EPA/600/R-20/137
www.epa.gov/ord

ORD Staff Handbook for Developing IRIS Assessments

Version 1.0

November 2020

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

Disclaimer

This document is distributed solely for the purpose of pre-dissemination public comment under applicable information quality guidelines. It has not been formally disseminated by the U.S. Environmental Protection Agency. It does not represent and should not be construed to represent any agency determination or policy.

ACKNOWLEDGMENTS

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

Primary Authors (currently or formerly EPA/ORD)
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
Kristina Thayer
Teneille Walker
George Woodall
Erin Yost

Additional Contributors (currently or formerly EPA/ORD)
Norman Birchfield
Johanna Congleton
Jeffry Dean
Ingrid Druwe
John Fox
Jason Fritz
John Lipscomb
Lucina Lizarraga
Roman Mezencev
Margaret Pratt
Susan Rieth
Paul Schlosser
Ravi Subramaniam

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

TABLES

FIGURES

ABBREVIATIONS

ADME	absorption, distribution, metabolism, and excretion
AEGL	acute exposure guideline level
AOP	adverse outcome pathway
ATSDR	Agency for Toxic Substances and Disease Registry
BMD	benchmark dose
BMDL	benchmark dose lower confidence limit
BMDS	Benchmark Dose Software
BMR	benchmark response
CASRN	Chemical Abstracts Service registry number
CI	confidence interval
COI	conflict of interest
CPHEA	Center for Public Health and Environmental Assessment
DNA	deoxyribonucleic acid
DTIC	Defense Technical Information Center
ECHA	European Chemicals Agency
EPA	U.S. Environmental Protection Agency
FIFRA	Federal Insecticide, Fungicide, and Rodenticide Act
GLP	Good Laboratory Practices
GRADE	Grading of Recommendations Assessment, Development, and Evaluation
HAWC	Health Assessment Workspace Collaborative
HEC	human equivalent concentration
HED	human equivalent dose
HERO	Health and Environmental Research Online
IAP	IRIS Assessment Plan
IARC	International Agency for Research on Cancer
IHAD	Integrated Hazard Assessment Database
IPCS	International Programme on Chemical Safety
IRIS	Integrated Risk Information System
LOAEL	lowest-observed-adverse-effect level
LOD	limit of detection
MeSH	Medical Subject Heading
MOA	mode of action
NAMs	new approach methodologies
NAS	National Academy of Sciences
NIEHS	National Institute of Environmental Health Sciences
NLM	National Library of Medicine
NMD	normalized mean difference
NOAEL	no-observed-adverse-effect level
NRC	National Research Council
NTP	National Toxicology Program
OCSPP	Office of Chemical Safety and Pollution Prevention
OECD	Organisation for Economic Co-operation and Development
OHAT	Office of Health Assessment and Translation
OPP	Office of Pesticide Programs
OR	odds ratio
PBPK	physiologically based pharmacokinetic
PC	partition coefficient
PECO	populations, exposures, comparators, and outcomes
PK	pharmacokinetic
POD	point of departure
PRISM	Pesticide Registration Information System
QA	quality assurance
QC	quality control
RfC	reference concentration
RfD	reference dose
RfV	reference value
RoB	risk of bias
ROBINS-I	Risk of Bias in Nonrandomized Studies of Interventions
SciRAP	Science in Risk Assessment and Policy
SD	standard deviation
SE	standard error
SOP	standard operating procedure
SR	Systematic Review
TCE	trichloroethylene
TK	toxicokinetics
Tox21	Toxicology in the 21st Century
TSCA	Toxic Substances Control Act
TSCATS	Toxic Substances Control Act Test Submissions
UF	uncertainty factor
Vd	volume of distribution
WHO	World Health Organization
WOS	Web of Science

PREFACE

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.
- IRIS assessments incorporate public input and expert peer review during development.
- The IRIS Program is multidisciplinary and decentralized, spanning multiple organizational divisions and geographic locations within ORD.
- The implementation of systematic review principles improves the rigor, transparency, and coherence of IRIS assessments.
- The IRIS Handbook provides operating procedures for developing assessments (Step 1 of the IRIS 7-step process: IRIS process) including incorporation of systematic review principles; it does not address later review steps of the IRIS process.
- The handbook does not supersede existing EPA risk assessment guidelines and does not serve as guidance for other EPA programs.
- This is intended to be a "living document"; updates will be based on emerging science and experience gained through its application.
- Ongoing assessments developed with previously established procedures may not reflect all the approaches or procedures as described in the handbook.

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. It also describes approaches for assessment development activities not explicitly addressed in current guidelines. The EPA guidelines have been developed over time and address the state of the science at the time they were developed. Thus, portions of the handbook may be updated as new science emerges, or when existing guidelines are updated.

The Integrated Risk Information System (IRIS) Program

The mission of the 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 evaluation of dose-response[2] for those potential hazards, the first two steps in the risk assessment paradigm. IRIS assessments are made available to Agency and Regional programs who complete the risk assessment process factoring in exposure and risk characterization. These assessments may also be used by state regulators, tribes, and international entities.

Systematic Review in Integrated Risk Information System (IRIS) Assessments

Systematic review plays an important role in enhancing scientific rigor and transparency. The principles of systematic review have been well developed in the context of evidence-based medicine (e.g., evaluating efficacy in clinical trials) and have recently been adapted for use across a more diverse array of scientific fields. IRIS assessments use the best available scientific information to answer the question(s) that are the focus of the review. It is important to recognize that EPA Cancer Guidelines describe approaches for drawing judgments regarding the "available data" (U.S. EPA, 2005b); that is, IRIS assessments 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 need for EPA customers to receive timely products from the IRIS Program.

The IRIS Program is helping to advance the science of systematic review by improving the application of methods to the types of studies typically available for IRIS assessments. Human studies 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 toxicokinetics and exposure-response relationships, and deriving toxicity values.

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

[2]The 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.

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 products[3]), and workshops involving expert practitioners of systematic review. In their 2014 review of the IRIS Program (NRC, 2014), the NAS recommended the explicit inclusion of the principles of systematic review as a sequential process during Step 1 of the IRIS process, as illustrated in **Figure i-1**. The IRIS Handbook has adapted this schematic recommendation for use as its underlying structural organization (see **Figure O-1** in the overview section). In addition to presenting stages ancillary to "systematic review," including scoping, problem formulation, and dose-response assessment, both figures highlight 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. This approach was further supported in a follow-up review by NAS in 2018 (NASEM, 2018).

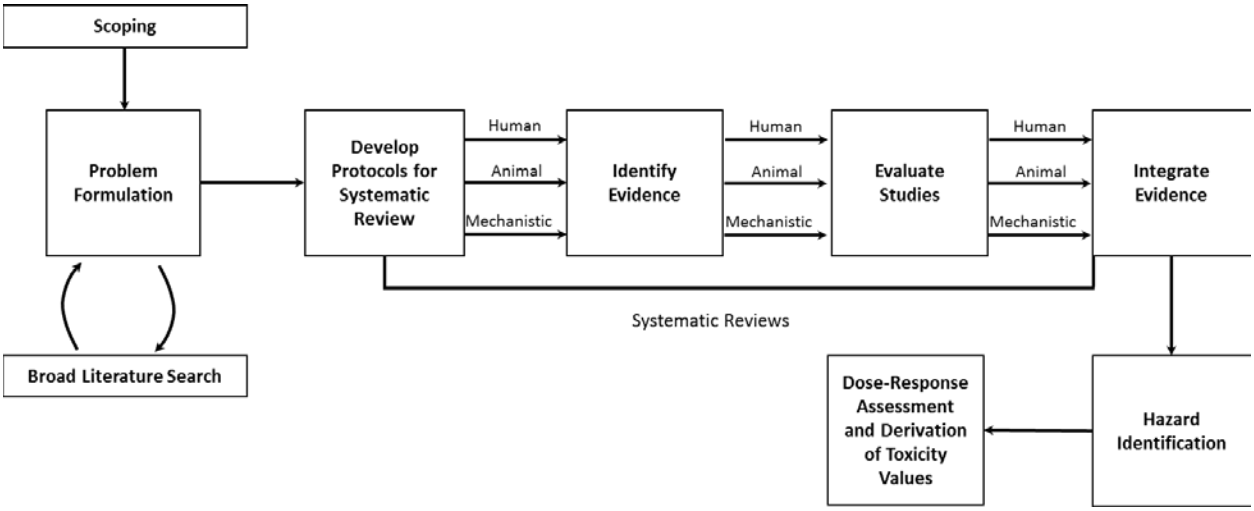


Figure i-1. National Academy of Sciences (NAS) illustration for considering systematic review in the context of the Integrated Risk Information System (IRIS) process. See Figure 1-2 in **NRC (2014)**, noting that although public input and peer review are not depicted, they are viewed as integral components of the IRIS process.

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 to retrieve, organize, evaluate, synthesize, integrate, and present scientific information in a more structured and transparent manner. Consistent with EPA's Framework for Human Health Risk

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

Assessment to Inform Decision Making (U.S. EPA, 2014a), IRIS assessment development begins with planning and scoping and problem formulation and the production of a conceptual model and analysis plan, which are described in the IRIS Assessment Plan (IAP) and protocol associated with the assessment. The systematic review approaches described in this handbook are used to develop the human health assessment (hazard and dose-response assessment), which is the core component of risk assessment addressed by IRIS assessments. These approaches include a literature identification strategy and evidence identification; evaluation of study methods; synthesis of the evidence from human, animal and mechanistic streams; integration of the evidence; and hazard identification. The IRIS assessment process also includes a systematic approach to the selection of studies for dose response to provide a transparent rationale for the decisions that guide the dose response assessment. However, most of the procedures described for conducting dose-response analyses are not amenable to the application of systematic review principles. An overarching goal of these procedures is to promote an efficient and productive IRIS Program. In alignment with the Framework's emphasis on tailoring risk assessments to inform the decision-making process in a meaningful way, the IRIS assessment development process is intended to be "fit for purpose." The specific needs of a particular assessment will determine which procedures are applicable based on the scoping and problem formulation activities to focus the assessment objectives to address the identified research question(s) which may include a modular approach (e.g., restrictions in scope or sequential development of specific health effects, such as cancer and other 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 IRIS ASSESSMENTS

OVERVIEW

Purpose

- The IRIS Handbook provides consistent procedures for each stage of draft development (Step 1 of the IRIS process).

Who

- Each IRIS assessment is developed by an assessment team.

What

- **Chapters 1–13** lay out the sequential stages for developing a complete draft assessment as Step 1 (Draft Development) of the IRIS process.

The IRIS Handbook is intended to provide operating procedures for the development of IRIS assessments to promote consistency and ensure that all contributors to IRIS assessments understand how the assessment components, including those that are part of systematic review, fit together, and at what points in the process the components are anticipated to occur. The 13 chapters in the IRIS Handbook describe each of the sequential stages that are 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 that 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 populations, exposures, comparators, and outcomes that the assessment will address.
  - 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 that are 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 are generally comprised of Office of Research and Development (ORD) scientists but can also include scientists from elsewhere in U.S. Environmental Protection Agency (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

The assessment development process is a sequential one. Although an initial, generalized systematic review protocol can be developed based on the health outcomes identified as a result of problem formulation, the detailed strategies for study evaluation and data extraction, in particular, are developed later and are informed by the preceding stages. The process also is iterative; new insights can require revision of the PECO(s), additional targeted literature searches and screening, reevaluation of studies or additional extraction of data to develop hazard conclusions. Revisions and additions to the protocol are documented during assessment development.

Figure O-1 summarizes the assessment development process from initial scoping through the derivation of toxicity values for identified hazards. It draws from the recommended process described in Figure 1-2 in the National Academy of Science (NAS) review (NRC, 2014) with some differences (in red). The IRIS process applies a systematic review approach from the literature identification stage through the selection of studies for dose-response assessment. In addition, while absorption, distribution, metabolism, and excretion (ADME) and mechanistic studies are identified in a comprehensive literature search process and serve to inform the refined evaluation plan, these studies are not the focus of the study evaluation strategies, which are developed for the health effects studies. ADME and mechanistic studies may have different levels of impact depending on assessment needs, and are therefore screened, categorized, and prioritized in a stepwise fashion, applying a greater focus to identify the most impactful studies for further

evaluation. These prioritization steps, undertaken through the assessment process, facilitate the analysis of mechanistic information to best inform the synthesis and integration of the evidence.

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 topic of each chapter with a brief description is provided in **Table O-1**.

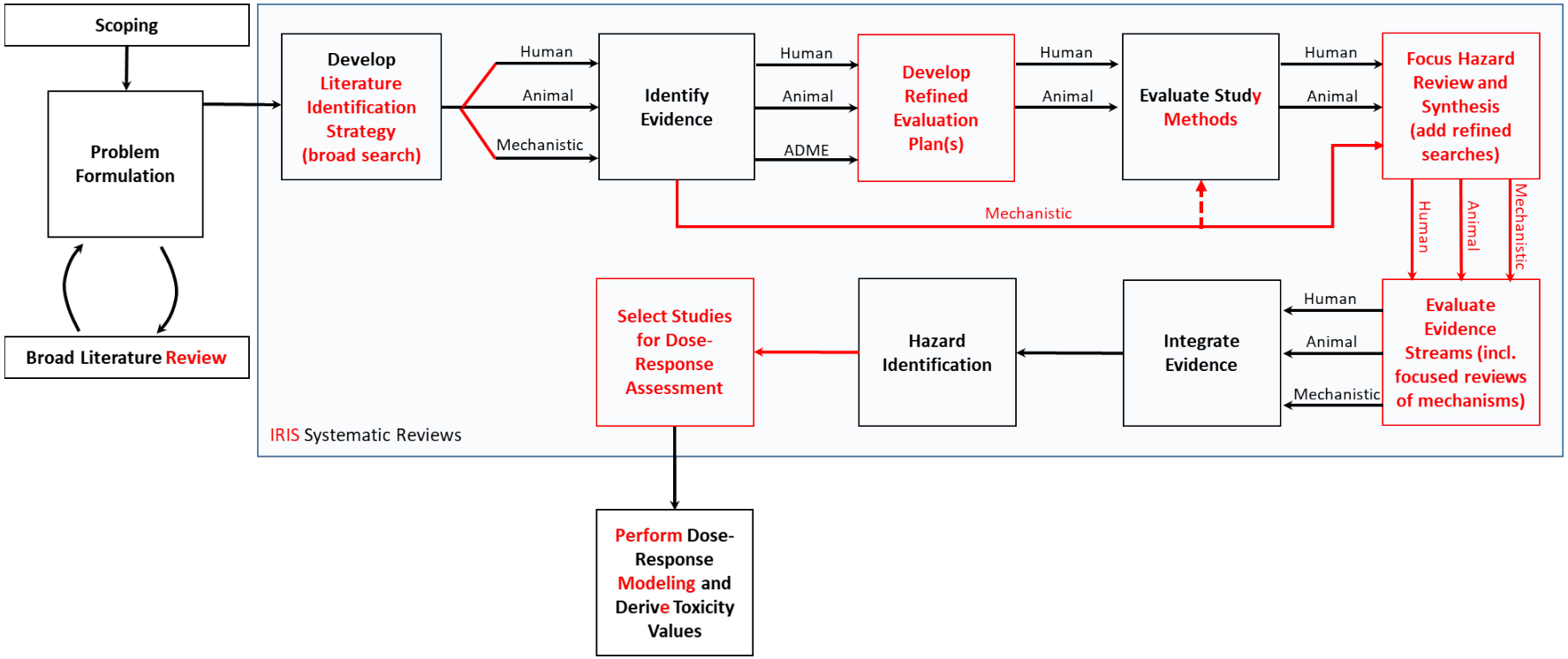


Figure 0-1. Integrated Risk Information System (IRIS) assessment draft development process.

Building upon the NAS illustration for considering systematic review in the context of the IRIS process [see Figure 1-2 in NRC (2014)], the IRIS draft 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 NAS 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 is incorporated at multiple stages of the process; this complexity is described in **Chapter 10**.

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	1	Define the parameters of the assessment. Develop Scoping Statement with EPA program offices.
Problem formulation	2	Preliminary literature survey. Describe health effects of potential interest and key science issues including pre-defined mechanistic analyses. Develop Assessment Plan containing draft PECO criteria. **Output: IRIS Assessment Plan (at least a 30-day public comment period).**
Systematic review protocol	3	Describes systematic review procedures for: PECO criteria, literature identification, study evaluation, data extraction/display. Record any updates in protocol history. **Output: Assessment Protocol (at least a 30-day public comment period).**
Literature search and screening • Identify health effect studies • Identify other informative studies relevant to evaluating potential health effects	4	Perform comprehensive literature search(es). May be overarching or specific to outcome or evidence stream. Use PECO criteria to identify relevant human and animal health effect studies. Identify ADME studies and pharmacokinetic (PK) and PBPK models from broad search or other, not necessarily systematic, searches. Identify mechanistic studies from broad search(es).
Literature inventory • Human, animal, mechanistic studies	4	Categorize studies as described in the protocol (e.g., by study type, health effect). Extract cursory information from relevant studies to allow for organization by study design/ mechanism.
Refined evaluation plan • Prioritize, refine PECO, define endpoint groupings	5	Summarize and interpret the impact of ADME data. Decide whether and how to prioritize and group sets of related endpoints into health effect categories for review, focusing on those most likely to inform hazard identification. Incorporate decisions into revised protocol.

Assessment development stage	Chapter	Purpose and other useful details
Study evaluation • Evaluate health effect studies for risk of bias and insensitivity • Evaluate PBPK Models and other information as needed	6	Evaluate individual human and animal health effect studies, considering bias and sensitivity. Evaluate PK and PBPK models and other information (e.g., mechanistic) as needed.
Organize hazard review • Presentation decisions	7	Finalize utility and organization of health effect categories and studies for hazard identification. Informed by study evaluation, toxicokinetic, and other mechanistic information (see below).
• Organize and prioritize relevant mechanistic information	7	Prioritize the identified mechanistic studies most relevant to the apical health effects under review. Determine presentation and focus areas for evaluation. Consider the need for additional, focused literature searches.
Data extraction and display • Human and animal health effects studies	8	Collect key health effect study information in a database and prepare graphical and tabular displays.
Synthesis • Human studies	9	Analyze results incorporating the strengths and weaknesses of the sets of human health effect studies by health effect or other selected grouping.
• Animal studies	9	Analyze results incorporating the strengths and weaknesses of the sets of animal toxicology studies by health effect or other selected grouping.
• Mechanistic information (data extraction, display, analysis, and synthesis)	10	Conduct focused, stepwise analyses of the most relevant mechanistic evidence and summarize results of the analyses by health effect or other selected grouping. More flexible approach compared to analyses of human and animal data on apical effects, dependent on the unique needs of the assessment. This step will be informed by considerations that arise from analyzing animal and human data (e.g., biological plausibility, human relevance, precursor data).

Assessment development stage	Chapter	Purpose and other useful details
Integration • 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)	11	Prepare evidence integration narrative for hazard identification and overall summary conclusions. Each narrative summarizes the strength of the evidence from the available human and animal health effect studies, incorporating mechanistic data (e.g., precursors) important for decisions regarding biological plausibility and coherence within evidence streams as well as consideration of human relevance, coherence of effects, and susceptibility across streams.
Hazard considerations and study selection for deriving toxicity values	12	Select the most informative studies and outcomes for dose-response analysis based on study confidence and other predefined considerations including hazard identification decisions and susceptibility.
Derive toxicity values • Cancer and/or noncancer	13	Model studies and develop a quantitative estimate for each hazard of concern. Consider uncertainty and susceptibility and describe confidence in the estimates. **Output: Draft Assessment (typically, a 60-day public comment period).**

PBPK = physiologically based pharmacokinetic; PK = pharmacokinetic.

Scoping and Problem Formulation: The Exploratory Phase

Scoping is the first stage in the development of an IRIS assessment. It involves early consultation and continued communication with clients in EPA program and regional offices to identify the information and level of detail required for the assessment to support EPA needs. **Chapter 1** provides an overview of the scoping process for IRIS staff to follow as they initiate an assessment.

The purpose of problem formulation is to identify potential health effects, type of studies, and science issues to be considered during assessment plan development. **Chapter 2** provides a description of the process used to develop the IRIS Assessment Plan (which describes what the assessment will cover) including the approach to information gathering, compilation of a preliminary literature survey, and the contents of the plan. This information is subsequently used to develop specific questions and considerations for the assessment's systematic review protocol (which describes how the assessment will be conducted). A summary of elements that are part of an IRIS assessment systematic review protocol are provided in **Chapter 3**.

PECO criteria, described in the assessment plan and subsequently the protocol, provide the framework for developing literature search strategies and inclusion/exclusion criteria, particularly with respect to exposure measures and outcome measures. The scope of the assessment and

research questions are reflected in the PECO, which may be revised either to broaden or narrow the assessment's scope as the available evidence becomes better understood. Note that PECOs identify the health effects that are the focus of the review; additional search strategies will be needed to include other important information, especially ADME and mechanistic studies.

Literature Search and Screening

The literature search is developed by IRIS Program staff working in conjunction with information specialist(s), either through a contractor or through EPA library services. **Chapter 4** describes the search and screening process for studies of health effects (i.e., animal toxicology or epidemiology studies) and for studies providing ADME and mechanistic information. The literature search strategy (including electronic database searches and other methods to identify studies, and specifying screening criteria), which draws from the decisions from the scoping and problem formulation stages, is developed, tested, and implemented. The references that result from the broad literature search strings are then screened using the selected criteria to compile a literature inventory of studies that will be included in the assessment.

IRIS Program staff (working with a contractor when necessary) develop an inventory of the identified studies, abstracting key elements (e.g., route of exposure, categorization of the exposure, or outcome measures). After the literature inventory is compiled and sorted by discipline and outcome, it is reviewed by the assessment team who, in consultation with other experts, may decide to conduct additional targeted searches and, for large databases, may need to systematically narrow the focus of the study evaluation process on a smaller number of the more informative studies. The rationale for decisions regarding grouping of outcomes, refining the set of health effects studies, and the evaluation strategies are developed as part of the refined evaluation plan, which is described in the protocol (see **Chapter 5**).

Study Evaluation

The considerations for evaluating epidemiology and animal toxicology studies reporting health effects data are developed by IRIS staff, working with additional subject matter experts as needed. This process is described in **Chapter 6**. Study descriptions and methods are assessed for risk of bias and sensitivity (the ability of the study to detect the potential effect in question), both assessed using several study design domains on an outcome-specific basis. Based on this evaluation, each study (or a specific analysis in a study) is classified as *high*, *medium*, *low*, or *uninformative* with respect to confidence in the results. Evaluation and analysis of ADME data and physiologically based pharmacokinetic (PBPK) models (sometimes denoted physiologically based toxicokinetic models) and mechanistic information also are described in **Chapter 6**. The evaluation of the quality of this evidence involves a different approach than that used to evaluate the health effects studies. A key part of this implementation is the documentation of the decisions made in the evaluation process.

At this point, the set of studies and outcomes is known, and conclusions about the confidence in these studies have been made. The organization of the synthesis needs to reflect this information, and to further elaborate the outcomes or groupings of outcomes that will contribute to the integration of evidence. Decisions about what study results to extract and how to display them follow from this prioritization and organization process. Approaches to organizing the synthesis and making decisions about what data to extract are provided in **Chapter 7**, and the development of a plan for data extraction and advice for the displays of study results are described in **Chapter 8**.

Synthesis of Human and Experimental Animal Data

For the purposes of IRIS assessments, evidence synthesis and integration are considered distinct but related processes. The first phase in the evaluation of potential hazards involves the analysis and synthesis of evidence within each of the human and animal evidence streams. This procedure is described in **Chapter 9**. The syntheses of separate evidence streams (i.e., human and animal evidence) described in this section and in **Chapter 10** (i.e., mechanistic evidence) will directly inform the integration across the evidence streams to draw overall conclusions for each of the assessed human health effects (described in **Chapter 11**). A key component of the synthesis is the analysis of variation in results and of factors that could contribute to this variability (e.g., specific methodological differences, exposure range, length of follow-up, type of exposure setting, age group, species, strain, age at dosing). PBPK models of internal dose may be useful in the comparison of results in different species and across routes of exposure. Where applicable, statistical methods for synthesizing evidence within a set of studies, such as meta-analysis, may be informative. The synthesis analyzes the evidence-specific factors that may increase or decrease the certainty that a chemical exposure poses a hazard, including evidence of consistency and coherence across related endpoints.

The review and synthesis of mechanistic evidence (see **Chapter 10**) occurs prior to and in conjunction with the integration of evidence within and across humans and animals. Some types of mechanistic evidence may be considered within the synthesis of human or animal effects evidence, and some types may be considered in a separate stage. Mechanistic information can be defined as any endpoint or experimental outcome that informs how toxicity results from exposure to the agent of interest, thereby lending support to conclusions of hazard based on human and animal evidence. Mechanistic information may be an observation or measurement of a molecular interaction or biological effect in humans, or experimental studies in animals, in vitro, or in silico. While there may be numerous mechanistic pathways to consider for any given chemical, these analyses primarily focus on determinations regarding biological plausibility (e.g., establishing the occurrence

of precursor events that are attributable to the agent), human relevance of effects in animals, and identifying susceptible human populations or lifestages.[4]

Developing Summary Hazard Judgments across Evidence Streams

The integration of evidence involves narrative summaries that bring together the findings from the analyses of the informative evidence relevant to each potential human health hazard, including summary judgments regarding the strength of the evidence (for or against an effect) from each evidence stream and as a whole (aka, a weight-of-evidence analysis of the totality of evidence). During evidence integration, a set of factors describing aspects of the evidence (e.g., consistency; dose-response) is evaluated for each assessed hazard using structured frameworks and predefined considerations across the sets of relevant studies (both positive and null). These evaluations of the available studies of exposed humans and experimental animals inform interpretations about the extent to which the data support a judgment that a human health hazard exists (or is unlikely to exist), given relevant exposure circumstances. The interpretations regarding the strength of the available human and animal evidence (including mechanistic evidence informing biological plausibility) are judged and then considered together with mechanistic information on the human relevance of the animal data, coherence of the findings across human and animal studies, and the available information on susceptible populations and lifestages. This culminates in a final judgment about the extent to which the available evidence supports that the chemical poses (or is unlikely to pose) each hazard in humans. This procedure is described in **Chapter 11**. Conclusions can be drawn for a broader outcome category (such as neurotoxicity or carcinogenicity), or finer levels of organization may apply. For example, a subgrouping for neurotoxicity could involve behavioral effects, or on a finer level, hyperactivity, depending on the scope of the assessment, size of the database, and specificity of the available evidence. **Figure O-3** displays the process of identifying and evaluating health effects studies and confidence in their results, and then making a final judgment about whether the chemical has the potential to cause a hazard in humans.

[4]The term susceptibility is used in this handbook to describe populations at increased risk, focusing on biological (intrinsic) factors, as well as social and behavioral determinants that can modify the effect of a specific exposure. Lifestage is defined as a distinguishable time frame in an individual's life that is characterized by unique and relatively stable behavioral and/or physiological characteristics that are associated with development and growth. Lifestage, along with other biological factors (e.g., genetic polymorphisms, gender, disease status, nutritional status) can confer differences in toxic response (i.e., sensitivity) to chemical exposure. U.S. EPA (2005a) provides information on age-related factors reflecting behavioral and physiological development among children.

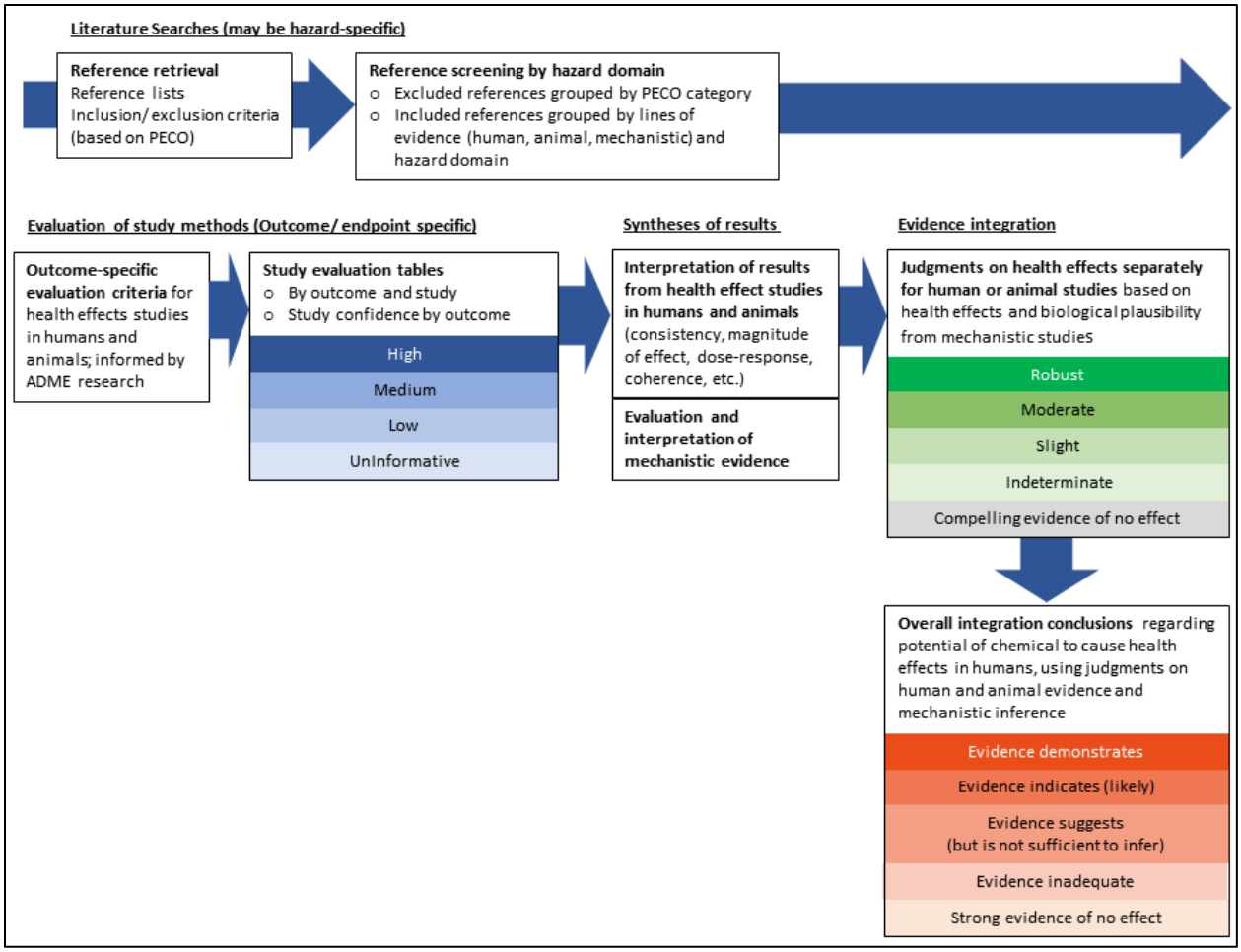


Figure O-3. Overview of process for evaluating evidence in Integrated Risk Information System (IRIS) assessments.

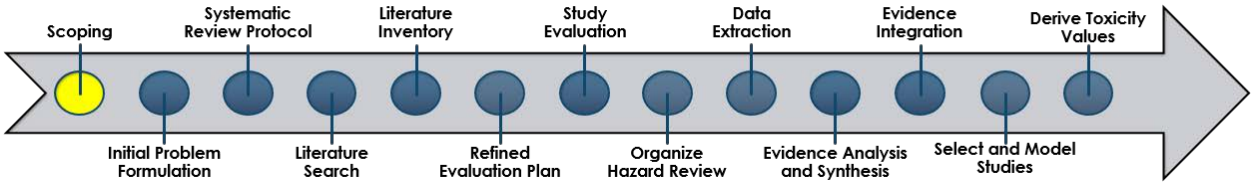
Selecting Studies for Dose-Response Modeling and Deriving Toxicity Values

In general, toxicity values are developed when the totality of the available evidence indicates that chemical exposure has the potential to cause the human health effect being evaluated (i.e., evidence integration judgments of **evidence demonstrates** or **evidence indicates (likely)** for noncancer hazards and descriptors of *carcinogenic to humans* or *likely to be carcinogenic to humans* for cancer hazards). The studies considered most informative for evaluating hazard are considered for deriving toxicity values; this would typically include *high* and *medium* confidence studies, with a generally increased emphasis on *high* confidence studies after considering the specific strengths and limitations of the medium confidence studies and the value of the results they may contribute. The process for systematically selecting studies for dose-response modeling is described in **Chapter 12**.

Chapter 13 describes procedures used in dose-response assessment to develop toxicity values (i.e., reference dose [RfD], reference concentration [RfC], oral slope factor, and inhalation unit risk). This process includes decisions regarding selection of data set(s), selection of

benchmark response (BMR) values, toxicokinetic modeling, modeling in the range of observations and extrapolation to lower exposures and response levels, and consideration of uncertainty and variability.

1. SCOPING OF IRIS ASSESSMENTS



SCOPING

Purpose

- To ensure that the IRIS assessment meets the toxicity assessment needs of EPA program and regional offices.

Who

- Scoping and problem formulation team including Assessment Manager(s) and other assigned staff, including senior liaison with EPA program and regional offices.

What

- Define scope of the IRIS assessment.

Scoping is the first stage in the development of an IRIS assessment. It involves early consultation and continued communication with clients in U.S. Environmental Protection Agency (EPA) program and regional offices to identify the information, timelines, and level of detail required for the assessment to support EPA needs. IRIS assessments, in contrast to those assessments developed by individual EPA programs (e.g., under TSCA section 6), are typically not focused on specific chemical uses or clean-up scenarios (e.g., intended use). The purpose of scoping is to ensure that the IRIS assessment meets the toxicity assessment needs of EPA program and regional offices, the primary users of IRIS information. Ongoing efforts or needs of other groups such as states and tribes are also factored into EPA's needs. This chapter provides a description of the scoping process and general points for IRIS staff to follow as they initiate an assessment.

1.1. OVERVIEW OF THE SCOPING PROCESS

The IRIS Assessment Manager takes an active role in scoping and works with other IRIS and Center for Public Health and Environmental Assessment (CPHEA) staff and contractors assigned to provide support in this process, including CPHEA's program liaison. Scoping involves consultation and close coordination with EPA programs (e.g., Office of Air and Radiation, Office of Water, Office of Land and Emergency Management, Office of Chemical Safety and Pollution Prevention [OCSPP])

and regions, and typically involves one or more meetings to discuss their expectations and the specific components of an IRIS assessment that are most important for addressing their needs (NRC, 2009). EPA programs and regions are also aware of other federal, state, and tribal needs related to chemical assessment, and relay this information to the IRIS Program during scoping. Other EPA offices (e.g., Office of Children's Health Protection, Office of Policy, Office of Environmental Justice) also provide information on biologically susceptible populations including lifestages, communities with potentially disproportionately high exposures, and provide additional perspectives on other useful information.

Regular follow-up communications throughout scoping, as well as the initial problem formulation process (see **Section 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 draft Assessment Plan (see **Section 2**) includes a summary of the Agency needs for the assessment (i.e., EPA program or regional office clients, required exposure route, statutory authority, anticipated uses).

1.1.1. Examples of Factors that Can Determine the Scope of an Assessment

The following are examples of questions that may be addressed during scoping communications with EPA programs and regions:

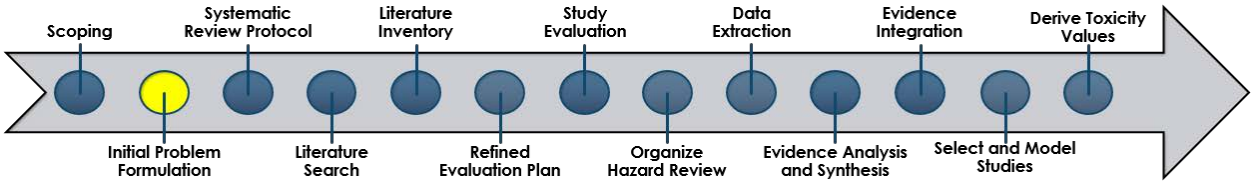
- What are the unique needs (including statutory authority) of the program/regional offices including the time frames for those needs?
- What are the exposure scenarios of primary concern or most immediate need? Is there a need for an assessment of particular routes (e.g., oral, inhalation, dermal) or 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 programs and regions (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 larger 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., a single chemical, vanadium pentoxide, vs. all vanadium compounds)?
- Are there communities, populations, or lifestages that are known or suspected to have disproportionately large exposure or are disproportionately sensitive to the chemical's toxicity?

1.1.2. Identification of Particular Concerns and Priorities of Agency/EPA Clients

For an assessment to address clients' needs, it is important to identify particular concerns and priorities, as illustrated by the representative questions that follow. These considerations help inform the project management timelines, specific aims and populations, exposures, comparators, and outcomes (PECO) for the assessment.

- What time constraints exist for decision making by EPA programs and regions or other stakeholders (e.g., statute or regulatory deadlines, court-ordered consent decrees)?
- What EPA actions are pertinent to this chemical and how might an IRIS assessment be useful (e.g., previous statutory or regulatory decisions, health effects of public concern)?
- Does the decision-making need or potential action pertain to a group of chemicals (such as chemicals used for similar purposes that may be considered alternatives or substitutes for each other)?
- Are there early indications that there may be greater risks to susceptible subpopulations or other issues that might affect dose-response (e.g., chemicals with a potential mutagenic 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?
- Is there a strong risk communication or decision-making need to characterize toxicity at exposures above a reference value?
- Do EPA's needs include occupational risks or other exposures that may be at ranges above typical environmental exposures?
- 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 2.1**.

2. PROBLEM FORMULATION AND DEVELOPMENT OF AN ASSESSMENT PLAN



PROBLEM FORMULATION

Purpose

- To identify potential health effects, types of studies, and science issues to be considered during assessment plan development.

Who

- Scoping and problem formulation team including Assessment Manager(s), Information Specialist (EPA staff or contractor), and other assigned staff.
- EPA programs and other offices.

What

- IRIS Assessment Plan.
- Preliminary literature survey ("systematic evidence map") and descriptions of potential endpoints and science issues to be addressed in the assessment.
- Public science meeting.

As part of problem formulation, which typically overlaps with scoping efforts, the Integrated Risk Information System (IRIS) Program identifies health effects that have been studied in relation to exposure to the chemical, as well as science issues that may need to be considered when evaluating its potential toxicity. Based on a preliminary literature survey (also referred to as a systematic evidence map) and the scope defined by U.S. Environmental Protection Agency (EPA), problem formulation activities are conducted to frame the scientific questions that will be the focus of the assessment. Problem formulation often includes the following activities which are performed primarily by the assessment team and contractors:

- A broad, preliminary literature survey is typically carried out to identify health effects or types of toxicity that have been studied in conjunction with exposure to the chemical or

substance as well as key toxicokinetic[5] and mode-of-action (MOA) issues, susceptible populations and lifestages, and differences in scientific interpretation or controversies that the assessment may need to address. **Chapter 4** describes approaches used to conduct literature surveys. Recent assessments from other state, federal, or international health agencies are also reviewed.

- The identified health effects are organized into relatively broad categories and summarized to describe the coverage of the evidence base. Summaries may be narrative or tabular, depending on the nature of the literature.

- The results of the literature survey are considered in the context of the needs identified by EPA during scoping (see **Chapter 1**) to prepare a draft Assessment Plan. This document provides a summary of the Agency need for the assessment; objectives and specific aims of the assessment; draft populations, exposures, comparators, and outcomes (PECO) criteria that outline the evidence considered most pertinent to the assessment; and identification of key areas of scientific complexity. Brief background information on uses and potential for human exposure is provided for context. The Assessment Plan is discussed in more detail in **Section 2.2**.

- The draft Assessment Plan is presented at a Public Science Meeting to solicit scientific and stakeholder input. The Public Science Meeting may be held in person or virtually. Any revisions to the specific aims and PECO resulting from the public comments will be reflected in the assessment's systematic review protocol (see **Chapter 3**), which also undergoes public comment. In cases where an assessment needs to be conducted under an expedited time frame, the Assessment Plan and protocol may not be released in advance for public comment. In these circumstances the systematic review protocol (which encompasses the assessment plan) would be released concurrent with the draft assessment (as a separate document or as part of the methods section of the draft assessment) or prior to release for informational purposes. **Figure 2-1** summarizes the purposes of the Assessment Plan and protocol.

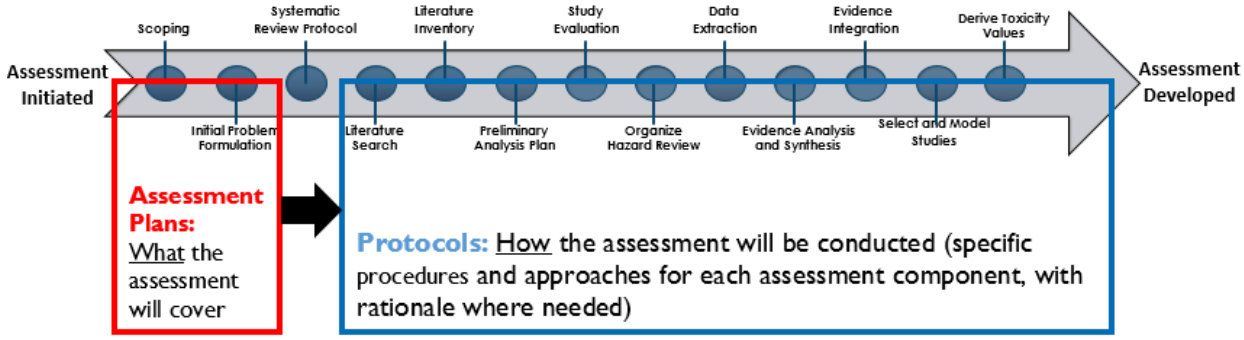


Figure 2-1. Integrated Risk Information System (IRIS) systematic review problem formulation and method documents.

[5]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.

2.1. PRELIMINARY LITERATURE SURVEY

The assessment team, contractors, and other Center for Public Health and Environmental Assessment (CPHEA) staff (as needed) with expertise in toxicology, epidemiology, and information science perform preliminary literature surveys. The availability of other assessments and reviews (especially when recent) may mitigate the need for conducting a preliminary literature survey using the processes described below, although the assessment team will still draw independent conclusions from the literature. Specialized software applications are used to conduct the preliminary literature surveys (see **Chapter 4**).

Preliminary literature surveys identify health assessments conducted by other federal, state, and international health agencies to help plan and focus the systematic review(s) to be conducted as part of developing the IRIS assessment. If previous assessments are unavailable or inadequate, the assessment team may conduct an alternative type of survey (e.g., search for recent review articles). The health assessments provide a source of previously evaluated health effects evidence for consideration in the development of the assessment. The surveys also include a search designed to identify relevant studies published after the cutoff dates used in other agency assessments. This search is used to identify recent studies with new data on previously evaluated health effects or on additional health outcomes that could be evaluated in the assessment, as well as review articles covering new topics or science issues not identified in previous assessments (note that review articles will be used to identify missed studies, health effects, or key areas of scientific complexity, not as primary studies themselves). These newly identified studies may also provide information on susceptible populations and lifestages. Searches should also be designed to identify available physiologically based pharmacokinetic (PBPK) models and reviews addressing mechanistic or MOA hypotheses. To identify studies not considered in previous assessments, a date of 2 years before the health assessment publication can be used if the cutoff date for new studies is not known.

The following examples of federal, state, and international health agencies that often produce assessments may be used as sources of health effect information in addition to any already existing EPA IRIS assessments (http://www.epa.gov/iris/index.html).

2.1.1. Federal

- Agency for Toxic Substances and Disease Registry (ATSDR), http://www.atsdr.cdc.gov/toxprofiles/.
- National Institute for Occupational Safety and Health, http://www.cdc.gov/niosh/.
- National Toxicology Program (NTP), https://ntp.niehs.nih.gov/.
- Occupational Safety and Health Administration, https://www.osha.gov/.

- EPA, Office of Chemical Safety and Pollution Prevention (OCSPP), https://www.epa.gov/aboutepa/about-office-chemical-safety-and-pollution-prevention-ocspp.
- EPA, Office of Water, http://water.epa.gov/.

2.1.2. State[6]

- California EPA, Office of Environmental Health Hazard Assessment, http://oehha.ca.gov/.
- New Jersey Department of Environmental Protection, http://www.nj.gov/dep/.
- Texas Commission on Environmental Quality, http://www.tceq.texas.gov.
- Minnesota Pollution Control Agency, https://www.pca.state.mn.us/

2.1.3. International

- European Chemicals Agency (ECHA), http://echa.europa.eu/.
- European Food Safety Authority, https://www.efsa.europa.eu/.
- German Federal Institute for Risk Assessment (Bundesinstitut für Risikobewertung; BfR), https://www.bfr.bund.de/en/home.html
- Health Canada, http://www.hc-sc.gc.ca/.
- International Agency for Research on Cancer (IARC), http://monographs.iarc.fr/.
- Joint Food and Agriculture Organization (FAO) of the United Nations/World Health Organization (WHO) Expert Committee on Food Additives, http://www.who.int/foodsafety/areas_work/chemical-risks/jecfa/en.
- Joint FAO/WHO Meeting on Pesticide Residues, http://www.fao.org/agriculture/crops/thematic-sitemap/theme/pests/jmpr/en/.
- National Industrial Chemicals Notification and Assessment Scheme, https://www.industrialchemicals.gov.au/transition-from-nicnas-to-aicis.
- Netherlands National Institute for Public Health and the Environment, http://www.rivm.nl/en.
- Public Health England, https://www.gov.uk/government/collections/chemical-hazards-compendium.
- WHO International Programme on Chemical Safety (IPCS) http://www.who.int/ipcs/en/.

[6]This is not a comprehensive list; for a list of health and environmental agencies of U.S. states and territories, please visit https://www.epa.gov/home/health-and-environmental-agencies-us-states-and-territories. Some states or territories may have developed chemical-specific exposure or toxicity materials that may be useful in assessment development.

2.2. ASSESSMENT PLAN

The assessment team compiles information obtained from federal, state, and international health agency assessments and additional literature searches to summarize health effects that have been studied. This information is used to help inform the level of effort and type of scientific expertise required to conduct the assessment. The assessment team or contractors will perform and summarize the results of the preliminary literature survey. These screening surveys can be considered as "systematic evidence maps" used to summarize literature characteristics to help identify available types of evidence and gaps (Miake-Lye et al., 2016). Summaries may be narrative or tabular, depending on the specific chemical. Additional supplemental information, including study designs, populations or test systems studied, exposure duration or design used, route of exposure, absorption, distribution, metabolism, and excretion (ADME) information, PBPK models, proposed MOAs, and susceptible groups will also be considered. The assessment team uses this information to develop the research question, specific aims, and draft PECO criteria for the systematic review. Based on the needs identified during scoping, the Assessment Plan 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). Finally, the Assessment Plan should indicate key science areas that will need to be considered. Examples of science issues may include topics such as human relevance of findings in animals, whether an endpoint is considered adverse or adaptive, conflicting studies or information, issues relating to toxicokinetics used to identify susceptible groups, exposure to a vulnerable group (e.g., populations that subsist primarily on wild caught fish), the existence and validity of PBPK models, or hypothesized MOAs that lack scientific consensus. Stakeholder input received during at least a 30-day comment period (this comment period may vary based on the scientific complexity of the issues associated with the chemical) on the draft Assessment Plan is considered as part of preparing the assessment's systematic review protocol. Any revisions to the specific aims and PECO are reflected in the assessment's systematic review protocol. The Assessment Plan contents become the initial portion of the protocol (see **Chapter 3**). Examples of public Assessment Plans 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 PECO, along with the supplemental tagging structure, is used to identify the evidence that addresses the specific aims of the assessment as well as to focus the search terms and inclusion/exclusion criteria in a systematic review. 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, and exposures (see **Table 2-1**).

In addition to the PECO criteria, studies containing supplemental material that are also potentially relevant to the specific aims are tracked during the literature screening process. **Table 2-2** presents major categories of supplemental material. The criteria are used to tag studies

during screening and to prioritize studies for consideration in the assessment based on the likelihood they will impact evidence synthesis conclusions for human health.

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

PECO element	Evidence
Populations	**Human:** Any population and lifestage (occupational or general population, including children and other sensitive populations). **Animal:** Nonhuman mammalian animal species (whole organism) of any lifestage (including preconception, in utero, lactation, peripubertal, and adult stages). *[Other species specific to the assessment (e.g., zebrafish and C. elegans when neurotoxicity is expected to be a primary health effect of concern).]*
Exposures	*[Example language that can be included if appropriate.]* Relevant forms: [chemical X] *(CAS number).* *Other forms of [chemical X] that readily dissociate (e.g., list any salts).* *Metabolites of interest, including.* *Measures of metabolites used to estimate exposures to [chemical X].* *Studies of the effects of exposure to the metabolites themselves.* *Indicate whether mixture studies are included.* *Others determined by the assessment team.* **Human**: Any exposure to [chemical X] (via [oral or inhalation] route[s] if applicable). *Specify if certain exposure assessment methods or metrics will NOT be included.* **Animal**: Any exposure to [chemical X] via [oral or inhalation] route[s]. *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.* Studies involving exposures to mixtures will be included only if they include exposure to [chemical X] alone. Other exposure routes, including [dermal or injection], will be tracked during title and abstract as "potentially relevant supplemental information."
Comparators	**Human**: A comparison or referent population exposed to lower levels (or no exposure/exposure below detection limits) of [chemical X], or exposure to [chemical X] for shorter periods of time, or cases vs. controls. However, worker surveillance studies are considered to meet PECO criteria even if no referent group is presented. Case reports describing findings in 1–3 people will be tracked as "potentially relevant supplemental information." **Animal**: A concurrent control group exposed to vehicle-only treatment or untreated control.
Outcomes	All health outcomes (both cancer and noncancer). *[State here if decisions have been made to limit to endpoints related to clinical diagnostic criteria, disease outcomes, histopathological examination, or other apical/phenotypic outcomes.]* May include the following statement, "EPA anticipates that a systematic review for health effect categories other than those identified (i.e., health effect 1, health effect 2…) will not be undertaken unless a significant amount of new evidence is found upon review of references during the comprehensive literature search."

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

Category	Evidence
Mechanistic	Studies reporting measurements related to a health outcome that inform the biological or chemical events associated with phenotypic effects, in both mammalian and nonmammalian model systems, including in vitro, in vivo (by various routes of exposure), ex vivo, and in silico studies. Studies where the chemical is used as a laboratory reagent generally do not need to be tagged (e.g., as a chemical probe used to measure antibody response). The identification and organization of these data (including, potentially, additional focused searches) is elaborated on in **Sections 4.1.3**, **4.3.3**, **6.6**, **10.1**, **and 10.5**.
Nonmammalian model systems	Studies in nonmammalian model systems (e.g., fish, birds, invertebrate species). *[unless included in Populations above.]*
Toxicokinetic (ADME)	Toxicokinetic (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. • These data are used to estimate the amount absorbed (A), distributed (D), metabolized (M), and/or excreted (E). • The most informative studies involve measurements over time such that the initial increase and subsequent concentration decline is observed, preferably at multiple exposure levels. • Data collected from multiple tissues or excreta at a single time-point also inform distribution. • ADME data can also be collected from human subjects who have had environmental or workplace exposures that are not quantified or fully defined. However, to be useful such data must involve either repeated measurements over a time-period when exposure is known (e.g., is zero because previous exposure ended) *or* time- and subject-matched tissue or excreta concentrations (e.g., plasma and urine, or maternal and cord blood). • 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. *Studies describing environmental fate and transport or metabolism in bacteria or model systems not applicable to humans or animals should not be tagged.

