TC; Churolov, L; Currie, GL; Antonic, A; Howells, DW; Macleod, MR. (2014). Meta-analysis of data from animal studies: A practical guide. *J Neurosci Methods* 221: 92-102. <http://dx.doi.org/10.1016/j.jneumeth.2013.09.010>
- Villeneuve, DL; Crump, D; Garcia-Revero, N; Hecker, M; Hutchinson, TH; Lalone, CA; Landesmann, B; Lettieri, T; Munn, S; Nepelska, M; Ottinger, MA; Vergauwen, L; Whelan, M. (2014a). Adverse outcome pathway (AOP) development I: Strategies and principles. *Toxicol Sci* 142: 312-320. <http://dx.doi.org/10.1093/toxsci/kfu199>
- Villeneuve, DL; Crump, D; Garcia-Revero, N; Hecker, M; Hutchinson, TH; Lalone, CA; Landesmann, B; Lettieri, T; Munn, S; Nepelska, M; Ottinger, MA; Vergauwen, L; Whelan, M. (2014b). Adverse outcome pathway development II: Best practices. *Toxicol Sci* 142: 321-330. <http://dx.doi.org/10.1093/toxsci/kfu200>
- VT ANR (Vermont Agency of Natural Resources). (2018). Air pollution control regulations. Montpelier, VT. https://dec.vermont.gov/sites/dec/files/aqc/laws-regs/documents/AQCD%20Regulations%20ADOPTED_Dec132018.pdf#page=127
- Washington State Legislature. (2009). Table of ASIL, SQER and de minimis emission values. Olympia, WA. <https://apps.leg.wa.gov/WAC/default.aspx?cite=173-460-150>
- Wasserstein, RL; Lazar, NA. (2016). The ASA's statement on p-values: Context, process, and purpose. *Am Stat* 70: 129-133. <http://dx.doi.org/10.1080/00031305.2016.1154108>
- West, RW; Piegorsch, WW; Pena, EA; An, L; Wu, W; Wickens, AA; Xiong, H; Chen, W. (2012). The impact of model uncertainty on benchmark dose estimation. *Environmetrics* 23: 706-716. <http://dx.doi.org/10.1002/env.2180>
- Wheeler, M; Bailer, AJ. (2009). Comparing model averaging with other model selection strategies for benchmark dose estimation. *Environ Ecol Stat* 16: 37-51. <http://dx.doi.org/10.1007/s10651-007-0071-7>
- Wheeler, MW; Abrahantes, JC; Aerts, M; Gift, JS; Davis, JA. (2022). Continuous model averaging for benchmark dose analysis: Averaging over distributional forms. *Environmetrics* 33: 1-11. <http://dx.doi.org/10.1002/env.2728>
- Wheeler, MW; Blessinger, T; Shao, K; Allen, BC; Olszyk, L; Davis, JA; Gift, JS. (2020). Quantitative risk assessment: Developing a Bayesian approach to dichotomous dose-response uncertainty. *Risk Anal* 40: 1706-1722. <http://dx.doi.org/10.1111/risa.13537>