Category	Evidence
Classical pharmacokinetic (PK) or physiologically based pharmacokinetic (PBPK) model studies	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, that are not a PBPK model. • The data are typically the concentration time-course in blood or plasma after oral and or intravenous exposure, but other exposure routes can be described. • A classical PK model might be elaborated from the basic structure applied in standard PK software, for example to include dermal or inhalation exposure, or growth of body mass over time, but otherwise does not use specific tissue volumes or blood flow rates as model parameters. • Such models can be used for extrapolation like PBPK models, although such use might be more limited. Note: ADME studies often report classical PK parameters, such as bioavailability (fraction of an oral dose absorbed), volume of distribution, clearance rate, and/or half-life or half-lives. If a paper provides such results only in tables with minimal description of the underlying model or software (i.e., uses standard PK software without elaboration), including "non-compartmental analysis," it should only be listed as a supplemental material ADME study. 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 elimination, and thereby estimate concentrations in blood or target tissues. . • Usually specific to humans or defined animal species; often a single model structure is calibrated for multiple species. • Some mechanistic dosimetry models might not be compartmental PBPK models but predict dose to the body or specific regions or tissues based on mechanistic data, such as ventilation rate and airway geometry. • 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); that is, data that are independent of in vivo ADME data that are otherwise used to estimate model parameters. • Chemical-specific information on metabolism (e.g., Vmax, Km) or other molecular processes (e.g., protein binding) might be obtained by fitting the model to in vivo ADME data or determined from in vitro experiments and extrapolated to in vivo predictions. • They allow extrapolation between species, routes of exposure, or exposure durations and levels; that is, they do not just quantify ADME for specific experiments to which they have been fitted.

Category	Evidence
Exposure characteristics (no health outcome)	Exposure characteristic studies include data that are unrelated to toxicological endpoints, but which provide information on exposure sources or measurement properties of the environmental agent (e.g., demonstrate a biomarker of exposure).
Mixture studies	Mixture studies that are not considered to meet the PECO criteria because they do not contain an exposure or treatment group assessing only the chemical of interest. This categorization generally does not apply to epidemiological studies where the exposure source might be unclear.
Routes of exposure not pertinent to PECO	Studies using routes of exposure that fall outside the PECO scope.
Case studies or case series	Case reports describing health outcomes after exposure will be tracked as potentially relevant supplemental information when the number of subjects is ≤3.
Acute duration exposures	For assessments that focus on chronic exposure, acute exposure durations (defined as animal studies of less than 1 d) are generally considered supplemental.
Records with no original data	Records that do not contain original data, such as other agency assessments, informative scientific literature reviews, editorials, or commentaries.
Abstract only (includes conference abstracts)	Records that do not contain sufficient documentation to support study evaluation and data extraction.
Others determined by assessment team	

It is important to emphasize that being tagged as supplemental material does not mean the study is excluded from consideration in the assessment. The initial screening level distinctions between a study meeting the PECO criteria and a supplemental study are often made for practical reasons and the tagging structure in **Table 2-2** is designed to ensure the supplemental studies are categorized for easy retrieval while conducting the assessment. Studies that meet the PECO criteria are those that are most likely to be used to derive toxicity values and will thus undergo individual level study evaluation (see **Chapter 6**) and data extraction (see **Chapter 8**). For evidence rich topics this is most likely to be evidence from animal bioassays and epidemiological studies. The impact on the assessment conclusions of individual studies tagged as supporting material is often difficult to assess during the screening phase of the assessment. Studies tagged as supplemental may (1) provide PBPK models supporting dose-response modeling; (2) become integral to the interpretation of other evidence at the level of needing individual level study evaluation (e.g., genotoxicity studies when conducting a cancer MOA analysis); (3) may be a single study that contributes to a well-accepted scientific conclusion and does not need to be evaluated and summarized at the individual study level (e.g., dioxin as an aryl hydrocarbon receptor [AhR] agonist); (4) provide key references for preparation of certain sections in an IRIS assessment (e.g., background information on sources, production, or use; overview of toxicokinetics); or (5) provide context for the rationale for conducting the assessment or assessment conclusions (e.g., information on pathways and levels of exposure). From a practical perspective, screening all

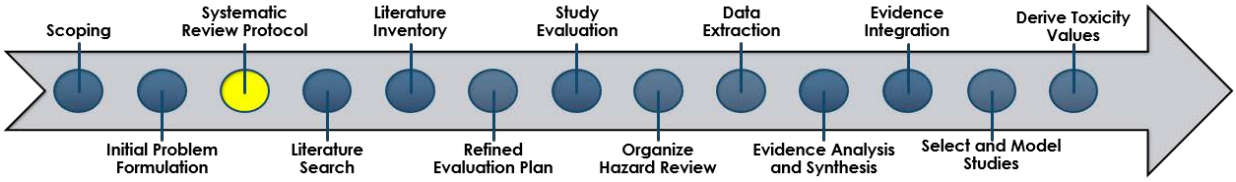
of these studies as meeting the PECO criteria at the title and abstract level means that the full-text needs to be obtained for full-text screening, which can be a very time and resource intensive process. Thus, the tagging strategy outlined below allows these studies to be identified at the title and abstract level so that the full-text can be retrieved as needed while conducting the assessment.

When chemicals assessed by the IRIS Program are known to have an abundance of animal and epidemiological evidence available, many of the available mechanistic studies are initially tagged as supplemental and their impact on the assessment is evaluated as described in **Section 4.3** (literature inventories), **Chapter 5** (refined evaluation plan), **Section 6.6** (study evaluation, when individual study evaluation is warranted), **Chapter 7** (organizing the hazard review), **Chapter 10** (analysis and synthesis of mechanistic information), and **Chapter 11** (evidence integration). A different PECO would be constructed for chemicals known to be data poor from scoping and problem formulation to include in silico, in vitro/ex vivo studies, and alternative animal model evidence in the PECO criteria.

ORGANIZATION OF THE ASSESSMENT PLAN

1. Introduction
2. Scoping and initial problem formulation
   - 2.1 Background (brief, provided for context)
     - Chemical and physical properties
     - Sources, production, and use
     - Environmental fate and transport for context
     - Environmental concentrations for context
     - Potential for human exposure for context
     - Populations and lifestages with potentially greater exposures and/or greater sensitivity to health outcomes
     - Summary of toxicokinetics
   - 2.2 Scoping summary—summarize Agency needs and anticipated uses in tabular format
   - 2.3 Problem formulation
     - Summarize health evaluation conclusions in recent assessments, especially those from other federal or international health agencies
     - Present any screening level literature survey results assembled to help identify priority health outcomes and lines of evidence
   - 2.4 Key science issues
3. Overall objective, specific aims, and draft populations, exposures, comparators, and outcomes (PECO)
   - 3.1 Assessment approach (if needed)
     - Describe any modular or interim product approach being taken for the assessment
   - 3.2 Specific aims
   - 3.3 Draft PECO
4. References

3. PROTOCOL DEVELOPMENT FOR IRIS SYSTEMATIC REVIEWS



DEVELOPMENT OF THE SYSTEMATIC REVIEW PROTOCOL

Purpose

- Establish a priori methods that will be used for assessment development.

Who

- Assessment team.

What

- An assessment-specific protocol that:
  - Describes the strategy for implementation of systematic review and serve as instructions and training material for individuals responsible for implementation.
  - Is revised as needed to provide documentation of decisions and changes made to assessment methods during draft development.

The protocol is a central component of a systematic review. It is intended to improve transparency and reduce bias in the conduct of the review by describing the review question and methods in advance (CRD, 2013; Higgins and Green, 2011a; IOM, 2011). The IRIS systematic review process involves the development and use of a protocol that presents the detailed methods for assessment development. Any adjustments made to the specific aims and populations, exposures, comparators, and outcomes (PECO) criteria in response to public input on the Assessment Plan are reflected in the protocol. The protocol, including the initial version released for public comment, should be as detailed as possible, with assessment-specific decisions and procedures (based on the IAP and subsequent work) described. However, it is expected that there will be stepwise refinements and greater specification to the protocol occurring before implementation of individual stages (e.g., development of outcome- and exposure-specific criteria, pilot testing of study evaluation workflows will likely lead to refinements; groupings of health outcomes will be informed by specific endpoints assessed in included studies) based on understanding of scientific/technical issues that arise during assessment development.

A template protocol containing the elements to be included in each chemical-specific protocol is available on the IRIS resource page in HAWC. The organization of the document is available in the text box below; further description of the protocol in this document would be redundant. The template includes the elements described in peer-reviewed systematic review protocol checklists (Haddaway et al., 2018; Moher et al., 2009).

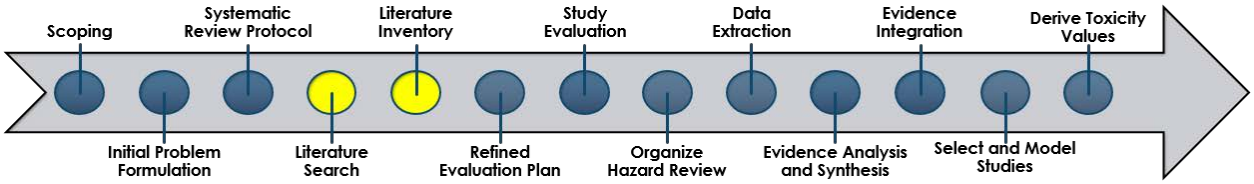
The protocol is typically released for a 30-day public comment process, ideally within 6 months of receiving input on the assessment plan. Public input is considered during preparation of the draft assessment. Refinements made while conducting the assessment (e.g., additional specificity on quantitative methods used for combining data sets and conducting dose-response) are acknowledged as updates to the protocol version released with the draft assessment. In rare cases where an assessment needs to be conducted under an expedited time frame, the protocol may be released concurrent with the draft assessment (as a separate document or as part of the methods section of the draft assessment) or just prior to release of the assessment for informational purposes (i.e., with no separate public comment period). The IRIS Program posts assessment protocols and protocol updates publicly on the IRIS website for a chemical (at www.epa.gov/iris).

ORGANIZATION OF PROTOCOL

1. Introduction*
2. Scoping and initial problem formulation*
3. Overall objectives, specific aims, and populations, exposures, comparators, and outcomes (PECO) criteria*
4. Literature search and screening strategies
   - 4.1. Use of existing assessments
   - 4.2. Literature search strategies
   - 4.3. Non-peer reviewed data
   - 4.4. Screening process
   - 4.5. Summary-level literature inventories
5. Refined evaluation plan
6. Study evaluation (reporting, risk of bias, and sensitivity) strategy
   - 6.1. Study evaluation overview for health effect studies
   - 6.2. Epidemiology study evaluation
   - 6.3. Experimental animal study evaluation
   - 6.4. Controlled human exposure study evaluation
   - 6.5. Physiologically based pharmacokinetic (PBPK) model descriptive summary and evaluation
   - 6.6. Mechanistic study evaluation
7. Organizing the hazard review
8. Data extraction of study methods and results
   - 8.1. Standardizing reporting of effect sizes
   - 8.2. Standardizing administered dose levels/concentrations
9. Synthesis of evidence
   - 9.1. Syntheses of human and animal health effects evidence
   - 9.2. Mechanistic information
10. Evidence Integration
    - 10.1. Evaluating the strength of the human and animal evidence streams
    - 10.2. Overall evidence integration judgments
    - 10.3. Hazard considerations for dose-response
11. Dose-response assessment: selecting studies and quantitative analysis
    - 11.1. Selecting studies for dose-response assessment
    - 11.2. Conducting dose-response assessments
12. Protocol History
13. References
14. Appendices (e.g., outcome-specific study evaluation considerations)

*From Assessment Plan, revised based on public input.

4. LITERATURE SEARCH, SCREENING, AND INVENTORY



LITERATURE SEARCH, SCREENING, AND INVENTORY

Purpose

- Identify the relevant studies for use in the assessment, document the search and screening process, and categorize studies.

Who

- Assessment team members, HERO Information Specialist, and
- Disciplinary workgroups (as needed); contractor support may also be used.

What

- Literature identification strategy (search and screening procedures).
- Literature inventory to categorize pertinent studies into broad subject areas.
- Implementation and documentation of any supplemental, topic-specific searches that may occur after initial search(es).

This chapter describes the elements and tasks involved in developing a literature search strategy, screening identified references, and creating an inventory of studies. The search, screening, and inventory strategies noted here can be employed in the initial search or in later targeted searches (e.g., searches for relevant mechanistic studies for targeted questions; see **Chapter 10**). The Health and Environmental Research Online (HERO; see **Section 4.1.1**) database is typically used to conduct and document literature searches. Increasingly, a variety of specialized software applications are used to facilitate the process of identifying and screening studies (see **Figure 4-1** for commonly used software applications for screening literature and developing inventories within IRIS and **Section 4.2.2** for additional information). The availability of specialized software applications for conducting literature assessments is growing rapidly and it is likely that the suite used within IRIS will evolve and expand over time. The Systematic Review (SR) Toolbox (http://systematicreviewtools.com/) is a comprehensive database of tools and has advanced search features to help find tools tailored to specific aspect(s) of systematic review.

Before using a tool from the SR Toolbox, the assessment team should be prepared to confirm its performance, capabilities, and support. Preferred software applications used within IRIS are publicly available, free (when possible), interoperable with other software applications used behind U.S. Environmental Protection Agency (EPA) firewalls and have access to technical support and documentation provided by the developer.

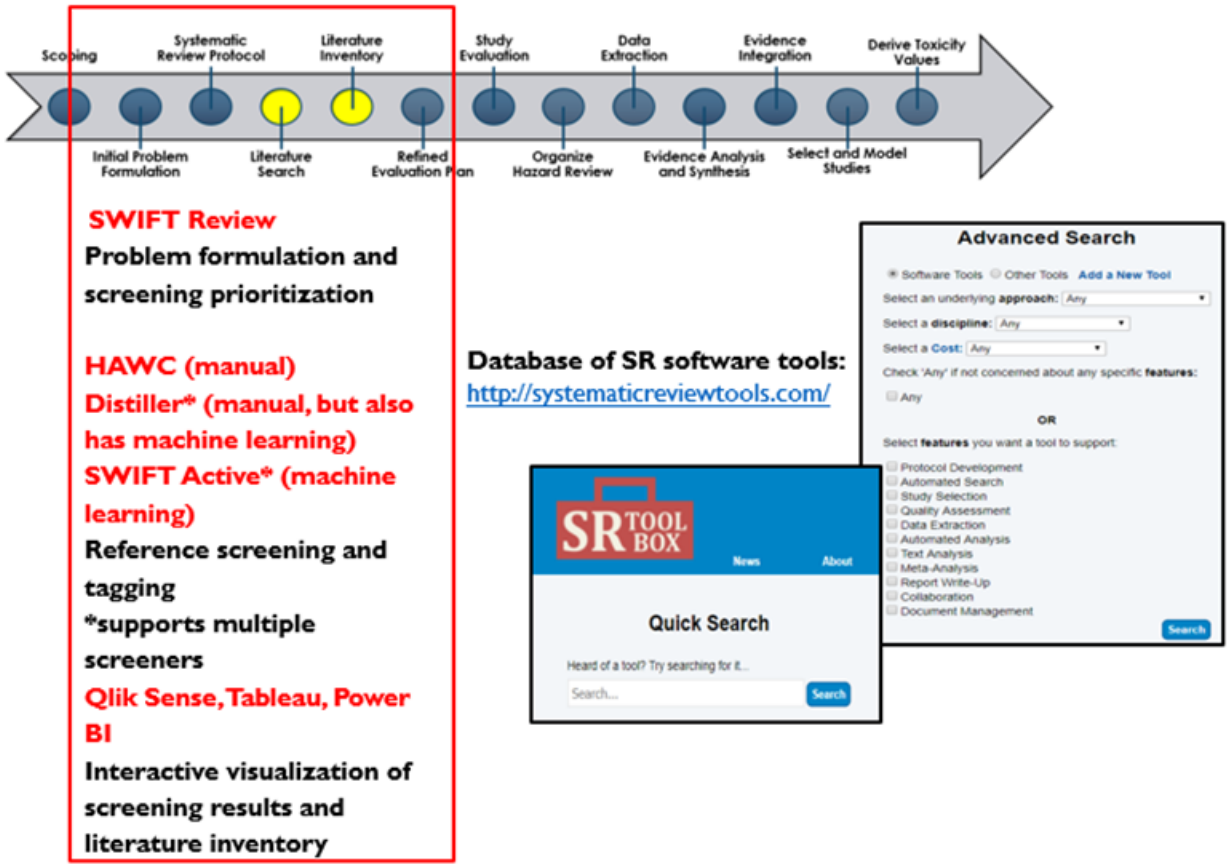


Figure 4-1. Commonly used software applications in the Integrated Risk Information System (IRIS) literature screening and inventory process.

4.1. LITERATURE SEARCH

The following sections discuss key components in a literature search process: using Health and Environmental Research Online (HERO; see **Section 4.1.1**); selecting databases (see **Section 4.1.2**); developing the literature search (see **Section 4.1.3**); documenting (see **Section 4.1.4**); and updating the literature search (see **Section 4.1.5**).

4.1.1. Health and Environmental Research Online (HERO)

HERO is a database of scientific studies and other references used to develop EPA's risk assessments aimed at understanding the health and environmental effects of pollutants and chemicals. It is developed and managed by staff in EPA's Office of Research and Development by

the Center for Public Health and Environmental Assessment (CPHEA). HERO staff include information scientists who specialize in developing and conducting literature searches as well as software programming experts who continually work to expand HERO's capabilities and interoperability with other software applications used in literature assessments, such as software tools used for screening studies (see **Section 4.2**) and data extraction/display (see **Section 8**). It is highly recommended to work closely with the HERO information specialists throughout the literature search process. Some useful tips and links for using HERO are described in **Figure 4-2**.

4.1.2. Selecting Databases

The assessment team is responsible for initiating the literature search request and working with information specialist(s) 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 conducts the search (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. This will promote interoperability between HERO and other software platforms used to help screen studies, 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.

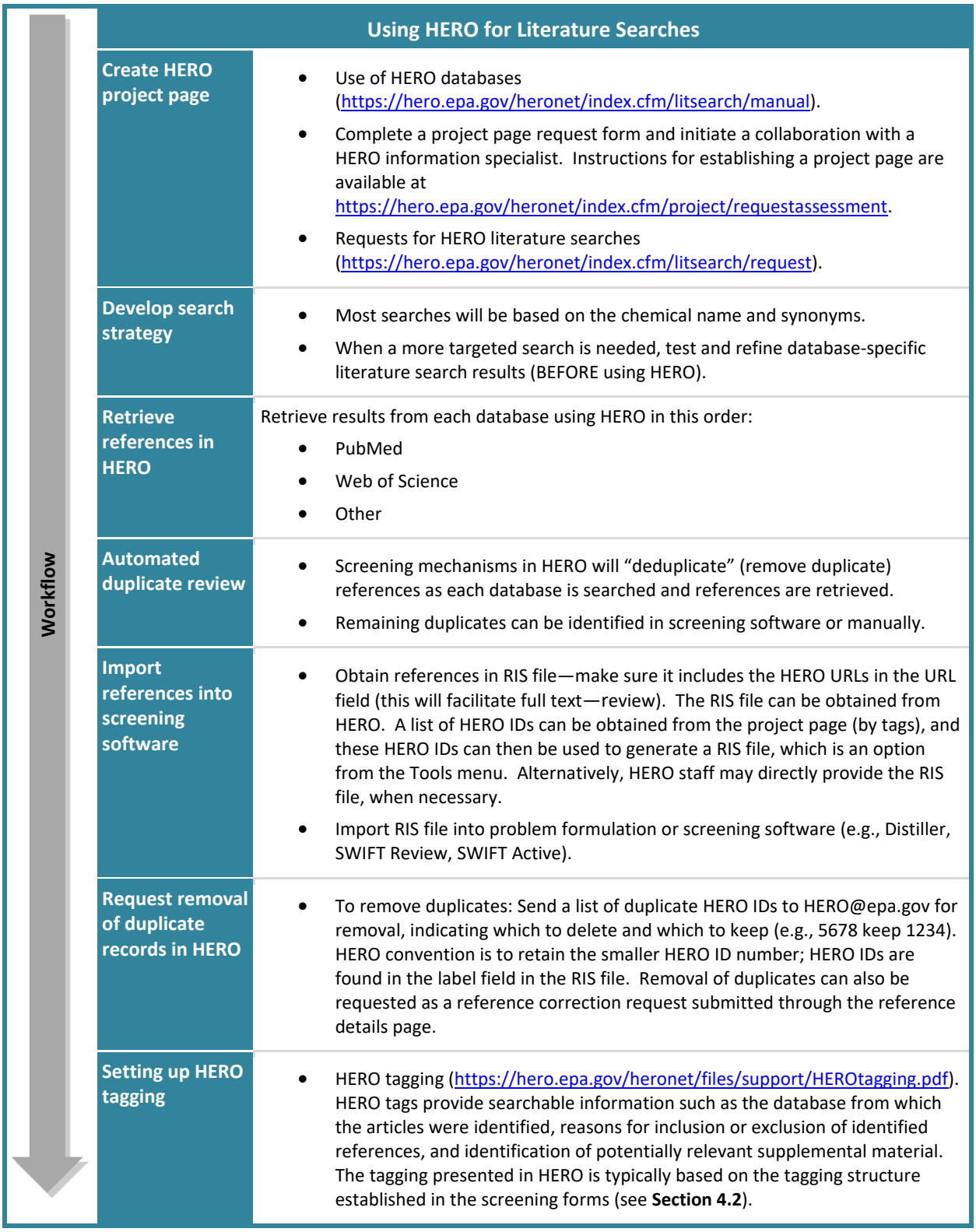


Figure 4-2. Workflow for Health and Environmental Research Online (HERO)—facilitated literature searches.

Core and Supplemental Databases

PubMed and **Web of Science** are the core sources that IRIS uses for published studies.[7] These three overlapping bibliographic databases cover a range of scientific disciplines including medical and life science, social science, and toxicology literature (see **Table 4-1**). Each is accessible through EPA's HERO database. **Figure 4-3** provides a detailed summary of recommendations for searching these databases. The EPA CompTox Dashboard (https://comptox.epa.gov/dashboard) is typically consulted to identify chemical and environmental fate properties (e.g., structure, solubility in water, bioaccumulation factors) and can be used during problem formulation and scoping to survey toxicity values and exposure limits developed by others under the "Hazard" tab. If needed for an assessment, ToxCast data can also be accessed from the CompTox Dashboard as part of identifying and analyzing mechanistic information. As new databases are evaluated for use in assessment development, they will be considered on a chemical-specific basis.

Additional databases can also be used to search for primary literature and summaries of primary literature that may not be available elsewhere. These include national research programs conducting standard 2-year animal bioassays (e.g., National Toxicology Program [NTP]). Studies from Japan and Europe can also be sought through several different databases (see **Table 4-1**). Another source of information are studies submitted to EPA under the Toxic Substances Control Act (TSCA), and as amended by the Frank R. Lautenberg Chemical Safety for the 21st Century Act. Under TSCA, companies that manufacture, process, or commercially distribute a chemical may 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. There is no requirement that these studies also be submitted for publication, so this database may be the only source of the data contained in these studies. Some information and studies can be found through the National Technical Reports Library (https://ntrl.ntis.gov/NTRL/). Search the EPA ChemView database at: (https://chemview.epa.gov/chemview). ChemView contains TSCA-related information, including data submitted to EPA [such as TSCA §8(e) notifications of substantial risk, unpublished health and safety studies, or test rule data], EPA assessments, EPA actions, and manufacturing, processing, use and release data submitted to EPA. See **Table 4-1** for details on relevant hazard information available through ChemView. EPA also maintains internal databases that contain submissions claimed to be confidential business information under relevant sections of TSCA, such as §5 and §8(e) if sufficient information on the studies is made public. Other databases may be useful for specific chemicals and may be included depending on attributes of the chemical under review (see **Table 4-1**).

[7]Other bibliographic databases were considered for inclusion in the core list but were not included for a variety of reasons, including budgetary constraints on obtaining subscription (SCOPUS, EMBASE). For ScienceDirect, prior experience has demonstrated searching of PubMed and Web of Scienceprovides sufficient coverage. TOXLINE was phased out in December 2019 and integrated into other NLM resources. Google Scholar is not a curated database, but an indexing service. In addition, there is no Application Programming Interface for Google Scholar, so direct download of search results is not feasible.

Table 4-1. Databases for primary literature

Database	Description and notes
Core databases	
PubMed (searched by HERO or contractors and added to HERO)	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 Medical Subject Headings (MeSH) terms. Can access through HERO. Test page for developing searches: http://www.ncbi.nlm.nih.gov/pubmed/advanced.
Web of Science (searched by HERO or contractors and added to HERO)	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.
CompTox Dashboard (searched by assessment team)	The CompTox Dashboard (https://comptox.epa.gov/dashboard) is designed to provide chemistry, toxicity, and exposure information for over 760,000 chemicals. Data and models within the dashboard also help with efforts to identify chemicals in most need of further testing and for reducing the use of animals in chemical testing. The dashboard can be searched by chemical identifiers (e.g., Name and CASRN), consumer product categories to view chemicals found in certain product types, and assays/genes associated with high-throughput screening data. Using high-throughput screening, living cells or proteins are exposed to chemicals and examined for subsequent changes that suggest potential biological responses. These data are compiled from sources including the EPA's computational toxicology research databases, and public domain databases such as the National Center for Biotechnology Information's PubChem database and EPA's ECOTox Knowledgebase.
ChemView (searched by assessment team)	This database may contain primary hazard studies and summaries such as the following: • 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). • EPA assessments such as hazard characterizations and risk-based prioritizations of high production volume (HPV) chemicals, alternative assessments and a list of safer chemical ingredients. 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. 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.

Database	Description and notes
NTP (searched by assessment team)	Database of 2-yr rodent bioassays and other toxicology studies (https://ntp.niehs.nih.gov/results/pubs/index.html).
Supplemental databases that may be searched by the assessment team depending on the topic	
AEGLs	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. 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.
Agricola	Use for U.S. Department of Agriculture-related compounds. Available through HERO. Test page for developing searches: http://agricola.nal.usda.gov/.
ChemIDPlus	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): https://chem.nlm.nih.gov/chemidplus/. 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/, listed as OECD HPV).
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/.
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 toxicology 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	Contains GLP guideline toxicology 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/ (may 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 Records; DTIC = Defense Technical Information Center; ECHA = European Chemicals Agency; FIFRA = Federal Insecticide, Fungicide, and Rodenticide Act; GLP = Good Laboratory Practice; IHAD = Integrated Hazard Assessment Database; OECD = Organisation for Economic Co-operation and Development; OPP = Office of Pesticides Program; PRISM = Pesticide Registration Information System; R&E = Research and Engineering; REACH = Registration, Evaluation, Authorisation and Restriction of Chemicals; SIAPs = SIDS Initial Assessment Profiles; SIARs = SIDS Initial Assessment Reports.

[a]Slight update to User's Guide: To 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.

[b]Contractors 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.

Search Recommendations for PubMed and Web of Science		
	PubMed (http://www.ncbi.nlm.nih.gov/pubmed)	**Web of Science (http://apps.webofknowledge.com)**
What fields are searched by default?	All fields.[a]	Topic, which includes title, abstract, and keyword fields.
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.
Can I limit by language?	Yes—for IRIS searches, it is helpful to import foreign-language results separately into HERO.	
Can I limit by publication type?	Yes—for IRIS searches, it is helpful to import reviews separately into HERO.	
Can I search by CASRN?	Use quotation marks around CASRNs; CASRNs not widely found in Web of Science records.	
Can I truncate terms?	Use with caution; truncated terms may explode to hundreds of terms and will not search in MeSH field. Truncated terms are treated as wildcards and will return up to 600 variations of the truncated term.	Yes.
Should I include synonyms in my search strategy?	Yes—include synonyms and alternate spellings; use ChemIDPlus (https://chem.nlm.nih.gov/chemidplus/chemidlite.jsp) to identify potential synonyms; use quotation marks around phrases that are not MeSH terms (http://www.ncbi.nlm.nih.gov/mesh).	
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]	
Can I search for cited references or related references?	Searches this as "links to similar articles." HERO does not use this feature as part of the literature search.	Search for cited references or related references; export available only for results that are found in Web of Science.
Other tips	Reviewing the search details window is highly recommended. Recently published articles may be in PubMed, but not indexed for Medline for several weeks or months.[c]	Use research areas to limit search results; recommend choosing research areas to include instead of excluding areas.

CASRN = Chemical Abstracts Service registry number; MeSH = Medical Subject Headings; RePORT = Research Portfolio Online Reporting Tool; TSCATS = Toxic Substances Control Act Test Submissions.

[a]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.

[b]The records previously available at TOXLINE, which was phased out in 2019, include citations to TSCATS records through approximately 2002; these records include health and safety studies, substantial risk notices, and voluntary information submitted to EPA under 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.

[c]To 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.

Figure 4-3. Summary of search strategies for commonly used databases.

4.1.3. Developing the Literature Search

All search strategies balance competing needs for sensitivity (the ability to identify all potentially relevant studies) and specificity (the ability to avoid identification of nonrelevant studies), using a process that is both manageable and reproducible. The efficiency of this process depends on optimizing the approaches used in the initial searching and screening stages. General recommendations for this include:

- When an existing assessment(s) is available from IRIS or another source (e.g., EPA, Agency for Toxic Substances and Disease Registry [ATSDR], NTP, or other federal, state, or international health agency), use it to serve as a starting point for the literature search. Although the possibility exists that the literature searches conducted for existing assessments may have missed important studies, the IRIS process provides overlapping mechanisms to ensure key literature is identified, including multiple opportunities for public input and use of additional search strategies such as citation mapping (see below). Indeed, journal reviews may also be used with some caution because the journal peer-review process does not provide the same opportunities for public engagement as most assessments conducted by governmental sources. This raises concern that important studies may have been missed, especially if the journal review article was not conducted using systematic review approaches. The search strategy may focus on updating the existing literature search and considering whether any refinements or supplemental searches are needed to address assessment needs. If the date of the last literature search is not known for the prior assessment, then the IRIS search should start at least 2 years before the release date. When the date is specified, then the IRIS search should start at the beginning of the calendar year (January 1) in the relevant year. Adjustments to database indexing (e.g., addition of PubMed Medical Subject Headings [MeSH] headings) that can make it difficult to exactly replicate the search results of the prior assessment, especially for studies newly added around the last literature search date.

- Studies cited in prior assessments can be used as "seed" studies when machine-learning software is used to help the screening process.

- For small databases, searches using just the chemical name(s) and Chemical Abstracts Service registry number (CASRN) may return a reasonable number of studies (e.g., <3,000) that is manageable for manual screening, and it is usually unnecessary to try and refine the search strings to identify fewer irrelevant studies. For assessments with very large databases, it may prove useful to rely on more advanced screening techniques (e.g., use of specialized software applications with machine-learning capabilities) to identify relevant studies. Alternatively, it may make sense to design more targeted search strings (e.g., to identify specific endpoints of interest) and incorporate supplemental searches for other informative materials such as mechanistic information (e.g., to identify studies relevant to a perturbed biological pathway that are not specific to the chemical).

- When targeted search strategies are used (e.g., to identify specific endpoints of interest), it is advantageous to use standard search strings across assessments when available. Most standard search strings used within the IRIS Program are available within SWIFT Review software (https://hawcprd.epa.gov/media/attachment/SWIFT-Review_Search_Strategies.pdf) and can be used to automatically "tag" studies by evidence type (human, animal, in vitro) and health outcome (Howard et al., 2016). The preset search strategies implemented with SWIFT Review were developed by information specialists at NTP/National Institute of Environmental Health Sciences (NIEHS) and ICF International but can also be customized by the user or HERO staff as needed within the software. When standard search strings are not available, literature search strategies are typically developed using key words related to the populations, exposures, comparators, and outcomes (PECO) criteria. Development of the search strategy can include identifying relevant search terms through (1) reviewing PubMed's MeSH, (2) extracting key word terminology from relevant reviews and a set of previously identified primary data studies

that are known to be relevant to the topic ("seed" studies), and (3) reviewing search strategies presented in other reviews.

- For some assessments, it may be useful to expand the chemical-specific search terms. 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' The EPA CompTox Chemicals Dashboard can be used to identify additional synonyms (https://comptox.epa.gov/dashboard, see the "synonyms" tab). SWIFT Review also has literature search strategies for identifying and tagging over 8,000 chemicals included in the Toxicology in the 21st Century (Tox21) chemical inventory (Howard et al., 2016). In brief, the searches were automatically constructed by using (1) the common name for the chemical as presented in the source reports listed above, (2) the CASRN, and (3) and retrieval of synonyms from the ChemIDPlus database which currently contains chemical names and synonyms for over 400,000 chemicals. Filters were applied to remove ambiguous terms, including short alphanumeric sequences that could be confused with arbitrary acronyms or abbreviations (e.g., "2VP" for "2-vinylpyridine"); English words that have been used as industrial trade names, street drug slang, etc.; or chemical formulas that do not unambiguously define a chemical. Because these chemical name searches were automatically generated, the search strategy should be manually reviewed prior to use in an IRIS assessment.

- If studies based in occupational settings are anticipated, expertise in industrial hygiene or occupational epidemiology should be sought to create a listing of industries, job categories, and titles that should be included in the search.

Note that 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]) require a high level of training and experience.

Primary Studies

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, which provide sufficient detail to allow evaluation of the study methods.

Grey Literature

The phrase grey literature refers to the broad category of studies, including primary studies, not found in standard, peer-reviewed literature databases (e.g., PubMed). This may include 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. Grey literature is typically identified during problem formulation, engagement with technical experts, and during solicitation of Agency, interagency, and public comment on assessment drafts during the defined steps of the IRIS process (as described at:

https://www.epa.gov/iris/basic-information-about-integrated-risk-information-system#process). Although grey literature may be more difficult to procure, it does not necessarily mean that these sources are of inferior quality. Note that while information from the grey literature can be used in IRIS assessments, if the results are unpublished and influential to the decisions made in the assessment (e.g., key for hazard characterization or used in dose-response modeling), the studies should be peer reviewed as described below.

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 (see **Sections 1 and 2**) may be 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 may 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, contractor-driven peer review would include an evaluation of the study similar to the peer review done for a journal publication. The contractor would identify and select two to 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 that the document underwent external peer review managed by the EPA, and the names of the peer reviewers would be identified. In certain cases, especially when the assessment is time sensitive, IRIS will conduct an assessment for utility and data analysis based on having access to a description of study methods and raw data that has 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).

Refining and Validating the Literature Search for Nonroutine Searches

The following process can be used to develop search string(s) when a nonstandard or targeted search strategy is needed:

- Identify a small set (10–20) of key validation set or test papers that the search would be expected to capture (e.g., papers identified in scoping and problem formulation). This study set can be used to test the sensitivity (error rate for missed studies) of the search. **The**

Assessment Manager and assessment team (or representatives of the appropriate disciplinary workgroups) should be involved in this process.

- Develop an initial search strategy, which can be informed by how the test studies are indexed in PubMed and other databases. Test search strings in each source database.

- If one of the seed or test studies was missed, determine the reason., i.e., was the paper not in the database (or is it incorrectly indexed in the database) or because of a limitation of the search string? All the test papers that could be found in a database need to be found with the search string; if any paper is missed, the search string should be reevaluated.

- Is the search identifying a high level of extraneous (off-topic) citations that could be eliminated through a change in the search string? Sometimes it can be challenging to develop a search strategy that removes the off-topic citations but still identifies the test studies. In these cases, machine-learning applications can be used to minimize the level of effort spent screening off-topic citations.

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 it is essential. 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 "Using HERO Literature Search Capabilities"). HERO uses five automated duplicate checking screens while importing references; however, some duplicates may persist and will require human review to identify and resolve. Many post-HERO software applications used to screen studies for relevance (e.g., Distiller SR) 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 RIS file.

Additional Search Strategies to Supplement Computerized Keyword Searches

Some publications will be missed with even the best-designed search strategy. Publications can be missed because they are not indexed correctly; because the databases searched do not include those journals; or because 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 in the initial screening process. The following strategies are approaches to identifying additional relevant literature; which of these should be used depends on the assessment needs. When used, these supplemental search strategies should be documented with enough detail to allow for replication. Steps taken should be

documented even if they do not produce additional citations. The results of all additional searches should be tagged to indicate the methods of identification.

- *Searching the reference list of pertinent reviews, assessments, and included primary (original) data papers*. The IRIS assessment team should review the reference list of key citations (review articles, other comprehensive documents, and articles with primary [i.e., original] data) to look for citations that may have been missed during database searching. Any potentially relevant citations should be screened for inclusion against the PECO and the source material that identified the reference should be documented during the literature screening process.

- *Citation mapping*. Additional search strategies that can be employed through a database (e.g., Web of Science) to include "forward" and "backward" searching from articles identified as key studies. Forward searching identifies articles that cite the key study, and backward searching identifies articles cited in the key study. Backward searches can be done manually by reviewing the cited references, typically in the introduction and discussion sections of a paper, for studies that were not identified in the database search. This type of searching is done on a case-by-case basis depending on factors such as whether the PECO has a targeted evidence type or health outcome focus, amount 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.

- *Searching of ToxCast/Tox21 high throughput screening data or bioinformatic databases for mechanistic evidence*. The CompTox Dashboard database (https://comptox.epa.gov/dashboard), search is a useful source for this type of search. Others include:

    - PubChem Bioassay database (https://pubchem.ncbi.nlm.nih.gov/), which partially overlaps with ToxCast/Tox21, but additionally includes biochemical and cell-based assay results from National Cancer Institute, National Institute of Neurological Disorders and Stroke, National Institute of Mental Health, European Structural Genomics Consortium, and commercial vendors.

    - BaseSpace Correlation Engine (Illumina, https://www.illumina.com/products/by-type/informatics-products/basespace-correlation-engine.html), which is a tool for meta-analysis of omics data (commercial).

    - Comparative Toxicogenomics Database, available at http://ctdbase.org/ (public).

    - Public data from omics experiments (Gene Expression Omnibus https://www.ncbi.nlm.nih.gov/geo/ or ArrayExpress https://www.ebi.ac.uk/arrayexpress/).

- Searching a number of the databases described in **Table 4-1** that include grey literature, such as ChemView and ChemID Plus.

- The public, stakeholders, and technical experts may provide additional publications.

Supplemental Literature Searches

In later stages of the assessment development process, more refined sets of focused searches may be required. The following bullets provide examples of possible scenarios.

- Targeted searches focused on 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 on potentially susceptible subpopulations and lifestages.

- Search strings to identify studies using descriptions of exposure to the agent of interest that do not include the chemical name (e.g., epidemiology studies of a broad chemical class or occupation may provide useful information).

- Targeted searches to identify absorption, distribution, metabolism, and excretion (ADME) and mechanistic studies, or studies of PBPK models; searches using the parent chemical name and CASRN alone may be too limiting for these types of data.

4.1.4. Documentation

Accurate documentation of the search strategy is essential to the systematic review process. Documentation of literature searches should include, at a minimum, the database(s) and date range covered by the search, search terms used and the filters (e.g., matching specific article types or MeSH terms in PubMed; matching topic areas in Web of Science) that were applied, and date(s) that the searches were performed (see an example template for documentation in **Table 4-2**).

Table 4-2. Example summary template of literature search results documentation

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

4.1.5. Updating the Literature Search

The literature search should be updated throughout draft development to identify literature published during the course of developing the review. The last full literature search update should be conducted less than 1 year before the planned release of the draft document for public comment. The team responsible for the assessment will manage the updating process with HERO information specialists, including identifying when and how often an update should be performed. The update should identify all relevant studies published since the last literature search update, and should be incorporated into the revised assessment in a manner consistent with the IRIS Stopping Rules (https://www.epa.gov/sites/production/files/2014-06/documents/iris_stoppingrules.pdf). 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 altered search string. Studies identified after peer review begins will only be considered for inclusion if they are directly applicable to the PECO eligibility criteria and fundamentally alter the conclusions in the assessment.

4.2. LITERATURE SCREENING

The literature screening process focuses on categorizing (or "tagging") studies into those that provide data that inform whether exposure to the chemical might cause toxicity (based on the PECO criteria and the supplemental tagging structure) and those that are irrelevant for the purposes of the assessment. 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 to facilitate this process, the Assessment Manager and assessment team should be directly involved in the literature screening process.

The literature screening results are released to the public in the protocol (or protocol update) for public comment. Any additional studies identified during public comment will be screened for adherence to the PECO criteria (see **Section 1.2.1**).

4.2.1. Determining Inclusion or Exclusion of Identified References

The PECO criteria are used to determine the inclusion or exclusion of identified references, focusing on capturing primary sources of health effects data. During the screening process, studies containing potentially relevant supplemental material will likely be identified and should be tagged as such as they may provide useful, and sometimes critical, information. Although they do not meet the PECO criteria, these studies are not necessarily excluded and often meet most, but not all, of the individual "P," "E," "C," "O" elements. Ultimately, they (1) may not be cited or considered in the assessment, (2) may be cited to provide context, or (3) may be carefully considered and cited in the assessment based on the results of analyzing the literature inventory, refining the evaluation plan (see **Chapter 5**) and organizing the hazard review (see **Chapter 7**). In many cases, these studies can be highly influential to specific assessment decisions. Studies to be categorized as "potentially relevant supplemental material" include the following:

- *Records that do not contain original data*: such as other agency assessments, informative scientific literature reviews, editorials, or commentaries. These materials can be used to cross-check for relevant records that might have been missed by database searches.

- *Mechanistic studies (including in vivo and in vitro studies and in silico models)*: Mechanistic information includes any measurement related to a health outcome that informs the biological or chemical events associated with phenotypic effects following exposure to a chemical but are not generally considered to be adverse outcomes (there are exceptions, e.g., hormone level changes are mechanistically relevant for many outcomes and may also be considered adverse outcomes themselves). The most relevant are those resulting from exposure to the agent of interest; however, information from studies that focus on the outcome/endpoint of interest are also of value. Mechanistic data are often observational and can inform key events responsible for biological effects. Upstream measurements of molecular, cellular, or physiologic interactions can implicate disruptions leading to adverse effects. In some cases, studies on related chemicals or of biological pathways or mechanisms that do not involve exposure to the agent of interest can provide useful mechanistic insights, potentially leading to additional targeted literature searches. Mechanisms for an outcome may vary by lifestage and data should be considered accordingly (U.S. EPA, 2006b).

- *Studies in nonmammalian model systems*: In most cases, studies in nonmammalian model systems (e.g., fish, birds, *C. elegans*) will be considered supplemental material. They may be considered key PECO studies in assessments where preliminary literature surveys (aka evidence or systematic maps) indicate studies in mammalian model systems are not available.

- *Toxicokinetic (ADME) studies*: The time course of the concentration of a chemical or its metabolite in biota (e.g., blood, liver) is determined by the rate and extent of ADME.

Relating adverse response to an appropriate internal tissue dose rather than administered dose or concentration is likely to improve the characterization of dose-response relationships (U.S. EPA, 2006a). Information and terms that are typically found in relevant ADME/toxicokinetic studies include the following:

- Absorption (systemic or local/portal-of-entry): Bioavailability, absorption rate(s), uptake rates, tissue location of absorption (e.g., stomach vs. intestine, nasal vs. lung), blood:air partition coefficient (PC), irritant/respiratory depression, overall mass transfer coefficient, gas-phase diffusivity, gas-phase mass transfer coefficient, liquid- (or tissue-) phase mass transfer coefficient, deposition fraction, retained fractions, computational fluid (airway) dynamics.

- Distribution: Volume of distribution (Vd) and parameters that determine Vd, including blood:tissue PCs (especially for the target or a surrogate tissue) or lipophilicity, tissue burdens, storage tissues or tissue components (e.g., serum binding proteins) and the binding coefficients, and transporters (active and passive). The fetus should also be considered as a potential site for distribution.

- Metabolism: Metabolic/biotransformation pathway(s); enzymes involved; metabolic rate; Vmax, Km; metabolic induction; metabolic inhibition, Ki; metabolic saturation/nonlinearity; key organs involved in metabolism; key metabolites (if any)/pathways; metabolites measured; species-, interindividual-, and/or age-related differences in enzyme activity or expression; site-specific activation (may be toxicologically significant, but little systemic impact); cofactor (e.g., glutathione) depletion.

- Excretion: Route(s)/pathway(s) of excretion for parent and metabolites; urine, fecal, exhalation, hair, sweat, lactation; elimination rate(s); mechanism(s) of excretion (e.g., passive diffusion, active transport).

- *Classical pharmacokinetic (PK) 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, that are not a PBPK model.

  - The data are typically the concentration time-course in blood or plasma after oral and or intravenous exposure, but other exposure routes can be described.

  - A classical PK model might be elaborated from the basic structure applied in standard PK software, for example to include dermal or inhalation exposure, or growth of body mass over time, but otherwise does not use specific tissue volumes or blood flow rates as model parameters.

  - Such models can be used for extrapolation like PBPK models, although such use might be more limited.

Note: ADME studies often report classical PK parameters, such as bioavailability (fraction of an oral dose absorbed), volume of distribution, clearance rate, and/or half-life or half-

lives. If a paper provides such results only in tables with minimal description of the underlying model or software (i.e., uses standard PK software without elaboration), including "non-compartmental analysis," it should only be listed as a supplemental material ADME study.

- *Physiologically based pharmacokinetic (PBPK) 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 elimination, and thereby estimate concentrations in blood or target tissues.

    - Usually specific to humans or defined animal species; often a single model structure is calibrated for multiple species.

    - Some mechanistic dosimetry models might not be compartmental PBPK models but predict dose to the body or specific regions or tissues based on mechanistic data, such as ventilation rate and airway geometry.

    - 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); that is, data that are independent of in vivo ADME data that are otherwise used to estimate model parameters.

    - Chemical-specific information on metabolism (e.g., Vmax, Km) or other molecular processes (e.g., protein binding) might be obtained by fitting the model to in vivo ADME data or determined from in vitro experiments and extrapolated to in vivo predictions.

    - They allow extrapolation between species, routes of exposure, or exposure durations and levels; that is, they do not just quantify ADME for specific experiments to which they have been fitted.

- *Exposure characteristics*: Exposure characteristic studies include data that are unrelated to toxicological endpoints, but which provide information on exposure sources or measurement properties of the environmental agent. While these data do not directly inform hazard interpretations, depending on the information, these data could inform the summary description of human exposure sources and environmental levels or provide insights related to the evaluation of individual studies (e.g., stability of the agent in solution, measurement of exposure biomarkers, precision of analytic detection methods).

- *Mixture studies*: These studies are generally considered to be supplemental unless they are epidemiological studies or contain an exposure or treatment group assessing only the chemical of interest, in which case they meet the PECO criteria.

- *Routes of exposure not meeting the PECO requirement*: Studies using routes of exposure that fall outside the PECO's exposure ("E") scope.

- *Human case reports or case series*: Studies without a comparison group (other than those specified in the PECO criteria, such as worker surveillance studies) when the number of subjects is ≤3.

- *Acute exposure duration*: Certain health effects (e.g., median lethal dose [LD_{50}] values) resulting from acute exposure (≤1 day) are tagged as supplemental during title and abstract screening. It is important to note that, in most assessments, studies of short-term duration (i.e., animal studies of less than ~30 days) are considered as meeting the PECO criteria during title and abstract screening and reviewed at the full-text level. Decisions on whether to include acute studies in the assessment are made as part of conducting the Literature Inventory (see **Section 4.3**) and Refined Evaluation Plan (see **Chapter 5**). For assessments that focus on chronic exposure, acute, and possibly short-term, exposure durations for health endpoints typically associated with very long durations of exposure (e.g., cancers that take a long time to develop) may be recategorized as supplemental if many longer-term studies are available.

- *Other*: There may be other types of information that are specific to an assessment question or issue identified during problem formulation.

A typical title and abstract screening form will have the following response options for assessing the PECO criteria: "yes, meets PECO criteria," "no, not relevant to PECO (aka exclude)," "tag as potentially relevant supplemental material," or "unclear" (see **Figure 4-4**). In many cases, a broad chemical name-based search is implemented to ensure that the mechanistic, ADME, and other types of supplemental evidence is fully identified and available for consideration. The IRIS Assessment Plan (IAP) should present decisions on how this information will be screened and evaluated. For example, in some cases, planned mechanistic analyses can be described in the specific aims of the assessment plan and the types of studies considered pertinent included in the PECO criteria. However, in most cases, it will not be possible to fully describe the analysis plan for mechanistic and other types of supplemental evidence until the assessment is further along. Thus, by necessity the approach is stepwise, and supplemental studies are tagged during screening as described below to allow for easy retrieval during the assessment.

During title and abstract screening, studies that meet the PECO criteria are tagged as "yes" and additional screening questions will ask about the type of evidence (human, animal, etc.). Teams will typically also want to inventory (using tags) the type of health outcomes assessed (e.g., hepatic, neurological). For studies identified as "supplemental" during the initial screening, including most mechanistic studies, it is similarly useful to categorize these studies for further consideration. For example, categorization of the mechanistic studies will often include capturing the relevant biological pathway affected or what health effects the data might inform and key properties that are examined in the studies (e.g., specific mechanisms of action, for example, receptor activation or binding activities; whether the endpoints inform key mechanistic characteristics identified for the health effect or toxicant), as well as information on the test system. However, these questions are often asked during a second phase of title and abstract screening conducted after the refined analysis plan is developed (see **Chapter 5**), as well as during full-text screening, when the scope of mechanistic analyses that need to be conducted becomes clearer. There is not a right or wrong approach for determining at which level (title and abstract or full-text) to tag studies, and often decisions of when to survey this information are made for pragmatic reasons. For example, the

time to screen studies at title and abstract level is increased when screeners are asked to apply more tags. So, for projects with many studies to screen, teams may want to wait and tag studies during a second phase of title and abstract screening or at the full-text level. In other cases, the title and abstract screeners may not have the content knowledge to do detailed tagging (e.g., on a particular mechanistic biological pathway). For assessment dissemination purposes, the categorization judgments are typically collapsed across title/abstract and full-text screening, but a record is maintained of where the tagging judgment was made (e.g., as a column in an Excel file created from Distiller or SWIFT Active output). Example screening forms are available in DistillerSR in the "IRIS Template Forms" project. Visual outputs of tagged categories are used to create literature inventory "heat maps" that can be presented in Word or Excel, interactive software applications such as Qlik or Tableau, or as dendrogram visualizations in Health Assessment Workspace Collaborative (HAWC).