- [White, RH; Fox, MA; Cooper, GS; Bateson, TF; Burke, TA; Samet, JM.](#) (2013). Workshop report: Evaluation of epidemiological data consistency for application in regulatory risk assessment. *Open Epidemiol J* 6: 1-8. <http://dx.doi.org/10.2174/1874297101306010001>
- [WHO](#) (World Health Organization). (2017). Guidelines for drinking-water quality, incorporating the 1st addendum (4th ed.). Geneva, Switzerland. <https://www.who.int/publications/i/item/9789241549950>
- [WHO](#) (World Health Organization). (2021). Online catalog for the Environmental Health Criteria (EHC) monographs [Database]. Geneva, Switzerland: World Health Organization (WHO). Retrieved from <http://www.who.int/ipcs/publications/ehc/en/>
- [Williams, AJ; Lambert, JC; Thayer, K; Dorne, J.](#) (2021). Sourcing data on chemical properties and hazard data from the US-EPA CompTox Chemicals Dashboard: A practical guide for human risk assessment [Review]. *Environ Int* 154: 106566. <http://dx.doi.org/10.1016/j.envint.2021.106566>
- [Wolffe, TAM; Whaley, P; Halsall, C; Rooney, AA; Walker, VR.](#) (2019). Systematic evidence maps as a novel tool to support evidence-based decision-making in chemicals policy and risk management. *Environ Int* 130: 104871. <http://dx.doi.org/10.1016/j.envint.2019.05.065>
- [Woodall, GM.](#) (2014). Graphical depictions of toxicological data. In P Wexler; M Abdollahi; A De Peyster; SC Gad; H Greim; S Harperk; VC Moser; S Ray; J Tarazona; TJ Wiegand (Eds.), *Encyclopedia of toxicology* (3rd ed., pp. 786–795). Waltham, MA: Academic Press. <http://dx.doi.org/10.1016/B978-0-12-386454-3.01051-4>
- [Woodall, GM; Goldberg, RB.](#) (2008). Summary of the workshop on the power of aggregated toxicity data. *Toxicol Appl Pharmacol* 233: 71-75. <http://dx.doi.org/10.1016/j.taap.2007.12.032>
- [Woodruff, TJ; Sutton, P.](#) (2014). The Navigation Guide systematic review methodology: A rigorous and transparent method for translating environmental health science into better health outcomes [Review]. *Environ Health Perspect* 122: 1007-1014. <http://dx.doi.org/10.1289/ehp.1307175>
- [Worksafe.](#) (2018). Workplace exposure standards and biological exposure indices (10th ed.). Wellington, New Zealand. <https://www.worksafe.govt.nz/topic-and-industry/work-related-health/monitoring/exposure-standards-and-biological-exposure-indices/>
- [Ziliak, ST.](#) (2011). *Matrixx v. Siracusano and Student v. Fisher*. *Significance* 8: 131-134.

APPENDIX A. GLOSSARY

Table A-1. Terms used in the Integrated Risk Information System (IRIS) Handbook

Term	Definition
Adverse outcome pathway (AOP)	An organizational framework providing a visual description of the sequential connections of causally linked key events between a single molecular initiating event and an adverse outcome. An AOP and a mode of action (MOA) both identify key events, but AOPs are not specific the agent initiating the pathway.
Analysis	See definition for <i>Study</i> .
Assessment team	Multidisciplinary team of IRIS Program staff working on the assessment. The team is led by 1 or 2 assessment managers.
Biological plausibility	A proposed association that is consistent with existing biological knowledge. This association can be strengthened by an understanding of the underlying mechanistic pathways involved in connecting the exposure to the adverse outcome.
Biological significance	A characterization made when the magnitude of the effect is considered to be biologically meaningful.
Chemical	As used in this Handbook, chemical is shorthand for environmental agents assessed within the IRIS Program, acknowledging that substances other than chemicals are also often assessed.
Citation	A record of scientific work, including journal publications and unpublished gray literature. A single citation can include multiple individual experiments, studies, or analyses. Alternatively, a single experiment, study, or analysis might be reported in multiple citations. The term citation can also be referred to as publication, record, or reference.
Coherence	The degree to which findings across different but biologically related endpoints are aligned. Coherence is a factor that is considered, in parts, during both evidence synthesis and evidence integration.
Consistency	Similarity of findings (i.e., similar direction) across independent studies or experiments for the same endpoint. Consistency is a factor that is considered during evidence synthesis.
Data	Retrieved, collected, or simulated quantitative or qualitative values (e.g., numbers, observations) that are generally attained from a single citation (e.g., peer-reviewed literature) or source (e.g., model, database).
Data extraction	The process of collecting information about study methods and results from individual studies. Also referred to as data collection or data abstraction.
Dose-response	The relationship between a quantified exposure (dose or concentration) and a quantified change in endpoint response.
Endpoint	An observable or measurable biological change used as an index of a potential health effect of a chemical exposure. Often, “endpoint” is used when describing animal toxicological findings while outcome” is used when describing human findings. Endpoint can also be referred to as effect.
Evaluation domains	The categories of attributes that are evaluated for each study (or outcome/exposure pair within study) during study evaluation

ORD Staff Handbook for Developing IRIS Assessments

Term	Definition
Evidence integration	Integration of animal and human evidence synthesis judgments to draw an overall conclusion(s) that incorporates inferences drawn on the basis of information on the human relevance of the animal evidence, cross-stream coherence across the human and animal evidence, susceptibility, and biological plausibility/MOA. This term is analogous to “weight of evidence” used in some other assessment processes.
Evidence profile table (EPT)	Structured tables summarizing the evidence synthesis and integration conclusions and their justification.
Evidence stream	Types of evidence (i.e., human, animal, mechanistic) that are used to inform evidence synthesis and integration judgments.
Evidence synthesis	A process leading to judgments regarding the certainty of the evidence for hazard from the available human and animal studies. These judgments are made in parallel, but separately. This term is analogous to “strength of evidence” used in some other assessment processes.
Experiment	See definition for <i>Study</i> .
Gray literature	The broad category of data or information sources not found in the standard, peer-reviewed literature databases such as PubMed and Web of Science. Common examples include industry or government reports.
Health Assessment Workspace Collaborative (HAWC)	An interactive web-based content management system for human health assessments that is intended to promote transparency, data usability, and understanding of the data and decisions supporting an environmental and human health assessment. Specifically, EPA HAWC is an application that allows the data and decisions supporting an assessment to be evaluated and managed in modules (e.g., study evaluation, summary study data) that can then be publicly accessed online.
Health effect category	Groups of related health outcomes or endpoints within a biological system (e.g., reproductive; cardiovascular). Each health effect category could have several units of analysis considered for evidence synthesis and integration.
Health and Environmental Research Online (HERO)	HERO is a database of scientific studies and other references cited in EPA assessments. In the assessment process, HERO staff perform the literature search, and the database serves as the repository for the identified references.
Indirectness	The outcome/endpoint being evaluated has an unclear linkage to the apical or clinical outcome of interest or is a surrogate that might not result in the outcome of interest.
IRIS Assessment Plan (IAP)	A planning document released at an early stage of the assessment development process. The IAP includes the rationale for conducting the assessment, scoping information, problem formulation analyses, and key science issues.
Key event	An empirically observable precursor step that is a necessary element of the mode of action or is a biologically based marker for such an element. It could represent a mechanism of action. A single key event is necessary, but not sufficient, for the health effect to occur.
Key event relationship	A sequential relationship between two key events.
Key science issues	Scientific questions or uncertainties that are important to address during the assessment process.
Lifestage	A distinguishable time frame in an individual’s life that is characterized by unique behavioral or physiological characteristics that are associated with development and growth.
Literature inventory	Summary level, sortable lists of the available studies that include additional basic study design elements (e.g., population, species/strain, exposure route/duration) to be used by the subject matter experts to organize and prioritize studies for further review.