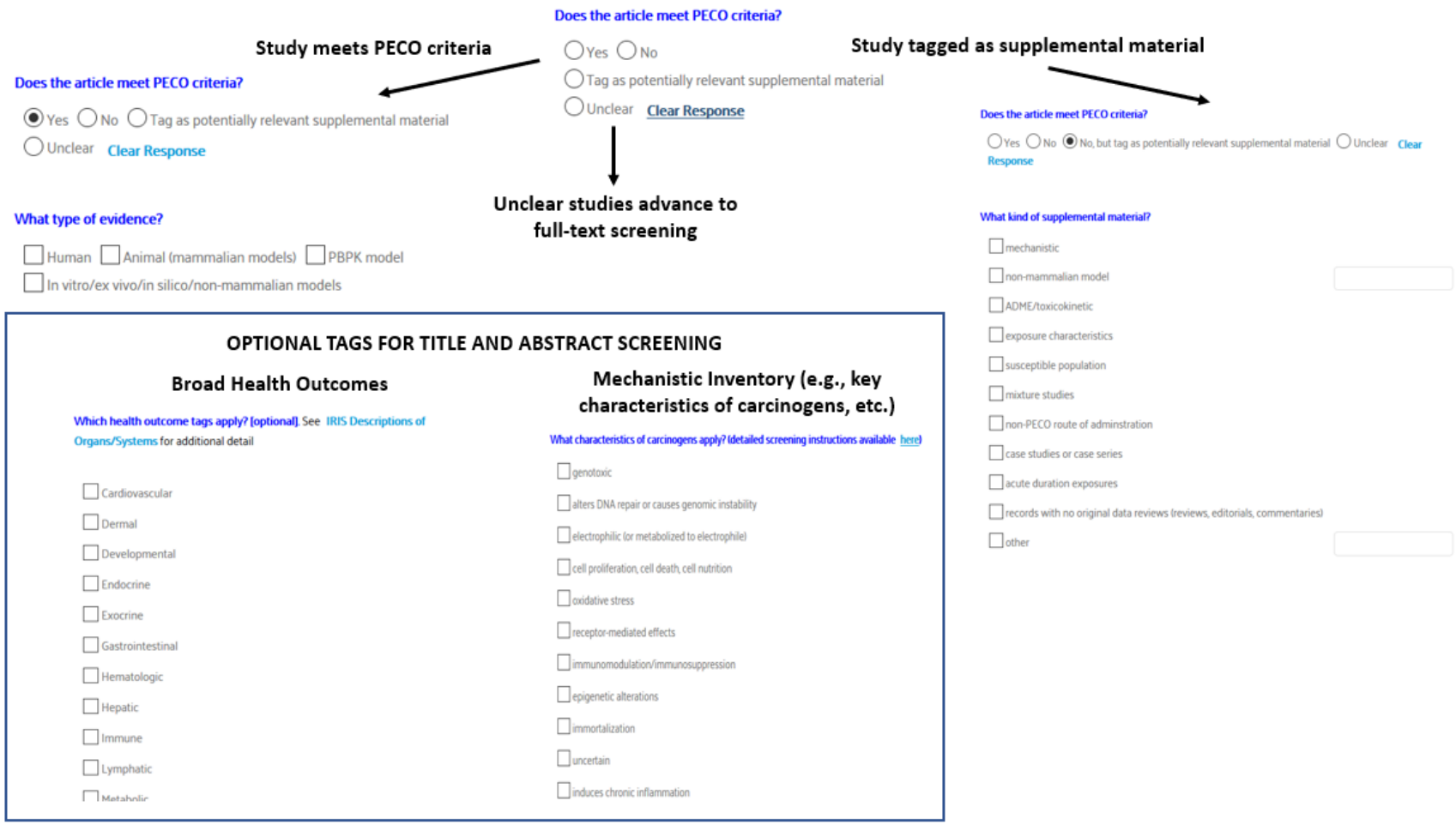


Figure 4-4. Common title and abstract screening and tagging questions.

4.2.2. Use of Machine-Learning Methods

The availability of specialized software applications for conducting literature assessments is expanding rapidly, especially for screening studies for relevance (Tsafnat et al., 2014). Machine-learning approaches (also referred to as natural language processing or artificial intelligence) can be used to efficiently prioritize large data sets to identify citations most likely to meet the PECO criteria or to identify information that may be used to conduct targeted searches. Use of machine-learning tools can typically reduce the screening burden by at least 50% (Howard et al., 2016). Machine learning may also prove useful after screening is complete to validate exclusion decisions based on included studies.

The SR Toolbox (http://systematicreviewtools.com/) is a comprehensive database of software tools and has advanced search features to help find tools tailored to specific aspect(s) of systematic review. Preferred software applications used within IRIS should be publicly available, free (when possible), interoperable with other software applications, and have technical support and methodological documentation provided by the developer. With respect to software applications that utilize machine-learning, when methodological documentation is not available from the developer then the performance is evaluated internally prior to routine usage. **Table 4-3** describes screening and other software applications commonly used for IRIS assessments, recognizing that this list is likely to expand over time. Users are encouraged to use training materials provided by the developer when using these tools. One-on-one or small group training sessions, including to other groups—both internal and external to EPA, can be organized upon request by contacting IRIS Program staff.

The use of machine-learning methods is documented in the assessment's systematic review protocol. Factors to consider include the number of studies that need to be screened and availability of seed studies for training. Manual screening at the title and abstract level is relatively fast, typically 10–20 seconds per study. For screening projects of <2,000 studies there may not be a significant time saving by using machine-learning approaches, particularly when a seed set is required. Machine-learning approaches work best when known "yes, meets PECO criteria" and "no, not relevant to PECO criteria" seed studies are available. If seed studies are not available, then active learning approaches such as SWIFT Active can be considered. Care should be taken when seed studies are used that they provide sufficient coverage when broad PECO questions are used.

Table 4-3. Summary of commonly used specialized software applications for literature screening

Software	Key features	Use in IRIS assessments
DistillerSR	• Web-based. • Not free, but competitively priced. Currently, no free program appears to be available that the features and extent of technical support as DistillerSR. • Artificial intelligence features added in 2018. • Easy to add screeners, including from outside EPA. • Detailed 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. • IRIS SOPs are available to transfer Distiller tagging decisions into HERO.	• Used in IRIS assessments for title-abstract screening, full-text screening, and to create literature inventories. DistillerSR is typically used for full-text screening and to create literature inventories even if other software is used to conduct title and abstract searching. • Compared to other applications, DistillerSR has more options for users to customize forms, including for use to create audit reports, study evaluation tools, and data extraction. IRIS typically uses HAWC for study evaluation and extraction for epidemiological and animal toxicology (see **Chapter 8**), but DistillerSR may be a suitable alternative for content that is not currently collected in HAWC. Complex data extraction like that done for epidemiological and animal toxicology studies is not easy to implement in DistillerSR, which is one reason why IRIS uses HAWC for these purposes. Also, DistillerSR does not have the visualization capabilities of HAWC.

Software	Key features	Use in IRIS assessments
SWIFT-Review	• Must be downloaded for installation. • Free. • Preset literature search filters for different types of study populations (human, animal, in vitro) and health outcomes (Howard et al., 2016). o The search strategies used in the filters were developed by professional information scientists and are available from within the software. The search strategies can be customized by the user. • Useful for problem formulation, topic modeling, data visualization, and document prioritization via machine learning. • Machine-learning 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. • Detailed help instructions available from within the software. • Interoperable with HERO and other software applications such as DistillerSR, SWIFT-Active Screener, and HAWC.	• Widely used in IRIS assessments during problem formulation and to prioritize records for screening in another software application.

Software	Key features	Use in IRIS assessments
SWIFT-Active Screener	• Web-based and free (upon request). • Easy to add screeners, including from outside EPA. • Incorporates "Active Learning" machine-learning methods that continuously update a prioritization model during screening, pushing the articles most likely to be relevant to the top of the list. • Incorporates a unique statistical model that estimates recall (percentage of relevant articles found so far), allowing users to make an educated decision about when to stop screening. • Machine-learning and recall-estimation models have been successfully validated using a large corpus comprising 26 systematic review data sets varying in size, percentage of relevant studies, and overall topic area. • Studies prioritized in SWIFT-Review can also be easily imported into SWIFT-Active Screener. • Detailed 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). • Users have direct access to a dedicated support and informatics team and user requests for new features, changes, and other customizations are actively considered and implemented.	• Widely used in IRIS assessments for title and abstract screening, especially when there are many studies to screen (e.g., 2,000+) and/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. • Active Screener supports multilevel screening projects and can be used for title-only, title/abstract, and full-text screening. Complex questionnaires are now supported, and the tagging and information extraction features are under active development with additional refinements anticipated in the near future.

Software	Key features	Use in IRIS assessments
HAWC	• Web-based and free. • EPA uses a derivative of the version of HAWC used by the NTP, which is free and open source. EPA HAWC is freely available for public use but cannot be customized except by HERO staff (or their contract support staff). • Interactive "click to see more" study flow diagrams. • Easy to add screeners, including from outside EPA. • Detailed help videos available at https://hawcprd.epa.gov/about/. • Full-text articles can be uploaded as attachments (individually). • Customizable tagging options. • No machine-learning or artificial intelligence capabilities.	• IRIS uses HAWC extensively for study evaluation and data extraction (see **Chapter 8**), but not for study screening because the software does not currently support multiple screeners and conflict identification/resolution tracking. IRIS SOPs are available to transfer screening decisions from other software applications into HAWC for subsequent study evaluation and data extraction.
Qlik Sense, Tableau, Power BI	• These are not screening tools but can be used to create web-based interactive study flow images and literature inventories. • Detailed help instructions available from within the software. • Easy to use and allows user to create many different visual displays.	• SOPs are being developed for template input Excel file formats to create web-based interactive study flow and literature inventories based on screening and tagging results. In most cases, the input Excel file will be based on screening conducted in DistillerSR, but the templates can be adapted when other screening applications are used.

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/; NTP version https://hawcproject.org/.
Qlik Sense: https://edap.epa.gov/public/hub/stream/aaec8d41-5201-43ab-809f-3063750dfafd.
Tableau: https://public.tableau.com/en-us/s/.
Power BI: https://powerbi.microsoft.com/en-us/.

4.2.3. Performing and Documenting the Screening Process

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 and/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 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 PECO relevance and tagged according to source. The time estimates in **Table 4-4** show a range of average times for experienced reviewers that can be used to estimate project timelines.

Table 4-4. 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 may 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).

Title and Abstract Screening

A structured form in literature screening software applications (e.g., DistillerSR, SWIFT Active) is created to assist in the literature screening process. Following a pilot phase to calibrate screening guidance, two screeners independently conduct a title and abstract screen of the search results to identify references that appear to meet the PECO criteria. 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 only require one screener to screen a study as "include" but two screeners required to screen a study as "exclude." This is acceptable because those studies marked as included would be confirmed relevant at the full-text level. References with no abstract may be screened based on title relevance or page numbers (articles two pages in length or less are likely to be conference reports, editorials, or letters). For references in a language other than English, online translation tools may be used to assess eligibility at the title and abstract level. Discussion among the primary screeners with consultation by a third reviewer or technical advisor (if needed) is used to resolve any screening conflicts at the screening level. Standards in the field do not require the rationale for excluding studies at the title and abstract level to be specifically annotated. Studies are often excluded because they do not meet multiple PECO criteria, and this becomes cumbersome to track in study flow diagrams. As discussed below, annotating rationales for exclusion at the full text level is recommended.

Title and abstract 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 based on title/abstract screening is unclear. Also, during title/abstract screening, studies not meeting the PECO criteria but identified as "potentially relevant supplemental material" may 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 tagged as such. For example, a study may examine health effect-related endpoints in exposed humans, but also test endpoints related to potential mechanisms as well as metabolism of the test agent. In this case, the study should be categorized as *human health effect studies*, *mechanistic studies*, and *ADME*. Conflict resolution is not required during the screening process to identify supplemental information (i.e., tagging by a single screener is typically sufficient to identify the study as potentially relevant supplemental material that will be inventoried and may be incorporated during draft development).

Full-Text Screening

Records that are not excluded based on the title and abstract advance to full-text review. Full-text copies of these potentially relevant records are retrieved, stored in the HERO database, and independently assessed by two screeners to confirm eligibility according to the PECO criteria. When the HERO URL link to the record is included in the reference file (e.g., in the URL field), then access to the full text can be obtained directly from the screening form, which makes the screening process go faster. It is critical to maintain the HERO ID as the primary means of identifying studies in the reference management file (i.e., RIS file) to maintain interoperability between HERO and screening software applications.

As was done during title and abstract 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. During this process, it is likely that for some references, the additional review of the full article will result in the realization that the reference is "not on topic,"

or is a background or "potentially relevant supplemental material," and should be tagged accordingly. In contrast to title and abstract screening, the reason for excluding studies at the full-text level is annotated during screening. For example, there may be references that initially seemed to meet the inclusion criteria, but this decision was changed after more careful review of the design or analysis. In these cases, it is important to document reasons that may not be initially obvious, particularly if the information cannot be found in the abstract. To screen full-text, non-English references, approaches for language translation may include engagement of a native speaker from within EPA, use of free web-based translation software, or fee-based translation services. Because use of fee-based services is expensive, free web-based tools can be used to help assess the likely impact of the study on the assessment. Use of fee-based translation services will focus on non-English studies that are likely to be impactful on hazard conclusions or dose-response analysis. Otherwise, the non-English study will generally be considered as supplemental material when creating a summary level literature inventory.

When there are multiple publications using the same or overlapping data, all publications on the research will be included, with one selected for use as the primary study; the others will be considered as secondary publications with annotation indicating their relationship to the primary record during data extraction. For epidemiology studies, the primary publication will generally be the one with the longest follow-up or the largest number of cases. For animal studies, the primary publication will typically be the one with the longest duration of exposure, or that assessed the outcome(s) most informative to the PECO. For both epidemiology and animal studies, EPA will include relevant data from all publications of the study, although if the same outcome is reported in more than one report, duplicative results will only be extracted once.

Documentation and Tagging

The results of the screening process are posted on the project page for the assessment in the HERO database with references tagged with appropriate category descriptors (e.g., included, excluded, tagged as potentially relevant supplemental material). Ideally, the tags used should be synchronized between the screening software and the HERO project page; modifications can be made but will need to be requested through HERO librarians. The included references (and sometime selected supplemental material, such as mechanistic data) advance to the next stage of assessment development, creating literature inventories (see **Section 4.3**), where a few key study characteristics will be extracted to help organize subsequent evaluations.

Figures 4-5 and 4-6 (showing use of machine-learning software) provide examples of the "literature flow diagram" that summarizes the literature search and screening results as outlined in **Sections 4.1 and 4.2** above. A variety of study flow formats are acceptable for use, depending on preferences of the assessment team and the nature of the study flow results. The study flow tags are also disseminated via the HERO database as well as in Excel file or interactive visualizations (e.g., Qlik Sense, Tableau, PowerBI). It may not be possible to present all the subtypes of evidence in the figure (e.g., specific types of supplemental material, types of health outcome assessed in

included studies). As described above, studies can be marked as excluded or supplemental material either during title and abstract or full-text review, and this information is also tracked and reported in public disseminations.

In general, targeted searches that fall outside the scope of the initial assessment search should be presented as separate study flow images. For example, targeted searches may be conducted to identify mechanistic data coming from upstream measurements of molecular, cellular, or physiologic interactions to implicate the key biological disruptions responsible for biological effects. These upstream measurement data are informed by studies of biological pathways that may not involve exposure to the specific agent of interest, thereby requiring additional targeted searches of the literature. It should be emphasized that the relevant references identified in **Figures 4-5 or 4-6** still require further analysis and decisions about their potential use in the synthesis of evidence (see **Chapters 5 and 7**). Thus, it is possible that study flow diagrams prepared early in the assessment as part of an evidence/systematic mapping or problem formulation process may be adjusted to reflect refined assessment judgments. These judgments (and rationales) should be described in the version of the assessment protocol released in conjunction with the draft assessment.

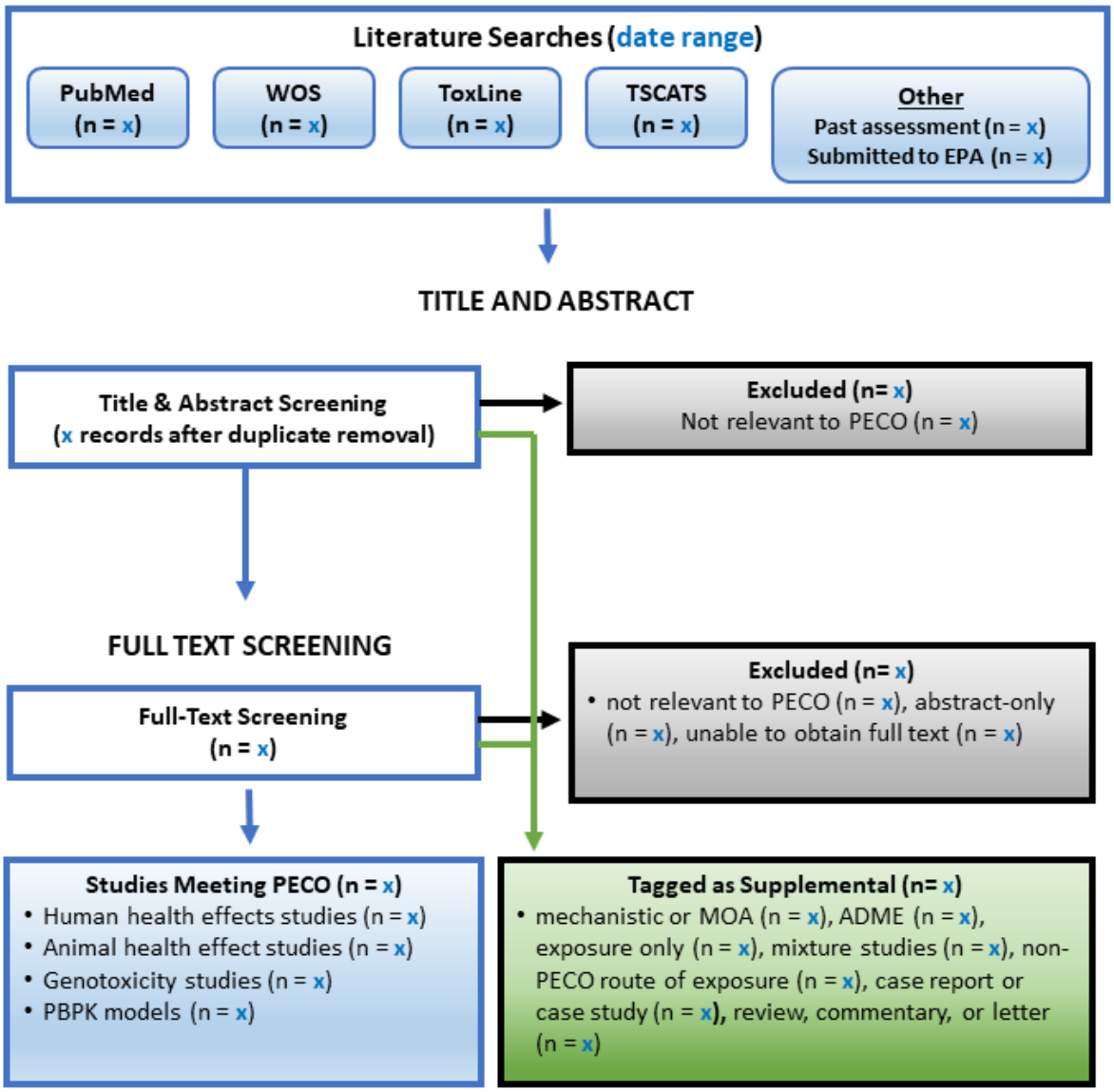


Figure 4-5. Example literature flow diagram.

IHAD = Integrated Hazard Assessment Database; OPP = Office of Pesticides Program; WOS = Web of Science.

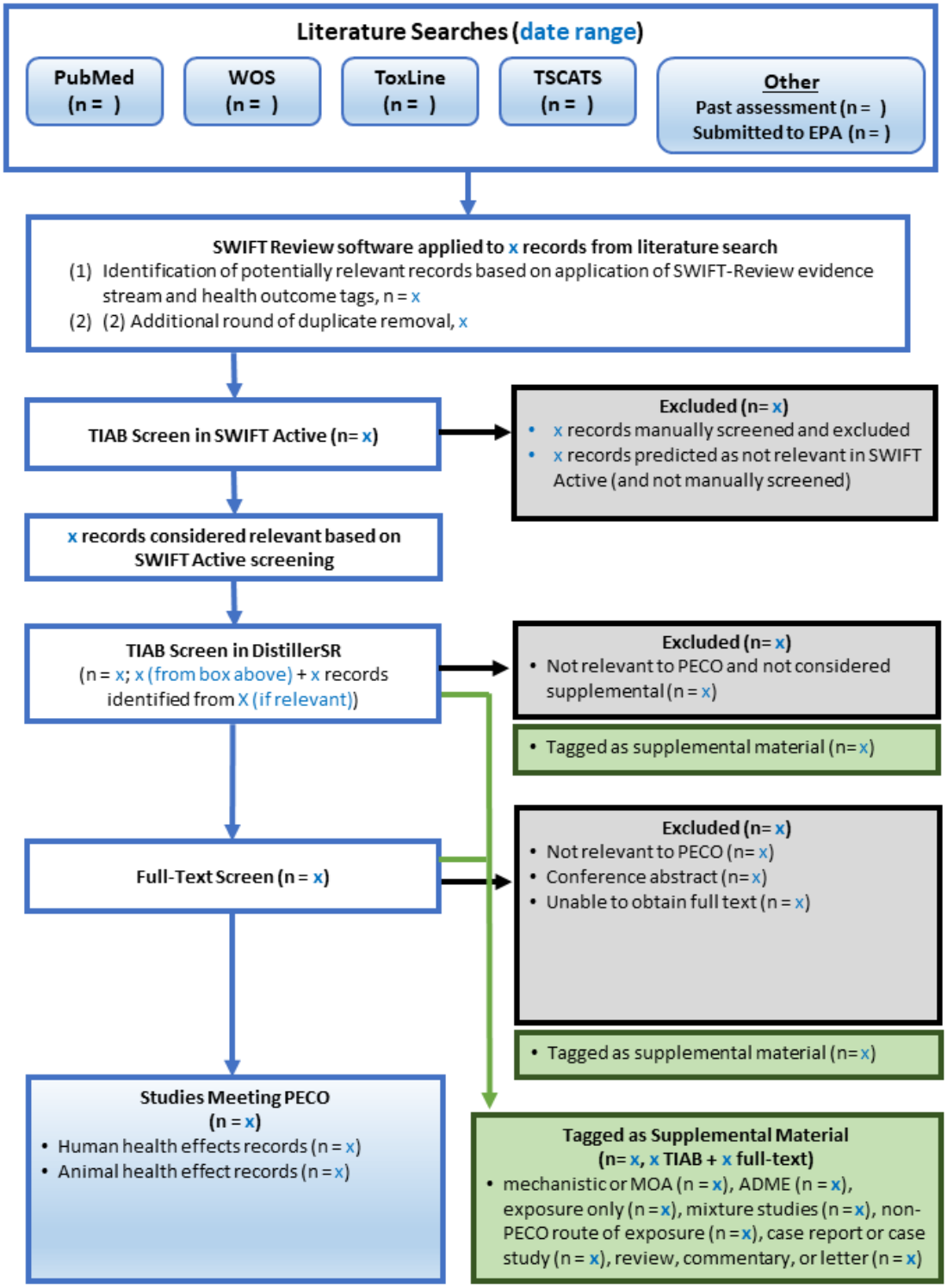


Figure 4-6. Example literature flow diagram when machine-learning software is used.

IHAD = Integrated Hazard Assessment Database; OPP = Office of Pesticides Program; WOS = Web of Science.

4.3. LITERATURE INVENTORIES

To facilitate subsequent review of individual studies or sets of studies by topic-specific experts or disciplinary workgroups, the relevant human and health-effect studies, as well as other informative mechanistic and ADME/PBPK studies, should be organized into literature inventories. These inventories build on tagging that was conducted during study screening to further characterize studies that meet the PECO criteria or certain types of supplemental material. The inventories are intended to serve as summary-level, sortable lists of the available studies and should include some basic study design elements to be used by the subject matter experts to organize and prioritize their review of the studies. Importantly, however, 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 and that are likely to provide insight into subsequent grouping or prioritization decisions (see **Chapter 5**, Refined Evaluation Plan) and to develop an understanding of what topic-specific study evaluation considerations may need to be developed (see **Chapter 6**, Study Evaluation). Template forms are available in DistillerSR in the "IRIS Template Form" project to facilitate the creation of literature inventories. Use of a standard form for creating literature inventories makes it easier to control file format and, thus, develop interactive literature inventories that are consistent across assessments.

IRIS assessments will typically include inventories of the following basic study types: epidemiology studies, animal health-effect studies by effect type (e.g., neurotoxicity, immunotoxicity, cancer), controlled human exposure studies, ADME or PBPK studies, and mechanistic studies (which may also be subdivided; see **Section 4.3.3**). Given their intended use in aiding subject matter experts in organizing their review of studies, inventories emphasize categorization of the health outcomes and/or endpoint measures included in the study. In some cases, it may be useful to expand some of the standard broad study type categories[8] into separate subcategories if the evidence base contains many studies investigating specific organs or systems, or if very specific health effects are known to be of concern for the agent under review. For example, a chemical with a large database for the single category of "reproductive toxicity" could be expanded to two groups: "male reproductive toxicity" and "female reproductive toxicity." These categories can then be subdivided according to the types of outcomes reported in the available studies (e.g., organ weights, histopathology, hormonal changes). Subcategorization is especially important for mechanistic studies, particularly for large databases; see **Section 4.3.3** for a detailed description of creating inventories for mechanistically relevant studies.

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.

[8]A list of standard terms for broad categories can be found at: http://www2.epa.gov/iris/iris-descriptions-organssystems; this list may be refined and revised.

4.3.1. Human or Animal Health Effects Study Inventories

The information extracted into the inventory should be minimal to maintain efficiency. It should generally comprise basic aspects of study design and the endpoints included in the study. The template literature inventory forms available in DistillerSR can be copied into chemical-specific projects and customized as needed by the assessment team. For epidemiology 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 toxicology studies, it includes information on exposure duration and timing (e.g., acute, chronic, developmental), administered exposure levels, route of exposure, species, strain, and sex. Both epidemiological and animal toxicology inventories collect information on broad categories of health outcomes (e.g., cancer, neurological, immune) and specific endpoints measured in each study. A brief description of key study findings may also be included in the literature inventory, especially in cases where a systematic evidence map analysis is anticipated. Extracting more detailed study information at this stage is typically not efficient because some of these studies may not be used in the assessment.

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

For ADME and PBPK studies, the range of exposures and time points studied, and, when available, the identification of parent compound and metabolites should be included. Regarding ADME data, almost all ADME studies provide information that is at least qualitatively useful, and it is rarely the case that there are competing mechanistic hypotheses for ADME. Because ADME studies vary quite widely in study design and details, flexible Microsoft Excel-based inventory table structures have been developed and are also available as a DistillerSR form. This inventory can help to abstract and organize specific information across the following study types: animal in vivo, human in vivo, and in vitro. These tables can also be used to summarize publications describing PBPK/pharmacokinetic (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 6.4**).

4.3.3. Mechanistic Information Inventories

Although the basic process for developing inventories is similar for mechanistic studies and human or animal health effect studies, the approach taken for analyzing mechanistic evidence has been adapted to include more steps for reconsidering the depth of the analyses that will be required for the assessment. As mentioned in **Section 4.2**, the initial literature screening will identify sets of other informative studies, including mechanistic studies, as "potentially relevant supplemental material," and not as a component of the PECO, which identifies studies presenting apical health effects that will be evaluated for reporting quality, risk of bias, and sensitivity.

Although existing mode of action (MOA) hypotheses are identified during problem formulation, at this early stage there still may be an incomplete understanding of the complex biological pathways involved in the toxic response to a chemical. For many chemicals, in vitro studies alone can outnumber human or animal health effect studies by orders of magnitude. In addition, because mechanistic studies possess a wide range of applicability to an assessment (e.g., they can suggest potential health effects that have not been examined in other study types, identify human biomarkers, explain conflicting findings, inform susceptibility, inform the relevance of effects observed in animals to humans), the questions and analyses applied to mechanistic studies will differ depending on the requirements for each assessment, requiring a multifaceted approach. To undergo a full reporting quality, risk of bias, and sensitivity evaluation of every identified study that may report mechanistic information before the relevant toxicity pathways have been identified or the needs of the assessment are better understood would not be an effective use of time. Therefore, to systematically process mechanistic studies, additional steps are taken to screen the mechanistic database, produce literature inventories, and narrow the focus of the analyses that will lead to the identification of mechanistic studies that will be evaluated.

The identification of studies that report mechanistic information is accomplished sequentially throughout title and abstract screening, full-text screening, and inventory extractions of human and animal studies. Although in vitro studies may be quickly tagged as mechanistic during early screening steps, human and animal studies reporting mechanistic information can be difficult to identify. Once a preliminary mechanistic database is completed, additional screening steps will further organize the mechanistic studies into categories. See **Table 4-3** for a summary of tools and templates for screening that can be customized for each assessment.

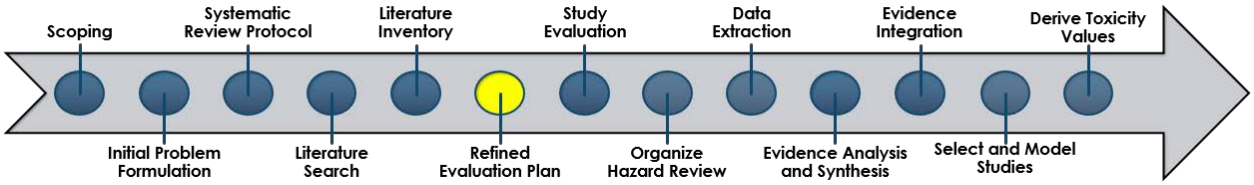
Developing an initial organizational scheme when screening mechanistic evidence is essential for efficiency in the analysis and synthesis stages, particularly when multiple health effects involve overlapping mechanistic pathways. This categorization will form the structure of the study inventories and facilitate a more effective review by subject-matter experts. Studies can initially be organized based on relevance to broad categories, e.g., organ system toxicity, immunotoxicology, cancer (including genotoxicity studies), neurotoxicity, reproductive and developmental toxicity, ADME/PK (note that studies may be added to more than one category). Categories corresponding to mechanistic events and/or key events (i.e., as part of an MOA or adverse outcome pathway [AOP]) or biological pathways can also be implemented for screening purposes. As introduced in **Chapter 2**, the problem formulation and assessment plan development stages of the IRIS process include identification of existing MOA hypotheses. The bibliographic information gathered from the literature survey during problem formulation can be used to develop screening strategies. Other approaches are useful; for example, for studies relevant to carcinogenesis, sorting by the ten key characteristics of human carcinogens (Smith et al., 2016) is an objective organizational approach that can facilitate the grouping of studies that report mechanistically related endpoints and assays; this concept has been extended to other toxic effects

[e.g., (Arzuaga et al., 2019)]. The most useful screening categories should be determined by the subject-matter experts. The template DistillerSR screening forms can be adjusted and refined during screening to accommodate new screening subcategories as trends and relationships in areas of experimentation and study design are identified. Further refinements with additional categorizations may be added after each screening step in response to the emergence of more focused areas of mechanistic study.

It is important during these initial screening stages to become familiar with chemical-specific issues and potential mechanisms of toxicity; discovery of new information in later stages may necessitate substantial revisions. This preliminary work includes clearly defining the chemical species of interest, active metabolites, and acceptable chemical formulations. It should capture any known issues relating to chemical purity and mixtures of isomers, valence or oxidation state (for metals), or concerns regarding solubility or volatility. As the screening progresses, the areas of research interest will indicate areas of mechanistic relevance, leading to the identification of networks of mechanistic events that will then inform the biological plausibility of the connection between exposure and apical effect (e.g., AOPs). Importantly, after incorporating information from the syntheses of human and animal evidence, the mechanistic inventories can help to gradually narrow the pool of evidence to the studies most relevant to informing hazard evaluations and assessment conclusions (see **Chapter 10**).

Once the screening steps have been completed, literature inventories can be produced to provide a synopsis of the data available for analysis. As with the human and animal health effects study inventories, the information extracted should, at first, involve only a minimal review and extraction of information. The information to be extracted from each study can be customized for each chemical. In general, the inventory should capture information that will help later stages of prioritization and analysis (see **Section 10.1**), e.g., test article, vehicle, and method of exposure (including exposure levels tested); whether a study was performed in vivo or in vitro; the species, strain, and sex of the experimental model; the tissue, region, and/or cell type studied; and the endpoints or outcomes measured, the assays used, and results. This step should include sufficient detail for subject-matter experts to develop a refined evaluation plan (see **Chapter 5**) that will guide the prioritization of mechanistic studies and the identification of studies for evaluation, but not an exhaustive capture of data or study details. By organizing and categorizing studies in the mechanistic database across models and endpoints into inventories, it will be possible to systematically analyze mechanistic study findings across diverse study designs from multiple angles and prioritize evidence depending on the hazard questions that arise. In addition, the screening tools and inventories provide a decision record that will increase transparency in the process for analyzing mechanistic information.

5. REFINED EVALUATION PLAN



REFINED EVALUATION PLAN

Purpose

- Develop a plan for evaluating outcomes from studies with respect to potential methodological considerations.

Who

- Assessment team members.

What

- Refined evaluation plan in the protocol.
- Refined inventory (if needed).

The purpose of the refined evaluation plan in the protocol is to describe any refinements to the set of studies meeting populations, exposures, comparators, and outcomes (PECO) criteria to be carried forward to study evaluation. This may be particularly necessary if a large number of studies meeting PECO were identified during the screening process. The process also helps determine which subset(s) of studies tagged during literature screening as "potentially relevant supplemental material" may need to be prioritized for consideration in the assessment.

In addition, the refined evaluation plan should identify and group the outcomes that will be the primary focus of study evaluation. This specification is needed for implementation of efficient study evaluation and data extraction because these processes are often the most resource intensive phases of conducting a systematic review and generally require the development of topic-specific outcome/endpoint considerations to guide the rating process. Even when a priori considerations are available, additional refinements and clarification should be expected when evaluating the available studies. Any refinements will be tracked in the updated/final assessment protocol.

Examples of questions that could be used to refine the evaluation plan include:

- If the resulting database **for each specific health effect** includes many studies that would **not** be expected to be key studies for hazard identification or dose-response, the assessment team may consider options for prioritizing a subset of the most relevant studies

for study evaluation and synthesis. Studies that are deprioritized are tagged as supplemental materials (typically at the full-text level) for the purposes of tracking study eligibility. Such considerations may include the following:

- Focusing on toxicity studies including exposures below a specified range, or employing an exposure route(s) that is the primary focus of the assessment, when many studies on the endpoint are available;

- Focusing on studies with more specific or objective measures of toxicity (e.g., functional endpoints) when a reasonable number of such studies are available, rather than studies with nonapical, broad, or nonspecific measures (e.g., self-reported symptoms); and/or

- Focusing on studies that address critical lifestage- or exposure duration-specific knowledge of the development of the health outcome (e.g., for endpoints relating to organ development or cancer, respectively), when many studies examine the same endpoint.

- Focusing on mechanistic events that are potentially the most impactful to the assessment (e.g., mutagenicity).

- Does absorption, distribution, metabolism, and excretion (ADME) information inform the evaluation plan?

  - Are there chemical moieties (parent chemical or metabolite[s]) found in test species and humans that can inform the specific test article(s) of primary interest?

  - What are the most informative test subjects (e.g., species, lifestage, disease state) based on information about metabolic pathways (including identification of responsible enzymes, knowledge of the functional maturation of key enzymes or related processes, and characterization of metabolic competition)?

  - Are there known lifestage-specific differences in absorption, distribution, metabolism (toxification or detoxification), or excretion germane to the assessed chemical?

  - Does information about ADME suggest additional endpoints of concern or required exposure durations?

  - Can information about biological persistence (e.g., half-life) and primary routes/methods of elimination shed light on the most informative timing of endpoint testing after exposure and reversibility, including evaluations across related chemicals (e.g., isomers; parent chemicals vs. metabolites), as well as the potential for use of exposure biomarkers?

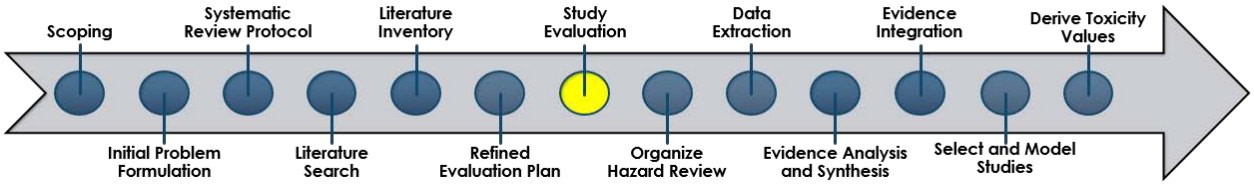
  - Notably, the aforementioned questions should be framed considering whether there are well-established ADME differences between the test species and humans that might affect the selected approaches.

- Is there a need for additional (targeted) searches (e.g., expanding to include a broader occupational categorization that could include the exposure of interest, exploration of a hypothesized mechanism of action)? If so, the assessment team will work with information

specialist(s) to propose additional search strategies and to test what is gained through their effect implementation.

- What issues will need to be considered when evaluating the study methods? Examples of these issues include reliability of various exposure measures or procedures for evaluating an endpoint of interest, or the sensitivity of a particular method used to evaluate a specific endpoint. Identification of these issues or considerations is based on an initial review of the methods used in the identified studies (or a subset of the studies, for large databases), background research (e.g., pertaining to sensitivity and specificity of a type of assay), review of secondary resources (e.g., review papers, commentaries), and consultation with technical experts. These considerations are incorporated into the specification of details of the study evaluation procedures. When feasible, details on how assessment-specific considerations will be addressed during study evaluation are indicated in the initial protocol release. However, it is also possible that additional adjustments will be made while implementing the protocol, i.e., during pilot work to calibrate rater responses for study evaluation. Such adjustments would be captured in a protocol update.

6. STUDY EVALUATION



EVALUATION OF INDIVIDUAL STUDIES

Purpose

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

Who

- Assessment team members and disciplinary workgroups (possibly with contractor support).

What

- Study evaluation considerations.
- Documentation of study evaluations.

6.1. STUDY EVALUATION OVERVIEW FOR HEALTH EFFECT STUDIES

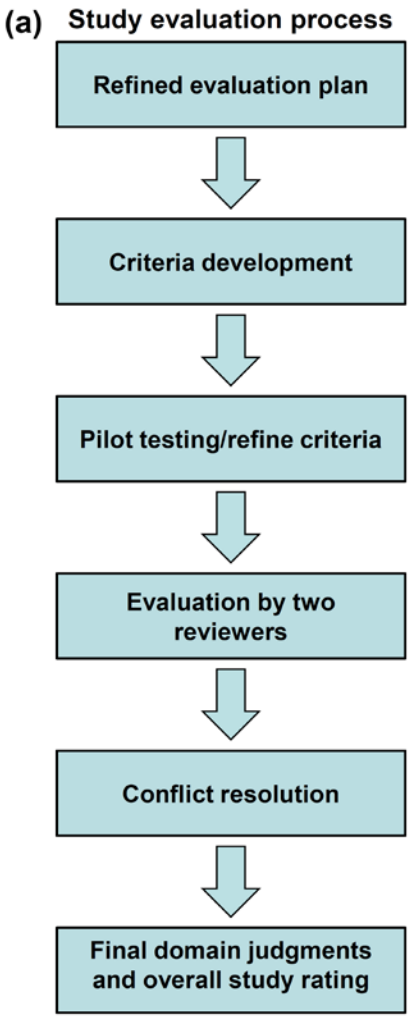
The purpose of this stage is to evaluate the studies for their validity and utility in assessing a potential change in the health effect, independent of the direction or magnitude of the study findings. Key concerns for the review of epidemiology, animal, and controlled human exposure studies are risk of bias, which is the assessment of internal validity (factors that may affect the magnitude or direction of an effect in either direction) and insensitivity (factors that limit the ability of a study to detect a true effect; low sensitivity is a bias towards the null when an effect exists). Reporting quality is evaluated to determine the extent the available information allows for evaluating these concerns. Additional detail on these concerns are provided in **Table 6-1**. Study evaluation, as defined herein, is a broad term encompassing interpretation of a variety of methodological features (e.g., study design, exposure measurement, study execution, data reporting). Study evaluation, as operationalized in the IRIS program, is analogous to other approaches that evaluate "study quality" or "utility" in that a wider set of issues are addressed in addition to risk of bias, including the rigor of study execution, study sensitivity, and reporting (Lynch et al., 2016; Rooney et al., 2016; NRC, 2014; Higgins and Green, 2011a). The study evaluations are aimed at discerning the expected magnitude of any identified limitations (focusing

on limitations 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 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.

Table 6-1. Key concerns for study evaluation of health effect studies

Key study evaluation concern	Aspect of the study design and conduct under evaluation
Reporting quality	Assesses whether enough information is provided to understand how the study was designed and conducted.
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 may bias the results.
Study sensitivity	Assesses whether there are factors in the design and conduct of the study that may reduce its ability to observe an effect, if present.

Study evaluation occurs before extracting results and characterizing hazards associated with exposure to the chemical of interest. The general approach (described in this section) of study evaluation for epidemiology, animal, and controlled human exposure studies is the same, but the specifics of applying the approach differ; thus, they are described separately (see **Sections 6.2 through 6.4**). The general approach for reaching an overall judgment is illustrated in **Figure 6-1**. Overall judgments should be assessed at the outcome level because different outcomes in the same study may have different strengths and limitations.



(b) Individual evaluation domains

Animal	Epidemiology
Selection and performance • Allocation • Observational bias/blinding	Participant selection
Confounding/variable control	Confounding
Selective reporting and attrition	Selective reporting
Exposure methods sensitivity • Chemical administration and characterization • Exposure timing, frequency, and duration	Exposure measurement
Outcome measures and results display • Endpoint sensitivity and specificity • Results presentation	Outcome ascertainment Analysis
Reporting quality	Other sensitivity

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 resulting bias or lack of sensitivity is unlikely to be of a notable degree.
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 unusable for hazard identification or dose response but may be used to highlight possible research gaps.

Figure 6-1. Overview of Integrated Risk Information System (IRIS) study evaluation approach. (a) An overview of the evaluation process. (b) The evaluation domains and definitions for ratings (i.e., domain and overall judgments, performed on an outcome-specific basis).

While this chapter describes a systematic and transparent process of determining confidence in a study, this process is inherently one of expert judgment. 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/endpoint combination is accomplished by applying prespecified considerations to a set of domains. These domains differ for epidemiology and animal studies (see **Figure 6-1**) and are discussed below in their respective sections. The core and prompting questions provided in this handbook for each domain are meant to guide the reviewer to seek out and think about relevant information pertaining to specific aspects of the study. Documentation of the important methodological features of a study is typically an iterative process, requiring refinement of an initial set of questions as specific features of the exposure setting or dosing regimen, endpoint(s), or study design(s) are discovered among the studies meeting populations, exposures, comparators, and outcomes (PECO) criteria. 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) guidance for carcinogenicity, neurotoxicity, reproductive toxicity, and developmental toxicity (U.S. EPA, 2005b, 1998, 1996b, 1991). Some prespecified considerations (e.g., the classification of the methods used to ascertain a specific outcome) may be used for an evaluation of that outcome in any assessment, whereas others may be assessment specific. For example, evaluation of exposure measures in epidemiology studies will need to be developed for each chemical. Development of additional considerations ideally includes a pilot phase to assess and refine the evaluation process (e.g., comparison of decisions and reaching consensus between reviewers, and when necessary, resolution of differences by discussion between the reviewers, the chemical assessment team, or technical experts). As reviewers examine a group of studies, additional chemical-specific knowledge or methodologic concerns may emerge and a second pass may become necessary. Once developed, the reviewers must ensure that each criterion is applied in a consistent fashion across all studies being evaluated.

Generally, 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). This helps ensure quality assurance. However, based on assessment needs, the assessment team should make decisions about how many reviewers are needed. While more than one reviewer is ideal, there may be rare instances when one reviewer is acceptable, such as when the assessment needs to be conducted under a rapid time frame and the outcome being reviewed is unlikely to be a driver for the assessment.

For studies that examine more than one outcome, the evaluation process should be outcome-specific, as the utility of a study may vary for the different outcomes. If a study examines multiple endpoints for the same outcome, evaluations may be performed at a more granular level if appropriate, but these measures may still be grouped for evidence synthesis. The evaluation can provide 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.

Authors may be queried to obtain missing critical information, particularly when there is missing reporting quality information or data (e.g., content that would be required to conduct a meta-analysis or other quantitative integration), or additional analyses that could address potential 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 should be documented (e.g., in HAWC) and considered unsuccessful if researchers do not respond to an email or phone request within 1 month of the attempt(s) to contact.

6.1.1. Evaluation Ratings

For each outcome in a study,[9] in each domain, reviewers will reach a consensus judgment of *good*, *adequate*, *deficient*, or *critically deficient*. 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. While study design features specific to the usability of the study for dose-response analysis are useful to inform those later decisions and can be noted, they do not contribute to the study confidence classifications. These categories are applied to each evaluation domain for each study as follows:

- *Good* represents a judgment that the study was conducted appropriately in relation to the evaluation domain, and any deficiencies, if present, are minor and would not be expected to influence the study results.

- *Adequate* indicates a judgment that there are methodological limitations relating to the evaluation domain, but that those limitations are not likely to be severe or to have a notable impact on the results.

- *Deficient* denotes identified biases or deficiencies that are interpreted as likely to have had a notable impact on the results or that may prevent reliable interpretation of the study findings.

- *Not reported* indicates that the information necessary to evaluate the domain was not available in the study. Generally, this term carries the same functional interpretation as

[9]Note: "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.

deficient for the purposes of the study confidence classification. Depending on the number and severity of other limitations identified in the study, it may or may not be worth reaching out to the study authors for this information.

- *Critically deficient* reflects a judgment that the study conduct introduced a serious flaw that makes the observed effect(s) uninterpretable. Studies with a determination of *critically deficient* in an evaluation domain are considered overall *uninformative* for the health outcome.

Once the evaluation domains have been rated, the identified strengths and limitations will be considered to reach a study confidence rating of *high*, *medium*, *low*, or *uninformative* for each specific health outcome(s). This will be based on the reviewer judgments across the evaluation domains for each health outcome under consideration and will include the likely impact the noted deficiencies in bias and sensitivity, or inadequate reporting, have on the results. Different outcomes within the same study can receive different ratings. The ratings, which reflect a consensus judgment between reviewers, are defined as follows:

- *High*: A well-conducted study with 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*: A satisfactory (acceptable) study where deficiencies or concerns are noted, but the limitations are unlikely to be of a notable degree. 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*: A substandard study where deficiencies or concerns were noted, and the potential for bias or inadequate sensitivity could have a significant impact on the study results or their interpretation. Typically, *low* confidence studies would 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 11.1, Tables 11-3 and 11-4**), 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. Studies rated as *low* confidence only because of sensitivity concerns about bias towards the null would require additional consideration during evidence synthesis. Observing an effect in these studies may increase confidence, assuming the study is otherwise well-conducted (see **Chapter 9**). This is one of the reasons it is important to document the primary rationale for decisions about confidence in the final rating.

- *Uninformative*: An unacceptable study where serious flaw(s) make the study results unusable for informing hazard identification. Studies with *critically deficient* judgments in any evaluation domain will almost always be classified as *uninformative* (see explanation above). Studies with multiple *deficient* judgments across domains may also be considered *uninformative*. *Uninformative* studies will not be considered further in the synthesis and

integration of evidence for hazard identification or dose-response but may be used to highlight possible research gaps.

For both the domain and overall study judgments, it is important to note that the designations are, by their nature, a categorization of what is essentially a continuous measure; thus, there is variation in quality and sensitivity within each level.

After the initial evaluation of the studies by level of overall confidence, the next stage is to examine each group (confidence level) of studies. In this stage, the reviewer rereads the studies and asks:

- 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?

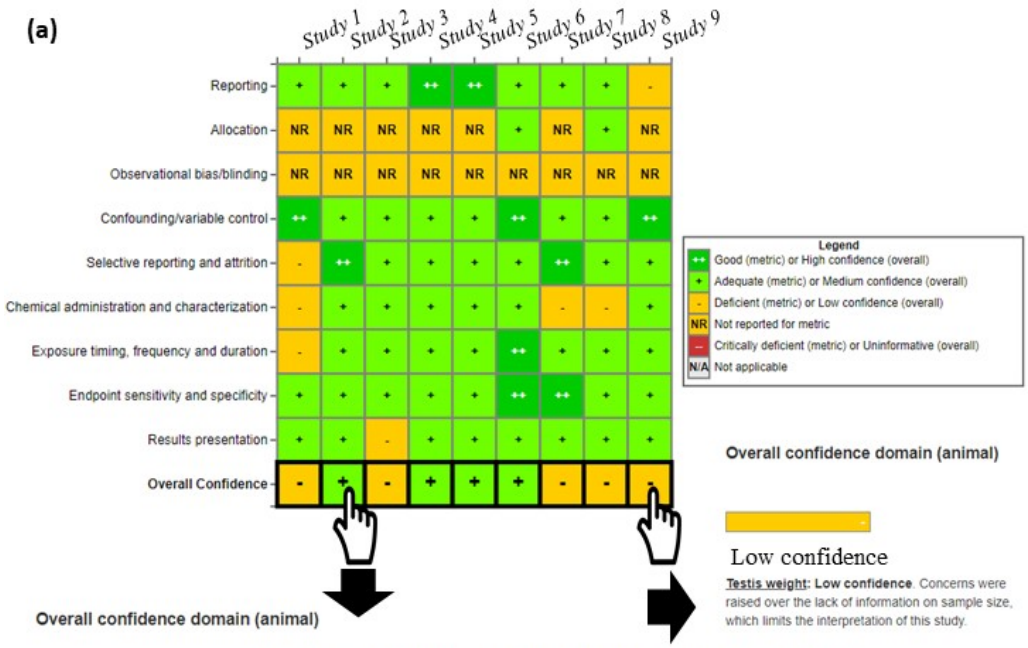
6.1.2. Documentation of Study Evaluations

Study evaluation determinations reached by each reviewer and the consensus judgment between reviewers are recorded in the EPA's version of Health Assessment Workspace Collaborative (HAWC) (https://hawcprd.epa.gov/) or documented in another format. Tutorials for using HAWC for study evaluation are available at https://hawcprd.epa.gov/about/ (Note: the tutorials are not IRIS specific). The final study evaluations reflect the consensus review and are housed in HAWC. They are made available when the draft is publicly released. There are several options for displaying study evaluation results in the assessment, using visualizations created automatically in HAWC or developed manually (see **Figures 6-2 and 6-3**). Note: All HAWC visualizations have "click to see more" functionality, where the user can click a domain to see the rationale—see **Figure 6-2 (c) and (d)**. The last two examples do not currently exist in HAWC and need to be created manually, so they do not have that functionality. The study confidence classifications and their rationales will be carried forward and considered as part of evidence synthesis (see **Chapter 9**), to aid in the interpretation of results across studies.