ORD Staff Handbook for Developing IRIS Assessments

Term	Definition
Literature inventory tree	Interactive visual display showing the inclusion/exclusion of citations and tagging organized by evidence stream or type of supplemental material. Literature inventory trees are developed in HAWC.
Literature flow diagram	Static visual display of the results from the screening process. These diagrams present the results of the search and selection process, from the number of records identified in the search to the number of studies included in the review.
Mechanism of action	Indicates a specific, critical interaction (e.g., the chemical interacting with a receptor; a secondary effect of exposure on a specific cell type) that is a primary driver of toxicity.
Mode of action (MOA)	A sequence of key events and processes, starting with interaction of a specific agent with a cell, proceeding through operational and anatomical changes, and resulting in the adverse health effect. Applicable to both cancer and noncancer outcomes.
New approach methodologies (NAMs)	Alternative test methods and strategies to reduce, refine, or replace vertebrate animals (e.g., in silico modeling, read-across).
Outcome	An observable or measurable biological change used as an index of a potential health effect of a chemical exposure. Often, “outcome” is used when describing human findings, while “endpoint” is used when describing animal toxicological findings. Outcome can also be referred to as effect.
PECO criteria	“Populations, exposures, comparators, outcomes (PECO)” criteria guide the assessment process. Most IRIS assessments have two sets of interrelated PECO criteria: (1) the problem formulation PECO, which are initial broad PECO criteria that could be refined based on specific priorities identified for the assessment; and (2) the assessment PECO, which are PECO criteria for the assessment that have been refined from the original problem formulation PECO based on information identified during scoping and problem formulation.
Pharmacokinetic (PK)	The study of the movement over time of parent chemical and its metabolite(s) in biological fluids, tissues, and excreta. The terms “toxicokinetic” and “pharmacokinetic” are often used interchangeably. Pharmacokinetic is more aptly used for pharmacologically active compounds, while toxicokinetic would cover toxic compounds. By convention, however, pharmacokinetic is commonly used in EPA, including in the description of physiologically based pharmacokinetic (PBPK) models.
Problem formulation	The process by which the IRIS Program identifies health effects that have been studied in relation to exposure to the chemical, as well as key science issues that might need to be considered for hazard evaluation or deriving toxicity values. Problem formulation is an iterative process that also considers stakeholder input received during the public comment process.
Publication	See definition for <i>Citation</i> .
Record	See definition for <i>Citation</i> .
Reference	See definition for <i>Citation</i> .
Risk of bias	Systematic errors or deviations from the truth, in either results or inferences that affect internal validity of a study; includes factors that might affect the magnitude or direction of an effect in either direction.
Scoping	Scoping is the first stage in the development of an IRIS assessment. The purpose of scoping is to ensure that the IRIS assessment meets the human health chemical toxicity assessment needs of EPA National Programs and Regional Offices.
Sensitivity	A measure of the ability of a study to detect a true effect. Sensitivity is an aspect of study evaluation.

ORD Staff Handbook for Developing IRIS Assessments

Term	Definition
Study	Discrete units of published citations. Also referred to as analyses or experiments. A single citation might contain data on multiple studies, and a single study might be published across multiple citations.
Study evaluation	The evaluation of issues related to risk of bias and sensitivity of studies that meet the assessment PECO. This process involves interpretation of a variety of methodological features (e.g., study design and conduct, exposure measurement or characterization, and selective reporting bias). The study evaluations are aimed at discerning bias that could substantively change a result presented in the study or the interpretation of that result, considering also the expected direction of the bias. The overall goal of the study evaluation is to evaluate the extent to which the results are likely to represent a reliable, sensitive, and informative presentation of a true response.
Strength of the association	Measure of the magnitude of effect (size of the association) observed.
Supplemental material	Information that does not meet PECO but is still potentially relevant to the specific aims of the assessment. This information is tracked during the literature screening process and might inform specific analyses later in the assessment process. Examples of potentially relevant supplemental material include physiologically based pharmacokinetic (PBPK) studies, mechanistic data, and information on susceptible populations.
Susceptible populations	Populations at increased risk from environmental exposures, focusing on biological (intrinsic) factors, as well as social and behavioral determinants that can modify the effect of a specific exposure. Populations affected by extrinsic factors resulting in higher exposures (e.g., proximity, occupation, housing) are not considered as part of the susceptible populations evaluations in IRIS assessments (IRIS assessments do not include exposure assessment).
Systematic Evidence Map (SEM)	A summary of the available evidence. SEMs do not seek to synthesize evidence or draw conclusions but instead to catalogue the evidence base, utilizing systematic search and selection strategies to produce a list of studies along with a high-level summary of key study design characteristics, results and study evaluation (optional). SEMs are often used by the IRIS Program as tools to refine the assessment PECO and identify data gaps.
Systematic review protocol	A structured and transparent procedure for completing the assessment. The initial sections of the protocol are identical to the IAP but have been revised to reflect any adjustments made in response to public input. The protocol also presents the assessment PECO, the units of analyses used for evidence synthesis, and any prioritized analyses of supplemental evidence. Finally, the protocol includes methodological details on the process that will be used for study evaluation, the structured frameworks used during evidence synthesis and integration, dose response, and toxicity value derivation.
Unit of analysis	An endpoint/outcome or group of related endpoints/outcomes within a health effect category that are considered together during evidence synthesis.
Variability	Variability refers to true heterogeneity or diversity, such as variations in risk from similar exposures, due to genetic or lifestyle differences within a population.