Figure 6-2. 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 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).

(a)



(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 low 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[a]				Developmental[a]		
				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 6-3. 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 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/. For clarification on how the overall confidence ratings are reached, see **Section 6.1.1**.

6.2. EVALUATION OF EPIDEMIOLOGY STUDIES

The principles and framework used for the evaluation of epidemiology 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 evaluation domains for IRIS's adapted approach are exposure measurement, outcome ascertainment, participant selection, confounding, analysis, study sensitivity, and selective reporting. For each domain, "core," "prompting," and follow-up questions are provided below, and are used to guide the development of assessment specific considerations. Reporting quality and risk of bias are considered during the evaluation of each domain, and the rating may be lowered when information needed to evaluate a domain is not available.

6.2.1. 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 as well as 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 (Sterne et al., 2016). The process used to develop these specific considerations will involve research into the issues identified in the set of studies; consultation with additional subject area experts may 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:

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

The evaluation strategy should define an "ideal" design (i.e., a study design with no risk of bias and high sensitivity) for the review question. This will be 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*) should be 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. The considerations should be pilot tested on three-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 into 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 6-2**.

A concern is how well the exposure measure represents the exposure in an etiologically relevant time window. IRIS does not make this evaluation strictly based on the 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 that 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 (i.e., 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 will be documented in the protocol.

Table 6-2. Example question specification for evaluation of exposure measurement in epidemiology studies

Domain and core question	Prompting questions	Follow-up questions	Considerations that apply to most exposures and outcomes
Exposure measurement 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?	For all: • 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)?	Is the degree of exposure misclassification likely to vary by exposure level? If the correlation between exposure measurements is *moderate*, is there an adequate statistical approach to ameliorate variability in measurements? 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)?	These considerations require customization to the exposure and outcome (relevant timing of exposure) **Good** • Valid exposure assessment methods used, which represent the etiologically relevant time period of interest. • Exposure misclassification is expected to be minimal. **Adequate** • Valid exposure assessment methods used, which represent the etiologically relevant time period of interest. • Exposure misclassification may 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
Exposure measurement 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)	For case-control studies of occupational exposures: • Is exposure based on a comprehensive job history describing tasks, setting, time period, and use of specific materials? For biomarkers of exposure and other analytic measures of exposure: • Is a standard assay used? Is the measure valid and precise? What are the intra- and interassay coefficients of variation? Is the assay likely to be 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?	Is the degree of exposure misclassification likely to vary by exposure level? If the correlation between exposure measurements is *moderate*, is there an adequate statistical approach to ameliorate variability in measurements? 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)	**Deficient** • Valid exposure assessment methods used, which represent the etiologically relevant time period of interest. Specific knowledge about the exposure and outcome raise 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. **Critically deficient** • 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 6-3**.

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 may vary in different populations or in 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 6-3. Example question specification for evaluation of outcome in epidemiology studies

Domain and core question	Prompting questions	Follow-up questions	Considerations that apply to most exposures and outcomes
<u>Outcome ascertainment</u> Does the outcome measure reliably distinguish the presence or absence (or degree of severity) of the outcome?	For all: • Is outcome ascertainment likely to be affected by knowledge of, or presence of, exposure (e.g., consider access to health care, if based on self-reported history of diagnosis)? For case-control studies: • 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? For mortality measures: • 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? For diagnosis of disease measures: • 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? For laboratory-based measures (e.g., hormone levels): • 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?	Is there a concern that any outcome misclassification is nondifferential, differential, or both? What is the predicted direction or distortion of the bias on the effect estimate (if there is enough information)?	These considerations require customization to the outcome **Good** • 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. **Adequate** • *Moderate* 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. **Deficient** • Outcome definition was not specific or sensitive. • Uncertainty regarding validity of assessment instrument. **Critically deficient** • Invalid/insensitive marker of outcome. • Outcome ascertainment is very likely to be affected by knowledge of, or presence of, exposure. Note: Lack of blinding should not be automatically construed to be *critically deficient*.

Participant Selection

This domain concerns the process through which participants are selected for (or leave) 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 6-4**.

In occupational cohort studies, the selection into the workforce (or into specific jobs within a work setting) may 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 and asthma), and to a lesser degree, cancer. Similarly, the decision to stay in a job or at a worksite may 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 may 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 may be jointly related to exposure and to disease. Directed acyclic graphs may 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, and time 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) may also be associated with likelihood to participate. A low participation rate of either or both groups does not in itself 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 6-4. Example question specification for evaluation of participant selection in epidemiology studies

Domain and core question	Prompting questions	Follow-up questions	Considerations that apply to most exposures and outcomes
Participant selection Is there evidence that selection into or out of the study (or analysis sample) was jointly related to exposure and to outcome?	For longitudinal cohort: • Did participants volunteer for the cohort based on knowledge of exposure and/or preclinical disease symptoms? Was entry into the cohort or continuation in the cohort related to exposure and outcome? For occupational cohort: • 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? • Could exposure produce symptoms that would result in a change in work assignment/work status ("healthy worker survivor effect")? For case-control study: • 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?	Were differences in participant enrollment and follow-up evaluated to assess bias? 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)? Were appropriate analyses performed to address changing exposures over time in relation to symptoms? Is there a comparison of participants and nonparticipants to address whether differential selection is likely?	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). **Good** • Minimal concern for selection bias based on description of recruitment process (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
Participant selection Is there evidence that selection into or out of the study (or analysis sample) was jointly related to exposure and to outcome? (continued)	For population based-survey: • Was recruitment based on advertisement to people with knowledge of exposure, outcome, and hypothesis?	Were differences in participant enrollment and follow-up evaluated to assess bias? 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)? Were appropriate analyses performed to address changing exposures over time in relation to symptoms? Is there a comparison of participants and nonparticipants to address whether differential selection is likely? (continued)	**Adequate** • 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. **Deficient** • Little information on recruitment process, selection strategy, sampling framework and/or participation OR aspects of these processes raises the potential for bias (e.g., healthy worker effect, survivor bias). **Critically deficient** • 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 it is that 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 based on 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, then 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 may 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 6-5**.

Table 6-5. Example question specification for evaluation of confounding in epidemiology studies

Domain and core question	Prompting questions	Follow-up questions	Considerations that apply to most exposures and outcomes
Confounding Is confounding of the effect of the exposure likely?	Is confounding adequately addressed by considerations in: • 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? 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)?	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)?	These considerations require customization to the exposure and outcome, but this may be limited to identifying key covariates. **Good** • Conveys strategy for identifying key confounders. This may 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 that are likely to be influential colliders or intermediates on the causal pathway. • Key confounders are evaluated appropriately and considered to be unlikely sources of substantial confounding. This often will include: o Presenting the distribution of potential confounders by levels of the exposure of interest and/or the outcomes of interest (with amount of missing data noted); o Consideration that potential confounders were rare among the study population, or were expected to be poorly correlated with exposure of interest; o Consideration of the most relevant functional forms of potential confounders; o Examination of the potential impact of measurement error or missing data on confounder adjustment; o 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
Confounding Is confounding of the effect of the exposure likely? (continued)	Is confounding adequately addressed by considerations in: • 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? 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)	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)	**Adequate** Similar to *good* but may not have included all key confounders, or less detail may be available on the evaluation of confounders (e.g., sub-bullets in *good*). It is possible that residual confounding could explain part of the observed effect, but concern is minimal. **Deficient** • Does not include variables in the models that are likely to be influential colliders or intermediates on the causal pathway. And any of the following: • 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]). **Critically deficient** • Includes variables in the models that are colliders and/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 of the 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, 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 may 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, and xylene in traffic emissions) due to confounding by these coexposures. While it may 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), then 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 may 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

This domain concerns the potential for distortion of results that can occur from inadequate or inappropriate statistical analysis. The core, prompting, and follow-up questions are provided in **Table 6-6**.

Table 6-6. Example question specification for evaluation of analysis in epidemiology studies

Domain and core question	Prompting questions	Follow-up questions	Considerations that apply to most exposures and outcomes
Analysis Does the analysis strategy and presentation convey the necessary familiarity with the data and assumptions?	• 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?	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)?	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). **Good** • 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 *p*-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 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 based only on a priori rationale with sufficient numbers. • No deficiencies in analysis evident. Discussion of some details may be absent (e.g., examination of outliers).

Domain and core question	Prompting questions	Follow-up questions	Considerations that apply to most exposures and outcomes
Analysis Does the analysis strategy and presentation convey the necessary familiarity with the data and assumptions? (continued)	• Does the study include additional analyses addressing potential biases or limitations (i.e., sensitivity analyses)?	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)	**Adequate** Same as *good*, except: • Descriptive information about exposure provided (where applicable) but may be incomplete; might not have discussed missing data, cut-points, or shape of distribution. • Includes analyses that address robustness of findings (examples in *good*), but some important analyses are not performed. **Deficient** • Does not conduct analysis using optimal characterization of the outcome variable. • Descriptive information about exposure levels not provided (where applicable). • Effect estimate and *p*-value presented, without standard error or confidence interval. • Results presented as statistically "significant"/"not significant." **Critically deficient** • Results of analyses of effect modification examined without clear a priori rationale and without providing main/principal effects (e.g., presentation only of statistically significant interactions that were not hypothesis driven). • Analysis methods are not appropriate for design or data of the study.

LOD = limit of detection.

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.

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 may have fewer than four levels of rating. The core and prompting questions are presented in **Table 6-7**.

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 (i.e., 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 6-7. Example question specification for evaluation of selective reporting in epidemiology studies

Domain and core question	Prompting questions	Follow-up questions	Considerations that apply to most exposures and outcomes
Selective reporting Is there reason to be concerned about selective reporting?	• 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?	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)?	These considerations generally do not require customization and may have fewer than four levels. **Good** • The results reported by study authors are consistent with the primary and secondary analyses described in a registered protocol or methods paper. **Adequate** • The authors described their primary (and secondary) analyses in the methods section and results were reported for all primary analyses. **Deficient** • 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 effect estimate as support for "no effect."

Some of the study features that can affect study sensitivity may 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 may not have been addressed and, therefore, should be included here. Examples include the exposure range (e.g., the contrast between the low- and high-exposure groups within a study), level or duration of exposure, and length of follow-up. In some cases (e.g., for very rare outcomes), sample size or number of observed cases may also be considered within this "sensitivity" domain. Although imprecision of estimates in some cases can be addressed through consideration of confidence intervals (CIs) or through calculation of a summary estimate from multiple studies, studies with no observed events present methodological challenges, particularly with respect to inclusion in meta-analyses. The age group under study may also be relevant within the context of study sensitivity, as the appropriate age group will depend on the outcome being examined; a population may be too young or too old to provide a meaningful analysis of the effect of interest. Information relevant to the evaluation of study sensitivity measures includes, but is not limited to, the exposure range spanned in the study, ages of participants (e.g., not too young in studies of pubertal development), length of follow-up (for outcomes with long latency periods), and choice of referent group and the level of exposure contrast between groups (i.e., the extent to which the "unexposed group" is truly unexposed, and the prevalence of exposure in the group designated as "exposed").

The core and prompting questions for this domain are presented in **Table 6-8**. 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 6-8. Example question specification for evaluation of sensitivity in epidemiology studies

Domain and core question	Prompting questions	Follow-up questions	Considerations that apply to most exposures and outcomes
Sensitivity Is there a concern that sensitivity of the study is not adequate to detect an effect?	• Is the exposure range adequate? • 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 related to risk of bias or otherwise that raise concerns about sensitivity?		These considerations may require customization to the exposure and outcome and may have fewer than four levels. Some study features that affect study sensitivity may have already been included in the other evaluation domains. Other features that have not been addressed should be included here. Some examples include: **Adequate** • The range of exposure levels provides adequate variability to evaluate primary hypotheses in study. • The population was exposed to levels expected to have an impact on response. • 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 adequately powered to observe an effect. • No other concerns raised regarding study sensitivity. **Deficient** • Concerns were raised about the issues described for *good* that are expected to notably decrease the sensitivity of the study to detect associations for the outcome.

6.2.2. Final Observations

As described in **Section 6.1**, 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 6.1.2**.

It is important to note that the confidence in the study may 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 may apply only to one outcome or one analysis of a study. Note that with a few exceptions, this 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 9**. Review of some of the results may 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).

6.3. EVALUATION OF EXPERIMENTAL ANIMAL TOXICOLOGY STUDIES

The approach to evaluating animal studies focuses on assessing aspects of study design and experimental conduct through the lens of three broad evaluation concerns: reporting quality, risk of bias, also referred to as internal validity, and study sensitivity. As part of study evaluation, IRIS first considers whether the study has reported sufficient details to conduct a RoB and sensitivity analysis. Studies that do not report basic information such as species are typically considered *uninformative*. The principles used to assess RoB (i.e., allocation, observational bias, confounding, selective reporting, attrition) are conceptually similar to those applied to randomized human clinical trials (Krauth et al., 2013; Higgins and Green, 2011b), but have been tailored for application to experimental animal studies. 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; (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. Some tools consider sensitivity in the RoB metrics (e.g., OHAT, Navigation Guide, SciRAP), but the IRIS approach considers it as a separate domain to better distinguish sensitivity considerations from RoB as commonly applied in systematic review.

The IRIS approach is organized around domains, which are issues or topics related to one of the evaluation concerns. As described in **Section 6.1**, each domain is associated with questions and considerations that guide the reviewer in judging the quality and informativeness of individual studies. The narrow set of domains employed in the current approach focuses the evaluation on the main issues related to quality and insensitivity that often arise in experimental animal studies used in IRIS human health assessments.

6.3.1. Development of Evaluation Considerations

An initial stage in the analysis of the animal studies is the development of evaluation considerations for the domains presented in **Table 6-9**. These considerations are used to describe the different levels of quality and informativeness from *good* to *critically deficient*, as defined in **Section 6.1.1**. The purpose of the evaluation considerations is to:

1) Specify attributes of the study that would impact your confidence in the study results, and

2) Provide a guide for evaluating each endpoint/outcome of interest that can be followed by others.

3) Ensure consistency across studies and reviewers.

The general considerations in **Table 6-9** are worded broadly and are not specific to any one endpoint/outcome or chemical and should serve as a starting point for developing the specific evaluation considerations. Assessment teams will consult with subject matter experts (e.g., IRIS Disciplinary Groups) to develop specific evaluation considerations based on needs of the assessment. Some domain considerations will need to be tailored to the chemical and endpoint/outcome while others are generalizable across assessments (e.g., considerations for reporting quality). Developing specific considerations requires a familiarity 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. Knowledge of issues related to the hazards and endpoints/outcomes (or groupings of endpoints/outcomes) identified in the revised evaluation plan to be assessed is also important to developing the specific evaluation considerations. Additionally, familiarity with issues regarding the chemical and exposure route is helpful.

Table 6-9. Domains, questions, and general considerations to guide the evaluation of animal studies

Evaluation concern	Domain—core question	Prompting questions	General considerations
Reporting quality	**Reporting quality** Does the study report information for evaluating the design and conduct of the study for the endpoint(s)/outcome(s) of interest? *Notes:* *Reviewers should reach out to authors to obtain missing information when studies are considered key for hazard evaluation and/or dose-response.* • This domain is limited to reporting. Other aspects of the exposure methods, experimental design, and endpoint evaluation methods are evaluated using the domains related to risk of bias and study sensitivity.	Does the study report the following? **Critical information** necessary to perform study evaluation: • Species, test article name, levels and duration of exposure, route (e.g., oral; inhalation), qualitative or quantitative results for at least one endpoint of interest. **Important information** for evaluating the study methods: • Test animal: strain, sex, source, and general husbandry procedures. • Exposure methods: source, purity, method of administration. • Experimental design: frequency of exposure, animal age and lifestage during exposure and at endpoint/outcome evaluation. • Endpoint evaluation methods: assays or procedures used to measure the endpoints/outcomes of interest.	• These considerations typically do not need to be refined by assessment teams, although in some instances the **important information** may be refined depending on the endpoints/outcomes of interest or the chemical under investigation. • A judgment and rationale for this domain should be given for the study. Typically, these will not change regardless of the endpoints/outcomes investigated by the study. In the rationale, reviewers should indicate whether the study adhered to GLP, OECD, or other testing guidelines. o **Good**: All **critical and important information** is reported or inferable for the endpoints/outcomes of interest. o **Adequate**: All **critical information** is reported but some **important information** is missing. However, the missing information is not expected to significantly impact the study evaluation. o **Deficient**: All **critical information** is reported but **important information** is missing that is expected to significantly reduce the ability to evaluate the study. o **Critically deficient**: Study report is missing any pieces of **critical information**. Studies that are *critically deficient* for reporting are *uninformative* for the overall rating and not considered further for evidence synthesis and integration.

Evaluation concern		Domain—core question	Prompting questions	General considerations
Risk of bias	**Selection and performance bias**	**Allocation** Were animals assigned to experimental groups using a method that minimizes selection bias?	For each study: • Did each animal or litter have an equal chance of being assigned to any experimental group (i.e., random allocation)?[a] • Is the allocation method described? • Aside from randomization, were any steps taken to balance variables across experimental groups during allocation?	These considerations typically do not need to be refined by assessment teams. A judgment and rationale for this domain should be given for each cohort or experiment in the study. • **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 *adequate*]). • **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). • **Not reported** (interpreted as *deficient*): No indication of randomization of groups or other methods (e.g., normalization) to control for important modifying factors across experimental groups. • **Critically deficient**: Bias in the animal allocations was reported or inferable.

Evaluation concern		Domain—core question	Prompting questions	General considerations
Risk of bias (continued)	**Selection and performance bias (continued)**	**Observational bias/blinding** Did the study implement measures to reduce observational bias?	For each endpoint/outcome or grouping of endpoints/outcomes in a study: • Does the study report blinding or other methods/procedures for reducing observational bias? • If not, did the study use a design or approach for which such procedures can be inferred? • What is the expected impact of failure to implement (or report implementation) of these methods/procedures on results?	These considerations typically do not need to 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 may strongly influence results prior to performing evaluations.) A judgment and rationale for this domain should be given for each endpoint/outcome or group of endpoints/outcomes investigated in the study. • **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] • **Adequate**: Methods for reducing observational bias (e.g., blinding) can be inferred or were reported but described incompletely. • **Not reported**: Measures to reduce observational bias were not described. o (Interpreted as *adequate*) The potential concern for bias was mitigated based on use of automated/computer driven systems, standard laboratory kits, relatively simple, objective measures (e.g., body or tissue weight), or screening-level evaluations of histopathology. o (Interpreted as *deficient*) The potential impact on the results is major (e.g., outcome measures are highly subjective). • **Critically deficient**: Strong evidence for observational bias that impacted the results.

Evaluation concern		Domain—core question	Prompting questions	General considerations
Risk of bias (continued)	**Confounding/variable control**	**Confounding** Are variables with the potential to confound or modify results controlled for and consistent across all experimental groups?	For each study: • Are there differences across the treatment groups (e.g., coexposures, vehicle, diet, palatability, husbandry, health status) that could bias the results? • If differences are identified, to what extent are they expected to impact the results?	These considerations may need to be refined by assessment teams, as the specific variables of concern can vary by experiment or chemical. 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. • **Good**: Outside of the exposure of interest, variables that are likely to confound or modify results appear to be controlled for and consistent across experimental groups. • **Adequate**: Some concern that variables that were likely to confound or modify results were uncontrolled or inconsistent across groups but are expected to have a minimal impact on the results. • **Deficient**: Notable concern that potentially confounding variables were uncontrolled or inconsistent across groups and are expected to substantially impact the results. • **Critically deficient**: Confounding variables were presumed to be uncontrolled or inconsistent across groups and are expected to be a primary driver of the results.

Evaluation concern		Domain—core question	Prompting questions	General considerations
Risk of bias (continued)	**Selective reporting and attrition bias**	**Selective reporting and attrition** Did the study report results for all prespecified outcomes and tested animals? *Note:* *This domain does not consider the appropriateness of the analysis/results presentation. This aspect of study quality is evaluated in another domain.*	For each study: Selective reporting bias: • Are all results presented for endpoints/outcomes described in the methods (see note)? Attrition bias: • Are all animals accounted for in the results? • If there are discrepancies, do authors provide an explanation (e.g., death or unscheduled sacrifice during the study)? • If unexplained results omissions and/or attrition are identified, what is the expected impact on the interpretation of the results?	These considerations typically do not need to be refined by assessment teams. A judgment and rationale for this domain should be given for each cohort or experiment in the study. • **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 is available from supplemental material. If results omissions or animal attrition are identified, the authors provide an explanation, and these are not expected to impact the interpretation of the results. • **Adequate**: Quantitative or qualitative results are reported for most prespecified outcomes (explicitly stated or inferred), exposure groups and evaluation time points. Omissions and/or attrition are not explained but are not expected to significantly impact the interpretation of the results. • **Deficient**: Quantitative or qualitative results are missing for many prespecified outcomes (explicitly stated or inferred), exposure groups and evaluation time points and/or high animal attrition; omissions and/or attrition are not explained and may significantly impact the interpretation of the results. • **Critically deficient**: Extensive results omission and/or animal attrition are identified and prevents comparisons of results across treatment groups.

Evaluation concern		Domain—core question	Prompting questions	General considerations
Sensitivity	**Exposure methods sensitivity**	**Chemical administration and characterization** Did the study adequately characterize exposure to the chemical of interest and the exposure administration methods? *Note:* *Consideration of the appropriateness of the route of exposure is not evaluated at the individual study level. Relevance and utility of the routes of exposure are considered in the PECO criteria for study inclusion and during evidence synthesis.*	For each study: • Are there concerns [specific to this chemical] regarding the source and purity and/or composition (e.g., identity and percent distribution of different isomers) of the chemical? If so, can the purity and/or composition be obtained from the supplier (e.g., as reported on the website)? • Was independent analytical verification of the test article purity and composition performed? • Did the authors take steps to ensure the reported exposure levels were accurate? • Are there concerns about the methods used to administer the chemical (e.g., inhalation chamber type, gavage volume)? For inhalation studies: • Were target concentrations confirmed using reliable analytical measurements in chamber air?	It is essential that these considerations are considered, and potentially refined, by assessment teams, as the specific variables of concern can vary by chemical (e.g., stability may be an issue for one chemical but not another). A judgment and rationale for this domain should be given for each cohort or experiment in the study. • **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. For inhalation studies, chemical concentrations in the exposure chambers are verified using reliable analytical methods. • **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; For inhalation studies, actual exposure concentrations are missing or verified with less reliable methods).

Evaluation concern		Domain—core question	Prompting questions	General considerations
Sensitivity (continued)	**Exposure methods sensitivity (continued)**	**Chemical administration and characterization** (continued)	For oral studies: • If necessary based on consideration of chemical specific-knowledge (e.g., instability in solution; volatility) and/or exposure design (e.g., the frequency and duration of exposure), were chemical concentrations in the dosing solutions or diet analytically confirmed?	• **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; levels of impurities are substantial or concerning; *deficient* administration methods, such as use of static inhalation chambers or a gavage volume considered too large for the species and/or lifestage at exposure). • **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).

Evaluation concern		Domain—core question	Prompting questions	General considerations
Sensitivity (continued)	**Exposure methods sensitivity (continued)**	**Exposure timing, frequency and duration** Was the timing, frequency, and duration of exposure sensitive for the endpoint(s)/outcome(s) of interest?	For each endpoint/outcome or grouping of endpoints/outcomes in a study: • Does the exposure period include the critical window of sensitivity? • Was the duration and frequency of exposure sensitive for detecting the endpoint of interest?	Considerations for this domain are highly variable depending on the endpoint(s)/outcome(s) of interest and must be refined by assessment teams. A judgment and rationale for this domain should be given for each endpoint/outcome or group of endpoints/outcomes investigated in the study. • **Good**: The duration and frequency of the exposure was sensitive, and the exposure included the critical window of sensitivity (if known). • **Adequate**: The duration and frequency of the exposure was sensitive, and the exposure covered most of the critical window of sensitivity (if known). • **Deficient**: The duration and/or frequency of the exposure is not sensitive and did not include most of the critical window of sensitivity (if known). These limitations are expected to bias the results towards the null. • **Critically deficient**: The exposure design was not sensitive and is expected to strongly bias the results towards the null. The rationale should indicate the specific concern(s).

Evaluation concern		Domain—core question	Prompting questions	General considerations
Sensitivity (continued)	**Outcome measures and results display**	**Endpoint sensitivity and specificity** Are the procedures sensitive and specific for evaluating the endpoint(s)/outcome(s) of interest? *Note:* • Sample size alone is not a reason to conclude an individual study is *critically deficient*. • Considerations related to adjustments/corrections to endpoint measurements (e.g., organ weight corrected for body weight) are addressed under results presentation.	For each endpoint/outcome or grouping of endpoints/outcomes in a study: • Are there concerns regarding the sensitivity, specificity, and/or validity of the protocols? • Are there serious concerns regarding the sample size? • Are there concerns regarding the timing of the endpoint assessment?	Considerations for this domain are highly variable depending on the endpoint(s)/outcome(s) of interest and must be refined by assessment teams. A judgment and rationale for this domain should be given for each endpoint/outcome or group of endpoints/outcomes investigated in the study. Examples of potential concerns include: • Selection of protocols that are insensitive or nonspecific for the endpoint of interest. • Evaluations did not include all treatment groups (e.g., only control and high dose). • Use of unreliable methods to assess the outcome. • Assessment of endpoints at inappropriate or insensitive ages, or without addressing known endpoint variation (e.g., due to circadian rhythms, estrous cyclicity). • Decreased specificity or sensitivity of the response due to the timing of endpoint evaluation, as compared to exposure (e.g., short acting depressant or irritant effects of chemicals; insensitivity due to prolonged period of nonexposure prior to testing).

Evaluation concern		Domain—core question	Prompting questions	General considerations
Sensitivity (continued)	**Outcome measures and results display (continued)**	**Results presentation** Are the results presented in a way that makes the data usable and transparent?	For each endpoint/outcome or grouping of endpoints/outcomes in a study: • Does the level of detail allow for an informed interpretation of the results? • Are the data analyzed, compared, or presented in a way that is inappropriate or misleading?	Considerations for this domain are highly variable depending on the outcomes of interest and must be refined by assessment teams. A judgment and rationale for this domain should be given for each endpoint/outcome or group of endpoints/outcomes investigated in the study. Examples of potential concerns include: • Nonpreferred presentation (e.g., developmental toxicity data averaged across pups in a treatment group, when litter responses are more appropriate; presentation of absolute organ-weight data when relative weights are more appropriate). • Failing to present quantitative results either in tables or figures. • Pooling data when responses are known or expected to differ substantially (e.g., across sexes or ages). • Failing to report on or address overt toxicity when exposure levels are known or expected to be highly toxic. • Lack of full presentation of the data (e.g., presentation of mean without variance data; concurrent control data are not presented).

Evaluation concern	Domain—core question	Prompting questions	General considerations
Overall confidence	**Overall confidence** Considering the identified strengths and limitations, what is the overall confidence rating for the endpoint(s)/outcome(s) of interest? *Note:* *Reviewers should mark studies that are rated lower than high confidence only due to low sensitivity (i.e., bias towards the null) for additional consideration during evidence synthesis. If the study is otherwise well conducted and an effect is observed, the confidence may be increased.*	For each endpoint/outcome or grouping of endpoints/outcomes in a study: • Were concerns (i.e., limitations or uncertainties) related to the reporting quality, 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?	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. 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 6.1.1**).

OECD = Organisation for Economic Co-operation and Development.

[a]Several 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).

[b]For 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 may 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).

6.3.2. Final Observations

As described in **Section 6.1**, once the specific considerations have been developed and pilot tested, 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*). Documentation of the ratings and the rationale behind their selection are essential to providing support and transparency for the reviewers' decision process. The study evaluation results are documented as described in **Section 6.1.2**. Finally, studies testing exposure levels that exceed the maximum tolerated dose [for example, see discussions on this topic in (U.S. EPA, 2005b)] are not excluded from the analysis described in **Table 6-9**, as such characteristics are considered during evidence synthesis.

6.4. EVALUATION OF CONTROLLED HUMAN EXPOSURE STUDIES

Controlled human exposure studies involve human subjects to test specific hypotheses about short-term exposures and biologic responses that inform potential mechanisms and understanding of exposure-response patterns. The exposures are generated in the laboratory to achieve predetermined concentrations for a period of minutes to hours. For study evaluation, a process incorporating aspects of the approaches used for epidemiology studies and experimental animal studies, as well as the ROBINS-I tool discussed in **Section 6.2** (Sterne et al., 2016), should be used to evaluate controlled exposure studies in humans. Reviewers should confirm that the authors included an explicit declaration that the study protocol was approved by an institutional review board. 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, inclusion of a clean air control exposure (if applicable), study sensitivity, and other aspects of the exposure protocol. Sample size and the process of recruitment, selection of study subjects, and differences in characteristics between groups should be considered as reflecting potential differences in sensitivity.

6.5. EVALUATION OF EXISTING COMPUTATIONAL PHYSIOLOGICALLY BASED PHARMACOKINETIC/PHARMACOKINETIC MODELS

For a specific target organ/tissue, it may be possible to employ or adapt an existing physiologically based pharmacokinetic (PBPK) model or develop a new PBPK model or an alternate quantitatively valuable approach for a PBPK model (e.g., a classical pharmacokinetic [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 may come from studies with animals or humans and may be in vitro or in vivo in design. It should be recognized that chemicals produce multiple toxicities, through different modes of

action (MOAs), which may vary by lifestage (U.S. EPA, 2006b), and with different dose-response functions. If data are available from studies evaluating susceptible lifestages (i.e., in utero/pregnant women, lactating women, growing child, adolescent), it should be considered as part of a PBPK model that reflects the potential absorption, distribution, metabolism, and excretion (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, 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 the EPA's *A Review of the Reference Dose and Reference Concentration Processes* (U.S. EPA, 2002b) 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, and

- 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 [human equivalent dose (HED)] or inhalation concentration [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 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, 2018b). 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 the Quality Assurance Project Plan for PBPK models (U.S. EPA, 2018b) and in the protocol template.

If no PBPK model exists or the existing PBPK models are determined to be 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, as well as 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. In addition to providing transparency through documentation, the process will confirm objectivity and scientific rigor.

6.6. EVALUATION OF INFORMATION RELEVANT TO MECHANISMS OF TOXICITY

As mentioned in **Chapter 4**, the initial literature screening will identify sets of other potentially informative studies, including mechanistic studies, as "potentially relevant supplemental material," and not as a component of the PECO, which identifies studies presenting apical health effects that will all be evaluated for reporting quality, risk of bias, and sensitivity. This is because, despite the early identification of existing mode of action (MOA) hypotheses during problem formulation, there still may be an incomplete understanding of the often staggeringly complex biological pathways involved in the toxic response to a chemical. For many chemicals, in vitro studies alone can outnumber human or animal health effect studies by orders of magnitude. In addition, because mechanistic studies possess a wide range of applicability to an assessment (e.g., they can suggest potential health effects that have not been examined in other study types, support findings of apical health effects, help to explain heterogeneous results across sets of studies, inform susceptibility, and inform the relevance of effects observed in animals to humans), the questions and analyses applied to mechanistic studies will differ depending on the

requirements for each assessment, requiring a multifaceted approach. To undergo a full reporting quality, risk of bias, and sensitivity evaluation of every identified study that may report mechanistic information before the relevant toxicity pathways have been identified or the needs of the assessment are better understood would not be an effective use of time. Therefore, individual study level evaluation of mechanistic endpoints will typically only be pursued when the interpretation of studies is likely to significantly impact hazard conclusions or assumptions about dose-response analysis (see **Section 4.3.3.** and **Chapter 10** for more information).

After the individual mechanistic studies have been identified and organized into sortable inventories, the specific analytical approach for the consideration of information from mechanistic studies (primarily in vitro, but also includes human and animal studies in vivo and ex vivo, and as well as in silico methods) is targeted to the assessment needs depending on the extent and nature of the human and animal evidence. In this way, the mechanistic synthesis might range from a high-level summary of potential mechanisms of action to specific, focused questions needed to address critical assessment issues (e.g., shape of the dose-response curve in the low dose region, applicability of the animal evidence to humans, addressing susceptible populations). The approach is intentionally flexible to allow for application to varied evidence bases and to accommodate the anticipated increased reliance on emerging technologies and methods, including new approach methodologies (NAMs), in the future. Regardless of the approach (see **Section 10.2.1**), the steps taken for the selective evaluation of mechanistic studies should be transparently described.

Similar to the evaluation of epidemiological and animal evidence, study evaluation considerations for individual mechanistic studies will differ depending on the type of endpoints, study designs, and model systems or populations evaluated. For human and animal studies reporting mechanistic endpoints, the same study evaluation considerations outlined in **Sections 6.2 and 6.3** may be used with outcome-specific criteria applied to the appropriate domains. It should be noted that because the evaluation process is outcome-specific, overall confidence classifications for human or animal studies that have already been determined will not automatically apply to mechanistic endpoints if reported in the same study; a separate evaluation of the mechanistic endpoints should be performed as the utility of a study may vary for the different outcomes reported. Developing specific considerations requires a familiarity with the studies to be evaluated and cannot be conducted in the absence of knowledge of the relevant study designs, measurements, and analytic issues. Knowledge of issues related to the hazards and the outcomes identified in the revised evaluation plan is also important for developing specific evaluation considerations. One challenge is that novel methodologies for studying mechanistic evidence are continuously being developed and implemented and often no "standard practices" exist.

For in vitro studies, the development of methods for assessing potential bias lags that of human and animal studies, though it is an active area of development in the field of systematic review. Historically, most tools used to evaluate these studies have focused on reporting quality; tools to assess risk of bias (internal validity) of mechanistic evidence are not well-established

(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 IRIS Program is in the pilot phase of testing approaches for arriving at study level judgments for in vitro studies based on the domains described for animal study evaluations described in **Section 6.3**, with modifications. This pilot approach for in vitro study evaluation is described and compared (differences between the approaches are explained in the right-hand column) with the approach for animal study evaluation in **Table 6-10**. The IRIS Program is aware of other tools and considerations for evaluating in vitro studies (Beronius et al., 2018; NASEM, 2018; OECD, 2018; U.S. EPA, 2018a) and will monitor developments through its engagements with the systematic review community. Existing tools tend to be general and designed for application to all in vitro studies, but it should be acknowledged that to be truly useful in evaluating the risk of bias, internal validity, and sensitivity of in vitro studies, additional evaluation considerations reflecting the specific model systems and assay(s) employed will likely need to be developed and applied, increasing the challenge of operationalizing a useful and practical, one-size-fits-all approach. Therefore, pilot testing will be key for refining these considerations to be useful and practical for all in vitro studies that will require evaluation.

Table 6-10. Pilot testing domains and criteria for in vitro study evaluation

Animal study evaluation domains and questions	Preliminary in vitro study considerations (to be further refined through pilot testing)
Reporting quality	**Reporting quality**
Critical information necessary for evaluation: Species; test article name; levels and duration of exposure; route (e.g., oral; inhalation); qualitative or quantitative results for at least one endpoint of interest	***Critical information*****:** *in vitro examples* Cell/tissue type(s) or test system; test material/chemical name; description of vehicle; concentration and duration of treatments; qualitative or quantitative results for at least one endpoint of interest *Differences:* description of vehicle is considered critical for in vitro studies due to potential nonspecific toxicity.
Important information: • Test animal: strain, sex, source, and general husbandry procedures • Exposure methods: source, purity, method of administration • Experimental design: frequency of exposure, animal age and lifestage during exposure and at endpoint/outcome evaluation • Endpoint evaluation methods: assays or procedures used to measure the endpoints/outcomes of interest	***Important information:*** *in vitro examples* • Test system: cell/tissue source (and verification of cell type, if demonstrated to be prone to contamination); cell passage number, cell counts or density/confluence at treatment and analysis; media composition (e.g., serum, antibiotics) and source; incubation conditions (e.g., temperature, CO_2/O_2 concentration, humidity level); measures taken to avoid contamination (e.g., mycoplasma testing). *Differences*: Some of these characteristics of the test system may be considered *critical information* for some experiments and not important for others. Specific considerations related to characterizing the test system will vary depending on the model used and will be refined through pilot testing. • Exposure and design: Purity and source of chemical and vehicle; method and timing of administration; timepoints of data collection. *Differences*: Because exposure and study design are closely linked in in vitro studies, these have been combined. • Endpoint evaluation methods: description of the endpoints measured and test assays used (sample size and replicates are considered under "outcome evaluation," paralleling what is done for in vivo studies).

Risk of bias—selection and performance	Risk of bias—observational bias/blinding
***Allocation*:** Were animals assigned to experimental groups using a method that minimizes selection bias?	***N/A*** *Differences*: "Allocation" removed for in vitro studies. Although an evaluation of allocation could be possible with a detailed plating layout, this information is rarely reported in published in vitro studies and it is unclear the extent to which this constitutes a systematic source of bias in in vitro studies. Allocation may be important to consider for more complex test systems (e.g., organotypic cultures; tissue-on-a-chip) and could potentially be considered under specificity based on the results of pilot testing.
Observational bias/blinding: Did the study implement measures to reduce observational bias? • Does the study report blinding or other methods/procedures for reducing observational bias? • If not, did the study use a design or approach for which such procedures can be inferred? • Were the assays evaluated using automated approaches that reduce concern for observational bias? • What is the expected impact of failure to implement (or report implementation) of these methods/procedures on results?	***Observational bias/blinding:*** Did the study implement measures to reduce observational bias? • Did the study take steps to minimize observational bias during analysis (e.g., blinding/coding of slides or plates for analysis; collection of data from randomly selected fields)? • If not, did the study use a design or approach for which such procedures can be inferred? • Were the assays evaluated using automated approaches (e.g., microplate readers) that reduce concern for observational bias? • What is the expected impact of failure to implement (or report implementation) of these methods/procedures on results? *Differences*: While this potential concern is considered relevant regardless of study type, based on experience many in vitro studies do not report these measures.

Risk of bias—confounding/variable control	Risk of bias—variable control and specificity
***Confounding/variable control*:** Are variables with the potential to confound or modify results controlled for and consistent across all experimental groups? • Are there differences across the treatment groups (e.g., coexposures, vehicle, diet, palatability, husbandry, health status) that could bias the results? • If differences are identified, to what extent are they expected to impact the results?	***Variable control*:** Are all introduced variables with the potential to affect the results of interest controlled for and consistent across experimental groups? • Are there concerns regarding the negative (untreated and/or vehicle) controls used? If known, do the results in the negative control groups differ significantly from expected background or historical incidence for the assay(s) of interest? • If applicable, was the assay signal normalized to account for non-biological differences across replicates and exposure groups? • Are there any known or presumed differences across treatment groups (e.g., coexposures, culture conditions, variations in reagent production lots) that could bias the results? If differences are identified, to what extent are they expected to impact the results? *Differences:* Although both can be related to confounding, given the increased heterogeneity of in vitro studies, this domain was made specific to variables under the experimenter's control and a separate domain below was added to consider features inherent to the chemical, test system or experiment that might affect results.

N/A	***Specificity*****:** Did the study address features of the chemical, test system or experiment that have the potential to affect the results for the endpoint(s) of interest independent of the effect of the test chemical on those endpoint(s)? • Did the test compound induce cytotoxicity (or were the levels used sufficient to induce cytotoxicity in related systems) to a degree that is expected to affect interpretation of results? • Are there concerns regarding the need for positive controls (e.g., concerns that the effects of interest may be inhibited or otherwise not manifest in the test system)? If one was used, was the selected positive test substance appropriate and was the intended positive response induced? If known, do the results in the positive control groups differ significantly from expected background or historical incidence? • Can the test article interfere with a given assay (e.g., auto-fluoresces or inhibits enzymatic processes necessary for assay signals)? *Differences:* It is expected that this domain will be test system specific. It will be refined through pilot testing, particularly to select the unique test system considerations most informative for judging this domain, and to reduce the potential for identifying the same issue across multiple domains (e.g., "endpoint sensitivity"). It may prove more appropriate to consider a specificity-type domain independently for the test system, chemical, and assay.
Risk of Bias—selective reporting and attrition	**Risk of bias—selective reporting**
Selective reporting**:** Did the study report results for all prespecified outcomes and tested animals? • Are all results presented for endpoints/outcomes described in the methods (does not consider the appropriateness of the analysis or results presentation)?	***Selective reporting:*** Did the study report results for all prespecified outcomes and replicates? • Are all results presented (quantitatively or qualitatively) for endpoints/outcomes described in the methods (does not consider the appropriateness of the analysis or results presentation)?

***Attrition*:** Are all animals accounted for in the results?	***N/A*** *Differences*: "Attrition" removed for in vitro studies. Generally, in vitro test methods are faster and more easily repeated than animal bioassays. Thus, loss of individual cells or tissues for nonspecific reasons during these short study durations is not a major concern and is largely addressed in other domains (e.g., specificity).
Sensitivity—exposure methods	**Sensitivity—exposure methods**
***Chemical characterization and administration*:** Did the study adequately characterize exposure to the chemical of interest and the exposure administration methods? • Does the study report the source and purity and/or composition (e.g., identity and percent distribution of different isomers) of the chemical? If not, can the purity and/or composition be obtained from the supplier (e.g., as reported on the website)? • Was independent analytical verification of the test article purity and composition performed? • Did the authors take steps to ensure the reported exposure levels were accurate? • For inhalation studies: were target concentrations confirmed using reliable analytical measurements in chamber air? • For oral studies: if necessary, based on consideration of chemical-specific knowledge (e.g., instability in solution; volatility) and/or exposure design (e.g., the frequency and duration of exposure), were chemical concentrations in the dosing solutions or diet analytically confirmed?	***Chemical characterization and administration*:** Did the study adequately characterize exposure to the chemical of interest and the exposure administration methods? • Are there concerns (specific to the chemical) regarding the purity and/or composition (e.g., identity and percent distribution of different isomers) of the test material/chemical? If so, can the purity and/or composition be obtained from the supplier (e.g., as reported on the website)? • Was independent analytical verification of the test article purity and composition performed? • Are there concerns about the stability of the test chemical in the vehicle and/or culture media (e.g., pH, solubility, volatility, adhesion to plastics) that were not corrected for (e.g., observed precipitate formation, enclosed chambers not used for testing volatile chemicals)? • Are there concerns about the preparation or storage conditions of the test substance?

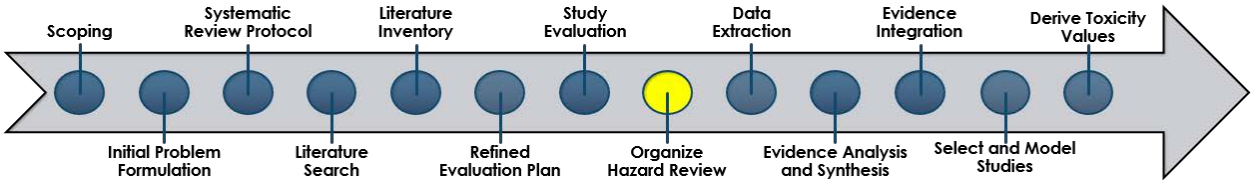
***Exposure timing, frequency, and duration*:** Was the timing, frequency, and duration of exposure sensitive for the endpoint(s)/outcome(s) of interest? • Does the exposure period include the critical window of sensitivity? • Was the duration and frequency of exposure sensitive for detecting the endpoint of interest?	***Exposure timing, frequency, and duration*:** Was the timing, frequency, and duration of exposure sensitive for the assay/model? Considerations will vary depending on the specific assay/model being used, but may include the following: • Were steps taken to determine the appropriate concentration range of the test article in the test system? Are there concerns that the amount of test article administered may not have reached a sufficient concentration to induce an effect? • Was the exposure duration sufficient to cause a measurable impact on the endpoint of interest (in the absence of a positive control)? • Was the doubling time considered in the frequency of dosing, timing of culture, or duration in culture at treatment? • Was the confluency at treatment appropriate? Are there concerns that the cells were quiescent/senescent, or growth inhibited due to confluence? *Differences*: Reworded to apply to unique aspects of cell/tissue cultures, where a "critical window of sensitivity" may more appropriately translate to, for example, a consideration of confluency and doubling times.

Sensitivity—outcome measures and results display	Sensitivity—outcome measures, results display, and analysis
***Endpoint sensitivity and specificity*:** Are the procedures sensitive and specific for evaluating the endpoint(s)/outcome(s) of interest? • Are there concerns regarding the specificity and validity of the protocols? • Are there serious concerns regarding the sample size? • Are there concerns regarding the timing of the endpoint assessment?	***Endpoint sensitivity*:** Are the procedures sensitive for evaluating the endpoint(s)/outcome(s) of interest? • Was the outcome assessment methodology consistent with accepted guidelines or established criteria for the assay(s)/endpoint measures used in the study? • If sensitivity was not determined to prioritize studies prior to evaluation, assay-specific considerations regarding sensitivity, specificity, and validity of the test methods will be described here (e.g., metabolic competency, antibody specificity). • Is the cell/tissue type selected for the study appropriate and sensitive (e.g., is it routinely used) for measuring the endpoints of interest for the target organ system of interest? Are there known variations in cellular signaling unique to the model system that could influence the possibility of detecting the effect(s) of interest? • Are there serious concerns about the number of replicates/sample size in the study? *Differences*: "Specificity" removed for in vitro studies, as a separate domain for assessing this has been created (above). In addition, the steps taken to prioritize in vitro studies for individual study evaluation may involve consideration of the sensitivity of the assay(s); any pre-evaluation considerations for prioritization will be transparently described elsewhere and will not be reconsidered during study evaluation.

***Results presentation*:** Are the results presented in a way that makes the data usable and transparent? • Does the level of detail allow for an informed interpretation of the results? • Are the data analyzed, compared, or presented in a way that is inappropriate or misleading?	***Results presentation and analysis*:** Are the results presented and analyzed in a way that makes the data usable and transparent? • Does the level of detail allow for an informed interpretation of the results? • Are the data analyzed, compared, or presented in a way that is inappropriate or misleading? Flag potentially inappropriate statistical comparisons for further review. *Differences*: "Analysis" added for in vitro studies. Although this is also considered for in vivo studies, this is emphasized for in vitro studies given the increased heterogeneity of potential study designs and comparisons that increases the possibility of a "skewed" presentation of findings.

1

7. ORGANIZING THE HAZARD REVIEW:



ORGANIZE AND PLAN HAZARD REVIEW

Purpose

- To focus the hazard evaluation on the most influential health effects and analyses, providing the basis for hazard conclusions and guiding dose-response analyses.

Who

- Assessment team members.

What

- Outline for the synthesis of evidence.

This section discusses the process of organizing and structuring the synthesis of the evidence to support the formulation of hazard conclusions and to guide the approaches to dose-response analyses. The organization and scope of the hazard evaluation is determined by the available evidence for the chemical regarding routes of exposure, metabolism and distribution, outcomes evaluated, and number of studies within each evidence stream pertaining to each outcome, as well as the results of the evaluation of sources of bias and sensitivity. Thus, for some databases, the available evidence may 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 may 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 may 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 and/or developmental studies in animal studies). Such decisions can sometimes be informed by specific mechanistic evaluations, for example analyses of the extent of the linkage between related outcomes. Note that during the literature screening process, many

studies are tagged as potentially relevant supplemental material. Not all studies will be cited or considered in the assessment. Understanding which of these studies merit further consideration typically happens during the process of constructing the literature inventory and organizing the hazard review.

Certain outcomes may be identified that were analyzed by a larger number of independent research teams, that are of greater concern because they are linked by a set of inter-related outcomes, or that were reported by studies concluded to be of higher confidence. These outcomes will then guide the order that the organ systems will be presented. Typically, the outcomes and the hazards with the strongest evidence (i.e., larger number of informative studies of higher confidence), should be presented first. Study results pertaining to outcomes for an organ system that are of lesser influence on hazard analysis, or only reported by studies with lower confidence, are generally presented in less detail compared to outcomes with stronger or more extensive evidence. At early stages of draft development, a careful review of the literature inventories (see **Section 4.3**) in the context of human and animal study evaluation decisions (see **Chapter 6**) can aid grouping and prioritization of health effects for synthesis, as well as decisions not to extract data on specific endpoints or health effects that are considered *uninformative*. In these latter cases, the literature inventory might be used to provide a brief summary of the available evidence in the assessment, but the study results may not undergo all the evidence synthesis and integration steps outlined in **Chapters 9–11**. When making such decisions based on confidence in the available studies, it is important to consider the specific nature of the limitations identified (e.g., if the studies are all *low* confidence due to reduced sensitivity, the outcome should probably be summarized). A decision to exclude certain outcomes or health effects from further review should not be biased by the direction of the study results (e.g., if a set of outcomes is not informative in the context of the hazard review, both positive and null studies should be excluded), and it should consider the potential for such evidence to support other synthesis decisions (e.g., to inform other potentially coherent endpoints, to flag important data gaps, or to identify potential susceptible groups). A rationale for all such decisions should be included in the assessment.

In addition to the evidence from health effects studies, there may be additional relevant information that guides the organization of the evidence syntheses. Absorption, distribution, metabolism, and excretion (ADME) information may be particularly influential. If absorption, distribution, and metabolism vary, or are expected to vary, by the route of exposure, then the study results should be discussed separately by route of exposure. Alternatively, if physiologically based pharmacokinetic (PBPK) models exist that allow presenting results in terms of an internal dose metric, the evidence might be synthesized using the internal dose metric allowing the comparison of effect estimates and relative severity across route of exposure. Even when ADME understanding is incomplete, it may make sense to apply additional levels of organization to the hazard review based on the available results, e.g., according to lifestage, animal strain, or sex if the available

studies suggest pronounced differences in susceptibility. A variety of organizational possibilities may make sense depending on the extent and nature of the available evidence.

Biologic understanding of disease also may be helpful to organizational decisions. If a mechanistic pathway is known to be pertinent to multiple outcomes, based on either information collected during problem formulation or on early indications from the mechanistic study inventory, then consideration might be given to organizing those related outcomes or hazards together. At this point, enough information may 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 may 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. Also, at this point, as some or all of the potential adverse health effects that will be evaluated have been identified, additional targeted searches for mechanistic information specific to those health effects and/or organ systems may need to be performed. These supplemental searches may involve new literature search strategies, and they may be health effect- or tissue-specific rather than chemical-specific.

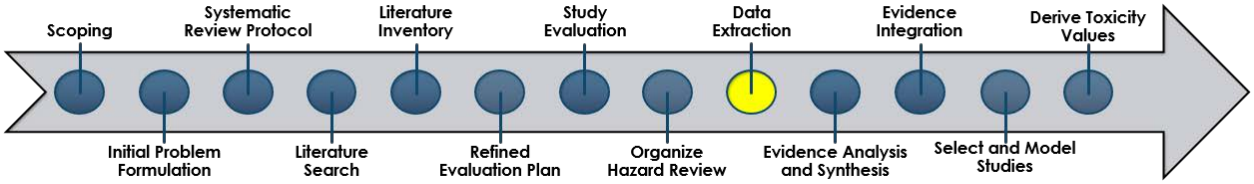
Table 7-1 lists some possible questions that may be asked of the evidence after pertinent studies have been identified, screened for relevance, and evaluated for confidence (e.g., after the literature search, screening and inventory, and study evaluations). These questions extend from considerations and decisions made during development of the refined evaluation plan to include review of the uncertainties raised during individual study evaluations as well as the direction and magnitude of the study-specific results. Resolution of these questions will then inform critical decisions about the organization of the hazard evaluation and what studies may be useful in dose-response analyses. The results of the literature inventories and organizing the level and grouping of hazard outcomes helps inform subsequent data extraction and visualization (see **Chapter 8**).

Table 7-1. Querying the evidence to organize syntheses for human and animal evidence

Evidence	Questions	Follow-up questions
ADME	Are absorption, distribution, metabolism, or excretion different by the route of exposure studied, lifestage when exposure occurred, or dosing regimens used?	Will separate analyses be needed by route of exposure or by methods of dosing within a route of exposure (e.g., are large differences expected between gavage and dietary exposures)? Which lifestages when exposure occurred, exposure durations, and frequencies are most applicable to the assessment?
	Is there toxicity information for metabolites that also should be evaluated for hazard?	What exposures will be included in the evaluation?
	Is the parent chemical or metabolite also produced endogenously?	
Outcomes	What outcomes are reported in studies? Are the data reported in a comparable manner across studies (similar output metrics at similar levels of specificity, such as adenomas and carcinomas quantified separately)?	At what level (hazard, grouped outcomes, or individual outcomes) will the synthesis be conducted? What commonalities will the outcomes be grouped by: • Health effect, • Exposure levels, • Functional or population-level consequences (e.g., endpoints all ultimately leading to decreased fertility or impaired cognitive function), or • Involvement of related biological pathways? How well do the assessed human and animal outcomes relate within a level of grouping?
	Are there inter-related outcomes? If so, consider whether some outcomes are more useful and/or of greater concern than others.	
	Does the evidence indicate greater sensitivity to effects (at lower exposure levels or severity) in certain groups (by age, sex, ethnicity, lifestage)? Should the hazard evaluation include a subgroup analysis?	
	Does incidence or severity of an outcome increase with duration of exposure or a particular window of exposure? What exposure time-frames are relevant to development or progression of the outcome?	
	Is there mechanistic evidence in the literature inventory that informs any of the outcomes and how they might be grouped together?	
	How complete is the evidence for specific outcomes? • What outcomes are reported by both human and animal studies and by one or the other? Were different animal species and sexes (or other important population-level differences) tested? • In general, what are the study confidence conclusions of the studies (*high*, *medium*, *low*, *uninformative*) for the different outcomes? What limitations that were identified may explain any inconsistencies in study results?	What outcomes should be highlighted? Should the others be synthesized at all? Would comparisons by specific limitations be informative?

Evidence	Questions	Follow-up questions
Dose-response	Did some outcomes include better coverage of exposure ranges that may be most relevant to human exposure than others?	What outcomes and study characteristics are informative for development of toxicity values?
	Does the study have multiple dose levels for which you can evaluate dose-response gradient? Are there outcomes with quantitative effect estimates (e.g., relative risk measures in epidemiology studies) that could allow examination or calculation of a combined measure of effect across multiple studies? Do the mechanistic data identify surrogate or precursor outcomes that are sufficient for dose-response analysis?	
	Are there groups that exhibit responses at lower exposure levels than others?	
	Are there findings from ADME studies that could inform data-derived extrapolation factors, or link toxicity observed via different routes of exposure, or link effects between humans and experimental animals?	Is there a common internal dose metric that can be used to compare species or routes of exposure?

8. EXTRACTION AND DISPLAY OF STUDY RESULTS OF HEALTH EFFECTS AND TOXICITIES FROM EPIDEMIOLOGY AND TOXICOLOGY STUDIES



DATA EXTRACTION AND DISPLAY

Purpose

- To support evidence synthesis and integration within the human and animal evidence.

Who

- Assessment team members in consultation with systematic review support staff and disciplinary workgroups as needed.

What

- Evidence tables and graphs.

At this stage in the assessment, hazards have been prioritized as explained in the refined evaluation plan and decisions have been made about grouping related outcomes. The next task is to extract and review relevant results from each study in the set, using both tables and graphs to organize and visualize findings in a way that facilitates drawing comparisons across the results and analyze the influence of relevant features (e.g., exposure range, study duration, study design) on the observed 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.

Not all studies that meet the populations, exposures, comparators, and outcomes (PECO) criteria will undergo data extraction based on decisions made during refining the evaluation plan (see **Section 5**). Studies considered *uninformative* during study evaluation are not considered further and, therefore, would not undergo data extraction. The same may be true for *low* confidence studies if enough *medium* and *high* confidence studies are available. The steps of study evaluation and data extraction may not always be strictly sequential, especially for animal toxicology 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), then it may be more efficient to conduct the data extraction in parallel to the study evaluation because much of the data extraction content directly informs the study evaluation judgments. In addition, outcomes or study designs that are determined to be less informative for dose-response and toxicity value derivation during organization of the hazard review may not go through detailed data extraction (e.g., single dose studies or studies with confounding exposures that cannot be controlled) but will be considered in the overall evaluation of evidence. Once an outcome/endpoint is selected for extraction, all informative studies that evaluated the outcome of interest should be captured in the extracted results regardless of statistical significance or null/negative findings. 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.

8.1. 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 for 1–4 hours of time per study, depending on the complexity of the study and number of outcomes to extract. 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 may 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. Efficient data extraction requires some knowledge of what and how information is presented in the set of studies to help make decisions on the extent of data extraction that is 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). Missing data from individual mechanistic (e.g., in vitro) studies will generally not be sought.

The Integrated Risk Information System (IRIS) Program commonly uses an 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 toxicology studies. Extracting into HAWC allows the creation of visuals that have interactive "click to see more" capabilities, which can reduce the number of summary tables that need to be developed and, therefore, make the assessment more concise. The visualization features of HAWC also make it easier to identify and present patterns of findings that support evidence synthesis and integration conclusions (see **Chapters 9 and 11**). Excel files created outside of HAWC for data extraction purposes can be imported into HAWC to create visuals, but these visuals will not have the "click to see more" functionality that requires direct extraction into HAWC. Benchmark dose

(BMD) modeling can also be done from within HAWC. Currently, HAWC is best suited for graphical displays of health outcome data. Tabular or narrative presentations or summarization of non-health outcome content (e.g., absorption, distribution, metabolism, and excretion [ADME]/toxicokinetic) 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 HAWC level of extraction 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).

8.1.1. Health Assessment Workspace Collaborative (HAWC)

Training tutorials are available at https://hawcprd.epa.gov/about/ and detailed instructions for summarizing specific data extraction elements are described within the HAWC extraction modules. A list of data extraction fields for animal bioassay, epidemiology, and in vitro studies in Excel format is available at the HAWC website (see "About," then "Downloads"). In addition to fields 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 to get information from figures. HAWC has a management dashboard to make tasks assignments and manage quality assurance (QA)/quality control (QC). Administered doses for animal studies can be presented in multiple dose metrics (e.g., mg/kg-day human equivalent dose [HED]) by adding new dose representations although the calculations are not automatic. Instead, the conversions are done outside of HAWC and manually entered. An Excel spreadsheet to guide conversions is available in the "IRIS Assessment Templates" HAWC project; however, it is highly recommended that the conversions are done (or reviewed) by someone with experience. Data extraction (and study evaluation) results in HAWC are available for download in Excel format. The HAWC project can be made viewable and downloadable to the public when a draft assessment is available for public comment. Static images of the HAWC figures should be used in the assessment document and a figure footnote can be used to provide the URL for readers who want to view the interactive web-based versions.

A frequently asked questions document is available in the "IRIS Assessment Templates" HAWC 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://hawcprd.epa.gov/media/attachment/HAWC_FAQ_for_assessment_readers.docx).

HAWC figure formats can be copied from existing assessments for use in new projects, and template figure formats are available in the "IRIS Assessment Templates" HAWC project. Currently, HAWC is best suited for graphical displays of health outcome data and tables are best constructed using other software applications, such as Microsoft Word. However, the downloadable Excel files

from HAWC can be used to create tables using Microsoft features such as mail merge. In addition, Excel files created outside of HAWC for data extraction purposes can be uploaded into HAWC to create visuals, but these visuals will not have the "click to see more" functionality that require direct extraction into HAWC. In addition to HAWC, R-based graphical scripts developed for use with other software tools (e.g., GraphPad Prism) also may be useful.

Certain aspects of data extraction can be done independently by support staff who are familiar with HAWC (e.g., contractors, student interns). Activities that are most amenable to delegation include uploading studies into HAWC and summarizing study design and methods for animal toxicology studies. However, extraction of results and creation of graphics should be done under close supervision by the assessment team. In addition, due to their nonstandard formats, epidemiological studies are more difficult to summarize and extract. Any delegation of extraction for epidemiological studies should be done under close supervision by epidemiologists on the assessment team.

The results that are extracted from each study are determined by the way the data have been presented by study authors and the needs of the assessment. When large amounts of quantitative analyses are presented in a paper, decisions will be needed to select the most informative effect estimates, as well as those that are more commonly presented in the set of papers. 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) may not be as informative as a more comprehensive analysis (e.g., analyses considering level of exposure) developed in only one or two studies. Thus, it may be necessary to extract more than one set of results. Statistical test results noted by study authors are recorded in the HAWC database with extracted data. Sometimes multiple approaches to evaluating statistical significance are possible and EPA may conduct other statistical analyses than were reported in the original papers. This gets recorded in the "Result notes" field in HAWC. 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 or CatReg, for analysis and visualization.

8.1.2. 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 and pilot period to orient new staff to the extraction process. Ideally, new staff should extract one study with review by someone experienced in data extraction/HAWC, followed by extraction of another two to three studies with an additional round of review.

- Ensure the extraction of study design and methods into HAWC or other formats 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 very time intensive to adjust.

- For consistent outcome/endpoint extraction, use the suggested terminology in the "HAWC Endpoint Terms" Excel file available in the IRIS Assessment Template project. This terminology has been suggested not only to promote consistency across assessments, but also interoperability with other databases (e.g., ToxRefDB, CEBS, Organisation for Economic Co-operation and Development [OECD] Harmonised Templates, and other ontologies) coded using the Unified Medical Language System (UMLS; https://www.nlm.nih.gov/research/umls/).

- 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/). In HAWC, when values are estimated, be sure to 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 accidental changes.

- Create HAWC visualizations early in the process to help QA/QC the extraction and aid the evidence synthesis process.

- Frequently monitor the consistency of extraction across studies using the data clean-up tool in HAWC.

- Use the management dashboard to track QA/QC.

8.1.3. Data Extraction into Tabular Format

As noted above, instructions for summarizing specific data extraction elements are described within the HAWC extraction modules. Below are several data extraction tips for studies when information will be summarized in tables without the use of HAWC to structure the data extraction (see **Section 8.3** for examples).

General Tips

- Abbreviate units of time within tables.

- 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 can get corrupted and cause problems with the Health and Environmental Research Online (HERO) plugin. 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 may be needed if additional columns (e.g., more detailed study design or results information) are presented. Consider including concurrent data in the same cell/row that may help explain or interpret findings, such as body weight when evaluating organ-weight effects, maternal toxicity indicators when interpreting toxicity in offspring, or mortality.

Epidemiological Evidence

- The organization of the information in the "Reference and study design" column is flexible but should be consistent throughout an individual table and should be as consistent as possible across tables.

- While several group numbers are reported in studies (e.g., total participants, numbers included in analysis), study size should reflect the number of participants in the primary analysis of interest.

- Description of the population may 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) and range (such as interquartile range).

- Include a summary of the study evaluation and the overall confidence conclusion (see **Chapter 6**).

- For prioritized outcomes, results across available studies on the outcome should be displayed regardless of statistical significance (see **Section 8.4**). When available, there should be some indication of the uncertainty in the result (e.g., 95% confidence interval [CI]), and it may 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.

- If multiple exposure measures are provided (e.g., cumulative and peak exposure), all may be presented in the table or selected metric(s) may 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 and then others as informative.

- If few or no quantitative results are reported, a qualitative description of results may 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

- 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 throughout a table and should be as consistent as possible across tables.

- Include a summary of the study evaluation and the overall confidence conclusion (see **Chapter 6**).

- Exposure levels should be presented in common units (e.g., mg/kg-day or mg/m^3) and be reported in the results column in line with the results corresponding to that 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 may 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 that are measured with different scales (discussed in **Section 8.2**). 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."

8.2. STANDARDIZING REPORTING OF EFFECT LEVELS AND SIZES

Approaches for designations of treatment-related findings or statistical significance provided by the study authors may 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), used historically to summarize study findings prior to the ready availability of dose-response modeling tools, have a number of limitations, particularly lack of consistency across studies.[10] When different approaches have been used across studies, EPA relies on consistent considerations to the extent possible (e.g., measurement scales, statistical testing methods) in order to reach an overall conclusion; all differing interpretations are captured transparently in the overall synthesis. The treatment-related findings presented in assessment text, tables, and graphs represent EPA conclusions. Differences from study author conclusions are annotated during the extraction process and may be discussed in the narrative of the assessment text, depending on degree of controversy and impact on assessment conclusions.

[10]EPA'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 (also see **Section 9.4.1**), 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.

In addition to providing quantitative outcomes in their original units for all study groups, results from outcome measures will be transformed, when possible, to a common metric to help compare distinct but related outcomes that are measured with different measurement scales. These standardized effect size estimates facilitate systematic evaluation and evidence integration for hazard identification, whether meta-analysis is feasible for an assessment (see **Section 9.1**). 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).

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.

- *Percentage of control response (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 upon the pooled variance for controls and treated units. Pooling variances may be problematic if variances differ substantially, in which case it may 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 may 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 percent change; a 0% can easily occur when the control incidence probability is small relative to sample size.

- *Extra risk*. Often used for defining toxicity values (see **Section 13**), 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 × k table at each dose. The metrics above can be applied to each control-treated comparison in a 2 × k table, resulting in k 2 × 2 metrics at each dose. One simplifying approach is to reduce the 2 × k table to a 2 × 2 table (e.g., severity rank ≤3 and >3). Statisticians and subject matter experts may suggest other approaches for reducing a 2 × k table to a single metric.

An additional approach for epidemiology studies is to extract adjusted statistical estimates when possible rather than unadjusted or raw estimates.

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).

8.3. STANDARDIZING ADMINISTERED DOSE LEVELS/CONCENTRATIONS

Exposures will be standardized to common units when appropriate. Exposure levels in oral studies will be expressed in units of mg/kg-day. Where study authors provide exposure levels in concentrations in the diet or drinking water, dose conversions will be made using study-specific food or water consumption rates and body weights when available. Otherwise, EPA defaults will be used (U.S. EPA, 1988), 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/m^3. Assumptions used in performing dose conversions will be documented. As discussed in **Section 8.1.1**, 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 manually entered. An Excel spreadsheet to guide conversions is available in the "IRIS Assessment Templates" HAWC project, however, it is highly recommended that the conversions are done (or reviewed) by someone with experience. 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.

8.4. 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. While the specific organization and level of detail may 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) as well as structural choices (e.g., types of information in columns, rows, axes, and symbols). Contextual information provided by peripheral analysis in a study or from supplemental material is often not extracted and may only be described in narrative form or table/graph notes.

There may be some results for an outcome that are more commonly reported across multiple studies, which could be presented graphically to evaluate consistency. Additional analyses (e.g., summary measures, trend tests) may 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 [e.g., motor activity], then by endpoint [e.g., horizontal activity, rearing]). Regardless of how the information is organized in the tables and graphics, a thorough quality assurance 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).

8.4.1. Determining the Level of Detail for Data Extraction

Data extraction at the level of summarizing effect size and variability information (e.g., mean, SD/standard error of the mean) is a laborious process and may take 30 minutes for simple studies with a single result to 4 or more hours for studies with multiple exposure metrics or outcomes/endpoints. Further, extraction time increases substantially if information is presented in figures (that must be digitized) compared with tables. Detailed extraction at the level of effect size information is generally pursued for key study findings, whereas questions arise on the extraction effort for 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 during evidence synthesis. The following strategies can be considered to minimize the data extraction burden while still presenting an accurate representation of all results (both null and exposure-related), as well as the information needed to provide context for interpretation of the primary outcomes.

- 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 may result in general observations for dams (body-weight gain, feed consumption, and 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 to be exposure-related, the data for these observations were not extracted.

- In the event dose-response data are not fully extracted, a user may "dummy code" the endpoint to generate exposure response arrays figures that display the direction of effect. This may be especially useful when the information is contained in figures that need to be digitized to obtain numbers. Dummy coding is not a significant resource saving when effect size information is presented in tabular format. To develop figures for animal studies in HAWC (described in **Section 8.5**), 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 as well as 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 is presented extensively in figures. The request does not have to 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 may be reasonable. Animal studies of learning may be especially challenging to summarize because they often include repeated measurements and judgments need to be made as to 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 may 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 may 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 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, then use of a tabular presentation or seeking copywrite permission to reproduce the original figure may be more appropriate.

8.5. GRAPHICAL AND TABULAR DISPLAY

Several graphical formats, notably exposure- or dose-response graphs, forest plots, and exposure response arrays, are routinely used in assessments. While these displays are useful for the presentation of human and animal health effect evidence, they are generally not as informative for the display of mechanistic data (see discussion in **Chapter 10**). The use of arrays and other types of graphical representations (both of raw data and analyses of those data) is a foundation of hazard identification and is also used in dose-response analysis. The display of data facilitates identification of patterns of response associated with chemical exposure and can aid in those evaluations as well as 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 BMDs corresponding to 10% responses.

The following sections discuss and provide examples for both graphical and tabular display (see **Figures 8-1, 8-2, 8-3, 8-4, 8-5**). 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. Alternatively, in many cases, only high and medium confidence results undergo full data extraction and confidence in those results would therefore not be a critical consideration. For data-poor chemicals or outcomes, however, low confidence results may need to be included and confidence should be included in the visualization to improve interpretability of the findings. Further discussion on incorporating confidence ratings into these graphics is included at the end of the discussion for each type of figure.

8.5.1. 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 an x-y graph of the level of the causative agent (drug, chemical, radiation, temperature, etc.) on the x-axis, versus the response level of the target (population, animal, organ, tissue) plotted on the y-axis. Dose-response curves are generally generated for a single effect. A dose-response graph can also be useful for epidemiology data, specifically in studies that examine

multiple exposure levels, or exposure as a continuous measure. 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 may be an increase (e.g., higher incidence) or a decrease (e.g., decrease in body-weight gain when 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).

In the example shown in **Figure 8-1**, where 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 where data are displayed for multiple studies (as in **Figure 8-2**), data for higher confidence studies should be somehow emphasized in the graphic. Examples for doing so is addition of an indicator line as a demarcation of where study confidence changes [**Figure 8-2(a) and (b)**] or added to the legend to indicate the quality of the studies as a parenthetical [**Figure 8-2(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 (see **Section 9.4**), which may 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 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 display, 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.

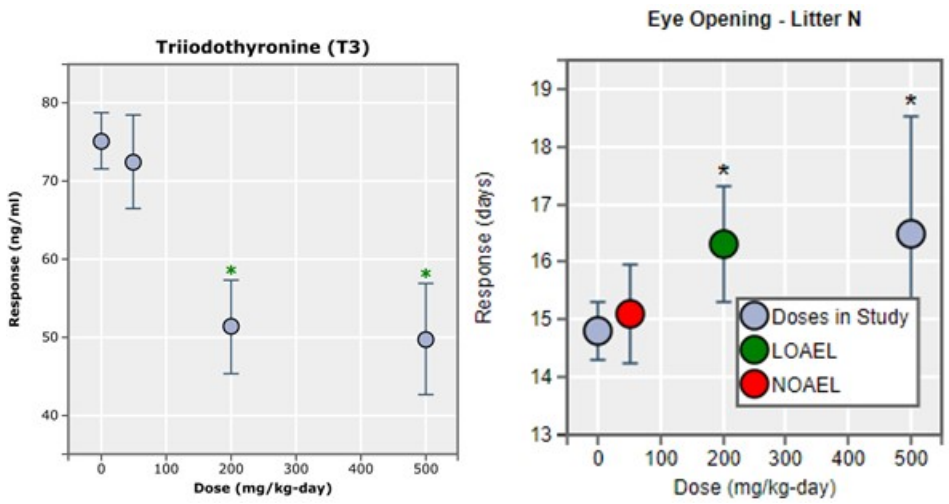


Figure 8-1. Examples of dose-response graphical displays for single endpoint created in Health Assessment Workspace Collaborative (HAWC) (for illustrative purposes only).

BMDS = Benchmark Dose Software.

The above visualizations can be 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.

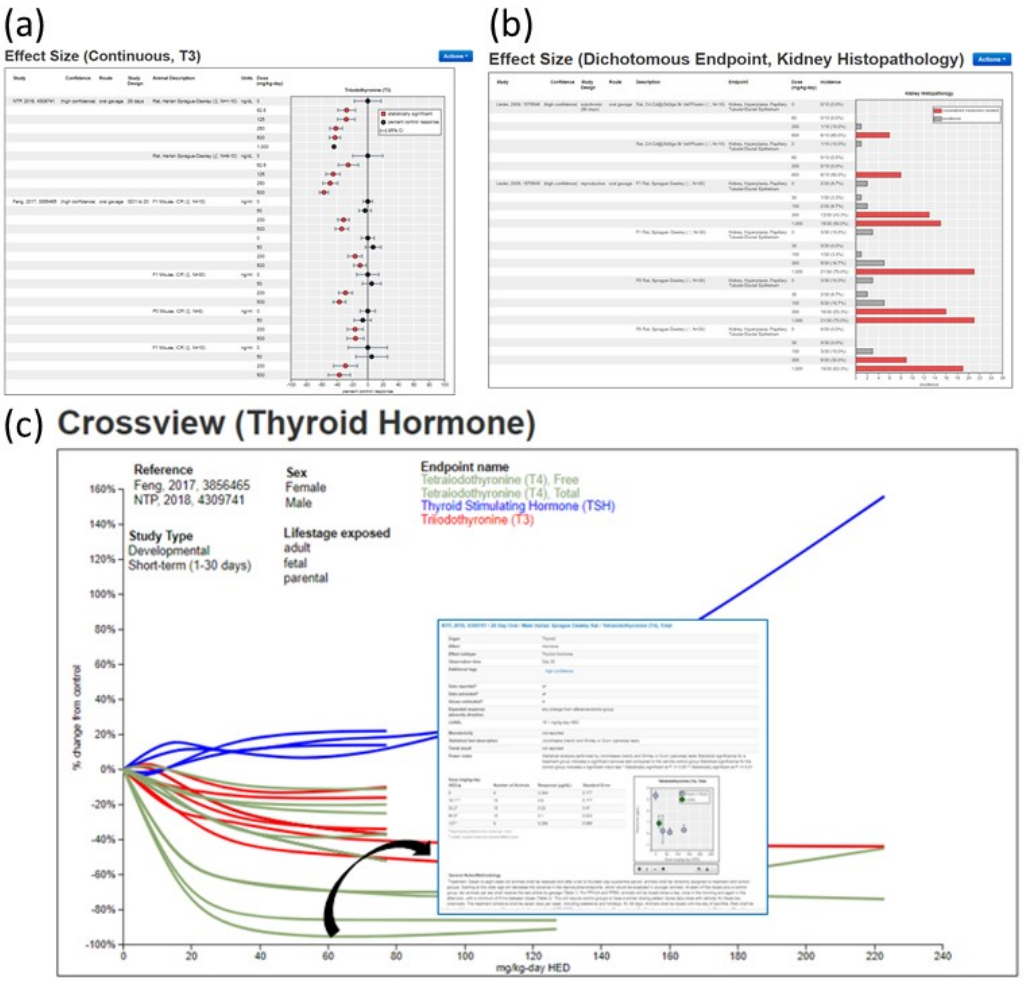


Figure 8-2. Examples of dose-response graphical displays across endpoints and studies created in Health Assessment Workspace Collaborative (HAWC) (for illustrative purposes only). (a) Data pivot (continuous variable). (b) Data pivot (dichotomous variable). (c) Animal bioassay endpoint cross view with 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/100000037/pfbs-estrous-cyclicity-effect-size-animal/; https://hawcprd.epa.gov/summary/data-pivot/assessment/100000039/pfbs-kidney-histopathology-effect-size-animal/; and https://hawcprd.epa.gov/summary/visual/100000087/.

8.5.2. Forest Plots

Forest plots are generally used to summarize epidemiologic data from a set of studies evaluating a specific health endpoint for the purposes of hazard identification. As commonly used, the underlying assumption is that all studies examined the same exposure contrast (e.g., "ever" vs. "never" exposed is comparable across studies). 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)].

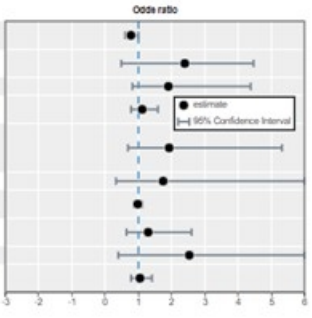
When applied to epidemiological data, forest plots typically array multiple point estimates of the effects of a specific exposure with a specific health endpoint (e.g., relative risks, odds ratios, hazard ratios) and their associated CIs (e.g., 95% CI) represented by lines from the lower bound of the CI to the upper bound with the point estimate clearly identified (see **Figure 8-3**). Additional details (e.g., design, numbers of cases, specific exposure metric, and study confidence evaluation) may be annotated as needed to transparently describe the available data. 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 where 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 of 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 8-3(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 8-3(b)**], ordering the studies from top to bottom by study confidence with labeled lines as demarcations of where confidence changes is a possibility.

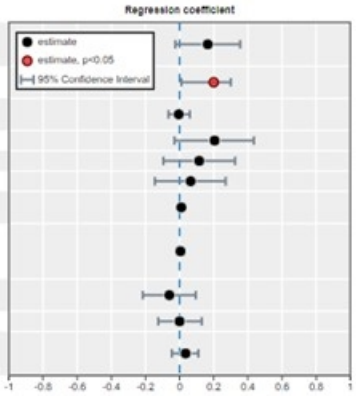
(a)

Outcome	Study	Population	Design	Exposure metric	Exposure central estimate	Sub-population	Exposure group	Odds ratio Units
Hyperthyroidism (subclinical)	Lopez-Espinosa et al. 2012	C8 Health Project (2005-2006) 10,725 children	Cross-sectional	Serum	1.5	Children 1-17 yrs (n=10725)	Continuous (Ln)	0.78 [IQR increase]
	Wen et al. 2013	NHANES (2007-2010), U.S., 1,181 adults (672 men, 509 women)	Cross-sectional	Serum	1.5	Men	Continuous (Ln)	2.41
						Women	Continuous (Ln)	1.91
Hypothyroidism	Lopez-Espinosa et al. 2012	C8 Health Project (2005-2006) 10,725 children	Cross-sectional	Serum	1.5	Children 1-17 yrs (n=10725)	Continuous (Ln)	1.11 [IQR increase]
	Lin C et al. 2012	Young Taiwanese Cardiovascular Cohort Study (2006-2008), Taiwan, 511 12-30 year olds (214 men, 337 women)	Cross-sectional	Plasma	1	Men and women 12-30 years old	T2 (60th-90th)	1.93
							T3 (>90th)	1.75
Hypothyroidism (subclinical)	Lopez-Espinosa et al. 2012	C8 Health Project (2005-2006) 10,725 children	Cross-sectional	Serum	1.5	Children 1-17 yrs (n=10725)	Continuous (Ln)	0.99 [IQR increase]
	Wen et al. 2013	NHANES (2007-2010), U.S., 1,181 adults (672 men, 509 women)	Cross-sectional	Serum	1.5	Men	Continuous (Ln)	1.3
						Women	Continuous (Ln)	2.54
Thyroid disease	Lopez-Espinosa et al. 2012	C8 Health Project (2005-2006) 10,725 children	Cross-sectional	Serum	1.5	Children 1-17 yrs (n=10725)	Continuous (Ln)	1.05 [IQR increase]

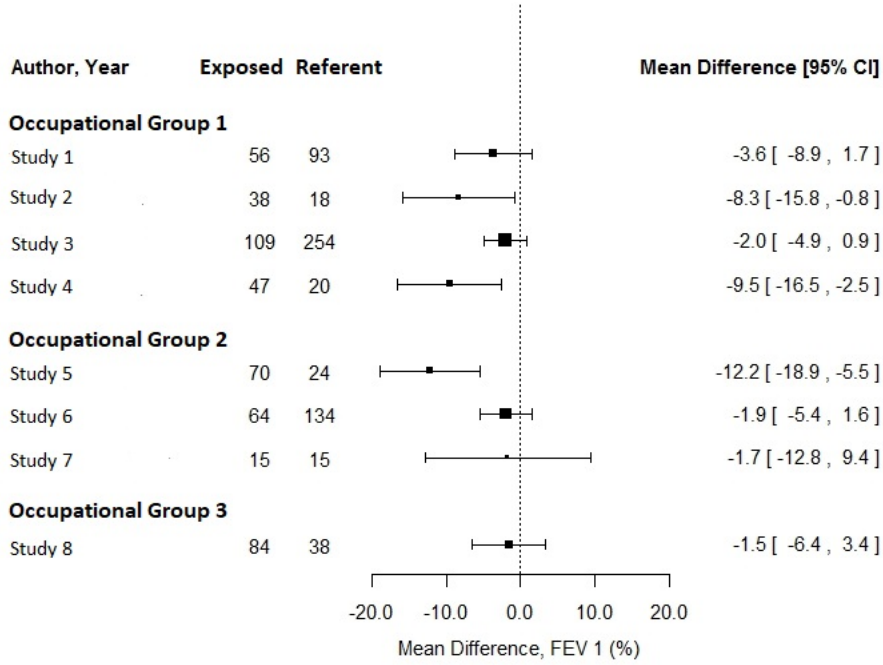


(b)

Study	Population	Design	Exposure central estimate	Sub-population	Outcome	Exposure group	Regression coefficient
Wang Y et al. 2013	Norwegian Mother and Child Cohort Study (2003-2004), Norway, 903 pregnant women	Cross-sectional	0.4	Pregnant women (n=903)	TSH (Ln)	Continuous	0.165
Webster et al. 2014	CHirP (2007-2008), Canada, 152 women	Cohort	0.6	Pregnant women (n=151)	TSH	Continuous	0.2
Preston E et al. 2018	Project Viva (1999-2002), U.S., 732 pregnant women and 480 neonates	Cohort (Prospective)	0.6	Pregnant women (n=718)	TSH (Ln)	Continuous	-0.0027
						Q2	0.204
						Q3	0.114
						Q4	0.063
Crawford et al. 2017	Time to Conceive study (2008-2009), U.S., 99 women	Cross-sectional	0.8	Women	TSH	Continuous (Log)	0.01
Lin C et al. 2012	Young Taiwanese Cardiovascular Cohort Study (2006-2008), Taiwan, 511 12-30 year olds (214 men, 337 women)	Cross-sectional	1	Men and women 12-30 years old	TSH (Ln)	Continuous	0.006
Wen et al. 2013	NHANES (2007-2010), U.S., 1,181 adults (672 men, 509 women)	Cross-sectional	1.5	Women	TSH (Ln)	Continuous (Ln)	-0.062
				Men	TSH (Ln)	Continuous (Ln)	0.002
Wang Y et al. 2014	Taiwan Maternal and Infant Cohort Study (2000-2001), Taiwan, 285 pregnant women and 116 neonates	Cohort (Prospective)	1.5	Pregnant women (n=283)	TSH	Continuous	0.033



(c)



(d)

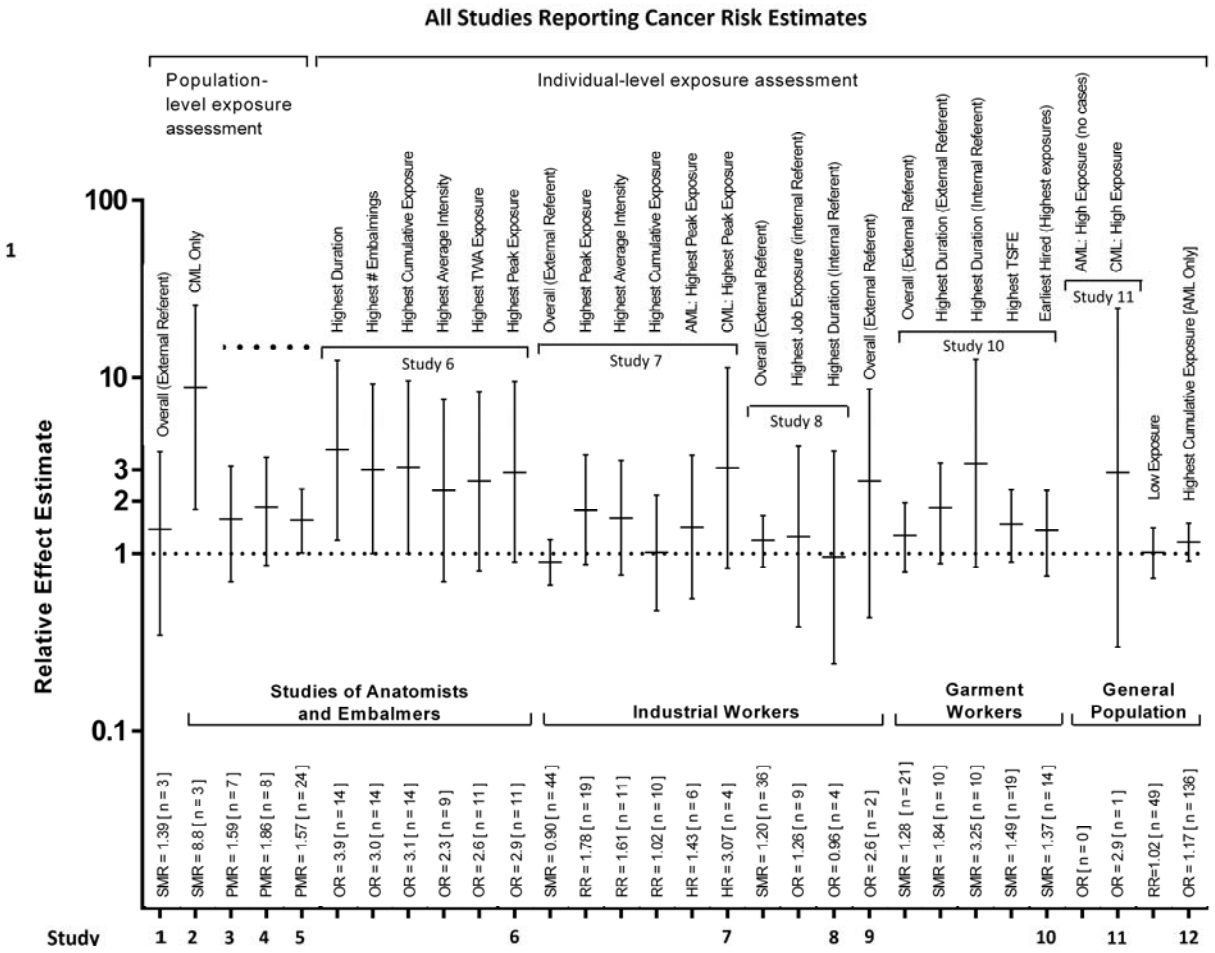


Figure 8-3. Examples of forest plots used for epidemiological evidence (for illustrative purposes only). (a) HAWC forest plot (odds ratio, null of 1), all medium confidence studies. (b) HAWC forest plot (regression coefficient, null of 0), all medium confidence studies. (c) R forest plot (mean difference, null of 0). (d) GraphPad Prism forest plot (null of 1).

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 a 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/example-forest-plot/
(currently limited to IRIS staff).

8.5.3. 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, and
- 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). Incorporation of the study confidence should be included as needed using the same techniques as described for other graphic formats discussed in this **Section 8.5**. **Figure 8-4(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 likely to be applicable to other types of graphical depictions of data as well.

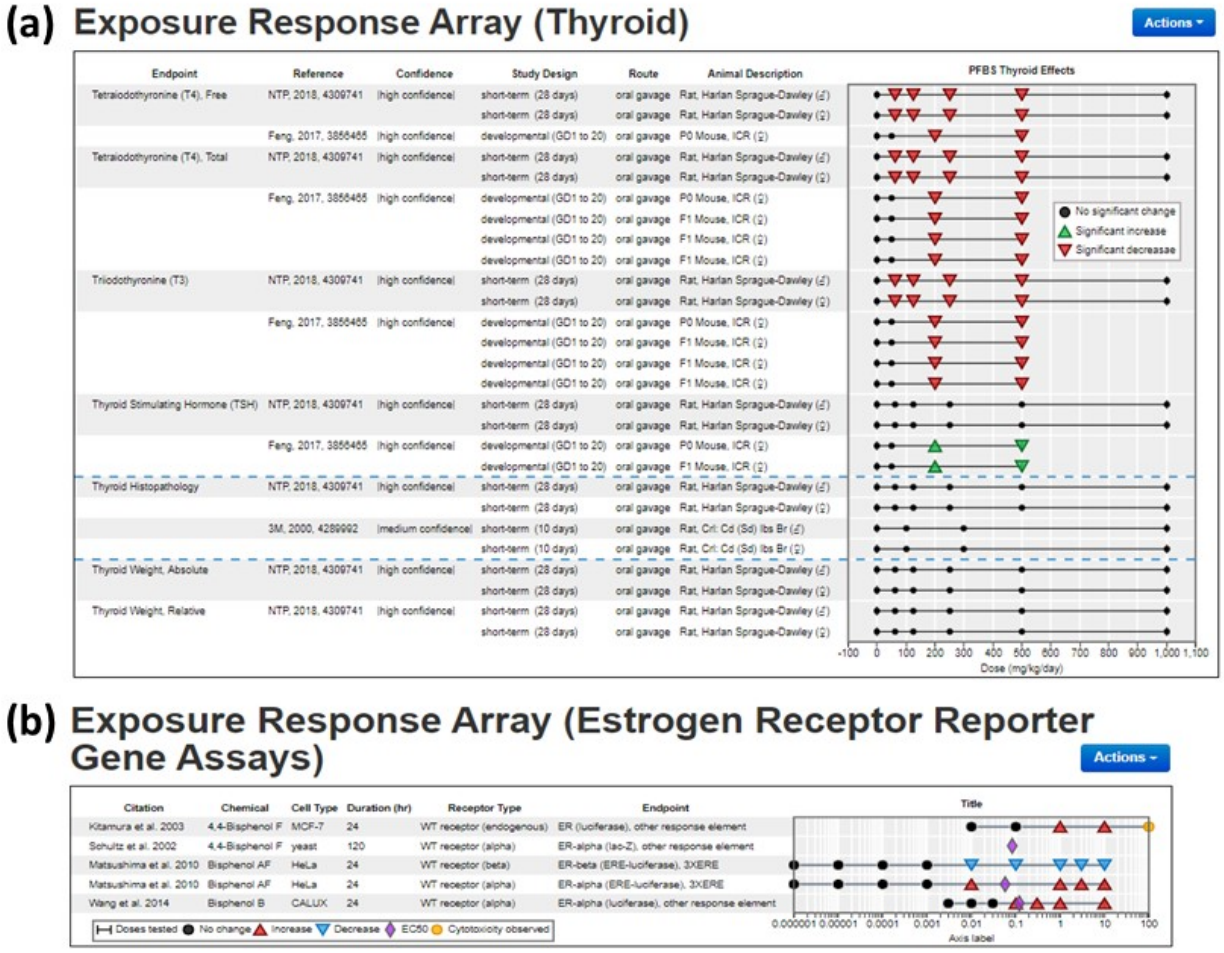


Figure 8-4. Examples of exposure response arrays. (a) HAWC exposure response array for animal studies. (b) HAWC exposure response array for in vitro studies.

Exposure response arrays can be created in HAWC for animal data using the "data pivot" visualization option. When effect size information has been extracted, then 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 "---," then no changes will be applied. Generally, "---" should be used.

The examples in HAWC are available at: https://hawcprd.epa.gov/summary/data-pivot/assessment/100000039/pfbs-thyroid-effects/ and https://hawcprd.epa.gov/summary/data-pivot/assessment/100000039/estrogen-receptor-reporter-gene-assays/.

8.5.4. Tables

While 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 may not be possible and would unnecessarily clutter the display. Tables can be used as stand-alone 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 not consistency in the effect estimates, units, or other factors across studies being reviewed, a tabular summary may be the most appropriate way to present the data.

Figure 8-5 shows examples of tables for epidemiology and animal toxicology studies. 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.

(a)

Reference, study confidence	Population	Median exposure (IQR) or as specified	Outcome	Unit change in exposure metric	OR (95% CI)
Prenatal exposure measure (maternal or cord blood samples)					
Study 1, medium	Birth cohort (enrolled 1992-93), Norway; 642 children (10 yrs)	0.2 (0.1-0.2)	Current asthma	doubling	1.05 (0.85,1.29)
			Ever asthma	doubling	0.96 (0.73,1.26)
Study 2, low	Birth cohort (enrolled 1997-2000), Faroe Islands; 559 children (5 and 13 yrs)	0.6 (0.5-0.8)	Ever asthma	doubling	1.03 (0.67,1.59)
Study 3, medium	Birth cohort (enrolled 2002-04), Greenland and Ukraine; 1024 children (5-9 yrs)	0.7 (0.3-2.0) (Greenland, 5^{th}-95^{th})	Ever asthma	1 SD change	0.90 (0.70,1.14)
Childhood exposure measure (concurrent with outcome ascertainment)					
Study 4, medium	Cross-sectional study (1999-08), U.S.; 1,877 adolescents (12-19 yrs)	0.8 (0.5-1.2)	Current asthma	ln-unit change	1.00 (0.76,1.33)
			Ever asthma	ln-unit change	0.99 (0.88,1.12)
Study 5, medium	Case-control study (2009-10), Taiwan; 456 children (10-15yrs)	0.8 (0.6-1.1)	Asthma diagnosed in last year	quartiles vs. Q1	Q2: 1.19 (0.68,2.09) Q3: 1.54 (0.86,2.76) Q4: 2.56 (2.11,6.93) p-trend: 0.04*
Study 2, low	Birth cohort (enrolled 1997-2000), Faroe Islands; 559 children (5 and 13 yrs)	1.0 (0.8-1.2)	Ever asthma	doubling	0.72 (0.44,1.18)

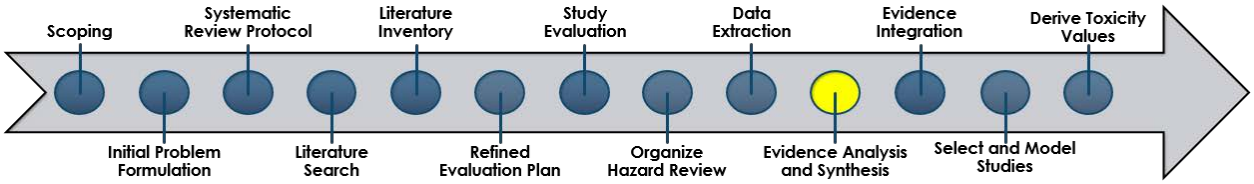
Studies are sorted by age at exposure measurement then median exposure level.
*p<0.05

Reference and study design	Results				
Serum thyroid hormones					
{Reference} Rats, Sprague-Dawley Gavage exposure 90-d exposure in adults Thyroid hormones (total T3/T4) measured by ELISA Low Confidence	**Doses** (mg/kg-d)	**0**	**100**	**300**	**1,000**
	TSH (ng/mL)				
	Male (n = 5–10)				
	Mean (SD)	0.46 (0.42)	3.29 (3.86)	2.65 (2.10)	3.88 (2.98)
	% of control[a]	–	615%	476%	743%
	Female (n = 5–10)				
	Mean (SD)	0.46 (0.31)	1.42 (1.11)	3.96 (5.15)	2.43 (1.74)
	% of control[a]	–	209%	761%	428%
	T4 (µg/dL)				
	Male (n = 9–10)				
	Mean (SD)	7.87 (1.22)	6.34* (1.22)	6.28* (1.03)	4.97* (0.76)
	% of control[a]	–	−19%	−20%	−37%
	Female (n = 9–10)				
	Mean (SD)	5.43 (0.86)	4.96 (0.62)	4.53* (0.88)	4.31* (0.76)
	% of control[a]	–	−9%	−17%	−21%
{Reference} Rats, Sprague-Dawley Dietary exposure F1: maternal exposure from GD 10 to PND 20 Thyroid hormones were measured by ELISA in male offspring only Medium Confidence	**Doses** (mg/kg-d)[c]	**0**	**15**	**146**	**1,505**
	TSH (ng/mL)				
	Male, F1, PND 20 (n = 10)				
	Mean (SD)	5.40 (0.62)	6.66 (1.24)	6.07 (1.41)	7.00* (1.31)
	% of control[a]	–	23%	12%	30%
	Male, F1, PNW 11 (n = 10)				
	Mean (SD)	4.74 (0.62)	5.81 (1.72)	5.36 (1.11)	4.96 (0.8)
	% of control[a]	–	23%	13%	5%
	T4 (µg/dL)				
	Male, F1, PND 20 (n = 10)				
	Mean (SD)	4.39 (0.93)	4.20 (0.77)	4.78 (0.49)	4.20 (0.52)
	% of control[a]	–	−4%	9%	−4%
	Male, F1, PNW 11 (n = 10)				
	Mean (SD)	4.77 (0.7)	4.84 (0.59)	5.21 (0.65)	5.20 (0.98)
	% of control[a]	–	1%	9%	9%

*Statistically significantly different from the control at $p < 0.05$ as reported by study authors.
[a]Percent change compared to control calculated as: (treated value − control value)/control value × 100.
[c]Time-weighted averages (TWAs) for each exposure group were calculated by multiplying the measured HBCD intake (mg/kg-day) reported by the study authors for GDs 10–20, PNDs 1–9, and PNDs 9–20 by the number of inclusive days of exposure for each time.
BW = body weight; GD = gestation day; PNW = postnatal week

Figure 8-5. Example tabular displays. (a) Table of epidemiology studies. (b) Table of animal bioassays.

9. ANALYSIS AND SYNTHESIS OF HUMAN AND EXPERIMENTAL ANIMAL DATA



ANALYSIS AND SYNTHESIS OF EVIDENCE

Purpose

- To summarize and interpret the results across all informative health effect studies within the human and animal evidence streams, with an emphasis on considerations pertinent to evidence integration.

Who

- Assessment team and disciplinary workgroups (as needed).

What

- Draft hazard synthesis sections describing human and animal toxicity data.

This chapter describes various approaches to synthesize the results of studies investigating links between exposure and outcome in either humans or animals. In IRIS assessments, evidence synthesis and integration are considered distinct, but related processes. The syntheses of separate evidence streams (i.e., human, animal, and mechanistic evidence) described in this chapter and **Chapter 10** will directly inform evidence integration within and across the evidence streams to draw overall conclusions for each of the assessed human health effects (as described in **Chapter 11**). This chapter also describes approaches to compare study results based on common or disparate metrics or analyses and encourages the consideration of whether additional analyses such as meta-analyses or statistical tests will add value. Evidence synthesis is an iterative process.

Approaches for analyzing mechanistic data on endpoints intended to characterize precursor events that lead to or modify the development of adverse health outcomes are described separately in **Chapter 10**.

Syntheses will be organized using the order and grouping levels that were established using the process described in **Chapters 5 and 7**. Briefly, the synthesis of the human or animal health effects evidence should preferably occur at the outcome level (e.g., asthma for human; lung histopathology for animal) if there is an adequate set of studies for analyses at this level. If studies

on a target system are sparse and varied, then the analyses may need to be conducted at a health effect or broader grouping (e.g., respiratory) level.

9.1. GENERAL CONSIDERATIONS FOR SYNTHESIZING THE HUMAN AND EXPERIMENTAL ANIMAL EVIDENCE

Each synthesis should summarize the available evidence relevant to assessing the extent to which chemical exposure is likely to cause a health effect (or not) based on considerations for causality adapted from those introduced by Austin Bradford Hill (Hill, 1965); these considerations include consistency, exposure response relationship, strength of the association, biological plausibility, coherence, and "natural experiments" in humans (U.S. EPA, 2005b, 2002a, 1994). Hill (1965) discusses nine considerations that could be used in the interpretation of epidemiology studies,[11] but notes that these are not offered as criteria or rules of evidence. Thus, although these considerations provide a framework for assessing evidence, they do not lend themselves to being used as a simple formula or checklist.

Most of the considerations discussed by Hill (1965) are applicable to health-effect studies in humans and animals, with some differences in terminology and definitions (see **Table 9-1**). This approach, taken for evidence synthesis within the IRIS Program, is informed by both Hill and another widely used approach, the Grading of Recommendations Assessment, Development, and Evaluation (GRADE) framework. The GRADE framework includes consideration of many of the Hill (1965) concepts but provides more details on how to evaluate and document the expert judgments embedded in the process of evidence synthesis (Guyatt et al., 2011a; Schünemann et al., 2011). Importantly, this section describes how the evidence syntheses consider and incorporate the conclusions from the individual study evaluations (see **Section 6**). **Table 9-1** provides the types of information that can be used in the synthesis of evidence for an outcome from either the human or animal health effects studies, including mechanistic information (see **Chapter 10**).

[11]One consideration specific to epidemiology studies—the temporal relationship between exposure and effect—is addressed during the evaluations of individual studies (see **Section 6.2**).