AOP = adverse outcome pathway; EPT = Evidence profile table; HAWC = Health Assessment Workspace Collaborative; HERO = Health and Environmental Research Online; IAP = IRIS Assessment Plan; MOA = mode of action; NAMs = new approach methodologies; PBPK = physiologically based pharmacokinetic; PECO = populations, exposures, comparators, outcomes; PK = pharmacokinetic; SEM = systematic evidence map. *Note:* Some terms in this table might be defined differently by other EPA programs for their specific contexts and needs (e.g., the Toxic Substances Control Act Systematic Review Protocol). Additional relevant terms can be found in the IRIS glossary (<https://www.epa.gov/iris/iris-glossary>).

APPENDIX B. SURVEY OF EXISTING ASSESSMENTS AND TOXICITY VALUES

Table B-1 lists organizations that disseminate toxicity values that can be queried to conduct a survey during IRIS problem formulation. In addition to these sources, the ToxVal database on the EPA CompTox Dashboard (https://comptox.epa.gov/dashboard/chemical_lists/TOXVAL_V5) can be searched for reference values, risk estimate values, and points of departure, as described in [Williams et al. \(2021\)](#). Any existing IRIS assessments for the chemical(s) of interest are also included in the survey (<http://www.epa.gov/iris/index.html>). If previous assessments are unavailable or inadequate, the assessment team may conduct an alternative type of survey (e.g., search for recent review articles).

Table B-1. Sources that can be queried for existing assessments and toxicity values, with example search results

Source	Example search results	Reference
American Conference of Governmental Industrial Hygienists (ACGIH)	e.g., See Table X	ACGIH (2007)
American Industrial Hygiene Association (AIHA)	e.g., No results found	AIHA (2016)
Agency for Toxic Substances and Disease Registry (ATSDR)		ATSDR (2021)
California Environmental Protection Agency (CalEPA)		CalEPA (2016)
Connecticut Department of Energy & Environmental Protection (CT DEEP)		CT DEEP (2015)
		CT DEEP (2018)
Deutsche Forschungsgemeinschaft, German Research Foundation (DFG)		DFG (2020)
Drinking Water Standards and Health Advisories (DWSHA)		U.S. EPA (2018a)
Acute Exposure Level Guidelines from the U.S. Environmental Protection Agency and National Research Council) (EPA/NRC AEGL)		U.S. EPA (2018b)
Health Canada		Government of Canada (2021)
		Health Canada (2020)
		Health Canada (1996)
Health and Safety Authority (HSA)		HSA (2020)
Health and Safety Laboratory (HSL)		HSL (2002)