Table 9-1. Important considerations for evidence syntheses

Consideration	Description of the consideration and its application in IRIS syntheses
Study confidence	Description: Incorporates decisions about study confidence within each of the considerations. Application: In evaluating the evidence for each of the causality considerations described in the following rows, syntheses will consider study confidence decisions. *High* confidence studies carry the most weight. Syntheses will consider specific limitations and strengths of studies and how they inform each consideration.
Consistency	Description: Examines the similarity of results (e.g., direction; magnitude) across studies. Application: Syntheses will evaluate the homogeneity of findings on a given outcome or endpoint across studies. When inconsistencies exist, the syntheses consider whether results were "conflicting" (i.e., unexplained positive and negative results in similarly exposed human populations or in similar animal models) or "differing" [i.e., mixed results explained by differences between human populations, animal models, exposure conditions, study methods or potential biases and degree of insensitivity; (U.S. EPA, 2005b)] based on analyses of potentially important explanatory factors such as: • Confidence in studies' results, including study sensitivity (e.g., some study results that appear to be inconsistent may be explained by potential biases or other attributes that affect sensitivity). • Exposure, including route (if applicable) and administration methods, levels, duration, timing with respect to outcome development (e.g., critical windows), and exposure assessment methods (i.e., in epidemiology studies). • Specificity and sensitivity of the endpoint for evaluating the health effect in question (e.g., functional measures can be more sensitive than organ weights). • Populations or species, including consideration of potential susceptible groups or differences across lifestage at exposure or endpoint assessment. • Toxicokinetic information explaining observed differences in responses across route of exposure, other aspects of exposure, species, or lifestages. The interpretation of consistency will emphasize biological significance, to the extent that it is understood, over statistical significance (see additional discussion in **Section 9.4**). Statistical significance from suitably applied tests (this may involve consultation with an EPA statistician) adds weight when biological significance is not well understood. Consistency in the direction of results increases confidence in that association even in the absence of statistical significance. It may be helpful to consider the potential for publication bias and to provide context to interpretations of consistency.[a]

Consideration	Description of the consideration and its application in IRIS syntheses
Strength (effect magnitude) and precision	Description: Examines the effect magnitude or relative risk, based on what is known about the assessed endpoint(s), and considers the precision of the reported results based on analyses of variability (e.g., confidence intervals; standard error). This may include consideration of the rarity or severity of the outcomes. Application: Syntheses will analyze results both within and across studies and may consider the utility of combined analyses as appropriate (e.g., meta-analysis, meta-regression). Note that a synthesis includes consideration of null (or negative) as well as positive results. While larger effect magnitudes and precision (e.g., $p < 0.05$) help reduce concerns about chance, bias, or other factors as explanatory, syntheses should also consider the biological or population-level significance of small effect sizes (see **Section 9.4**).
Biological gradient/ dose-response	Description: Examines whether the results (e.g., response magnitude; incidence; severity) change in a manner consistent with changes in exposure (e.g., level; duration), including consideration of changes in response after cessation of exposure. Application: Syntheses will consider relationships both within and across studies, acknowledging that the dose-response (e.g., shape) can vary depending on other aspects of the experiment, including the biology underlying the outcome and the toxicokinetics of the chemical. Thus, when dose-response is lacking or unclear, the synthesis will also consider the potential influence of such factors on the response pattern.
Coherence	Description: Examines the extent to which findings are cohesive across different endpoints that are related to, or dependent on, one another (e.g., based on known biology of the organ system or disease, or mechanistic understanding such as toxicokinetic/dynamic understanding of the chemical or related chemicals). In some instances, additional analyses of mechanistic evidence from research on the chemical under review or related chemicals that evaluate linkages between endpoints or organ-specific effects may be needed to interpret the evidence. These analyses may require additional literature search strategies. Application: Syntheses will consider potentially related findings, both within and across studies, particularly when relationships are observed within a cohort or within a narrowly defined category (e.g., occupation; strain or sex; lifestage of exposure). Syntheses will emphasize evidence indicative of a progression of effects, such as temporal- or dose-dependent increases in the severity of the type of endpoint observed. If an expected coherence between findings is not observed, possible explanations should be explored including the biology of the effects as well as the sensitivity and specificity of the measures used.

Consideration	Description of the consideration and its application in IRIS syntheses
Mechanistic evidence related to biological plausibility	Description: There are multiple uses for mechanistic information (see **Section 9.2**) and this consideration overlaps with "coherence." This examines the biological support (or lack thereof) for findings from the human and animal health effect studies and becomes more impactful on the hazard conclusions when notable uncertainties in the strength of those sets of studies exist. These analyses can also improve understanding of dose- or duration-related development of the health effect. In the absence of human or animal evidence of apical health endpoints, the synthesis of mechanistic information may drive evidence integration conclusions (when such information is available). Application: Syntheses can evaluate evidence on precursors, biomarkers, or other molecular or cellular changes related to the health effect(s) of interest to describe the likelihood that the observed effects result from exposure. This will be an analysis of existing evidence, and not simply whether a theoretical pathway can be postulated. This analysis may not be limited to evidence relevant to the PECO but may also include evaluations of biological pathways (e.g., for the health effect; established for other, possibly related, chemicals). The synthesis will consider the sensitivity of the mechanistic changes and the potential contribution of alternate or previously unidentified mechanisms of toxicity.
Natural experiments	Description: Specific to epidemiology studies and rarely available, this examines effects in populations that have experienced well-described, pronounced changes in chemical exposure (e.g., lead exposures before and after banning lead in gasoline). Application: Compared to other observational designs, natural experiments have the benefit of dividing people into exposed and unexposed groups without them influencing their own exposure status. During synthesis, associations in *medium* and *high* confidence natural experiments can substantially reduce concerns about residual confounding.

PECO = populations, exposures, comparators, and outcomes.

[a]Publication 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). When evidence of publication bias is present for a set of studies, less weight may be placed on the consistency of the findings for or against an effect during evidence synthesis and integration (see **Section 11.1**). Publication bias is discussed in more detail in **Section 9.4.2**.

In addition, to the extent the data allow, the syntheses will discuss analyses relating to potential susceptible populations,[12] based on knowledge about the health outcome or organ system affected, demographics, genetic variability, lifestage, health status, behaviors or practices, social determinants, and exposure to other pollutants (see **Table 9-2**). This information will be used to

[12]Various terms have been used to characterize populations that may be at increased risk of developing health effects from exposure to environmental chemicals, including "susceptible," "vulnerable," and "sensitive." Further, these terms have been inconsistently defined across the scientific literature. The term susceptibility is used in this Handbook to describe populations at increased risk, focusing on biological (intrinsic) factors, as well as social and behavioral determinants that can modify the effect of a specific exposure. However, certain factors resulting in higher exposures to specific groups (e.g., proximity, occupation, housing) may not be analyzed to describe potential susceptibility among specific populations or groups.

describe potential susceptibility among specific lifestages, populations, or subgroups (see **Section 12.1**) summarizing across evidence streams and hazards to inform hazard identification and dose-response analyses.

Table 9-2. Individual and social factors that may increase susceptibility to exposure-related health effects

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

DNA = deoxyribonucleic acid.

9.1.1. Analysis and Synthesis of Evidence Requires Scientific Judgment

It is important to stress that the process of developing a synthesis of evidence does not involve counting the number of "positive" and "negative" studies, nor is this a paragraph by paragraph summary of each study. This chapter is designed to provide the reviewer with the basic principles to systematically consider and discuss the influence of the risk of bias and sensitivity factors identified in the evaluation of individual studies (see **Chapter 6**), in conjunction with the results observed within a set of studies, to draw interpretations regarding the evidence pertaining to the health effect under review (see **Table 9-1**). Thus, the results of individual studies are interpreted, considering specific study limitations, including the direction of potential biases if they can be reasonably anticipated.

Generally, based on the evaluation of individual studies (see **Chapter 6**), the synthesis should be based primarily on studies of *medium* and *high* confidence (when available) regardless of the reported results (i.e., null, negative or positive results), with *high* confidence studies receiving the most weight and *low* confidence studies occupying a supporting role; *uninformative* studies are not discussed. *Low* confidence studies may be considered if few or no studies with higher confidence are available, or if the study designs of the *low* confidence studies address notable uncertainties or understudied aspects (e.g., developmental lifestages) in the set of *high* or *medium* confidence studies on a given health effect. If *low* confidence studies are used, then a careful

examination of risk of bias and sensitivity with potential impacts on the evidence synthesis conclusions should be included in the discussion.

As previously described, these syntheses will articulate the strengths and the weaknesses of the available evidence organized around the applied Bradford Hill considerations described in **Table 9-1**, as well as issues that stem from the evaluation of individual studies (e.g., concerns about bias or sensitivity). When possible, results across studies should be compared using graphs and charts or other data visualization strategies. Visualizations should generally include information on study confidence. The analysis will typically include examination of results stratified by any or all of the following: study confidence classification (or specific issues within confidence evaluation domains, such as low vs. high sensitivity), population or species, exposures (e.g., level, patterns [intermittent or continuous], duration, intensity), and other factors that may have been identified in the refined evaluation plan (e.g., sex, lifestage, or other demographic). The number of studies and the differences encompassed by the studies will determine the extent to which specific types of factors can be examined to stratify study results. Additionally, if supported by the available data, additional analyses across studies (such as meta-analysis) may also be conducted for both the human and animal evidence syntheses.

9.2. ANALYSIS AND SYNTHESIS OF HUMAN (PRIMARILY EPIDEMIOLOGY) 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. A suggested strategy is to compare results by the degree of sensitivity and potential direction of bias.

Grouping studies by the level and variation or range of exposure experienced by the study populations may explain a set of seemingly inconsistent results or provide evidence of a biologic gradient or exposure-response relationship. Associations among populations exposed to lower levels may be null or highly variable with wide confidence intervals (CIs), while associations from studies at higher levels may 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 based on 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 9-1** 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 may 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)]. It is important to keep in mind that a nonmonotonic curve in an individual study may be biologically plausible and informative (Vandenberg et al., 2012; Wigle and Lanphear, 2005).

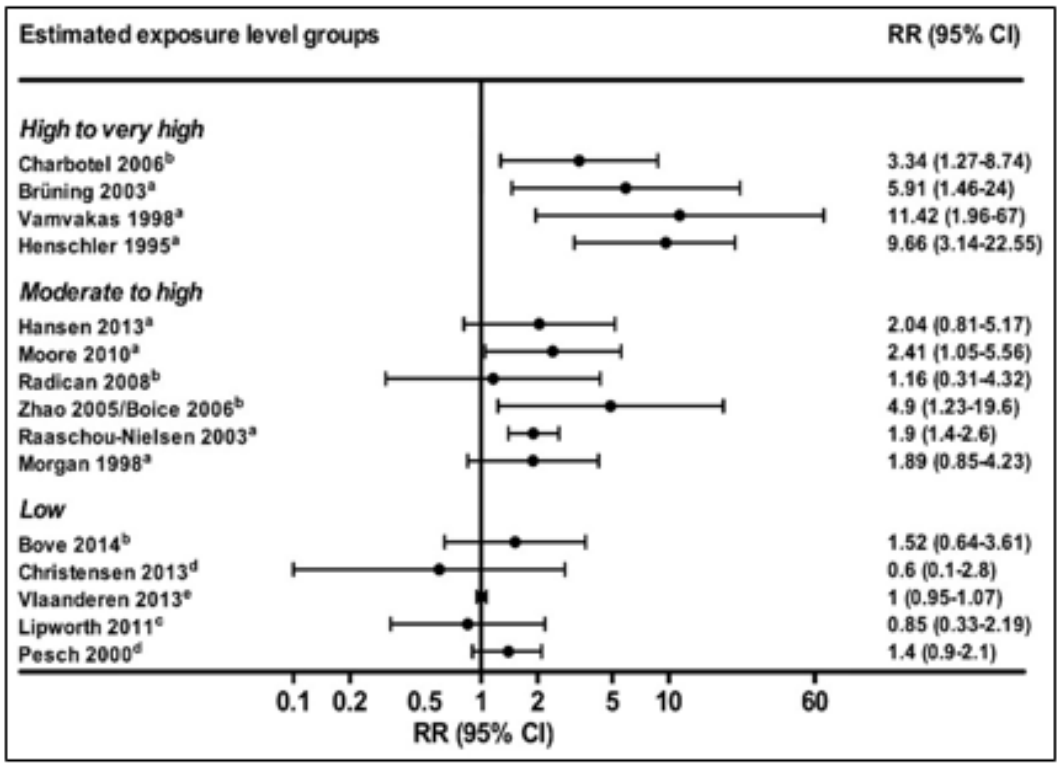


Figure 9-1. Trichloroethylene (TCE) and kidney cancer: stratification by exposure level (U.S. EPA, 2011b).

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 strength or magnitude 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. "Strength" encompasses not only magnitude of the association, but also precision in the effect measure estimates. Higher precision, as reflected by narrow confidence bounds or smaller standard errors (SEs), also adds confidence in the observed association; as described previously, however, precision of individual studies may 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) reduces 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. While 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 was 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 > 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.

In general, syntheses should include a discussion of outstanding questions or data gaps in the evidence at their conclusion.

9.3. ANALYSIS AND SYNTHESIS OF ANIMAL EVIDENCE

Paralleling the approach for human evidence, the syntheses of the available animal evidence incorporates the evaluations of confidence in study methods, considering specific concerns regarding reporting quality, risk of bias, or sensitivity in individual studies (see **Section 6.3**), as well as across the set of studies on an outcome(s). The study confidence is combined with analyses of the results from individual studies and sets of studies to assess and describe the evidence most relevant to the considerations summarized in **Table 9-1**. Additional analyses may also be conducted, such as a summary estimate across studies (see expanded discussion in **Section 9.4**).

In addition to the considerations common across the human and animal evidence (see **Section 9.1**), some examples of questions more pertinent to the animal evidence synthesis include:

- *Exposure range*: 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 toxicokinetics 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?

- *Toxicokinetics*: Can differences in response be explained by differences in toxicokinetics (e.g., metabolism) across different animal species?[13] (This factor may also be considered within the context of differences in response seen by route of exposure.) The discussion of the evaluation of absorption, distribution, metabolism, and excretion (ADME) information (see **Section 5**) can be considered in this analysis.

- *Study evaluation:* specifically, the sensitivity of individual studies, including the timing and duration of exposure, as well as the timing and conduct of endpoint evaluations (see

[13]Although toxicokinetics may also differ due to differences in age, sex, or strain, chemical-specific data describing such differences are rarely available.

Section 6.3): When the results for a specific outcome differ across studies, are the differences reasonably explained by the timing or duration of exposure based on what is known about the outcome of interest, or by the sensitivity of the specific methods used to evaluate the outcome?

- *Endpoint comparisons*: Are there notable differences in the specific endpoints evaluated across studies, or in the way those endpoints were assessed (study evaluations may highlight some of the latter differences; see **Section 6.3**)? For some health effects, the relevant endpoints evaluated in animal studies can be highly heterogeneous. 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).

The analysis of the animal evidence emphasizes interpretations regarding the consistency of the findings across studies, the magnitude and dose-response dependency of the results, and the coherence of related effects across the database.

The consistency of results considers if the results were replicable across studies performed by different laboratories, as well as whether similar results were observed across studies of different design (e.g., species, strain and/or sex; age at exposure or endpoint analysis; exposure route, administration method, or surrogate measurement).[14] Consistent results across species or routes of exposure substantially increase confidence that similar results would occur in all experimental animals and experimental paradigms, increasing confidence that the findings are not attributable to chance. It is important to emphasize that the study evaluations (see **Section 6.3**), including the expected impact of the limitations identified, are considered in the evaluation of consistency.

Another important consideration in the analysis of the experimental animal evidence is the evaluation of the pattern (e.g., dose-response) and strength of the observed effects. Trend tests (conducted by U.S. Environmental Protection Agency [EPA] if an appropriate test is not reported by authors) are preferred for use in assessing the dose-dependency of results within studies (and possibly, across closely related studies, if appropriate). Note that consideration should be given to the exposure spacing in studies providing information related to understanding the potential dose-response relationship. Dose-response patterns do not necessarily need to exhibit monotonicity; however, a lack of monotonicity should be discussed and examined in the context of the data available from studies of similar design (e.g., endpoints; exposure timing) and, possibly, from related chemicals or established knowledge of the biological changes associated with the observed effects (aka, "biological understanding").

[14]Physiologically based pharmacokinetic (PBPK) models, if available, may facilitate comparing studies that used different exposure routes (inhalation vs. oral) or measures of exposure, such as biomarkers that might be back-calculated to environmental exposures (see **Section 6.5**). Toxicokinetic differences (e.g., expression or activity of important enzymes) may exist across sexes and ages, and this should be considered in the analysis of consistency, when applicable.

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 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 as well as the sensitivity and specificity of the measures used.

9.4. ADDITIONAL CONSIDERATIONS AND ANALYSES THAT INFORM CONSISTENCY

9.4.1. Role of Tests of Statistical Significance in Analyzing Evidence

Statistical significance testing is an important tool for supporting a decision that there is a demonstrable effect, especially when biological significance (U.S. EPA, 2002b) of an outcome is uncertain or unclear (e.g., no suitable normal range). A pattern of statistically significant results for an effect (or related effects), of similar size, across comparable, well-designed studies generally increases confidence that the effect is associated with the exposure. Whenever a database includes other comparable, well-designed studies without statistically significant results, the evaluation of consistency across all results must also be part of the overall weight of evidence. This section highlights aspects of statistical significance relevant to this evaluation, especially that the lack of statistical significance in the presence of an elevated effect estimate does not necessarily rule out an association. 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 using the approaches described in **Sections 9.1–9.3** should include both effect estimates 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 connection to biological significance [e.g., Rothman (2010)].

- *P*-values[15] by themselves 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 may not show statistical significance for this effect. Support for the observation can come from examining patterns in results across all studies that report data for the same endpoint, considering differences in methods (e.g., relative exposure ranges, duration of exposure, age of test animals), variability of effects, and coherence with related evidence (also see **Sections 9.1–9.3**).

    - In addition, 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 not infrequent 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 may 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 used[16]—are also important to address when distinguishing between "conflicting" and "differing" evidence (see **Table 9-1**).

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 can increase confidence in findings for an outcome (see **Section 9.4.2** for further information).

9.4.2. Additional Statistical Analyses: Individual Studies and Meta-Analysis

Additional statistical analyses of individual studies or across a set of studies may increase precision in estimating the magnitude of the association and provide support that either an association does exist across study results or that an association is not supported across all study

[15]A *p*-value is the probability under a specified statistical model that a statistical summary of the data would be greater than or equal to that observed [Wasserstein and Lazar (2016)], such as for the difference in incidence between a treated and a control group, assuming binomial variability.

[16]Sometimes it may 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.

results; such decisions would generally apply only to *medium* and *high* confidence studies and would be thoroughly reviewed for the value added to the assessment before proceeding. One example that tends to be overlooked by many investigators who generate individual studies is trend testing to evaluate response patterns across treated groups. In general, detection of a dose-response trend across all treated groups directly addresses this component of causality, while multiple pairwise comparisons with control responses less efficiently consider each dose group one at time. When trend tests are not presented in published studies, such as those provided in National Toxicology Program (NTP) bioassay reports (or details of the trend test used are not provided), EPA calculates trend tests (using summary statistics in published studies, such as means and variance estimates) as necessary. Overall, a variety of other statistical analyses may be more suitable than those reported by the original authors, but that may vary by study design (e.g., repeated measures) or complexity of the dose-response (e.g., competing toxicity, pharmacokinetic considerations in different dose ranges) and are beyond the scope of the handbook to list.

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. Where applicable, the preferred statistical method for synthesizing evidence within a set of studies that may report positive as well as null (or negative) results is some type of statistical meta-analysis. This may 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.[17] Meta-analyses also may help to demonstrate that all the studies considered, rather than just one influential study, contributed to the evidence synthesis conclusions. Metaregression, examining the influence of various factors on results across studies, could be used in some circumstances (e.g., with sufficient numbers of studies; see **Section 12.2** for further discussion). For evidence synthesis, however, a single effect estimate may not be needed, as the focus is on examining patterns and variability (consistency) across studies. The decision to conduct a meta-analysis of a specific outcome in a set of studies is based on an evaluation of the potential contribution of such an analysis (e.g., explicit weighting of studies based on variance, or estimation of a more precise estimate than can be seen in a single study).

If a meta-analysis or other comprehensive analysis is conducted by EPA or by others, 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). Study confidence, exposure levels, exposure route, species, lifestage, and numerous other considerations may contribute to the observed results and to heterogeneity among studies. Statistical significance or other criteria based on the study results

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

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). If a meta-analysis is conducted, the synthesis must also include a discussion of the results from studies that did not contribute to the combined analysis (because, for example, their results could not be converted into the necessary form).

The validity of a meta-analysis depends on decisions regarding inclusion and exclusion of studies, evaluation of study methods, and decisions regarding data extraction. Additional considerations for conducting a meta-analysis include:

- What could the analysis contribute to the synthesis of the evidence?
- What factors, if any, should be used to stratify a meta-analysis?
- 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.

9.4.3. Reporting or Publication Bias

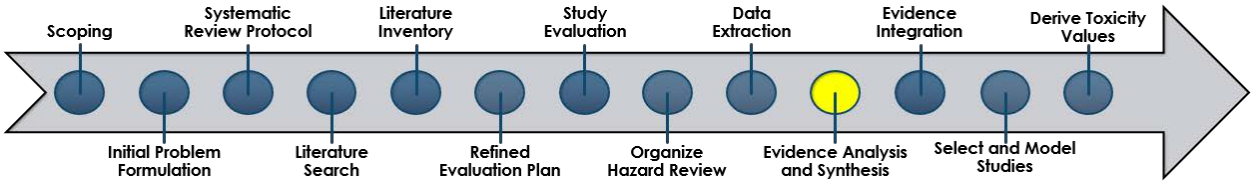
The potential influence of publication bias is another point that may be considered during evidence synthesis. Publication bias involves the influence of the direction, magnitude, or statistical significance of the results on the likelihood of a paper being published; it can result from decisions made, consciously or unconsciously, by study authors, journal reviewers, and journal editors (Salami and Alkayed, 2013; Matthews et al., 2011). Of concern are results from small studies; small "positive" studies are more likely to be published than small "negative" studies. Thus, an evaluation of publication bias can sometimes provide particularly useful context to evaluations of consistency when the evidence on an outcome is "weak."

There are approaches to minimize the impact of publication bias or detect its presence (Parekh-Bhurke et al., 2011). The identification of intended study outcomes that were not subsequently reported in publications may be accomplished by searching registries of planned or ongoing studies. Publication bias may be minimized if a comprehensive, thorough literature search strategy is designed to identify unpublished or gray literature (e.g., meeting abstracts and proceedings, technical reports) and to include foreign language articles. Finally, there are some albeit imperfect analytical tools that may help to detect the presence of publication bias, including tests of small study effects, selection model approaches, and tests of excess significance (Ioannidis et al., 2014).

A potential conflict of interest (COI) by one or more authors of a study may contribute to reporting or publication bias (Guyatt et al., 2011b). While IRIS does not formally include COIs as a component in the evaluation of bias and sensitivity of study outcomes, funding source and a report of a COI by the authors can be noted for a study in Health Assessment Workspace Collaborative (HAWC). When there is evidence that a conflict of interest is may be present, a more careful

assessment of the consistency of study results, publication and reporting bias may be merited for a health effect.

10. ANALYSIS AND SYNTHESIS OF MECHANISTIC INFORMATION



ANALYSIS AND SYNTHESIS OF MECHANISTIC INFORMATION

Purpose

- To consider the available mechanistic data in light of other identified hazard-specific information to inform evidence integration conclusions.

Who

- The assessment team, in consultation with appropriate disciplinary workgroup(s) and subject matter experts.

What

- Draft mechanistic synthesis sections for selected health effects describing the assessment-specific mechanistic questions or issues, as well as the interpretations drawn from the mechanistic data.

IRIS assessments evaluate mechanistic data to inform hazard identification determinations regarding the biological plausibility of human and animal data, to identify susceptible populations and lifestages, and to inform dose-response relationships. Mechanistic studies include a variety of designs (i.e., in vitro, in vivo using various routes of exposure, ex vivo, and in silico) and report measurements that inform the biological or chemical events associated with toxic effects but are not generally considered adverse outcomes on their own (there are exceptions; for example, hormone level changes are mechanistically relevant for many outcomes and may also be considered adverse outcomes themselves). The IRIS Program considers mechanistic information to be important to assessing the potential human health hazards and dose-response relationships of chemicals found in the environment. As such, consideration of mechanistic data is incorporated throughout assessment development, as described in this Handbook. Nevertheless, incorporating mechanistic studies into a systematic review framework remains challenging. Challenges include screening large numbers of diverse studies efficiently; developing transparent and reproducible criteria for identifying the most informative mechanistic studies; the lack of well-developed systematic review tools to assess the internal validity of in vitro and in silico studies; transparently

comparing and judging (large) sets of highly heterogeneous models, exposure paradigms, and outcomes relevant to a given mechanistic topic during evidence synthesis; and underdeveloped structured frameworks to guide integration of mechanistic information with human and animal health effects evidence. This chapter presents important concepts and example approaches for organizing mechanistic evidence to inform hazard identification. This chapter is based largely on *The Guidelines for Carcinogenic Risk Assessment* (U.S. EPA, 2005b), which contains material generally applicable to both cancer and noncancer health effects, particularly Section 2.3 ("Analysis of Other Key Data") and Section 2.4 ("Mode of Action—General Considerations and Framework for Analysis") of that document. These sections should be reviewed before reading further. It is expected that the approaches discussed below will be further clarified and refined based on application to specific chemical assessments (and consideration of review comments) and broader discussions among experts in environmental health and systematic review, for example, at workshops held at the National Academy of Sciences that focused on the systematic review of mechanistic data (http://dels.nas.edu/Upcoming-Workshop/Strategies-Tools-Conducting-Systematic/AUTO-5-32-82-N) and evidence integration (http://dels.nas.edu/Upcoming-Event/Evidence-Integration-Workshop/AUTO-0-96-15-Q).

10.1. PREPARATION FOR THE MECHANISTIC ANALYSIS

Determining the areas of focus for the mechanistic analysis is a stepwise process and continues throughout assessment planning and development, as described in **Section 1.1** (overview of the scoping process), **Section 2.2** (assessment plan), **Section 4.3** (literature inventories), **Chapter 5** (refined evaluation plan), **Section 6.6** (study evaluation, when individual study evaluation is warranted), **Chapter 7** (organizing the hazard review), and **Chapter 11** (evidence integration). At the end of this chapter, a quick reference outline has been provided to summarize the steps involved in considering mechanistic data, many of which are performed concurrently with other sections of the assessment.

10.1.1. Identification and Screening of Mechanistic Studies

Decisions on whether and how to conduct specific mechanistic evaluations begin during scoping and problem formulation analyses performed as part of preparing the IRIS Assessment Plan (IAP) (see **Chapter 2**, Problem Formulation and Development of an Assessment Plan). It is important to review and assess the likely impact of potentially controversial mechanistic issues (e.g., evidence a chemical is mutagenic, the human relevance of α2u globulin) on assessment conclusions early in the process. This involves an initial review of existing mechanistic analyses as well as information regarding the absorption, distribution, metabolism, and excretion (ADME)/toxicokinetics (TK) of the chemical and possibly other related chemicals in the same class (read-across). The early identification of pre-defined mechanistic analyses will help to frame the approach used for conducting and organizing a preliminary literature survey ("evidence mapping").

These steps are already described in **Chapters 2 and 4**, but a brief review is provided here with additional considerations to ensure an efficient process for tagging and screening these potentially large mechanistic databases.

Literature Identification

To review, typically, a broad chemical-name-based search is implemented to ensure that the mechanistic evidence is fully identified and available for consideration, although other approaches may be used (e.g., preliminary surveys based on comprehensive reviews or prior assessments). Regardless, the IAP should present decisions on how mechanistic information will be surveyed. When specific mechanistic analyses are identified as critical for an assessment during scoping, these analyses can be described in the specific aims of the assessment plan, and the types of studies considered pertinent can be included in the PECO criteria. In most cases, however, it will not be possible to fully describe the analysis plan for mechanistic evidence until the assessment is further along. Thus, mechanistic studies are most commonly tagged as "potentially relevant supplemental material" during screening (described in **Section 4.2.1**) and organized into literature inventories (described in **Section 4.3.3**) to allow for straightforward access at later stages of assessment development. As described previously, these inventories typically include the type of relevant health outcomes (e.g., hepatic, neurological) and some details on the model and experimental design, and they may categorize the mechanistic studies by relevant biological pathway affected (e.g., receptor activation/binding activities) or sort them using an organizational construct (e.g., key events for a mode of action [MOA] and/or adverse outcome pathway [AOP]; key characteristics of carcinogens). In addition, some assessments may add supplemental searches for capturing information nonspecific to the chemical being assessed (e.g., on relevant mechanisms, biology, or related chemicals) that were not identified when the original literature search was conducted. These decisions often occur during a second phase of screening conducted after the initial literature search and further refinements to the assessment analysis plan (see **Chapter 5**), when the full scope of mechanistic analysis that needs to be conducted is clearer.

Literature Screening: Tips for Tagging Mechanistic Evidence

A typical title and abstract (TIAB) screening form will have the following response options for assessing PECO relevance: a "yes," "no," "tag as potentially relevant supplemental material," or an "unclear" tag. During TIAB screening, mechanistic studies that meet the PECO criteria (when applicable) are tagged as "yes," and additional screening questions will ask about the evidence type (human, animal, in vitro/ex vivo/in silico). More typically, during TIAB screening, mechanistic studies are tagged as "potentially relevant supplemental material," with additional screening questions asked to further categorize the supplemental material (see **Table 2-2** and **Figure 4-4**).

Screening questions that categorize mechanistic information into a given construct (e.g., key characteristics; key events for an MOA and/or AOP) may also be asked at the full-text level when more complete study content information is available. There is not a right or wrong approach for

when to conduct a detailed inventory of mechanistic information and often the decisions of when to survey this information are made for pragmatic reasons. For example, the time to screen studies at TIAB level is increased when screeners are asked to apply more tags. So, for projects with many studies to screen, teams may want to wait and tag studies during a second phase of TIAB screening or at the full-text level. In other cases, the TIAB screeners may not have the content knowledge to do detailed tagging.

Refining the Scope and Purpose of the Mechanistic Analyses

Decisions on whether and how to conduct mechanistic evaluations will depend not only on scoping and problem formulation, but also on hazard characterization signals from the human and animal evidence streams (see **Chapter 9**). While mechanistic analyses can provide critical information for hazard identification and dose-response, a comprehensive mechanistic evaluation (which may include an MOA analysis) is not necessarily conducted for every potential hazard discussed in the assessment. The scope, complexity, and depth of the mechanistic analyses will vary with the level of emphasis placed on the health effect for evidence synthesis. For some health effects, it may become apparent that a high-level survey of mechanistic information (possibly limited to prominent reviews or existing assessments) will be sufficient and a detailed study-by-study analysis would have limited influence on assessment conclusions and would therefore be an inefficient use of resources. For example, effort spent on an in-depth analysis of mechanisms associated with a health effect that is 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 may inform the shape of the dose-response curve (i.e., if the available human data have substantial quantitative uncertainties). The same may be true for animal and human outcomes with well-accepted mechanistic associations (e.g., dioxin as an aryl hydrocarbon receptor agonist), where a broad overview can provide the appropriate context. The literature inventories (see **Section 4.3**) can 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 effects literature where only mechanistic studies are available to inform hazard.

Table 10-1 summarizes the stages of the assessment workflow where mechanistic information will be identified and some examples of key questions. In addition, some areas of uncertainty in the overall hazard identification and dose-response assessment that may be addressed by mechanistic information are summarized in **Table 10-2**. These considerations may provide rationale for focusing the mechanistic analyses on key areas specific to the assessment. It is important to note that none of the approaches presented here represent rules or criteria for prioritization; every database will have unique considerations, and generalities should not be interpreted as immutable rules.

Table 10-1. Preparation for the analysis of mechanistic evidence

Assessment stages of identifying mechanistically relevant information	Examples of evidence to review and key considerations
Scoping and problem formulation	• For the chemical under review, identify existing chemical-specific analyses and MOAs from other agency assessments or review articles. If summary information is lacking, are there structurally similar chemicals that are better studied mechanistically? • Are there indications that a specific mechanistic analysis will be warranted? For example, are there areas of scientific controversy or predefined assessment questions that will require a mechanistic evaluation (e.g., a potential mutagenic MOA)? o If so, consider whether additional, targeted literature searches would be informative. • What is the active moiety of the agent? Are there metabolites that should be considered? Are there indications that the purity is critically important? Is the chemical endogenously produced?
Literature inventory of toxicokinetic, ADME, and physicochemical information	• Based on ADME differences across species, is there evidence that suggests a lack of relevance of the animal exposure scenarios to human situations? Is there evidence that the active moiety would not be expected to reach the target tissue(s) in some species? • Are there metabolic pathways involved that may indicate greater sensitivity at a particular lifestage or in susceptible human populations (see **Table 9-2** for examples)? • If a validated PBPK model is available, revisit any decisions to focus on specific routes of exposure and consider the use of alternative exposure markers.
Literature inventories of human, animal, and mechanistic information (including all in vitro and in silico studies)	• Which human health hazards (both cancer and noncancer) appear to be well studied in the mechanistic inventory? For cancer, which key characteristics of carcinogens are indicated by the database? • Are there mechanistic endpoints identified from human and animal studies meeting PECO criteria that could be added to the mechanistic inventory?

Assessment stages of identifying mechanistically relevant information	Examples of evidence to review and key considerations
Human and animal evidence syntheses	• Evidence that may be used to explain or resolve specific uncertainties includes: o Effects that differ across populations (e.g., species; sex; strain) o Evidence (e.g., biological precursors) in humans or animals that may provide a mechanistic link between the exposure and the observed outcome o Animal effects that may not be relevant to humans o Susceptible populations and lifestages (e.g., animal strain; human demographic)

PBPK = physiologically based pharmacokinetic.

Table 10-2. Example considerations that can focus the mechanistic analysis and synthesis

Areas of uncertainty that may be addressed by mechanistic information	Considerations and examples of areas for mechanistic focus
Database incompleteness based on literature inventories of human, animal, and mechanistic information	• If there are mechanistic toxicity data on organ systems or health hazards that were not examined by human or animal studies meeting the PECO criteria, evidence mapping or similar approaches can highlight these knowledge gaps and help determine whether a separate synthesis of this evidence is necessary.
Inconsistency within the human and animal evidence	• For the health effects of potential concern, a mechanistic evaluation may be warranted to inform questions regarding the consistency of the available human or animal studies. • Heterogeneous results across different animal species or human populations might be explainable by evidence that a mechanism is only relevant in certain species (e.g., saccharin exposures causing bladder calculi and cancer only in male rats), or that multiple mechanisms are operant (e.g., evidence demonstrating that certain populations cannot metabolize a reactive metabolite; evidence that variability in gene expression correlates with variability in response).
Questions regarding biological plausibility and coherence	Mechanistic information can strengthen or weaken the evidence for an association between exposure and the health effect based on existing knowledge of how the health effect develops (biological plausibility) and the relatedness of outcomes within and across health effect categories (coherence). • Observations of mechanistic changes that are associated with the health outcome in question can increase the strength of the judgments, particularly when the changes are observed in the same exposed population presenting the health effect. • Biological understanding (general knowledge of biological changes associated with the observed effects) or strong mechanistic support (e.g., a shared key event) for linkages across outcomes can increase the strength of the evidence when changes are related. 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). • The plausibility of an association observed in human or animal studies may be diminished if expected findings are not apparent in mechanistic evidence, or an expected pattern among biologically linked health effects is not observed. • If the mechanistic evidence is conflicting or is otherwise insufficient to provide a mechanistic explanation for an association (or lack thereof), this will not change the interpretation of the results from sets of human and/or animal studies.

Areas of uncertainty that may be addressed by mechanistic information	Considerations and examples of areas for mechanistic focus
Questions regarding the human relevance of findings in animals	• Note that in the absence of sufficient MOA information, effects in animal models are assumed to be relevant to humans (U.S. EPA, 2005b, 1998, 1991). • Observations of mechanistic changes in exposed humans that are coherent with mechanistic or toxicological changes in experimental animals (and that are interpreted to be associated with the health outcome under evaluation) strengthen the human relevance of the animal findings. o Evidence of biological precursors that link the exposure to the observed outcome in humans and animals strengthens human relevance. • 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, this can support the view that the animal response is not relevant to humans. o Focusing on health effects that differ across populations (e.g., by species, sex, strain) can provide mechanistic explanations for these differing effects that can strengthen relevance or a lack of relevance to humans.
Potential susceptibility	Mechanistic understanding of how a health outcome develops, even without a full MOA, can help to identify susceptible population groups. Hazard Identification: • 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 (see **Section 11.2**). • Differences in susceptibility may be explained by an analysis of toxicokinetic or toxicodynamic differences across lifestages or populations (e.g., animal strain; human demographic). Dose-Response Analysis: • Evidence indicating the presence of a sensitive population or lifestage in humans can inform selection of studies for quantitative analysis, e.g., by prioritizing studies including these populations (see **Chapter 12**). • If studies directly addressing the identified susceptibilities are unusable for quantitative analysis, susceptibility data may still support refined human variability uncertainty factors or probabilistic uncertainty analyses (see **Section 13.4.1**).

Areas of uncertainty that may be addressed by mechanistic information	Considerations and examples of areas for mechanistic focus
Questions regarding understanding of mechanism(s) that may affect dose-response decisions	Chemical-specific mechanistic information or established biological understanding that describes how effects develop may help clarify the exposure conditions expected to result in these effects. This can not only increase the strength of the hazard conclusions, it can also optimize dose-response decisions, particularly if the selection of critical parameters for dose-response modeling is uncertain, or the data amenable to dose-response analysis are weak or only at high exposure levels. MOA inferences can support the use of: • Specific dose-response models, e.g., o Models integrating data across several related outcomes o Models that incorporate toxicokinetic knowledge • Proximal measures of exposure, e.g., external vs. internal metrics • Improved characterization of responses, e.g., o The use of well-established precursor events linked qualitatively or quantitatively to apical health effect(s) in lieu of direct observation of apical endpoints o The combination of related outcomes, such as benign and malignant tumors in the same tissue or tumors in different tissues that operate through the same MOA

Section 10.4.2 introduces several approaches for both organizing and synthesizing mechanistic data (i.e., MOAs, AOPs, ten key characteristics). They are not mutually exclusive, and one or more approaches may be used in an assessment. They can be used in concert to identify, organize, analyze, and synthesize mechanistic information in a way that increases the transparency of the assessment. It is important to keep in mind that the evaluation of mechanistic evidence is a phased process and the specifics of the approaches will nearly always differ across chemical assessments. Because the analysis and application of mechanistic information differs for each assessment, it is possible that only a subset of information in this chapter will be applicable to a given assessment. Therefore, a general approach to synthesizing mechanistic data is outlined in the following sections, as shown in **Figure 10-1**.

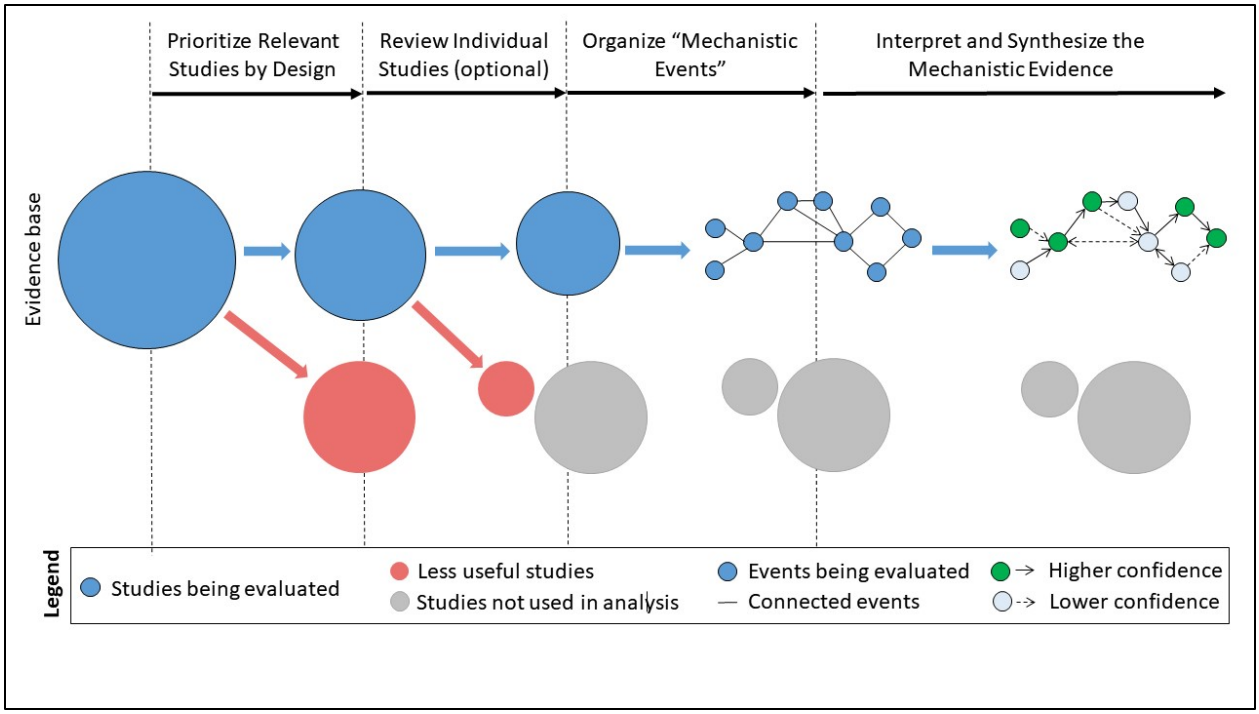


Figure 10-1. Schematic overview of the process for evaluating mechanistic evidence from a large evidence base.

10.2. PRIORITIZATION AND EVALUATION OF MECHANISTIC STUDIES

10.2.1. General Considerations for Prioritization

Once the general purpose and scope of a review of the mechanistic information has been determined and indicates the need for an analysis beyond a summary of secondary sources, the next stage is to develop a plan on prioritizing the most relevant mechanistic evidence in the mechanistic information inventory (see **Section 4.3.3**) for evaluation. 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. For many chemicals, the number of mechanistic studies is quite large and pragmatic approaches need to be taken to narrow the scope of studies that require detailed summarization and evaluation at the individual study level. Some events may be well-accepted scientifically and do not require a detailed analysis of individual studies. A subset of the most relevant individual studies may require detailed summarization and evaluation for chemicals with little or no evidence from epidemiological studies or animal bioassays when (1) the science is less established, (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.

As introduced above and summarized in **Table 10-1**, the prioritization of mechanistic information begins with decisions made on which health effects and associated outcomes and endpoints, exposure levels, and lifestage(s) are included in the hazard synthesis. The next step of

prioritization is to identify the most mechanistically relevant studies based on the extent to which the reported endpoints, as well as the experimental models, assays, and study designs used to experimentally evaluate these endpoints, inform the identified hazard questions of interest. The considerations in the list below can help further refine that set of studies to those best suited to answering these questions.

- Key biological pathway(s) of interest (e.g., key events identified during problem formulation or based on the mechanistic literature inventory). For example, experiments that challenge the essentiality of a biological pathway of interest or presumed key event(s) are typically high priority. Examples of such experiments include studies in knockout mice, experiments introducing chemical inhibitors of target receptors, and animal studies incorporating a surgical blockade of signaling events.

- Studies evaluating effects in target tissues versus nontarget tissues. There are notable exceptions (e.g., analysis of endocrine activity), and in some cases, the critical target tissues may be difficult to pinpoint (e.g., multisite carcinogenicity; widespread immune dysfunction).

- Model systems that are better outcome predictors (e.g., species, sex, or culture systems known to be representative models for the health effect). Given differences in biological complexity across models, in vivo exposures are prioritized over in vitro exposures, and primary cells are generally favored over immortalized cell lines. However, it should be noted that this is not a rule and may change depending on context; many in vitro assays are designed to be sensitive for detecting an endpoint that is otherwise difficult to observe in vivo or in primary cells in vitro. Special consideration may be given to assays with unique modifications that increase sensitivity to a particular effect or more closely mirror human biology. For example, the Ames assay uses bacterial strains engineered to be sensitive to mutagens, particularly with the addition of a rat liver microsomal fraction to enable metabolic activation.

- The exposure paradigms used, including route and dose or concentration level tested. For in vivo studies, routes may be prioritized based on relevancy to the assessment-specific scope and exposure scenarios in humans and animals. However, the literature inventory can also identify important mechanistic information obtained by other routes of exposure that may not be similar to environmental exposures (e.g., intratracheal instillation, subcutaneous injection) but can help establish biological plausibility. The toxicokinetic knowledge and the exposure route will be considered in the context of whether effects occur at the portal of entry or systemically. For in vitro studies, the toxicological relevance of the treatment levels might be informed by emerging approaches such as IVIVE (see **Section 10.2.3**) or other extrapolations.

- Lifestage(s) or population(s) (e.g., sex or another demographic) known to be most susceptible.

- Appropriateness of a study design or assay to measure the selected endpoint. For example, assays that directly evaluate mutations induced by an agent (e.g., by incidence, frequency, or type of mutation) are generally considered to be more predictive of mutagenic potential (depending on the model system) than an indirect measure of genetic damage, such as

sister chromatid exchanges [see, for example, Eastmond (2017) and Eastmond et al. (2009)].

- After considering the bulleted list of factors above, well-accepted assay designs are favored over test methods that may improve methods of detection but have not been adequately validated.

10.2.2. Conducting a More Detailed Review of Individual Experiments

As described in **Section 6.6**, an exhaustive analysis of individual studies reporting mechanistic endpoints is not always an effective or efficient way to consider mechanistic data, particularly with larger databases. However, when critical uncertainties exist, it becomes more important to rigorously evaluate a subset of the mechanistic evidence that will be most impactful to hazard and dose-response decisions in the assessment. The following scenarios describe cases where individual study-level assessment may be most warranted:

- When a single study (or very small set of studies) is likely to drive influential mechanistic conclusions for human health hazard identification and/or dose-response.

- When notable heterogeneity in results exists among similar studies/endpoints/test systems. For mechanistic events that appear to be of critical importance, a more intensive review of study methods may help to highlight the results that can be interpreted with greater confidence. Unexplained heterogeneity may reduce confidence in the mechanistic event. However,

    - if the studies are well designed for evaluating the mechanistic event in question and no or minimal heterogeneity is present, it is indicative of reproducibility. Reproducibility strengthens the confidence in the mechanistic event, and further evaluation of these individual studies is likely unnecessary.

    - when results for important mechanistic events appear to differ across populations (e.g., species, sexes), exposure paradigms (e.g., duration, route, test article), exposure levels, or other study characteristics critical to mechanistic interpretations, a more detailed review of the studies may be needed to determine whether there is an underlying cause or explanation for the disparate results.

- When studies are identified that experimentally challenge potential relationships between key events or the necessity of individual events for developing health effects. These experiments may increase or decrease confidence in a hypothesized mechanism. For some MOAs, the necessity of a specific mechanistic event leading to a downstream event or an adverse outcome can be tested by inhibiting that mechanistic event (e.g., pharmacologically, genetically, surgically) and observing whether the incidence or degree of the downstream event or adverse outcome has been affected.

- Importantly, when a decision is made to perform a more detailed review of individual mechanistic studies, considerations regarding study evaluations for a set of related studies (e.g., all reporting a similar assay or test system) should be identified. These considerations can help further prioritize studies and aid in the overall evaluation of the mechanistic evidence for an endpoint or outcome. The approach is intentionally flexible to allow for

application to varied evidence bases and to accommodate the anticipated increased reliance on emerging technologies and methods, including new approach methodologies (NAMs), in the future. Regardless of the approach (see **Section 10.2.1**), the steps taken for the selective evaluation of mechanistic studies should be transparently described.

10.2.3. Use of Emerging Mechanistic Data Types

Extensive efforts are underway to expand the use of in vitro and other nontraditional toxicity information in hazard determination and risk assessment, both within the IRIS Program as well as the wider Agency. In particular, ToxCast™/Toxicology in the 21st Century (Tox21) high-throughput screening (HTS) data and in vitro or in vivo toxicogenomic studies are increasingly used as resources to understand the mechanistic profiles that are enriched in response to chemical exposure. ToxCast™/Tox21 HTS bioactivity data are generated by cell-free (biochemical) and cell-based assays in human and rodent primary cells or cell lines that characterize a wide spectrum of biological responses to specific chemical exposures, including cell proliferation, cell death, and activities of enzymes, ion channels, receptors, or transcription factors (Judson et al., 2010). Assays frequently employed within the field of toxicogenomics, broadly defined as the study of genomic structure and function as it responds to exposure to foreign agents, currently rely on the quantification of gene expression products and methodologies designed to fit individual gene-level changes into ontological pathways to elucidate molecular responses to chemical exposure; gene expression microarrays and RNA-Seq are two such assays that have gained wide acceptance and serve as the backbone of most toxicogenomic-based studies. These approaches have required a shift in paradigm to a systems biology approach wherein gene expression changes must be interpreted as complex molecular signaling events that take place in an evolving cellular background where apical outcomes are not likely to be the result of a single genetic change.

Analysis of HTS and transcriptomic data can support evaluation of the plausibility of exposure-outcome associations found by epidemiological studies or help to establish the human relevance of apical outcomes observed in the exposed animal models. Likewise, toxicogenomic approaches can inform human relevance through correlated expression between tissues of exposed animals and cells or tissues of humans with corresponding health conditions. Further, the ability to query the enrichment of relevant signatures against an experimental gene expression data set suggests the potential to use transcriptomic data qualitatively (e.g., Is this chemical genotoxic?). In addition, comparative toxicogenomics can identify other chemicals that induce changes in gene expression similar to the chemical under assessment. Depending on chemical-specific circumstances, these data may additionally provide support to resolve some concerns in chemical risk assessment, such as concerns over conflicting results of different assays, or the relevance of effects observed in animal studies at higher doses to low doses more typical for environmental exposures.

Methodologies such as in vitro to in vivo extrapolation (IVIVE) and "high-throughput" benchmark dose modeling are being developed and adapted to support point-of-departure (POD)

calculations from nontraditional toxicity endpoints (Wambaugh et al., 2018; Dean et al., 2017; Farmahin et al., 2017; Thomas and Waters, 2016; Wetmore et al., 2015; Wetmore et al., 2014; Thomas et al., 2013). These transcriptomic values can increase confidence in PODs estimated in chemical risk assessment using traditional approaches from dose-response modeling of the occurrences or intensities of apical endpoints. Higher sensitivities of gene responses to environmentally relevant exposures can help assess the relevance of apical endpoints selected for benchmark dose (BMD) modeling. By BMD modeling of the gene expression data associated with the identified biologically relevant gene expression signatures, these studies indicate that transcriptional BMD values can closely predict the BMD values of known and unknown apical endpoints. In the future, these ongoing research efforts may inform the determination of potential human hazards by employing endpoints and/or models not previously considered as "adverse effects" suitable for human health risk assessment. Additional considerations for using HTS and transcriptomic data to identify and prioritize chemicals with limited databases (e.g., in the absence of apical human or animal data) are currently being developed and applied.

As these assays routinely result in dense and highly complex data sets, much effort has been directed at interpretation of such data and its application to both toxicology and risk assessment. Assessment teams that plan on conducting analyses of HTS and toxicogenomic data need to ensure they have access to experts to provide insight on the bioassays included in an HTS platform (e.g., assay methods, performance), biological pathways relevant to the health conditions under consideration, and the bioinformatics knowledge to construct and interpret HTS activity profiles or gene expression profiles. The major challenge in applying toxicogenomic data to human health risk assessment lies in understanding how to best distill complex gene expression data sets into easily comprehensible statements of actionable information. Processing raw microarray data to biologically interpretable data requires additional expert input from microarray or next-generation sequencing (NGS) bioinformatics. Some examples of recent activities and recommendations on the use of transcriptomics data at federal agencies are presented in **Table 10-3**.