ORD Staff Handbook for Developing IRIS Assessments

Source	Example search results	Reference
Indiana Department of Environmental Management (IDEM)		IDEM (2019)
Idaho Department of Environmental Quality (ID DEQ)		Idaho DEQ (2019)
Institut für Arbeitsschutz, The Institute for Occupational Safety and Health (IFA)		IFA (2020)
Integrated Risk Information System (IRIS)		U.S. EPA (2021b)
International Toxicity Estimates for Risk (ITER)		TERA (2021)
Japan Society for Occupational Health (JSOH)		JSOH (2017)
Massachusetts Department of Environmental Protection (MassDEP)		MassDEP (2019)
Minnesota Department of Health (MDH)		MDH (2019)
Michigan Department of Environment, Great Lakes & Energy (MI EGLE)		Michigan DEQ (2016)
National Air Toxics Information Clearinghouse (NATICH)		U.S. EPA (1993)
North Carolina Department of Environmental Quality (NC DEQ)		NC DEQ (2014)
Nevada Division of Environmental Protection (NDEP)		NDEP (2017)
National Institute for Occupational Safety and Health (NIOSH)		NIOSH (2021)
New Jersey Department of Environmental Protection (NJ DEP)		NJ DEP (2020)
New York State Department of Environmental Conservation (NY DEC)		NYSDEC (2006)
Office of Air Quality Planning and Standards (OAQPS)		U.S. EPA (2020a)
Ontario Ministry of Labour		Ontario Ministry of Labour (2020)
Office of Pesticide Programs (OPP)		U.S. EPA (2021c)
Oregon Department of Environmental Quality (OR DEQ)		Oregon DEQ (2018)
Occupational Safety and Health Administration (OSHA)		OSHA (2019)
		OSHA (2020a)
		OSHA (2020b)
Protective Action Criteria (PAC) Database		DOE (2018)
Publications Quebec		Légis Québec (2020)
Rhode Island Department of Environmental Management (RI DEM)		RI DEM (2008)
		Tiesjema and Baars (2009)

ORD Staff Handbook for Developing IRIS Assessments

Source	Example search results	Reference
Rijksinstituut voor Volksgezondheid en Milieu (RIVM), The Netherlands Institute for Public Health and the Environment		Dusseldorp et al. (2011)
		RIVM (2001)
Safe Work Australia		Safe Work Australia (2019)
Southwest Clean Air Association (SWCAA)		SWCAA (2021)
Texas Commission on Environmental Quality (TCEQ)		TCEQ (2021)
		TCEQ (2018)
United States Army Public Health Center (U.S. APHC)		U.S. APHC (2013)
Vermont Department of Environmental Conservation (VT DEC)		VT ANR (2018)
Washington State Dept. of Ecology		Washington State Legislature (2009)
Worksafe		Worksafe (2018)
World Health Organization (WHO)		WHO (2017)
		WHO (2021)

APPENDIX C. EXAMPLE ISSUES FROM EXISTING INTEGRATED RISK INFORMATION SYSTEM (IRIS) ASSESSMENTS

Table C-1. Examples of key science issues in Integrated Risk Information System (IRIS) assessments

Key science issue topic area	Examples
Human relevance of findings in animals	<ul style="list-style-type: none"> Human relevance of effects in animals that involve peroxisome proliferator-activated receptor alpha (U.S. EPA, 2020d) Interspecies differences in metabolism (U.S. EPA, 2018d, 2017b)
Whether an endpoint is considered adverse or adaptive	<ul style="list-style-type: none"> Hepatic effects such as increased liver weight, cellular hypertrophy, and single cell necrosis/apoptosis (U.S. EPA, 2020d)
Issues where conflicts in the evidence are known, including hypothesized modes of action that lack scientific consensus	<ul style="list-style-type: none"> Modes of action for mouse lung tumors (U.S. EPA, 2018d, 2017b)
Identification of published physiologically based pharmacokinetic models that have no or limited in vivo validation data	<ul style="list-style-type: none"> Physiologically based pharmacokinetic model for chloroprene (U.S. EPA, 2010)
Complex chemistry issues that can affect toxicity, pharmacokinetic, or applicability of toxicity values to different forms of the chemical	<ul style="list-style-type: none"> Pharmacokinetic differences across sexes and species (U.S. EPA, 2021d, 2020d) Consideration of speciation and toxicity/pharmacokinetic differences across different metal salts (U.S. EPA, 2021d, e)
Confounding by co-exposures in epidemiological studies	<ul style="list-style-type: none"> Correlation of exposure among a group of similar chemicals, e.g., PFAS (U.S. EPA, 2020d) and phthalates (Radke et al., 2018) Coexposure to beneficial nutrients (e.g., selenium, polyunsaturated fatty acids) and harmful contaminants (e.g., polychlorinated biphenyls) in populations exposed to methylmercury in fish (U.S. EPA, 2020c)
Issues where chemical-specific information is missing	<ul style="list-style-type: none"> Application of analogue-based read-across methods (U.S. EPA, 2021d) Using well-studied PFAS (e.g., PFOA, PFOS) to characterize data gaps for less studied PFAS (U.S. EPA, 2020d)

PFAS = per-and polyfluoroalkyl substances; PFOA = perfluorooctanoic acid ; PFOS = perfluorooctane sulfonic acid.

APPENDIX D. SUPPLEMENTAL DATABASES

Table D-1. Supplemental databases that may be searched by the assessment team depending on the topic

Supplemental databases	Description
AEGLs	<p>AEGLs represent threshold exposure limits of airborne concentrations for the general public applicable to emergency exposures ranging in duration from 10 min to 8 h. AEGL-1 is the concentration above which individuals could experience notable discomfort, irritation, or certain asymptomatic nonsensory effects. AEGL-2 is the concentration above which individuals could experience irreversible or other serious, long-lasting adverse health effects. AEGL-3 is the concentration above which individuals could experience life-threatening adverse health effects or death.</p> <p>AEGLs and their technical support documents are available from the following website: https://www.epa.gov/aegl/access-acute-exposure-guideline-levels-aegls-values#chemicals.</p>
Agricola	<p>Use for U.S. Department of Agriculture-related compounds. Available through HERO. Test page for developing searches: http://agricola.nal.usda.gov/.</p>
ChemIDPlus	<p>Includes links to resources from a variety of sources in the United States (e.g., ATSDR; Registry of Toxic Effects of Chemical Substances) and other countries (OECD member country assessments of HPV chemicals, summaries of studies submitted to ECHA under REACH, International Uniformed Chemical Information database, IUCLID): http://chem.sis.nlm.nih.gov/chemidplus/.</p> <p>Note that although IUCLID houses similar data, the OECD HPV assessments, or SIAPs and SIARs, do have some government review/oversight. IUCLID summaries can simply house study summaries provided by industry without review by government. OECD SIARs/SIAPs are available through the eChemPortal (https://www.echemportal.org/echemportal/index.action, listed as OECD HPV).</p>
DTIC	<p>Contains government-funded (primarily Department of Defense) research, studies, and other materials relevant to the defense community. Advance search options available through the R&E gateway. Requires government sponsor to access advanced search options: https://www.dtic.mil/DTICOnline/.</p>
ECOTOX (Optional)	<p>Review of the list of references in the ECOTOX database for the chemical(s) of interest (https://cfpub.epa.gov/ecotox/).</p>
Japan CHEmicals Collaborative Knowledge database (J-CHECK)	<p>Japan CHEmicals Collaborative Knowledge database (J-CHECK, http://www.safe.nite.go.jp/jcheck/top.action) is a database developed to provide the information regarding "Act on the Evaluation of Chemical Substances and Regulation of Their Manufacture, etc." (CSCL) by the authorities of the law, Ministry of Health, Labour and Welfare, Ministry of Economy, Trade and Industry, and Ministry of the Environment. J-CHECK provides the information regarding CSCL, such as the list of CSCL, chemical safety information obtained in the existing chemicals survey program, risk assessment, etc. in cooperation with eChemPortal by OECD.</p>

ORD Staff Handbook for Developing IRIS Assessments

Supplemental databases	Description
OPP, EPA^a IHAD	Contains DERs (reviews of toxicological study reports), memoranda, cancer reports, metabolism reports, etc. for all of OPP. Accessible to any EPA employee with FIFRA confidential business information access authorization.
OPP, EPA^a PRISM Documentum	Contains GLP guideline toxicological study reports for all pesticides from 1996 to present. Study reports older than 1996 can be acquired within a few days. Accessible to any EPA employee with FIFRA confidential business information access authorization. Go to: OPP@Work— http://intranet.epa.gov/opp00002/ (might require permission). OPP Applications (under popular sites in green box on left). e-Registration Workflow (Documentum Login).