Table 10-3. Activities and recommendations on the use of transcriptomics data at EPA and other agencies

Activity	Key points	Reference
EPA draft interim guidance for microarray-based assays: data submission, quality, analysis, management, and training considerations	• Recommendations on performance of transcriptomic experiments for use in risk assessments. Suggests compliance with MIAME standards (Brazma et al., 2001). • Criteria for accepting data in a risk assessment (assay validity and biologically meaningful response).	U.S. EPA (2007)
FDA Microarray/Sequencing Quality Control (MAQC) and Sequencing Quality Control (SEQC) Project	• International consortium for evaluating microarray and next-generation sequencing platforms (i.e., RNA-Seq) used to quantify changes in global gene expression. • Assesses and compares various sequencing platforms and data analysis methods. • Establishes best practices for reproducibility. • Evaluates the advantages and limitations of these technologies for use in safety assessments.	https://www.fda.gov/science-research/bioinformatics-tools/microarraysequencing-quality-control-maqcseqc
NTP approach to genomic dose-response modeling	• Recommendations for considering transcriptomic studies in dose-response assessments. • Experimental design recommendations for dose-response evaluation. • Signal detection filter to ensure adequacy and confidence in exposure-related effects. • Effect size and trend tests to identify biologically plausible and reproducible responses. • Parametric dose-response models that identify biological potency estimate. • Identification and grouping of gene ontologies that are responsive to treatment. • Provides biological and mechanistic interpretation of omics analyses.	NTP (2018)

Activity	Key points	Reference
An approach to using toxicogenomic data in EPA human health risk assessments: a dibutyl phthalate case study	• Presents a case study using toxicogenomic data to evaluate DBP-induced male reproductive effects and makes recommendations on the use of toxicogenomic data in risk assessment. • Evaluates consistency across studies and datasets. • Conducts dose-response modeling of gene(s) anchored to the MOA or outcome. • Performs additional pathway analysis according to outcomes of interest and/or critical time windows considered in the assessment.	U.S. EPA (2009)

10.3. SYNTHESIS OF MECHANISTIC EVIDENCE

This section introduces several important concepts and example approaches to organizing evidence to facilitate mechanistic analyses. It also includes important information to consider when drafting the mechanistic evidence synthesis.

10.3.1. General Considerations for Synthesizing the Mechanistic Evidence

Specific information across evidence streams should be identified, considered, and documented when organizing the synthesis. As previously described (see **Section 10.1**), **Table 10-1** identifies the main sources of mechanistic evidence and should serve as a starting point for organizing the evidence to be analyzed. In addition to reviewing the summaries of ADME understanding and health effect-related findings in humans and animals, it is essential to review the wider scientific literature for other relevant information. For example, the mechanistic literature may include examples of biologically plausible MOAs, systems, or biological processes. As the information is assembled, it is useful to begin considering the evidence for mechanistic events in the context of the modified Bradford Hill considerations, particularly biological plausibility and coherence. These concepts are discussed later in this chapter. Once these data have been assembled, determinations can be made regarding which mechanistic categories have sufficient information to be considered in the assessment (i.e., some biologically plausible effects were observed in studies evaluating in vitro models or tissue systems relevant to the hazards being assessed). The following sections discuss approaches for analyzing and synthesizing the evidence to form a coherent narrative of mechanistic events.

10.3.2. Approaches for Organization and Analysis

When an in-depth analysis is warranted, the MOA approach described in the EPA Cancer Guidelines (U.S. EPA, 2005b) is perhaps the most well developed and thorough demonstration of

what has become an accepted framework for the analysis of mechanistic data to inform hazard identification. The EPA 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 (WHO/IPCS, 2007a) and noncancer (WHO/IPCS, 2007b) outcomes from chemical exposures by establishing an MOA framework based on modified Bradford Hill considerations for causality.

As described in EPA Cancer Guidelines, an MOA is "defined as a sequence of key events and processes, starting with interaction of an agent with a cell, proceeding through operational and anatomical changes, and resulting in cancer formation" (U.S. EPA, 2005b). An MOA analysis is a tool for judging whether the available data provide mechanistic support for carcinogenicity—or for other toxicities as applied here—by drawing on information to help explain the underlying mechanism(s) behind the apical health effects observed in humans or animals. Frequently, the terms "MOA" and "mechanism" are used interchangeably; however, an "MOA" describes the (often general) process for how a chemical induces a toxic effect, whereas a "mechanism" 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. As described above, an MOA is typically composed of a sequence of key events, where a "key event is an empirically observable precursor step that is itself a necessary element of the mode of action or is a biologically based marker for such an element" (U.S. EPA, 2005b). In this context, a "mechanistic event," or what might be considered a "mechanism," is likely to be captured as a "key event."

An MOA analysis involves a critical review of the key events and the empirical evidence or biological understanding of the relationships between those key events. Competing explanations or well-supported, alternative MOAs should also be included in the analysis. Key events are part of the pathway from exposure to effect, with each being necessary, but not sufficient, for the health effect to occur. This conceptually distinguishes mechanistic events from an MOA. For example, an MOA for carcinogenesis caused by exposure to an agent may involve oxidative stress, increased cytokine production, inflammation, cytotoxicity, and cell proliferation. Any single one of these events would not be considered a complete MOA because it alone would not be sufficient to cause cancer, but together, some or all of these may be key events in an MOA. Support for an MOA may be strengthened by a more complete understanding of the biological interactions, including, for example, if a temporal- and/or dose-dependent progression can be pieced together from the evidence found in mechanistic studies.

An analysis of mechanistic events is part of an MOA analysis, but the MOA analysis is broader in scope in that it uses a modified version of the Bradford Hill considerations to determine whether the available data for a chemical's effects, including epidemiologic and experimental animal studies on apical outcomes, can support a proposed MOA(s) for the toxic effect(s) of an agent. Consideration of the evidence strength, consistency, specificity of association, dose-response concordance, temporal relationship, biological plausibility, and coherence are described in the

cancer guidelines (U.S. EPA, 2005b) and can be very useful for constructing an effective narrative of evidence linking exposure to toxic effects. In general, it can be useful to provide summary statements for each key event and for the modified Bradford Hill considerations. However, these considerations are not a checklist; no one aspect is either necessary or sufficient for drawing inferences of causality (U.S. EPA, 2005b). Rather, these considerations should be used, when helpful, to emphasize strength (or the lack thereof) in the mechanistic evidence.

The MOA analysis draws information on the toxicity of a chemical from many diverse sources, and as such, can be an exceedingly complex endeavor that is difficult to document in a clear narrative. Within the cancer guidelines MOA framework, there are useful concepts and organizational approaches that may provide structure for assessing the confidence in, as well as the limitations and uncertainties in, whether a given mechanism is associated with a toxic effect. These approaches include a review of the data using pathway-based conceptual frameworks such as the AOP approach, use of logic-based analyses such as counterfactual reasoning or hypothesis-based testing (Rhomberg et al., 2013), and the application of clustering approaches to prioritize and group subsets of large mechanistic databases [e.g., (Chiu et al., 2018)]. Approaches to the analyses may be customized depending on the size and complexity of the database and the current extent of scientific understanding regarding the mechanisms of toxicity of the chemical. For all MOA analyses, it is useful to create an analysis summary table displaying the evidence for how each key event in an MOA has been established in relation to each modified Bradford Hill consideration that has been evaluated. If there is evidence for more than one MOA for any health outcome, additional summary evidence analysis charts may be similarly prepared to allow for a direct comparison of the relevant evidence between and across MOAs.

Adverse Outcome Pathways (AOPs) have become functional and versatile tools for use in the risk assessment workflow. AOPs describe the sequential connections of causally linked key events between a single molecular initiating event and an adverse outcome, and are not chemical specific (Villeneuve et al., 2014a, b). AOPs share some similarities with MOAs, in that they are composed of the same modular components [i.e., the sequence of key events leading to the adverse outcome; (U.S. EPA, 2005b)]. As such, they may provide a simplified visual representation and organizational framework for the more complex relationships and associations described in an MOA. For example, MOA information for a chemical may be overlaid onto an AOP to aid risk assessors in organizing the available data (and identifying research needs) for a particular health hazard within pathways of biological responses to external insults that lead to adverse outcomes of regulatory interest. Thus, the outcomes from other exposures with similar molecular initiating events or key events may be predicted from measurable upstream events. The Adverse Outcome Pathways Knowledge Base (AOP-KB) is a useful centralized resource to access publicly available, crowdsourced information on AOPs and their development (https://aopkb.oecd.org/index.html).

AOPs are not directly equivalent to the MOA framework; MOA analyses are chemical-specific and include a structure for assessing causality and a more detailed consideration of the

adverse outcome. However, AOPs may become more informative and efficient tools for hazard risk assessments when they are coupled with quantitative information to better inform dose response decisions. To this end, a number of methodologies are being developed to enable quantification of AOPs, including empirical dose-response modeling, Bayesian networks (BN), and systems biology (SB) modeling approaches. Limitations to AOP quantification remain, however, in that there are a limited number of completed AOPs, as well as insufficient dose response information for key events in a given AOP and a lack of fluidity in the translation between biological processes and mathematical methodologies to quantify key events. The development and utilization of methodologies and tools to enable the quantification of AOPs continues to be an active area of interest for ORD.

Within the MOA framework, establishing the biological plausibility of an association between key events in a pathway from exposure to effect is inherently dependent on the current state of the knowledge (Fedak et al., 2015; Hill, 1965). If the current scientific understanding of biological pathways is underdeveloped, it could lead to an uneven focus, or bias, on particular MOAs or mechanistic relationships that may not tell the full story. Instead of identifying and organizing the mechanistic evidence according to predefined MOAs, a more objective approach is to categorize the literature from a broad search for chemical-specific mechanistic information according to commonly recognized properties of carcinogens. These properties, which have been grouped into 10 key characteristics of carcinogens (Smith et al., 2016), provide a systematic method for identifying, organizing, and summarizing the available mechanistic studies for analysis and interpretation. The key characteristics approach does not provide a framework for assessing causality, but when used to summarize mechanistic evidence within an MOA framework, helps to survey the mechanistic landscape of evidence and identify areas of focused research relevant to mechanistic events that may not have been previously recognized. This concept is currently being expanded to other health effects beyond cancer. Certainly, there are other variations of approaches to organizing, analyzing, and synthesizing mechanistic information that have similarities to those discussed here, and additional examples will be developed as the field advances.

The effective presentation of findings from mechanistic analyses can greatly assist in developing a concise mechanistic evidence synthesis that is transparent and enhances the reader's comprehension. Ultimately, the mechanistic support for the hazard conclusions will be summarized in the evidence profile table (see **Chapter 11**). The analysis and synthesis of mechanistic data performed prior to this can be presented in many ways. Tabular displays such as those used for human and animal evidence may be useful for presenting mechanistic evidence (see **Section 8.5**), as well as other types of tables and figures.

In most assessments, it will be useful to present at least a subset of the data in mechanistic data tables, although for small or simple databases, a narrative summary of findings across the relevant studies may suffice. Studies pertaining to relevant mechanistic events may be reports of endpoints measured in humans or in experimental settings in animals or in vitro (or in silico).

When prioritizing results from mechanistic studies (e.g., in a table), studies or data most relevant to the hypothesized MOAs or most pertinent to evaluating the adverse outcome will be emphasized. It is important to organize the available data in a way that will complement the synthesis and to document the rationale for these decisions.

10.4. FOCUSING THE MECHANISTIC EVIDENCE SYNTHESIS TO INFORM EVIDENCE INTEGRATION AND DOSE-RESPONSE ANALYSIS

The mechanistic analyses can inform the integration of evidence within and across evidence streams (**Chapter 11**) and dose-response analyses (**Chapters 12–13**). Examples of how mechanistic information can inform these steps are summarized in **Table 10-4**.

Table 10-4. Examples of how mechanistic information can inform evidence integration and dose-response analysis, and questions relevant to focusing the mechanistic synthesis

Assessment step informed by the mechanistic synthesis	Questions and considerations for focusing the mechanistic synthesis
Interpreting the consistency, coherence, and biological plausibility of the human and animal health effect evidence (see Section 11.1)	• Are the hypothesized MOAs biologically plausible, considering known toxicokinetic processes and the biological or experimental support for connections between mechanistic events? Consider consistency with established MOAs for related agents. • Are there notable uncertainties in the sets of human or animal health effect studies for which related mechanistic information is available? An understanding of mechanistic pathways (e.g., by identifying mechanistic precursor events linked qualitatively or quantitatively to apical health effect[s]) can influence the strength of the evidence integration conclusions, providing either support for or against biological plausibility (see additional discussion on this consideration in the bullets below and in **Table 11-2**). • Are there mechanistic key events that appear to be related to the health effects of interest? Consider whether these findings might serve as precursors informing an association between exposure and effect. If there are notable uncertainties in the set of available human and animal studies most relevant to the health effect of interest (e.g., they are all *low* confidence), consider a focused analysis of mechanistic precursors to inform strength of evidence determinations. • 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 and 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 may indicate that the agent produces effects other than those hypothesized, and/or that other pathways are being activated. This may warrant separate evaluations. • Do independent studies and different experimental hypothesis-testing approaches, perhaps from different model systems, identify key events in the MOAs 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, or laboratories strengthen confidence in the validity of an MOA.

Assessment step informed by the mechanistic synthesis	Questions and considerations for focusing the mechanistic synthesis
Interpreting the consistency, coherence, and biological plausibility of the human and animal health effect evidence (see Section 11.1) (continued)	• Are there proposed events in the biological pathway or AOP (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. Conversely, the appearance of unanticipated effects that, upon further review, are associated with upstream mechanistic events in the MOA can increase confidence. • 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 α2u-globulin MOA being solely operative in rodent tumorigenesis limited to male rats)? • Are there other key uncertainties or data gaps that were identified during the analyses of the sets of available human or animal health effect studies? If so, does the literature inventory of mechanistic studies indicate that there are likely to be a reasonable number of studies on the topic? If yes, a focused analysis of these studies may be informative. If no, consider whether an additional focused search of mechanistic information might be worthwhile (i.e., to identify other informative studies that were not captured by the initial PECO).
Considering the human relevance of animal findings (see Section 11.2)	• What is known about the human relevance of key events (note that, at this stage, this does not refer to whether the studies employed typical human exposure levels, but rather focuses on critical differences 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 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 directly relevant to humans. • The analysis of human relevance will focus on evaluations of the following issues. The extent of the analysis will vary depending on the anticipated impact of the animal evidence to the overall evidence integration judgment. o 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). o Coherence of mechanistic changes observed in exposed humans with animal evidence of mechanistic or toxicological changes. o Evidence for a plausible mechanistic pathway or MOA, within which the key events and relationships are evaluated regarding the likelihood of similarities (e.g., in presence or function) across species.

Assessment step informed by the mechanistic synthesis	Questions and considerations for focusing the mechanistic synthesis
Characterizing potential susceptible populations or lifestages to inform integration across evidence streams (see Section 11.2) and study selection for dose-response analysis (see Chapter 12)	• Do the results from the human and animal health effect studies appear to differ by categories that indicate the apparent presence of susceptible populations (e.g., across demographics, species, strains, sexes, or lifestages)? Consider analyses to better characterize the sources and impact of potential susceptibilities that might be explained by mechanistic information (e.g., due to genetic polymorphisms or metabolic deficiencies). • Do the mechanistic events provide information suggesting populations or lifestages that might be particularly susceptible to the MOA, including cumulative risk scenarios and toxicokinetic differences? 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, can clarify characteristics of important events (e.g., their presence or sensitivity across lifestages or across genetic variations) and helps identify susceptible populations. • Identification of lifestages or groups likely to be at greatest risk can clarify hazard descriptions and identify key data gaps including whether the most susceptible populations and lifestages have been adequately tested. If a proposed mechanistic pathway or MOA indicates a sensitive population or lifestage in humans, consider whether the appropriate analogous exposures and populations or lifestages were adequately represented in the human or animal database. • When there is evidence of susceptibilities, but specific studies addressing these susceptibilities are unavailable for quantitative analysis, susceptibility data may support refined human variability UFs or probabilistic uncertainty analyses.
Evaluate biological understanding, including the identification of precursor events, to optimize dose-response analysis (see Chapter 13)	• A biological understanding of linkages within or across mechanistic events/MOAs, including the identification of precursor events in humans and the exposure conditions expected to result in these effects, can inform the use of: o Particular dose-response models (e.g., models integrating data across several related outcomes or incorporating toxicokinetic knowledge), o Proximal measures of exposure (e.g., external vs. internal metrics), o Surrogate endpoints (e.g., use of well-established precursors in lieu of direct observation of apical endpoints), and o Improved characterization of responses (e.g., combination of related outcomes, such as benign and malignant tumors resulting from the same MOA). o If the human and animal health effect data amenable to dose-response analysis are weak or only at high exposure levels, consider evaluating the precursor data for quantitative analysis.

UF = uncertainty factor.

10.4.1. Information to Include in the Mechanistic Evidence Synthesis

As previously discussed, the MOA weight-of-evidence framework described in the cancer guidelines (U.S. EPA, 2005b) is perhaps the most well developed and accepted framework for the synthesis and integration of mechanistic evidence to inform hazard identification. Regardless of the framework or presentation style employed, the synthesis text describing mechanistic information relevant to a particular health effect(s) should include summary interpretations of the mechanistic analyses, similar to the human and animal syntheses (see **Chapter 9**). Namely, the goal is to summarize the available mechanistic evidence in a manner that informs the evidence integration conclusions, including both qualitative and quantitative decisions, into a narrative format. Typically, this involves evaluation of modified Bradford Hill considerations (Hill, 1965) focused on specific questions informing how the mechanistic evidence can be applied to address key assessment issues, noting that the various considerations might be applied to a specific mechanistic event. In the future, as this process is applied more systematically to mechanistic data, it may be possible (and useful) to characterize judgments of strength and sufficiency during evidence integration using a standardized set of conclusions. The EPA Cancer Guidelines (U.S. EPA, 2005b) provide guidance on the process for developing MOA conclusions (applicable to cancer and noncancer MOA evaluations), emphasizing three conclusions that should be addressed in every MOA analysis: (1) Is the hypothesized MOA sufficiently supported in the test animals? (2) Is the hypothesized MOA relevant to humans? And (3) Which populations or lifestages can be particularly susceptible to the hypothesized MOA? Thus, when such analyses are warranted, the mechanistic evidence synthesis should summarize the evidence available relevant to each of these key conclusions and the rationale for all resultant conclusions. Some key considerations for synthesizing the evidence relevant to these conclusions are briefly summarized in **Table 10-4**. Notably, evaluations of the strength, consistency, and specificity of the association between key events and health effects should include consideration of the dose-response and temporal association between key events and hazard endpoints, as well as the concordance of mechanistic events across different experimental models, exposure paradigms, or types of investigations.

Additional details on these considerations are provided in EPA Cancer Guidelines (U.S. EPA, 2005b), including processes for evaluating whether and how MOA information might influence quantitative estimates and dose-response variability, some examples of which are summarized in **Table 10-4**. Importantly, because the evaluations of MOAs are typically focused on a particular health effect, it can be important to consider whether MOAs (or specific key events) might be applicable to other health effects, possibly in other tissue systems (e.g., health effects or tissue systems with less mechanistic data). Although much emphasis in this chapter has been placed on the cancer guidelines (U.S. EPA, 2005b), similar concepts are available in EPA guidance on assessing noncancer health effects and should be consulted [e.g., (U.S. EPA, 1998, 1996b, 1991)]. As previously described, the mechanistic synthesis conclusions from the analyses described above will inform the overall evidence integration narrative, a process that is analogous to the

weight-of-evidence narrative described in the cancer guidelines (U.S. EPA, 2005b). This is described in **Chapter 11**.

10.5. SUMMARY OF WORKFLOW FOR ANALYSIS AND SYNTHESIS OF MECHANISTIC EVIDENCE

This outline provides an abridged view of the process of considering mechanistically relevant information throughout assessment development. Because the process does not always follow the order as laid out in this handbook, the corresponding sections have been noted.

1. **Problem Formulation and Development of an Assessment Plan (Chapter 2, Sections 2.1 and 2.2)**

   Goal: To the extent possible, assess the likely impact of mechanistic information on assessment conclusions during scoping and problem formulation. This will help frame the approach used for conducting and organizing a preliminary literature survey ("evidence mapping")

   Prepare the preliminary literature survey of mechanistic information (see **Sections 2.1** and **Chapter 4**):

   - Identify authoritative reviews and existing chemical assessments from other agencies reporting relevant MOAs
   - Identify reviews of ADME/TK information that may be relevant for mechanistic considerations
   - As time allows (e.g., for assessments with smaller or less complex mechanistic databases), proceed to Step 2 and provide these screening and tagging results in Step 4
   - Optional: If possible, building from the citations identified, generate preliminary visual outputs (e.g., heat maps) of tagged categories of mechanistic information to create literature survey evidence maps. These may be very broad if screening has not been extensively conducted. Visualizations can be created using, for example:
     - Word or Excel
     - Interactive software applications such as Qlik or Tableau
     - Dendrogram visualizations in Health Assessment Workspace Collaborative (HAWC)

2. **Literature Identification and Refined Evaluation Plan (see Chapters 4 and 5)**

   Identify primary mechanistic evidence and create a basic literature inventory (see **Section 4.3.3**; note that some steps below may have already been completed for the IAP)

- First-level title/abstract (TIAB) screening (e.g., in Distiller) from the literature search results

    - Identify and tag studies captured in the broad literature search that contain "potentially relevant supplemental material"

    - Identify and tag studies within "potentially relevant supplemental material" that should be included in the mechanistic inventory (tag: "mechanistic")

    - These high-level screening steps may be done simultaneously with some or all steps of the second-level screening, depending on size and complexity of database

- Second-level screening (at either TIAB or full-text level) categorizes studies identified as "mechanistic" within "potentially relevant supplemental material" (note that this step may be performed after the refined evaluation plan is developed, see below); potential categories include:

    - Designation of "deprioritization" criteria to identify studies of effects beyond the assessment scope (e.g., a health effect not determined to be a focus for the assessment, a coexposure study, a study in a nonrelevant species); assigning categories to deprioritized studies enables easy retrieval later if the mechanistic analyses indicate a revised prioritization

    - Type of health effects or outcomes investigated (e.g., hepatic, neurological, cancer)

    - Authoritative reviews, other agency assessments, or other types of studies for further consideration

    - Mechanistic studies may also be categorized as pertinent to other typical supplemental material content (reference the supplemental table in the handbook) to identify, for example, susceptible populations and lifestages, studies of metabolism or kinetics at the cellular level, or studies in nontraditional model systems

    - As needed, additional levels of organization within each hazard category may be designated by chemical team, for example, by relevant biological pathway affected, receptor activation/binding activities, key characteristic of carcinogen or toxicant

    - It may not always be possible to categorize studies based on TIAB content, thus the tagging will continue at the full-text level (see below)

        - For assessment purposes, the categorization judgments are typically collapsed across TIAB and full-text screening, but a record is maintained of where the tagging judgment was made (e.g., as a column in an Excel file created from Distiller or SWIFT Active output)

        - Example screening forms are available in DistillerSR in the "IRIS Template Forms" project

- Optional: Update the preliminary evidence map from initial literature survey described in Step 1 with a more detailed evidence map reflecting further screening to provide a visualization of the categorized mechanistic literature

Review other evidence informing the potential impact of specific mechanistic analyses

- Summarize major findings related to ADME/TK (if further developed since IAP)

- Review literature inventories from human and animal health effects studies

Based on preliminary findings from the mechanistic, animal, human, and ADME/TK evidence, further refine main areas of focus for potential mechanistic analyses, create literature inventory, and identify additional topic areas to be searched and reviewed

- As decisions are made on which mechanistic topics to prioritize, the full-text versions of the studies can be uploaded into the screening program for more extensive tagging to develop the mechanistic literature inventories:

    - Inventories (e.g., in Distiller, Excel) extract information not captured during screening, including information on endpoints evaluated, assay(s) used, model system, exposure route and levels tested, and direction of results, in a sortable format (customized extraction forms can be created in Distiller)

    - In some cases, it can be useful to inventory additional study details that were identified in the initial screening phases (e.g., use of the chemical as a positive control, testing for effects as part of a mixture, inclusion of specific pathway inhibitors or use of knockout models in the experiment)

    - Continue to add details and study design features to the mechanistic literature inventory as the specific areas of focus for mechanistic analyses are refined.

- Review the mechanistic literature inventory or evidence map to determine whether additional literature searches are warranted

    - Identify data gaps (e.g., mechanistic data relevant to effects not reported in human or animal health effect evidence inventories)

    - For topic areas expected to be important, consider whether there is a need for targeted literature searches focused on a particular mechanistic event, or on mechanisms operating in related chemicals

        - Work with Health and Environmental Research Online (HERO) staff to create search string for customized search

        - Consider using machine learning (e.g., SWIFT Review, SWIFT Active) to identify additional studies on key mechanistic events

- Summarize any additional, focused literature searches, as well as any decisions to refine the scope or prioritization of specific mechanistic topics, in an update to the protocol (see **Section 3**).

3. Prioritization of mechanistic topics for analysis and synthesis (Section 10.1 and **Chapter 11)**

Potential mechanistic topics informing evidence integration within human and animal evidence streams:

- Consider the application of the ADME summaries to identify/prioritize the most "functional" systems for predicting the health effect (see **Chapter 5** and **Section 6.5**)

- Identify mechanistic information in exposed humans or humanized models that is likely to impact the interpretation of the human evidence; for example:

    - Mechanistic information from exposed humans (e.g., identification of human biomarkers)

    - Studies in model systems possibly more relevant to the human health effect in question than other available models

        - This would include studies in human cells or tissues, after considering whether they are expected to represent the necessary human complexity (e.g., appropriate receptors and normal-function cell types). In general, immortalized cell lines from humans would be considered less informative (see Step 4)

        - It would likely also include some manipulated systems (e.g., animal cells expressing the appropriate human receptors, in vivo human cell transfer models)

- Identify whether there is mechanistic information in experimental animals or other test systems that is likely to impact the interpretation of the animal evidence; for example:

    - Mechanistic studies in exposed animals, as well as studies in animal cells or tissues, applying similar considerations to those described above for human models

    - Evidence to explain differences in key results across species or strains

- Determine whether there are mechanistic data informing potential susceptible populations and lifestages (also important for integration across streams, see below)

Potential mechanistic topics informing evidence integration across evidence streams:

- Determine as early as possible whether mechanistic information is likely to impact the overall evidence integration conclusion (e.g., evidence that the animal response is unlikely to be relevant to humans)

- Determine whether there are mechanistic data informing potential susceptible populations and lifestages. Consider whether these potential susceptibilities appear

to be adequately addressed in the available human and animal studies (e.g., studies not encompassing a likely susceptible group might be viewed as less able to address the hazard question), which helps to inform whether such mechanistic analysis are likely to be more or less impactful.

4. Prioritization of mechanistic evidence addressing topic areas of interest, and considerations for study-level evaluations (see Sections 6.6 and 10.2)

Review major groups of studies in the inventory; identify (based on inventories, human and animal evidence, and ADME/TK information) potential "key" mechanistic events (e.g., consistent with cancer guidelines, existing biological pathways, AOPs)

Option 1: If mechanistic evidence is unlikely to impact a health effect judgment or dose-response quantification of that health effect (e.g., if a complex analysis to establish mechanistic understanding would not further increase or decrease the certainty in the human or animal evidence); see **Table 10-4** for considerations related to this option:

- Provide a concise summary of mechanistic information for the health effect synthesis
    - The mechanistic summary may be based primarily on studies in the inventory, but may also include information from informative reviews or other agency assessments
- If some mechanistic conclusions would prove useful (e.g., establishing a dose-range for upstream events), a table with an overview of selected studies and relevant details may be provided in lieu of reviewing these study details in the synthesis text

Option 2: If it has been determined that a subset of mechanistic evidence is likely to be impactful for the assessment (e.g., a cancer MOA with conflicting evidence), the following stepwise process should be continued until an informed scientific judgment can be made for the mechanistic event, to be documented transparently for the assessment:

1) Begin by identifying, to the extent possible:
    - Endpoints most sensitive for predicting effect or mechanistic event
    - Assay systems with the most accepted methods for evaluating endpoints based on sensitivity, specificity, and relevance to endpoint
    - Other aspects that may need to be considered, e.g., conditions that most closely predict human exposures based on ADME/TK; model system selected; exposure method; purity, solubility, or volatility of chemical; whether cytotoxicity was measured; results reported at subtoxic concentrations; this should be determined by the assessment team

2) Sort and rank studies in the group based on the identified considerations. If an expert judgment can be made at this point, for example, if there are many studies that conform to the above selected considerations, and the results are consistent and reproducible, you may stop here. However, this will not often be the case, and further evaluation and documentation will be needed to decide

- Identify existing considerations for methods used to measure the selected endpoints and/or specific identified assays; for example:
    - Organisation of Economic Co-operation and Development (OECD) guidance, informative reviews, or other definitive sources with scientific consensus
    - Considerations for specific assays developed at EPA for previous assessment analyses
- Consider using an existing tool for evaluating mechanistic studies; for example:
    - Toxic Substances Control Act (TSCA) (U.S. EPA, 2018a)
    - Science in Risk Assessment and Policy (SciRAP) (Beronius et al., 2018)
    - MIAME (Brazma et al., 2001) and/or Systematic Omics Analysis Review (SOAR) [(McConnell et al., 2014); for microarray studies]
- Continue to formulate and refine considerations for evaluating this set of experiments/assays until an expert judgment can be reached, including possibly the same examples provided in 1.b and 1.c, with further detail provided (e.g., the required number of experimental replicates, appropriate positive and negative controls)

3) Provide a summary of study level judgments in a table to clearly convey results of any evaluations conducted

5. Synthesis of mechanistic evidence informing evidence integration decisions (see Sections 10.3 and 10.4 and **Chapter 11)**

Use the results of the prioritization decisions and judgments to frame the analysis

- Identify potential key events (necessary, but not necessarily sufficient, for health effect to occur)
- Identify relevant biological pathways that may provide context
- If potential key events or relevant biological pathways cannot be reasonably identified based on the available mechanistic information, summarize the evidence briefly, noting gaps in data/areas of research

Develop mechanistic syntheses for the specific mechanistic question(s) relevant to each assessed health effect, highlighting the mechanistic evidence most informative to the evidence integration narrative

- Summarize understanding of chemical and physical properties and ADME/TK
- Characterize potential susceptible populations and lifestages, as well as data gaps in understanding, based on the available mechanistic evidence

- Describe the evidence informing potential key events (necessary, but not sufficient, for the health effect to occur following exposure), including any judgments drawn regarding each key event and the supporting rationale for all such judgments

- Summarize the overall evidentiary support (or lack thereof) for potential MOAs (e.g., for integrating evidence using the adapted Bradford Hill considerations described in the cancer guidelines), with greater weight given to results of evaluated, higher confidence studies, also taking into account:

    - other existing MOAs, including those for structurally related chemicals

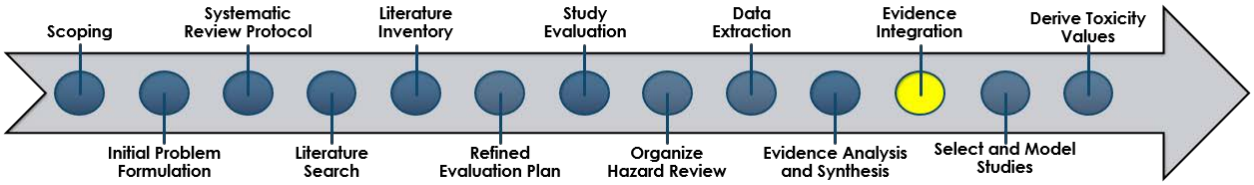
    - data gaps, and whether these are likely to indicate understudied areas or unpublished null results

- Specifically summarize the strength of the mechanistic evidence, if available and necessary, informative to the:

    - independent human and animal evidence summaries (e.g., mechanistic biomarkers informing biological plausibility; see Step 3)

    - overall evidence integration conclusion (e.g., data informing the human relevance of findings in animals; see Step 3)

11. EVIDENCE INTEGRATION



DEVELOPING SUMMARY HAZARD JUDGMENTS

Purpose

- To draw integrated judgments across human, animal, and mechanistic evidence to assess the potential that a substance is hazardous to humans.

Who

- Assessment team.

What

- Provide an evidence integration narrative with summary judgments and supporting rationale documented using an evidence profile table for each health effect.

This chapter describes the process for integrating the human, animal, and mechanistic evidence to develop an *evidence integration narrative*.[18] This narrative, which is separate from the syntheses of the human, animal, and mechanistic evidence, is a short (up to a few pages) summary of the assessment hazard identification judgments for each assessed health effect (i.e., each noncancer health effect and specific type of cancer, or other grouping of related outcomes). The evidence integration narratives serve to summarize the judgments regarding the evidence on a chemical's carcinogenic or toxic potential to humans and the conditions of its expression in the available studies (U.S. EPA, 2005b). These decisions directly inform the dose-response analyses (see **Chapter 13**) that provide estimates of the conditions of its expression (and the associated

[18]The phrase "evidence integration" used here is analogous to the phrase "weight of evidence" used in some other assessment processes (EFSA, 2017; U.S. EPA, 2017; NRC, 2014; U.S. EPA, 2005b). The IRIS Program has adopted the term *evidence integration*, recommended in the National Research Council (NRC) review of IRIS (NRC, 2014) as more descriptive of the process that is employed: "The present committee found that the phrase *weight of evidence* has become far too vague as used in practice today and thus is of little scientific use. In some accounts, it is characterized as an oversimplified balance scale on which evidence supporting hazard is placed on one side and evidence refuting hazard on the other. The present committee found the phrase *evidence integration* to be more useful and more descriptive of what is done at this point in an IRIS assessment—that is, IRIS assessments must come to a judgment about whether a chemical is hazardous to human health and must do so by integrating a variety of evidence."

uncertainties) more broadly. The goal of the evidence integration approach is not to describe the amount or "completeness" of the evidence base, but to critically assess and judge the evidence supporting a causal association (or lack thereof) between a chemical exposure and a specific health effect(s).

Evidence integration combines decisions regarding the strength of the animal and human evidence, incorporating information on biological plausibility, with decisions regarding: information on the human relevance of the animal evidence (including mechanistic evidence in animals, especially in cases where phenotypic animal evidence- i.e., from apical endpoints, is lacking) and relevance of the in vitro mechanistic evidence to exposed humans (considering toxicokinetics and other biological considerations specific to the health effect); coherence across bodies of evidence; and information on susceptible populations and lifestages. As previously discussed in **Chapter 10**, the approach to evaluating the mechanistic evidence relevant to each assessed health effect follows a stepwise approach beginning during assessment scoping and continuing throughout assessment development; it is expected to vary depending on the nature and impact of the uncertainties identified within each evidence base, as well as the specific mechanistic information available to address those uncertainties. Thus, this chapter builds upon the analysis and synthesis of the human (predominantly epidemiology) and experimental animal toxicology studies (see **Chapter 9**) and incorporates the available mechanistic data as appropriate to inform decisions (see **Chapter 10**).

The specific decision frameworks for the structured evaluation of the strength of the human and animal evidence streams and for drawing the overall evidence integration judgment are described in **Sections 11.1 and 11.2**, respectively. This process is informed by the Grading of Recommendations Assessment, Development, and Evaluation [GRADE; (Morgan et al., 2016; Guyatt et al., 2011a; Schünemann et al., 2011)], which arrives at an overall conclusion for each body of evidence based on a structured set of considerations.

During evidence integration, a two-step, sequential process is used, as follows (and depicted in **Figure 11-1**):

- Step 1: judgments regarding the strength of the evidence from the available human and animal studies are made in parallel, but separately, using a structured evaluation of an adapted set of considerations first introduced by Sir Austin Bradford Hill (Hill, 1965). **Table 11-2** describes the structured application of these considerations and the explicit incorporation of study confidence within each evaluation domain, and **Tables 11-3 and 11-4** present the structured frameworks for drawing these judgments. Based on the approaches and considerations previously described in **Section 9.2**, these summaries incorporate the relevant mechanistic information that informs the biological plausibility and coherence within the available human or animal health effect studies. Note that at this stage, the animal evidence judgment does not yet consider the human relevance of that evidence. The separate judgments documenting the strength of the available human or animal evidence prior to integrating them with other considerations to draw assessment conclusions about the potential for human health effects are interim steps included to increase the transparency of the decision process, but they are not final assessment

conclusions themselves. To add transparency and improve clarity in the systematic process, a standardized set of terms is used to describe the strength of the human and animal evidence for each assessed health effect. The terms associated with the different strength of evidence judgments are *robust*, *moderate*, *slight*, and *indeterminate*, which are differentiated by the quantity and quality of information available to rule out alternative explanations. Additionally, a judgment of *compelling evidence of no effect* may be used in rare instances.

- Step 2: the animal, human, and mechanistic evidence judgments are combined to draw an overall judgment(s) that incorporates inferences drawn based on information on the human relevance of the animal evidence and mechanistic evidence in animals or in vitro, coherence across the separate bodies of human and animal evidence, and other important information (e.g., judgments regarding susceptibility). Note that without evidence to the contrary, the human relevance of animal findings is assumed.[19] The output of step 2 is a summary judgment of the evidence base for each potential human health effect (see **Table 11-5** for the structured framework used to draw this overall judgment).

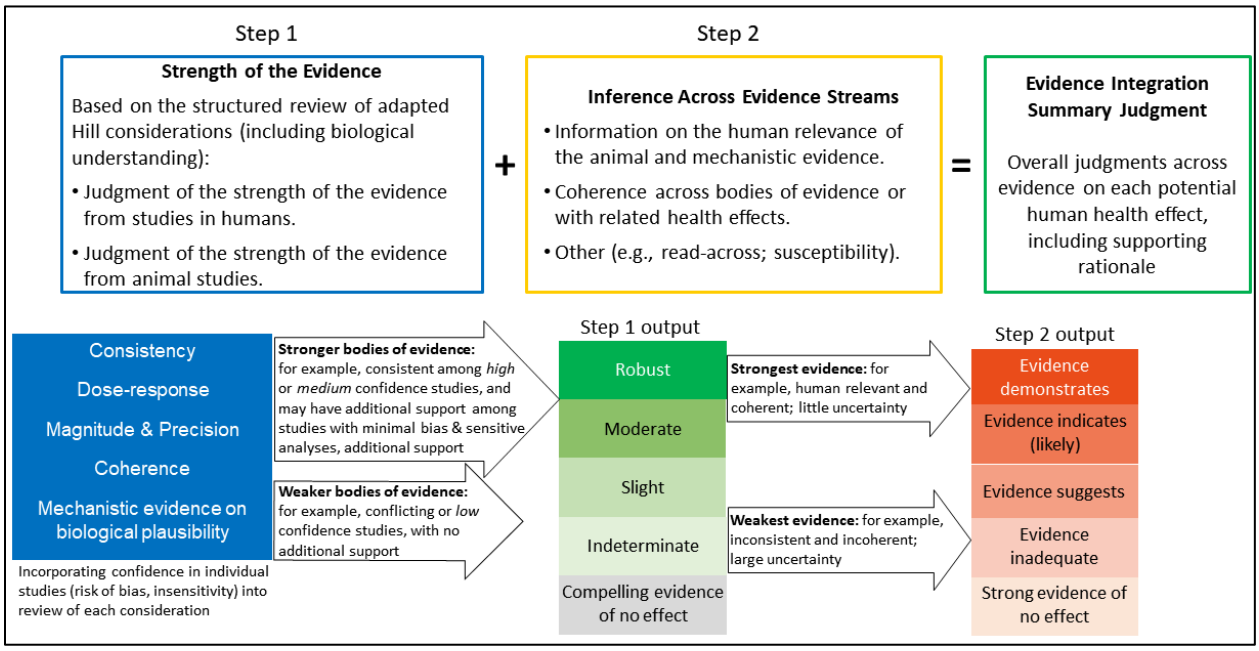


Figure 11-1. Process for evidence integration. Note: for carcinogenicity, the judgments described here for different cancer types are used to inform the evidence integration narrative for carcinogenicity and selection of one of EPA's standardized cancer descriptors, following the methods described in the EPA Cancer Guidelines (U.S. EPA, 2005b) see **Section 11.2**.

[19]As described in the EPA reference dose (RfD)/reference concentration (RfC) technical report (U.S. EPA, 2002b), "one of the major default assumptions in EPA's risk assessment guidelines is that animal data are relevant for humans[e.g., U.S. EPA (1998, 1996a, 1991)]. Such defaults are intended to be used in the absence of experimental data that can provide direct information on the relevance of animal data." This default assumption, as well as the analysis of evidence directly informing the relevance of animal evidence (when it exists), is consistent across EPA and other national and international agencies [e.g., ATSDR (2018); NIEHS (2015); NTP (2015); IARC (2006); U.S. EPA (2005b)].

The decision points within the structured, two-step evidence integration process should be summarized in an evidence profile table for each health effect category, or grouping of related outcomes, in support of the evidence integration narrative (see **Table 11-1** for an example template that is being applied to draft IRIS assessment products in 2019-2020; it is expected that the format will continue to evolve and may be modified on an assessment-specific basis). As described in **Chapters 9 and 10**, the human, animal, and mechanistic syntheses serve as inputs providing a foundation for the evidence integration decisions; thus, the major conclusions from these syntheses (including the key findings and inferences from any MOA analyses; see **Section 10.3.2**) should be summarized in the evidence profile table. The evidence profile tables summarize not only the judgments for each step of the structured evidence integration process, but also the evidence that supports them. Separate sections are included for summarizing the human and animal evidence judgments, inference drawn across evidence streams, and for the overall evidence integration judgment, each of which should present the key information from the different bodies of evidence that provided the primary support for that decision. As described in **Section 11.2**, judgments drawn using this process regarding the evidence relevant to each cancer type are used to inform the evidence integration narrative for carcinogenicity and selection of one of EPA's standardized cancer descriptors using the methods described in the EPA Cancer Guidelines (U.S. EPA, 2005b).

It is preferable that at least two reviewers independently draw evidence integration judgments using the considerations and frameworks described in **Sections 11.1 and 11.2**, with any differences resolved by discussion to reach consensus. Although Health Assessment Workspace Collaborative (HAWC) currently is limited in its ability to document evidence integration judgments, it is anticipated that in the future HAWC will be updated to include capabilities allowing for multiple reviewers to independently summarize their evidence integration judgments via evidence profile tables using a process similar to that used for study evaluation (see **Chapter 6**).

Table 11-1. Evidence profile table template (example[a])

Evidence Integration Summary Judgment (see **Table 11-5**)
Describe judgment (e.g., **evidence demonstrates**[b]) regarding the chemical exposure evidence relevant to each potential human health hazard, providing the primary interpretations from the human, animal, and mechanistic evidence streams (with priority in the table given to the evidence streams having the most impact on the overall judgment), as well as a summary of the models and range of dose levels in the studies upon which the overall judgment is primarily reliant.

Summary of Human, Animal, and Mechanistic Evidence					Inferences across evidence streams
Evidence from Studies of Exposed Humans [may be separated by exposure route or other study design characteristic][c]					Human relevance of findings in animals, including short rationale Cross-stream coherence (e.g., biologically related outcomes affected in both humans and animals), including short rationale Summary of potential susceptible populations or lifestages Other inferences: o Other MOA analysis inferences (e.g., judgments relevant to dose-response analysis) o Relevant information from other sources (e.g., read across)
Studies, outcomes, and confidence	**Factors that increase certainty[d]**	**Factors that decrease certainty[d]**	**Key findings**	**Summary strength of evidence judgment**	
<u>May be separate rows by outcome</u> References • Study confidence • Possibly, study design description	Consistency (e.g., across studies or populations) Dose-response gradient Coherence of observed effects Effect size (may relate to size or severity) Mechanistic evidence providing plausibility *Medium* or *high* confidence studies[e]	Unexplained inconsistency Imprecision *Low* confidence studies[e] Evidence demonstrating implausibility or a lack of expected coherence	Description of the primary findings, as interpreted in the evidence synthesis If sensitivity issues were identified, describe the impact on reliability of the reported findings	Describe the strength of the evidence from human studies based on the factors at left, including the primary evidence basis and considering: Results across human epidemiological and controlled exposure studies Interpretations regarding any human mechanistic evidence informing biological plausibility (e.g., precursor events linked to adverse outcomes) Judgments are summarized as one of the following (see **Table 11-3**): • ⊕⊕⊕ *Robust* • ⊕⊕⊙ *Moderate* • ⊕⊙⊙ *Slight* • ⊙⊙⊙ *Indeterminate* • −−− *Compelling evidence of no effect*	
Evidence from In vivo Animal Studies [may be separated by exposure route or other study design characteristic][c]					
Studies, outcomes, and confidence	**Factors that increase certainty[d]**	**Factors that decrease certainty[d]**	**Key findings**	**Summary strength of evidence judgment**	
<u>May be separate rows by outcome</u> References Study confidence	Consistency and/or Replication (e.g., across species, studies, or laboratories) Dose-response gradient	Unexplained inconsistency Imprecision *Low* confidence studies[e]	Description of the primary findings, as interpreted in the evidence synthesis	Describe the strength of the evidence for an effect in animal studies based on the factors at left, including the primary evidence basis and considering: Results across animal toxicological studies	

Possibly, study design description	Coherence of observed effects Effect size (may relate to size or severity) Mechanistic evidence providing plausibility *Medium* or *high* confidence studies[e]	Evidence demonstrating implausibility or a lack of expected coherence	If sensitivity issues were identified, describe the impact on reliability of the reported findings	Interpretations regarding any animal mechanistic evidence informing biological plausibility (e.g., precursor events linked to adverse outcomes) Judgments are summarized as one of the following (see **Table 11-4**): • ⊕⊕⊕ *Robust* • ⊕⊕⊙ *Moderate* • ⊕⊙⊙ *Slight* • ⊙⊙⊙ *Indeterminate* • ——— *Compelling evidence of no effect*

Mechanistic Evidence or Supplemental Information [may include separate summaries for each focused topic of analysis][f]			
Biological events or pathways (or other information category)	**Primary evidence evaluated**	**Key findings, interpretation, and limitations**	**Evidence stream summary**
<u>May include separate rows by biological/key events or other feature of the analysis approach</u> Generally, studies are not listed, but will cite the synthesis section	<u>May be multiple rows emphasizing evidence most informative to the mechanistic event or pathway(s) analyzed</u> Typically includes: Evidence type(s) (e.g., study designs, assays, endpoints) Species (may include life stage- or sex-specific description, if important to interpretation) System (in vivo; in vitro; in silico) Range of exposure levels and durations tested May summarize information that is not chemical-specific (e.g., for use in read-across)	*Key findings*: Summary of findings in the body of evidence (may focus on or emphasize highly informative study designs, endpoints or findings). *Interpretation:* Summary of expert interpretation from the synthesis of the body of evidence and supporting rationale. Generally, the evidence that informs analyses of biological events or pathways includes one or more endpoints. Factors that increase or decrease certainty in the individual events or pathways may be applied in a similar manner as for the human and animal evidence streams or using an alternative analysis approach for a biological event/pathway. *Limitations:* summary of key sources of uncertainty or limitations of the study designs tested (e.g., for the biological event or pathway being examined)	Overall summary of expert interpretation across the assessed set of biological events, potential mechanisms of toxicity, or other analysis approach (e.g., AOP). Includes the primary evidence supporting the interpretation(s) May inform within-stream judgments for other evidence streams (e.g., in vitro assay results supporting limited evidence from experimental animals) Describes and substantiates the extent to which the evidence influences inferences across evidence streams (e.g., establishing a biological linkage between animal findings and outcomes observed in humans). Characterizes the limitations of the evaluation and highlights uncertainties and data gaps.

[a]As this represents an evolving template, it is anticipated that modified evidence profile table templates may be applied on an assessment-specific basis (noting that the format used to support the evidence integration narratives within a given assessment should be consistent across health effects).

[b]For the strongest conclusion, the evidence integration narrative and summary judgment here would be described as: "The currently available **evidence demonstrates** that [chemical] causes [health effect] in humans [in some assessments, these conclusions might be based on data specific to a particular life stage of exposure, sex, population, or other specific group] under relevant exposure circumstances. This conclusion is primarily based on studies of [humans, animals, and/or mechanistic evidence] that assessed [range of doses/concentrations or specific cutoff-level dose/concentration and exposure duration, or

other exposure design summary]." The other evidence integration judgment levels that could be drawn here are **evidence indicates (likely)**, **evidence suggests**, **evidence inadequate**, and **strong evidence supports no effect** (see **Table 11-5**).

[c]To add clarity and emphasize the most influential evidence for decision-making, the rows for the different evidence streams within the table may be reordered (e.g., animal or mechanistic evidence streams may be summarized first within the table when those data are most impactful to the evidence integration judgments). Likewise, when data within an evidence stream are lacking or otherwise not informative to the evidence integration decisions, the summary for that evidence stream may be collapsed or abbreviated. Lastly, in addition to exposure route, the summaries of each evidence stream may include multiple rows- e.g., by study confidence, outcome, population, or species, if this informed the analysis of heterogeneity in results or other features of the evidence.

[d]See **Table 11-2** for a description of how these factors, which are explicitly tied to the adapted Hill considerations, are evaluated to judge increases or decreases in certainty.

[e]Study confidence, based on evaluation of risk of bias and study sensitivity (**Chapter 6**), and information on susceptibility are considered when evaluating each of the other factors that increase or decrease certainty in the evidence (e.g., consistency). Notably, lack of findings in studies deemed insensitive neither increases nor decreases certainty.

[f]It is expected that there will be a large amount of heterogeneity in the critical uncertainties requiring mechanistic analyses across assessments and health effects, as well as differences in the relative impact of the analyses on evidence integration judgments. Thus, separate rows or sets of rows (if for example, the question-specific judgment requires analyses of different MOAs or key events, or across exposure routes or species) will typically be used to transparently illustrate the different judgments. Examples of assessment-specific uncertainties that may be separately addressed in different rows include an evaluation of the sufficiency of the evidence informing a hypothesized MOA, the strength of the evidence for specific key events or identified precursors (in humans or animals), or concerns regarding the human relevance of findings in animals (these potential uncertainties are elaborated on elsewhere, including **Table 10-4**).

1

11.1. INTEGRATING WITHIN THE HUMAN AND ANIMAL EVIDENCE STREAMS

As previously described, prior to drawing overall evidence integration judgments about whether chemical exposure has the potential to cause certain health effect(s) in humans given relevant exposure circumstances, the strength of evidence from the available human and animal studies is summarized and judged. Concurrently, for each assessed health effect or health effect grouping, the influential mechanistic findings (including in vitro and in silico models and relevant NAMs) relating to understanding the underlying biological changes leading to the observed effects in exposed humans and animals and synthesized using the approaches and considerations outlined in **Chapter 10**, will be considered and incorporated. For the human and animal evidence streams, the considerations previously outlined in **Chapter 9** (the different features of the evidence considered and summarized during evidence synthesis) should be evaluated within the context of how they affect judgments of the strength of evidence (see **Table 11-2**).