AEGL = acute exposure guideline level; ATSDR = Agency for Toxic Substances and Disease Registry; CASRN = Chemical Abstracts Service registry number; CSCL = Chemical Substances Control Law; DER = data evaluation record; DTIC = Defense Technical Information Center; ECHA = European Chemicals Agency; FIFRA = Federal Insecticide, Fungicide, and Rodenticide Act; GLP = Good Laboratory Practice; HERO = Health and Environmental Research Online; HPV = high production volume; IHAD = Integrated Hazard Assessment Database; IUCLID = International Uniformed Chemical Information Database; J-Check = Japan CHEMicals Collaborative Knowledge database; OECD = Organisation for Economic Co-operation and Development; OPP = Office of Pesticide Program; PRISM = Pesticide Registration Information System; R&E = research and engineering; REACH = Registration, Evaluation, Authorisation and Restriction of Chemicals; SIAP = SIDS Initial Assessment Profile; SIAR = SIDS Initial Assessment Report.

^aContractors do not have access to PRISM Documentum or IHAD; other pesticide databases, such as the National Pesticide Information Retrieval System through Purdue University, can also be assessed for relevance.

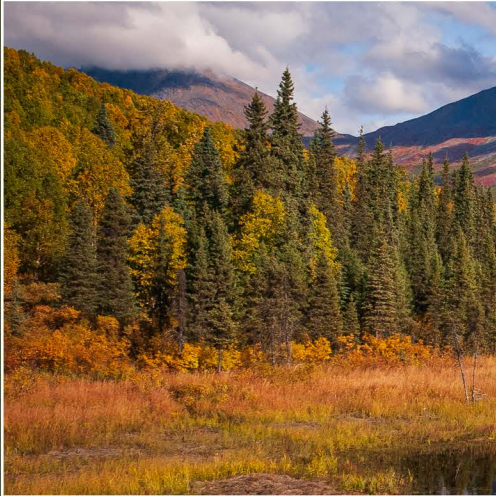
APPENDIX E: ESTIMATING TIME TO CONDUCT THE ASSESSMENT

In general, two screeners (ideally including at least one from the assessment team) should perform the literature screening using screening software. Screening is first done at the title or abstract level with subsequent screening at the full-text level. All decisions regarding tagging during the screening process should be tracked in the screening software and made available through the Health and Environmental Research Online (HERO) literature database upon public release of assessment-related documents, including assessment plans, protocols, and draft assessments. Disseminated content includes the list of all studies considered, categorized by those that were included, those that were excluded, and those marked as supplemental material. When studies cited in prior assessments need to integrate with a new analysis, the studies from the prior assessment should be reviewed for populations, exposures, comparators, and outcomes (PECO) relevance and tagged according to source. The time estimates in Table E-1 show a range of average times for experienced reviewers that can be used to estimate project timelines.

Table E-1. Time estimates per study

Phase	Average time estimate per study
Title and abstract review	10–20 sec (180–360 per h)
Title and abstract screening + characterization of relevant studies by type of study population (human, animal, in vitro, in silico), type of health outcome, or as supplemental material	30 sec (120 per h)
Full-text screening + reason for exclusion, characterization of relevant studies by type of study population (human, animal, in vitro, in silico), type of health outcome, or supplemental material	3–5 min (12–20 per h, depending on study complexity)
Literature inventory	5–15 min (4–12 per h, depending on study complexity)
Study evaluation	0.5–2.5 h (depending on study complexity and type)
Data extraction	1–4 h (depending on study complexity)

Note: Time estimates are after the pilot phase and assume familiarity with screening software platforms. During the pilot phase, time estimates for each step could double. Pilot testing study number estimates: title and abstract review (100 studies), full-text review (10–20 studies), and study evaluation and data extraction (2–5 studies, depending on diversity of studies).



PRESORTED STANDARD
POSTAGE & FEES PAID
EPA
PERMIT NO. G-35

Office of Research and Development (8101R)
Washington, DC 20460

Official Business
Penalty for Private Use
\$300



Recycled/Recyclable Printed on paper that contains a minimum of 50% postconsumer fiber content processed chlorine free