To add transparency and improve clarity in the systematic process, a standardized set of terms is used to describe the strength of the human and animal evidence for each assessed health effect (**Figure 11-1**). The terms associated with the different strength of evidence judgments are *robust*, *moderate*, *slight*, and *indeterminate*, which are differentiated by the quantity and quality of information available to rule out alternative explanations. Additionally, a judgment of *compelling evidence of no effect* may be used in rare instances where the evidence indicates that a hazard is unlikely. The evaluation of these factors is used within structured frameworks to make the strength of evidence judgments, as described in **Tables 11-3 and 11-4**, and thus, directly informs the overall evidence integration judgment (see **Section 11.2** and **Table 11-5**). In general, consistent and/or coherent observations of effects across independent studies examining various aspects of exposure or response (e.g., different exposure settings, dose levels or patterns, populations or species, related endpoints) will result in a stronger evidence integration judgment.

Evidence integration is typically performed separately for each major class of health effects (e.g., neurotoxicity, reproductive toxicity). In many cases, however, the development of several evidence integration narratives and associated evidence integration judgments may be necessary to describe a single major health hazard. In practice, it often makes sense to draw inferences at a finer level of specificity of effect (e.g., learning and memory, pregnancy outcomes) and then use these inferences to draw conclusions about the broader health effect categories. Thus, the evaluation of the strength of the human or animal health effects evidence (i.e., applying the considerations in **Table 11-2** within the frameworks in **Tables 11-3 and 11-4**) will preferably occur at the most specific health outcome level possible (e.g., an analysis at the level of decreased pulmonary function is generally preferable to an analysis of the broader category of respiratory system effects), if there is an adequate set of studies for analyses at this level and considering the interrelatedness of the outcomes studied in that evidence base. If studies on a target system are sparse or varied, or if the

interpretation of evidence strength relies largely on the consideration of coherence across related outcomes, then the analyses may need to be conducted at a broader health effect level. The factors judged to increase or decrease the strength of the evidence are documented in tabular format using the evidence profile table, as previously described.

For human and animal evidence, the analyses of each consideration in **Table 11-2** are used to judge the strength of evidence for the separate evidence streams (see **Tables 11-3** and **11-4**). While the application of the criteria in these tables is mostly straightforward, it is important to emphasize the difficult situation of addressing inconsistent results. One of the more common scenarios in IRIS assessments involves evidence consisting of a handful of well-conducted studies, with several studies showing a relatively consistent effect (e.g., on the same endpoint; on closely related endpoints) with a comparable or even greater number of studies of similar design not demonstrating effects on those endpoints ("null" studies). In such scenarios, it is important to not only look at study confidence and the specific parameters of the study design (e.g., comparability of the exposure levels or durations, or animal ages at exposure) to evaluate whether the data likely represent "differing results" (U.S. EPA, 2005b) or are in fact "conflicting evidence," but to also weigh other parameters related to causality before making a decision. For example, if the several positive studies exhibit effects of relatively small magnitude, with no evidence of a dose-response relationship or mechanistic linkage, a conclusion of *slight* might be more appropriate than a conclusion of *moderate* (or, in a scaled-up scenario with a larger number of *high* or *medium* confidence studies, *moderate* rather than *robust*), unless there is an explanation such as study conduct or sensitivity for the differing results. For these borderline decisions, it is helpful to ensure that the evidence integration narrative discusses the relative merits of both possibilities and justifies the ultimate decision.

This decision process and the explanatory rationale for the decisions are described in the evidence integration narrative and associated evidence profile table for each health effect. **Section 11.2** provides the criteria that guide how these within-stream judgments inform development of an overall evidence integration judgment for each health effect, and the terms used to summarize those evidence integration judgments.

Table 11-2. Considerations that inform evaluations and judgments of the strength of the evidence

Consideration	Increased evidence strength or certainty (of the human or animal evidence)	Decreased evidence strength or certainty (of the human or animal evidence)
The structured criteria in this table will guide the application of strength of evidence judgments for an outcome or health effect (see **Tables 11-3** and **11-4**). Evidence scenarios that do not warrant an increase or decrease in evidence strength will be considered "neutral" and are not described herein (and, in general, are not captured in the assessment-specific evidence profile tables).		
Risk of bias; sensitivity (across studies)	• An evidence base of *high* or *medium* confidence studies increases strength.	• An evidence base of mostly *low* confidence studies decreases strength. An exception to this is an evidence base of studies in which the issues resulting in *low* confidence are related to insensitivity. This may increase evidence strength in cases where an association is identified because the expected impact of study insensitivity is towards the null. • Decisions to increase strength for other considerations in this table should generally not be made if there are serious concerns for risk of bias.
Consistency	• Similarity of findings for a given outcome (e.g., of a similar magnitude, direction) across independent studies[a] or experiments increases strength. The increase in strength is larger when consistency is observed across populations (e.g., geographical location) or exposure scenarios in human studies, and across laboratories, populations (in particular, species), or exposure scenarios (e.g., route; timing) in animal studies.	• Unexplained inconsistency [i.e., conflicting evidence; see (U.S. EPA, 2005b)] decreases strength. Generally, strength should not be decreased if discrepant findings can be reasonably explained by study confidence conclusions; variation in population or species, sex, or lifestage; exposure patterns (e.g., intermittent or continuous); levels (*low* or *high*); or exposure duration. A health effect evidence base of a single or a few studies does not, on its own, decrease evidence strength.

Consideration	Increased evidence strength or certainty (of the human or animal evidence)	Decreased evidence strength or certainty (of the human or animal evidence)
Strength (effect magnitude) and precision	• Evidence of a large magnitude effect (considered either within or across studies) can increase strength. Rare or severe effects, even if they are of a small magnitude, may also increase evidence strength. • Precise results from individual studies or across the set of studies increases strength, noting that biological significance is prioritized over statistical significance.	• Strength may be decreased if effect sizes that are small in magnitude are concluded not to be biologically significant, or if there are only a few studies with imprecise results.
Biological gradient/ dose-response	• Evidence of dose-response increases strength. Dose-response may be demonstrated across studies or within studies and it can be dose- or duration-dependent. It may also not be a monotonic dose-response (monotonicity should not necessarily be expected, e.g., different outcomes may be expected at low vs. high doses due to activation of different mechanistic pathways or induction of systemic toxicity at very high doses). • Decreases in a response after cessation of exposure (e.g., symptoms of current asthma) also may increase strength by increasing certainty in a relationship between exposure and outcome (this is most applicable to epidemiology studies because of their observational nature).	• A lack of dose-response when expected based on biological understanding and having a wide-range of doses/exposures evaluated in the evidence base can decrease strength. • In experimental studies, strength may be decreased when effects resolve under certain experimental conditions (e.g., rapid reversibility after removal of exposure). However, many reversible effects are of high concern. Deciding between these situations is informed by factors such as the toxicokinetics 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, as well as the exposure context focus of the assessment (e.g., addressing intermittent or short-term exposures). • In some cases, and typically only in toxicology studies, the magnitude of effects at a given exposure level might decrease with longer exposures (e.g., due to tolerance or acclimation). Like the discussion of reversibility above, a decision about whether this decreases evidence strength depends on the exposure context focus of the assessment and other factors. • If the data are not adequate to evaluate a dose-response pattern, then strength is neither increased nor decreased.

Consideration	Increased evidence strength or certainty (of the human or animal evidence)	Decreased evidence strength or certainty (of the human or animal evidence)
Coherence	• Biologically related findings within an organ system, or across populations (e.g., sex) increase strength, particularly 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.	• An observed lack of expected coherent changes (e.g., well-established biological relationships) will typically decrease evidence strength. However, certainty in 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 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.
Mechanistic evidence related to biological plausibility	• Mechanistic evidence of precursors or health effect biomarkers in well-conducted studies of exposed humans or animals, in appropriately exposed human or animal cells, or other relevant human, animal, or in silico models (including NAMs) increases strength, particularly when this evidence is observed in the same cohort/population exhibiting the phenotypic health outcome. • Evidence of changes in biological pathways, or support for a proposed MOA in appropriate models also increases strength, particularly when support is provided for rate-limiting or key events or is conserved across multiple components of the pathway or MOA.	• Mechanistic understanding is not a prerequisite for drawing a conclusion that a chemical causes a given health effect; thus, an absence of knowledge should not be used a basis for decreasing strength (NTP, 2015; NRC, 2014). • Mechanistic evidence in well-conducted studies that demonstrate that the health effect is unlikely to occur can decrease certainty in the evidence from human or animal health effect studies. A decision to decrease certainty depends on an evaluation of the strength of the mechanistic evidence 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). Likewise, based on evaluating the relative strengths of the opposing evidence, it may be determined that the mechanistic evidence demonstrates that the observed health effect(s) are only likely to occur under certain scenarios (e.g., above certain exposure levels).

[a]Publication bias has the potential to result in strength of evidence judgments that are stronger than would be merited if the entire body of research were available. However, the existence of publication bias can be difficult to determine (see **Section 9.4.3** for additional discussion). If strong evidence of publication bias exists for an outcome, the increase in evidence strength resulting from considering the consistency of the evidence across studies may be reduced.

Table 11-3. Framework for strength of evidence judgments (human evidence)

Strength of evidence judgment	Description
Robust (⊕⊕⊕) …evidence in human studies (strong signal of effect with little residual uncertainty)	A set of *high* or *medium* confidence independent studies reporting an association between the exposure and the health outcome, 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; and an exposure response gradient is demonstrated. Supporting evidence, such as associations with biologically related endpoints in human studies (coherence) or large estimates of risk or severity of the response, may help to rule out alternative explanations. Similarly, mechanistic evidence from exposed humans may serve to address uncertainties relating to exposure-response, temporality, coherence, and biological plausibility (i.e., providing evidence consistent with an explanation for how exposure could cause the health effect based on current biological knowledge) such that the totality of human evidence supports this judgment.
Moderate (⊕⊕⊙) …evidence in human studies (signal of effect with some uncertainty)	• Multiple studies showing generally consistent findings, including at least one *high* or *medium* confidence study and supporting evidence, but with some residual uncertainty due to potential chance, bias, or confounding (e.g., effect estimates of low magnitude or small effect sizes given what is known about the endpoint; uninterpretable patterns with respect to exposure levels). Associations with related endpoints, including mechanistic evidence from exposed humans, can address uncertainties relating to exposure response, temporality, coherence, and biological plausibility, and any conflicting evidence is not from a comparable body of higher confidence, sensitive studies.[a] • A single *high* or *medium* confidence study demonstrating an effect with one or more factors that increase evidence strength, such as: a large magnitude or severity of the effect, a dose-response gradient, unique exposure or outcome scenarios (e.g., a natural experiment), or supporting coherent evidence, including mechanistic evidence from exposed humans. There are no comparable studies of similar confidence and sensitivity providing conflicting evidence, or if there are, the differences can be reasonably explained (e.g., by the populations or exposure levels studied (U.S. EPA, 2005b)).

Strength of evidence judgment	Description
Slight (⊕⊙⊙) …evidence in human studies (signal of effect with large amount of uncertainty)	One or more studies reporting an association between exposure and the health outcome, where considerable uncertainty exists: • A body of evidence, including scenarios with one or more *high* or *medium* confidence studies reporting an association between exposure and the health outcome, where either (1) conflicting evidence exists in studies of similar confidence and sensitivity (including mechanistic evidence contradicting the biological plausibility of the reported effects),[a] (2) a single study without a factor that increases evidence strength (factors described in *moderate*), OR (3) considerable methodological uncertainties remain across the body of evidence (typically related to exposure or outcome ascertainment, including temporality), AND there is no supporting coherent evidence that increases the overall evidence strength. • A set of only *low* confidence studies that are largely consistent. • Strong mechanistic evidence in well-conducted studies of exposed humans (*medium* or *high* confidence) or human cells (including NAMs), in the absence of other substantive data, where an informed evaluation has determined that the data are reliable for assessing the health effect of interest and the mechanistic events have been reasonably linked to the development of that health effect.[b] This category serves primarily to encourage additional study where uncertain evidence does exist that might provide some support for an association.
Indeterminate (⊙⊙⊙) …evidence in human studies (signal cannot be determined for or against an effect)	• No studies in humans or well-conducted studies of human cells. • Situations when the evidence is highly inconsistent and primarily of *low* confidence. • May include situations with *medium* or *high* confidence studies, but unexplained heterogeneity exists (in studies of similar confidence and sensitivity), and there are additional outstanding concerns such as effect estimates of low magnitude, uninterpretable patterns with respect to exposure levels, or uncertainties or methodological limitations that result in an inability to discern effects from exposure. • A set of largely null studies that does not meet the criteria for *compelling evidence of no effect*, including evidence bases with inadequate testing of susceptible populations and lifestages.
Compelling evidence of no effect[c] (———) …in human studies (strong signal for lack of an effect with little uncertainty)	Several *high* confidence studies showing null results (for example, 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 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.

[a]Different strength of the evidence judgments are possible in scenarios with unexplained heterogeneity across sets of studies with similar confidence and sensitivity. Specifically, this judgment considers the level of support (or lack thereof) provided by evaluations of the Hill considerations, including the magnitude or severity of the effects (larger or more severe effects can provide stronger evidence; see **Table 11-2**), coherence of related findings (including mechanistic evidence), dose-response, and biological plausibility, as well as the comparability of the supporting and conflicting evidence (e.g., the specific endpoints tested, or the methods used to test them; the

specific sources of bias or insensitivity in the respective sets of studies). The evidence-specific factors supporting either of these evidence integration judgments are clearly articulated in the evidence integration narrative.

[b]This determination is based on expert judgment dependent on the state-of-the-science at the time of review. Scientific understanding of toxicity mechanisms 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 and in silico testing and other NAMs) will continue to increase. Thus, the sufficiency of mechanistic evidence alone for identifying potential hazards is expected to increase as the science evolves. As NAMs and efforts to validate non-traditional evidence for use in human health assessment mature, it is expected that such evidence scenarios will be sufficient for a determination of *moderate* in the future. The understanding of such evidence scenarios at the time of handbook development is consistent with a determination of (the upper end of) *slight*.

[c]The criteria for this category are intentionally more stringent than those justifying a conclusion of *robust*, consistent with the "difficulty of proving a negative" [as discussed in (U.S. EPA, 1996b, 1991, 1988)].

Table 11-4. Framework for strength of evidence judgments (animal evidence)

Strength of evidence judgment	Description
Robust (⊕⊕⊕) …evidence in animals (strong signal of effect with little residual uncertainty)	A set of *high* or *medium* confidence experiments with consistent findings of adverse or toxicologically significant effects across multiple laboratories, exposure routes, experimental designs (e.g., a subchronic study and a two-generation study), or species; and the experiments reasonably rule out the potential for nonspecific effects to have caused the effects of interest. Any inconsistent evidence (evidence that cannot be reasonably explained based on study design or differences in animal model) is from a set of experiments of lower confidence or sensitivity. To reasonably rule out alternative explanations, multiple additional factors in the set of experiments exist, such as: coherent effects across biologically related endpoints; 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. Similarly, mechanistic evidence (e.g., precursor events linked to adverse outcomes) in animal models may exist to address uncertainties in the evidence base such that the totality of animal evidence supports this judgment.
Moderate (⊕⊕⊙) …evidence in animals (signal of effect with some uncertainty)	• At least one *high* or *medium* confidence study with supporting information increasing the strength of the evidence. Although the results are largely consistent, notable uncertainties remain. However, in scenarios when inconsistent evidence or evidence indicating nonspecific effects exist, it is not judged to reduce or discount the level of concern regarding the positive findings, or it is not from a comparable body of higher confidence, sensitive studies.[a] The additional support provided includes either consistent effects across laboratories or species; coherent effects across multiple related endpoints; 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. Mechanistic evidence in animals may serve to provide this support or otherwise address residual uncertainties. • A single *high* or *medium* confidence experiment demonstrating an effect in the absence of comparable experiment(s) of similar confidence and sensitivity providing conflicting evidence, namely evidence that cannot be reasonably explained (e.g., by respective study designs or differences in animal model) (U.S. EPA, 2005b).
Slight (⊕⊙⊙) …evidence in animals (signal of effect with large amount of uncertainty)	Scenarios in which there is a signal of a possible effect, but the evidence is conflicting or weak: • A body of evidence, including scenarios with one or more *high* or *medium* confidence experiments reporting effects but without supporting or coherent evidence (see description in *moderate*) that increases the overall evidence strength, where conflicting evidence exists from a set of sensitive experiments of similar or higher confidence (including mechanistic evidence contradicting the biological plausibility of the reported effects).[d] • A set of only *low* confidence experiments that are largely consistent. • Strong mechanistic evidence in well-conducted studies of animals or animal cells (including NAMs), in the absence of other substantive data, where an informed evaluation has determined the assays are reliable for assessing the health effect of interest and the mechanistic events have been reasonably linked to the development of that health effect.[b] This category serves primarily to encourage additional research in situations where uncertain evidence does exist that might provide some support for an association.

Strength of evidence judgment	Description
Indeterminate (⊙⊙⊙) …evidence of the effect under review in animals (signal cannot be determined for or against an effect)	• No animal studies or well-conducted studies of animal cells. • The available models (not considering human relevance) or endpoints are not informative to the hazard question under evaluation. • The evidence is inconsistent and primarily of *low* confidence. • May include situations with *medium* or *high* confidence studies, but there is unexplained heterogeneity and additional concerns such as small effect sizes (given what is known about the endpoint) or a lack of dose-dependence. • A set of largely null studies that does not meet the criteria for *compelling evidence of no effect*.
Compelling evidence of no effect[c] (———) …in animals (strong signal for lack of an effect with little uncertainty)	A set of *high* confidence experiments examining a reasonable spectrum of endpoints relevant to a type of toxicity 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. The experiments were designed to specifically test for effects of interest, including suitable exposure timing and duration, post exposure latency, and endpoint evaluation procedures, and to address potentially susceptible populations and lifestages. Mechanistic data in animals (in vivo or in vitro) that address the above considerations or that provide information supporting the lack of an association between exposure and effect with reasonable confidence may provide additional support such that the totality of evidence supports this judgment.

[a]Different strength of the evidence judgments are possible in scenarios with unexplained heterogeneity across sets of studies with similar confidence and sensitivity. Specifically, this judgment considers the level of support (or lack thereof) provided by evaluations of the Hill considerations, including the magnitude or severity of the effects (larger or more severe effects can provide stronger evidence; see **Table 11-2**), coherence of related findings (including mechanistic evidence), dose-response, and biological plausibility, as well as the comparability of the supporting and conflicting evidence (e.g., the specific endpoints tested, or the methods used to test them; the specific sources of bias or insensitivity in the respective sets of studies). The evidence-specific factors supporting either of these evidence integration judgments are clearly articulated in the evidence integration narrative.

[b]This determination is based on expert judgment dependent on the state-of-the-science at the time of review. Scientific understanding of toxicity mechanisms 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 and in silico testing and other NAMs) will continue to increase. Thus, the sufficiency of mechanistic evidence alone for identifying potential hazards is expected to increase as the science evolves. As NAMs and efforts to validate non-traditional evidence for use in human health assessment mature, it is expected that such evidence scenarios will be sufficient for a determination of *moderate* in the future. The understanding of such evidence scenarios at the time of handbook development is consistent with a determination of (the upper end of) *slight*.

[c]The criteria for this category are intentionally more stringent than those justifying a conclusion of *robust*, consistent with the "difficulty of proving a negative" [as discussed in (U.S. EPA, 1996b, 1991, 1988)].

11.2. OVERALL EVIDENCE INTEGRATION JUDGMENTS

For each health effect or specific cancer type of potential concern, the first sentence of the evidence integration narrative includes the summary judgment (see description in **Table 11-5**)

and, for evidence integration narratives on potential carcinogenicity, includes the cancer descriptor (U.S. EPA, 2005b), as described below. As outlined previously, the narrative also provides a summary of the strength of each evidence stream, including the within-stream judgments drawn and the evidentiary support for those judgments, and the inferences and overall judgments across the evidence streams, with the supporting study-specific design and exposure context provided.

More specifically, for each health effect, the evidence integration narrative should include:

- A descriptive summary of the primary judgments about the potential for health effects (or lack thereof) in exposed humans, based on the following analyses:
  - Judgments regarding the strength of the available human and animal evidence (see **Section 11.1**);
  - Consideration of the coherence of findings (i.e., the extent to which the evidence for health effects and relevant mechanistic changes are similar) across human and animal studies;
  - Other information on the human relevance of findings in animals (e.g., conclusions from analyses of toxicokinetic differences across species; inferences based on related chemicals); and
  - Conclusions drawn based on the focused mechanistic analyses, including evaluations identified during preliminary stages of assessment scoping and problem formulation, as well as well as those based on analyses identified during stepwise consideration of the health effect-specific evidence during draft development (see **Section 10.1**). This should typically include discussion of biological understanding (general knowledge of biological changes associated with the observed effects) and potential mechanisms of toxicity (chemical-specific interactions and alterations leading to the observed effects), including whether and to what extent the mechanisms are likely to be conserved across species.
- A summary of key evidence supporting these judgments, highlighting the evidence that was the primary basis for these judgments and any notable issues (e.g., data quality; coherence of the results), and a narrative expression of confidence (a summary of strengths and remaining uncertainties) for these judgments. Typically, an evidence profile table will be used to summarize the key evidence and decision rationale (see example in **Table 11-1**).
- Information on the general conditions for the expression of these health effects based on the available evidence (e.g., exposure routes and levels in the studies that were the primary drivers of these judgments), noting that these conditions of exposure will be clarified during dose-response analysis (see **Chapter 13**).
- Indications of potentially susceptible populations or lifestages (i.e., an integrated summary of the available evidence on potential susceptible populations and lifestages drawn across the syntheses of the human, animal, and mechanistic evidence).

- A summary of key assumptions used in the analysis, which are generally based on EPA guidelines. Note that the key assumptions for drawing evidence integration judgments are captured in the systematic review protocol.

- Strengths and limitations of the evidence integration judgments, including key uncertainties and data gaps, which evidence was most impactful to the overall judgment, as well as the limitations of the systematic review.

As noted above, integrating the human and animal evidence includes consideration of the available information regarding the potential biological processes involved in chemical-specific toxicity based on the available mechanistic evidence and general biological knowledge about the outcomes. Evidence integration also includes evaluations of the coherence of effects observed across species, the relevance of the animal and in vitro mechanistic evidence to humans, and a summary of the current knowledge on potentially susceptible populations and lifestages. Reiterating considerations described in earlier sections (e.g., **Chapter 10**; **Table 11-2**), some notable examples of how focused mechanistic analyses inform the evidence integration judgments include:

- When there is a strong animal response, evidence establishing that the mechanism(s) underlying the animal response do not operate in humans indicates that the animal response is irrelevant to humans (see considerations in **Table 10-4**). By itself, this scenario provides neither support for, nor support against, the identification of a human hazard. However, it is possible that an extensive body of such animal evidence can meet the criteria for **strong evidence supports no effect** (see **Table 11-5**) for other health effects evaluated across the set of studies that were without an observed response (i.e., if there is experimental support that the animal models are relevant to humans for the other health effects and no conflicting human evidence exists).

- When it is widely accepted (e.g., a large body of high-quality consistent mechanistic literature) that an animal model(s) is inappropriate for the evaluation of a specific human health outcome, then the animal evidence may be considered uninformative for evaluating human health effects and provides neither evidence for, nor evidence against, the identification of a human hazard.

- Observing effects that are known to be biologically related (coherent) across human and animal studies can strengthen judgments of the relevance of animal data to humans, and the evidence overall. This includes observed changes in biological precursors or key mechanistic events. In many cases, mechanistic information is not useful to the separate judgments of the animal and human evidence, but it does inform the interpretation of the evidence when integrating across the animal and human evidence. For example, results indicating changes in precursors or key mechanistic events in humans that relate to apical effects observed in animals but not in humans would increase confidence in the animal evidence. It can be useful to incorporate information on related chemicals or endpoints to help clarify the nature and plausibility of the cross-species association(s). This support for coherence can also be provided even when the effects are on different organ systems (e.g., thyroid effects and neurodevelopment). As previously discussed, it is not necessary (or expected) that effects manifest in humans are identical to those observed in animals

[e.g., U.S. EPA (1991)], although this typically provides stronger evidence. For example, tumors in one animal species can be predictive of carcinogenic potential in humans or other species, but not necessarily at the same site (U.S. EPA, 2005b). This might be due, for example, to cross-species differences in distribution or metabolism, where a carcinogenic metabolite might be formed or distributed to different anatomical sites in different species or in different animal strains. Similarly, malformations at one anatomical site in animals can suggest the potential for developmental toxicity that could appear in another form in humans (U.S. EPA, 1991). Differences in response might be attributable to species differences in critical periods, metabolism, developmental patterns, or mechanisms of action. When effects appear dissimilar, an evaluation of the relatedness of the findings can inform the evidence integration conclusions. In such cases, without sufficient evidence for an alternative approach, specific findings in humans (e.g., breast cancer, kidney malformations) would typically be integrated with animal inferences across a broader level of specificity (e.g., cancers observed at any site; any malformation or other manifestations of developmental toxicity). However, for some health outcomes, there might be sufficient toxicological understanding of MOA to anticipate site concordance; for example, a hormonal MOA operating in endocrine or reproductive organs (U.S. EPA, 2005b).

- A response seen across multiple animal species increases confidence that a relevant mechanistic pathway is conserved and is operating in humans. These interpretations might be further informed by considering mechanistic information on related chemicals and endpoints.

- Even in the absence of experimental evidence demonstrating differences in responses across populations or lifestages, MOA understanding can support the conclusion that there are likely to be pronounced differences in susceptibility (e.g., across humans and animals; across animal species; across sexes or lifestages).

- Other information not previously considered when drawing the within-stream judgments may be useful to incorporate (e.g., consideration and discussion of toxicokinetic data or structure-activity relationships informative to drawing inferences across the available human, animal, and mechanistic evidence).

To draw overall evidence integration judgments, the second stage of evidence integration combines the judgments regarding the animal and human evidence while also considering mechanistic information on the human relevance of the findings in animals, relevance of the mechanistic evidence to humans, coherence across bodies of evidence, and information on susceptible populations and lifestages. **Table 11-5** describes the five evidence integration judgment levels, the summary language associated with each level, and the types of evidence that fit each level. The five evidence integration judgment levels reflect the differences in the amount and quality of the data that inform the evaluation of whether exposure may cause the health effect(s) under specified exposure conditions. Notably, as these categories overlay a continuum of potential overall evidence strength, borderline judgments should be more fully characterized within the evidence integration narrative. Likewise, as the summary judgment levels within the evidence integration narrative reflect a high-level summary characterization, it is important to include the

primary evidentiary basis for each judgment, including the experimental model or observed population, and exposure levels tested or estimated.

The strength of the evidence for each hazard (or lack thereof) is important to characterize in the evidence integration narrative for several reasons, the foremost of which is its use in selecting studies and outcomes for use in dose-response analysis (see **Section 12.2** and **Chapter 13**). Each summary judgment should be provided alongside information on the characteristics of exposure in the studies providing the primary supporting evidence, which will then be refined to better estimate the necessary conditions of exposure during dose-response analysis (**Chapter 13**). Consistent with EPA noncancer and cancer guidelines, a judgment that **strong evidence supports no effect** should only be used when the available data supporting an apparent lack of an effect are extensive and considered to be reliable and complete (as described in **Tables 11-3, 11-4, and 11-5**); lesser levels of evidence suggesting a lack of an effect are characterized as **evidence inadequate**.

Table 11-5. Evidence integration judgments for characterizing potential human health hazards in the evidence integration narrative

Overall 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] under relevant exposure circumstances. 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[d]].	**Evidence demonstrates**	A strong evidence base demonstrating that [chemical] exposure causes [health effect] in humans. • This conclusion level is used if there is *robust* human evidence supporting an effect. • This conclusion level could also be used with *moderate* human evidence and *robust* animal evidence if there is strong mechanistic evidence that the findings in animals are anticipated to occur and progress in humans. Most notably, an MOA interpreted with reasonable certainty would rule out alternative explanations. The evidence integration narrative should further characterize this judgment by discussing whether there was adequate testing of potentially susceptible populations and lifestages, based on the assessed health effect and chemical knowledge (e.g., toxicokinetics).
The currently available **evidence indicates** that [chemical] likely causes [health effect] in humans under relevant exposure circumstances. 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]**	An evidence base that indicates that [chemical] exposure likely causes [health effect] in humans, although there may be outstanding questions or limitations that remain, and the evidence is insufficient for the higher conclusion level. • This conclusion level is used if there is *robust* animal evidence supporting an effect and *slight*-to-*indeterminate* human evidence, or with *moderate* human evidence when strong mechanistic evidence is lacking.[f] • This conclusion level could also be used with *moderate* human evidence supporting an effect and *slight* or *indeterminate* animal evidence, or with *moderate* animal evidence supporting an effect and *slight* or *indeterminate* human evidence. In these scenarios, any uncertainties in the *moderate* evidence are not sufficient to substantially reduce confidence in the reliability of the evidence, or mechanistic evidence in the *slight* or *indeterminate* evidence base (e.g., precursors) exists to increase confidence in the reliability of the *moderate* evidence.

Overall evidence integration judgment[a] in narrative	Evidence integration judgment level	Explanation and example scenarios[b]
The currently available **evidence suggests** that [chemical] may cause [health effect] in humans under relevant exposure circumstances. 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 but is not sufficient to infer**[g]	An evidence base that suggests that [chemical] exposure may cause [health effect] in humans, but there are very few studies that contributed to the evaluation, the evidence is very weak or conflicting, and/or the methodological conduct of the studies is poor. • This conclusion level is used if there is *slight* human evidence and *indeterminate*-to-*slight* animal evidence. • This conclusion level is also used with *slight* animal evidence and *indeterminate*-to-*slight* human evidence. • This conclusion level could also be used with *moderate* human evidence and *slight* or *indeterminate* animal evidence, or with *moderate* animal evidence and *slight* or *indeterminate* human evidence. In these scenarios, there are outstanding issues regarding the *moderate* evidence that substantially reduced confidence in the reliability of the evidence, or mechanistic evidence in the *slight* or *indeterminate* evidence base (e.g., null results in well-conducted evaluations of precursors) exists to decrease confidence in the reliability of the *moderate* evidence. • Supplemental evidence (e.g., read-across) supporting a general scientific understanding of mechanistic events that result in the health effect could also be used if the mechanistic evidence is sufficient to highlight potential human toxicity[h]—in the absence of informative conventional studies in humans or in animals (i.e., *indeterminate* evidence in both).
The currently available **evidence is inadequate** to assess whether [chemical] may cause [health effect] in humans under relevant exposure circumstances.	**Evidence inadequate**[i]	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). • This conclusion level is used if there is *indeterminate* human and animal evidence. • This conclusion level is also used with *slight* animal evidence and *compelling evidence of no effect* human evidence. • This conclusion level could also be used with *compelling evidence of no effect* animal evidence and *indeterminate* human evidence if the database lacks experimental support that the models are relevant to humans for the effect of interest.

Overall evidence integration judgment[a] in narrative	Evidence integration judgment level	Explanation and example scenarios[b]
		• This conclusion level could also be used with *slight*-to-*robust* animal evidence and *indeterminate* human evidence if strong experimental evidence (e.g., a MOA interpreted with reasonable certainty) indicates the findings in animals are 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 indicates that the available evidence is insufficient to reach conclusions.
Strong evidence supports no effect in humans under relevant exposure circumstances. This conclusion is based on studies of [humans or animals] that assessed [exposure or dose] levels of [range of concentrations].	**Strong evidence supports no effect**	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 *high* 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 heath effect of interest. • This conclusion level is used if there is *compelling evidence of no effect* in human studies and *compelling evidence of no effect* to *indeterminate* in animals. • This conclusion level is also used if there is *indeterminate* human evidence and *compelling evidence of no effect* animal evidence in models with experimental support that the models are relevant to humans for the effect of interest. • This conclusion level could also be used with *compelling evidence of no effect* in human studies and *moderate*-to-*robust* animal evidence if strong mechanistic information indicates that the animal evidence is unlikely to be relevant to humans.

[a]Evidence 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**. These determinations regarding confidence in the evidence supporting hazard are useful for other assessment decisions, including prioritizing studies and outcomes in quantitative analyses and characterizing assessment uncertainties (see **Sections 12.2 and 13.4**). Thus, for all evidence scenarios, but particularly for those in the lower end of this range, it is important to characterize the uncertainties in the evidence base within the evidence integration narrative and convey the evidence strength to subsequent steps, including toxicity values developed based on those effects.

[b]Terminology of "is" refers to the default option; terminology of "could also be" refers to situational options dependent on mechanistic understanding.

[c]In 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.

[d]If concentrations cannot be estimated, an alternative expression of exposure level such as "occupational exposure levels," will be provided. This applies to all conclusion levels.

Overall evidence integration judgment[a] in narrative	Evidence integration judgment level	Explanation and example scenarios[b]

[e]For 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.

[f]The strength of the evidence is neither increased or decreased due to a lack of experimental information on the human relevance of the animal evidence or mechanistic understanding (mechanistic evidence may exist, but it is inconclusive); in these cases, the animal data are judged not to conflict with current biological understanding (general knowledge of biological changes associated with the observed effects) and thus are assumed to be relevant, while findings in humans and animals are presumed to be real unless proven otherwise.

[g]Health effects characterized as having **evidence demonstrates** and **evidence indicates (likely)** (and, in some cases, **evidence suggests**) are evaluated for use in dose-response assessment (see **Chapter 12**). When the database includes at least one well-conducted study and a judgment of **evidence suggests** is drawn, quantitative analyses may still be useful for some purposes (e.g., providing a sense of the magnitude and uncertainty of estimates for health effects of potential concern, ranking potential hazards, or setting research priorities), but not for others [see related discussions in (U.S. EPA, 2005b)]. It is critical to transparently convey the extreme uncertainty in any such estimates.

[h]This determination is based on expert judgment dependent on the state-of-the-science at the time of review. As previously discussed (see **Section 11.1**), scientific understanding of toxicity mechanisms 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 and in silico testing and other NAMs) will continue to increase. Thus, the sufficiency of mechanistic evidence alone for identifying potential hazards is expected to increase as the science evolves. The understanding of such evidence scenarios at the time of handbook development is consistent with a determination of **evidence suggests**.

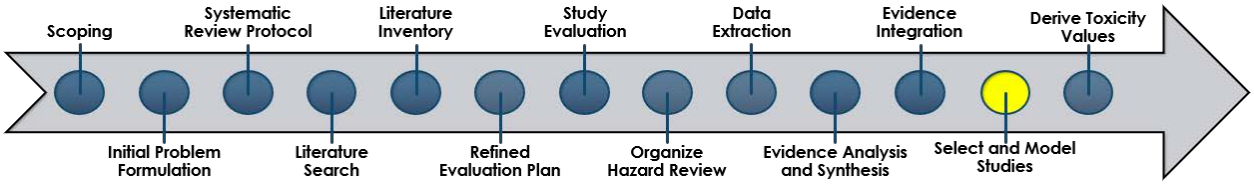
[i]Specific narratives for each of the health effects with an evidence integration judgment of **evidence inadequate** may be deemed unnecessary.

For evaluations of carcinogenicity, consistent with EPA Cancer Guidelines (U.S. EPA, 2005b), one of EPA's standardized cancer descriptors is used as a shorthand characterization of the evidence integration narrative, describing the overall potential for carcinogenicity across all potential cancer types. These 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***. In some cases, mutagenicity should also be evaluated (e.g., when there is evidence of carcinogenicity) because it influences the approach to dose-response assessment and subsequent application of adjustment factors for exposures early in life (U.S. EPA, 2005b, c).

An appropriate cancer descriptor is selected as described in EPA Cancer Guidelines (U.S. EPA, 2005b). For each assessed cancer subtype, an evidence integration narrative and summary judgment should be provided, as described above (see **Table 11-5**). Separately, the evidence integration narrative and cancer descriptor for potential carcinogenicity consider the interrelatedness of cancer types potentially related to chemical exposure, consistency across the human and animal evidence for any cancer type [noting that site concordance is not required (U.S. EPA, 2005b)], and the uncertainties associated with each assessment-specific conclusion. In general, however, if a systematic review of more than one cancer type was conducted, then the overall judgment and discussion of evidence strength in the evidence integration narrative for the cancer type(s) with the strongest evidence for hazard should be used as the driver (considering the totality of evidence on carcinogenicity) to inform selection of the cancer descriptor, with each assessment providing a transparent description of the decision rationale. The cancer descriptor and evidence integration narrative for potential carcinogenicity, including application of the MOA framework, 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, 2005b, c).

For both noncancer health effects and carcinogenicity, it is important to transparently and succinctly convey the evidence integration judgments, the supporting rationale, and the key data supporting those decisions. More than one judgment can be made when a chemical's effects differ by dose or exposure route; if the database supports such analyses, this decision should be clarified based on a focused review of the mechanistic evidence or a more detailed dose-response analysis (see **Chapter 13**). Throughout this expert judgment-driven decision process, there can be instances where it may make sense to lay out both sides of a controversial argument (as well as the implications of each) before drawing evidence integration conclusions. In such instances, the evidence integration narrative should be clear and transparent in articulating the rationale for the final decision(s). These judgments and justifications are carried forward to inform decisions for dose-response analysis, including study and endpoint selection, as well as model selection and characterizations of uncertainty during the derivation of toxicity values (see **Chapters 12 and 13**).

12. HAZARD CONSIDERATIONS AND STUDY SELECTION FOR DERIVING TOXICITY VALUES



HAZARD CONSIDERATIONS AND SELECTING STUDIES 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.

Who

- Assessment team and disciplinary workgroups.

What

- For each health outcome under consideration, a dose-response analysis plan that weighs the strengths and weaknesses of the relevant data for deriving 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 (**Chapter 11**). 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)** as well as for carcinogenicity descriptors of ***carcinogenic to humans*** or ***likely to be 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 may be useful for some purposes when the evidence includes a well-conducted study (particularly when that study may 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 cancer guidelines 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 may 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, 2005b)." Toxicity values should not be developed for other evidence integration judgments.

As discussed in **Section 12.1**, 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 13**). 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 may differ among outcomes (generally for animal studies). **Section 12.2** builds upon these considerations, as well as general principles of dose response analysis, to prioritize the studies most appropriate for use in deriving toxicity values.

12.1. HAZARD CONSIDERATIONS FOR DOSE-RESPONSE

This section of the assessment at the end of the hazard identification chapter 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 toxicokinetic 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 assessment of a given health effect. Health effects that were analyzed in relation to 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 may be selected as the most representative endpoint to consider for dose-response analysis. This section may 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 that are 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. The existence and validity of physiologically based pharmacokinetic (PBPK) models or toxicokinetic information that may allow the estimation of internal dose for route-to-route extrapolation should be presented. 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).

This 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 may 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 may be conferred by lifestage, then this should be explicitly discussed. Differences in absorption, distribution, metabolism, and excretion (ADME) may 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, may 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 may confer susceptibility are typically summarized and evaluated with respect to patterns across studies pertinent to consistency, coherence, and the magnitude and direction of effect measures. Relevant factors may 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). **Table 12-1** provides an incomplete list of examples that could define a susceptible population or lifestage.

There may 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 may 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 12-1. Individual and social factors that may 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

12.2. 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. In some cases (generally, when more evidence is available), this strength of the evidence characterization (see **Section 11.2**) 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 (see **Chapter 11** for discussion on how different studies can influence the overall judgments) are considered for deriving toxicity values; thus, this should focus almost exclusively on *high* or *medium* confidence studies. However, there are additional considerations specific to their use in quantitative analyses, as discussed in **Section 12.2.1.** 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.

12.2.1. 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 discussed within the narrative hazard summaries, 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 12-2** 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 demonstrate many of the preferred considerations, but in different combinations, so that it is not clear that one data set 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 may be relevant (see **Section 13.4**).

Table 12-2. Attributes used to evaluate studies for derivation of toxicity values

Study attributes		Considerations	
		Human studies	**Animal studies**
Study confidence		*High* or *medium* confidence studies (see **Section 6**) are highly preferred over *low* 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 *high* and *medium* confidence studies are further differentiated based on 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 toxicodynamics, 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 to be 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 toxicokinetic 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 *Benchmark Dose Technical Guidance*, §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 may 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 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.

[a]In epidemiology 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 **Section 6**), some aspects (e.g., covariate-adjusted effect estimates) are important to reconsider for developing more informative quantitative estimates.

[b]Power 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 (Hoenig 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 13.5**). 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 to 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.[20] 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 may also be possible to derive a toxicity value by combining suitable studies in an endpoint-specific database in a metaregression 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 12.2.2**.

In addition to the more general considerations described above, specific statistical issues may 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* (U.S. EPA, 2012b). Several important considerations from the guidance concerning the levels and patterns of response observed across treatment groups are highlighted here:

- 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 benchmark response (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 point of departure (POD) and minimizing the need for low-dose extrapolation.

- The minimum data set to be used for estimating the BMD and BMDL should show a biologically or statistically significant dose-related trend in response for the selected endpoint(s) [see §2.1.5, (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

[20]Note 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).

studies showing stronger 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 12.2**).

- In cases where the biologic 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, biologic significance may be inferred from other data on the same chemical and endpoint [see §2.1.5, (U.S. EPA, 2012b)].

- Dose-response analysis may 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). If the one elevated response is near the BMR, however, adequate benchmark dose (BMD) and benchmark dose lower confidence limit (BMDL) computation may result (Kavlock et al., 1996). Also, fitting multiple models to the dataset 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, may 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 addition to providing a thorough rationale for the studies selected for dose-response analysis, the dose-response chapter of the assessment outlines the various reasons for not analyzing particular studies quantitatively and considers the impact on the overall toxicity value derivation of excluding any data sets judged not suitable for dose-response analysis.

12.2.2. 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. (Note that this type of analysis is distinct from meta-analysis, described in **Section 9.4.2**.) The evaluation of study strengths and similarities described above (see **Section 12.2.1**) 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 may be desirable to combine the data to obtain more precise estimates of PODs [see the 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 may 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 and/or strain studied, sex, and route of exposure. It may 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 may increase the precision of the estimated POD and may 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 12.2.1**)*?* 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 may 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 may be useful for estimating a common (internal) dose measure, particularly across routes of exposure.

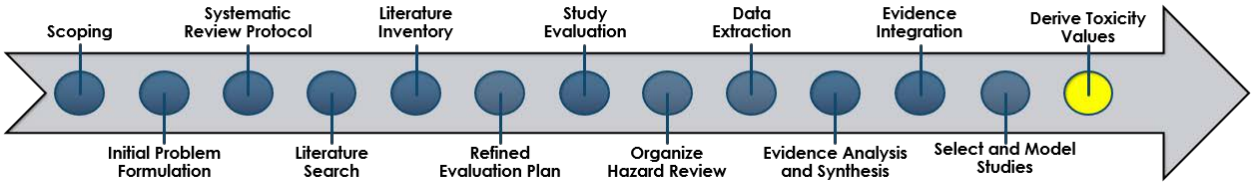
- *Is there evidence of homogeneous responses to exposure?* Species, sexes, and lifestages often differ in dose-response, so convincing evidence of similar responses would be needed to consider combining the data from these groups. For example, a hypothesis test of no difference across groups can be 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, then 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 may be considered among inclusion criteria, but a lack of statistical significance may 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 may 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, and categorical regression).

13. DERIVATION OF TOXICITY VALUES



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.

Who

- Assessment team with statistician and other disciplines as needed for each outcome.

What

- Toxicity values that represent current scientific understanding, including transparent evaluation of associated uncertainties.

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 *Guidelines for Carcinogen Risk Assessment* (U.S. EPA, 2005b), and the EPA *Supplemental Guidance for Assessing Susceptibility from Early-Life Exposure to Carcinogens* (U.S. EPA, 2005c). Some familiarity with the development and use of these toxicity values is presumed. 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 that they typically occur in this process, and include selecting benchmark response (BMR) values (see **Section 13.1**), dose characterization and dose-response modeling (see **Section 13.2**), developing candidate toxicity values (see **Section 13.3**), characterizing uncertainty and confidence (see **Section 13.4**), and selecting final toxicity values (see **Section 13.5**). These topics build from the selection of hazards, studies, and outcomes for dose-response analyses, as discussed in **Chapter 12**.

13.1. SELECTING BENCHMARK RESPONSE VALUES FOR DOSE-RESPONSE MODELING

When dose-response modeling is feasible and appropriate (see **Chapter 12**), the BMR that determines the point of departure (POD) for each toxicity value is selected, irrespective of the particular dose-response models under consideration (e.g., multistage), prior to modeling. 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, 2005b), p 1–11, Footnote 3], as discussed further below.

When linear low-dose extrapolation is used, 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 epidemiologic cancer data (U.S. EPA, 2012b, 2005b). 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 biologic considerations are taken into account when selecting the BMR. For deriving an RfD or RfC, the objective is to determine an exposure level that is "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.[21] The following recommendations for BMR selection for nonlinear low-dose extrapolation (for both human and animal effects) focus on biologic 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 benchmark doses (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 biologic 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,

[21]The 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 lowest-observed-adverse-effect level (LOAEL). In either case, an adjustment for extrapolating to a lower exposure, such as a LOAEL-to-no-observed-adverse-effect level (NOAEL) uncertainty factor (UF), 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].

13.2. 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 that are generally needed to develop toxicity values [(U.S. EPA, 2012b, 2005b), §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 13.1**.

2) To derive toxicity values, extrapolation below the POD is typically necessary. This step is described further in **Section 13.3**, 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 12**). Key practices are described in **Section 13.2.1** for modeling human data and in **Section 13.2.2** for modeling animal data.

13.2.1. Exposure-Response Modeling of Human Data

Observational epidemiology studies require evaluation of several attributes, as described in **Sections 6.2 and 12.1**, before conducting exposure-response modeling. If multiple human studies are suitable for exposure-response modeling and if no single study is judged to be appreciably better than the others for the purposes of deriving toxicity values, data or results from multiple studies may be combined where justified, or toxicity values may 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 may include a lag period, if warranted. Additionally, data on incident cases are generally preferred over mortality data (U.S. EPA, 2005b), 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, 2005b).

The modeling of cancer epidemiology data typically involves relative risk models. For grouped or categorical exposure data, results may not be sufficiently precise to discern the shape of the exposure-response relationship, and a linear model is often used (U.S. EPA, 2005b). 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 13.1**, a BMR of 1% extra risk is typically used for relatively common cancers; a lower BMR, for example for less common cancers, may be more suitable for establishing a POD near the lower end of the observed range [(U.S. EPA, 2005b), §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, 2005b); see §2.2.1.1., §3.2.1, §3.3.5.; also see **Section 13.2.3**].

Noncancer Data

Grouped epidemiological data for noncancer effects may be modeled by Benchmark Dose Software (BMDS) models, in the same way as grouped laboratory animal data (see **Section 13.2.2**). Some situations, such as the need to account for covariates, may 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 13.1** for general recommendations).

In some circumstances with adequate human epidemiological data for noncancer effects, the output of the dose-response analysis may be dose-response functions and associated risk-specific doses, rather than a BMD and reference value (NRC, 2013).

13.2.2. 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 follow 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 biologic effect or to a reliable surrogate measure. The active agent may 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, 2005b), §3.1; (U.S. EPA, 1994), §4.3]. See **Section 6.5** for more information.

Use of 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, that result in the same internal dose POD, thereby completing the interspecies extrapolation. This approach may be preferred if the data being modeled are in a nonlinear PBPK range, as it may 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 much below the range of exposures needed for the extrapolation, the PBPK results may 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 UF for human variation (see **Section 13.3.2**) to the animal-based POD prior to application of the human PBPK model (doing some prior to PBPK-based dosimetric adjustments may allow the PBPK model to do those adjusts in a dose-range that is calibrated for, although this is not always the case).

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. Before it is accepted for such use, a PBPK model considered for route-to-route extrapolation 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, to pass a quality review (metabolism and elimination 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, or 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 toxicodynamics differences, including those closely related to toxicokinetics are not accounted for by the model. For example, Oshiro et al. (2014) observed that inhaled ethanol exposures in rats did not induce the same neurodevelopmental outcomes that would be expected following oral exposure, despite similarly high internal doses. This discrepancy may be due to toxicodynamic aspects not considered in the model, such as how different internal exposure time-profiles may impact outcomes. For example, a degree of acclimation can occur from a slow increase in blood and tissue concentration that occurs during inhalation exposure, that does not occur during the rapid concentration increase that occurs

after a bolus oral exposure, while the two exposures may result in similar AUCs. Further, if one only had the Oshiro et al. (2014) inhalation bioassay as a measure of ethanol's developmental effects, and used a PBPK model with the dose metric that is presumed to best estimate risk (AUC) for an inhalation-to-oral extrapolation, the result would be a significant underprediction of response or risk from oral exposure to ethanol.

Therefore, in the case of non-cancer assessments when a PBPK model is required and used for route-to-route extrapolation, the potential added uncertainties from this application may 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). The same choice may be appropriate for a cancer risk assessment where a "nonlinear" mode of action; i.e., when there are sufficient mechanistic data to establish that mode of action. For cancer risk assessment, if there is a clear mutagenic mode of action, the expectation that cumulative risk is proportional to AUC as a predictor of total genetic damage is stronger. Hence it may then be appropriate to use a PBPK model for route-to-route extrapolation, even though there is not a formal mechanism to account for the increased uncertainty, because there is greater general confidence in use of AUC as a measure of risk.

The methodology for route-to-route extrapolation differs depending on the availability of human PBPK models and requires the PBPK models to simulate both routes of exposure. (Potential added uncertainty that occurs for route-to-route extrapolation vs. more typical animal-to-human extrapolation is discussed at the end of this section.) For simplicity, the process is outlined for the linear case above, in which the relationship for test animals between internal dose and external exposure is linear and the extrapolations to internal dose and then to humans can be adequately accomplished following dose-response modeling.

Given human and rodent PBPK models capable of simulating both oral and inhalation exposure, an internal POD is derived based on exposure conditions of the rodent bioassay, generally from dose-response modeling of the internal dose or external exposure. The human PBPK model is then used to estimate either the daily HEC or HED required to achieve an internal dose equivalent to the rodent internal POD. The inhalation human PBPK model should simulate continuous 24-hour/day exposure, while the oral human PBPK model simulation may need to account for dietary or drinking water exposure profiles because first-pass metabolism saturation may occur for episodic ingestion and significantly impact internal doses (see **Figure 13-1A**).

If only a rodent model is available, then as outlined in **Figure 13-1B**, the rodent PBPK model is used to estimate the rodent equivalent alternate exposure required to achieve the same internal dose POD from the rodent bioassay, as above, but on a daily basis, in contrast to the application of the rodent PBPK model to obtain a POD reflecting an exposure profile consistent with the bioassay. A default methodology for extrapolation to humans is then applied to the daily rodent equivalent exposure (e.g., body weight$^{3/4}$ scaling for oral exposure). See **Section 13.2.2** for details of the default extrapolation methods.

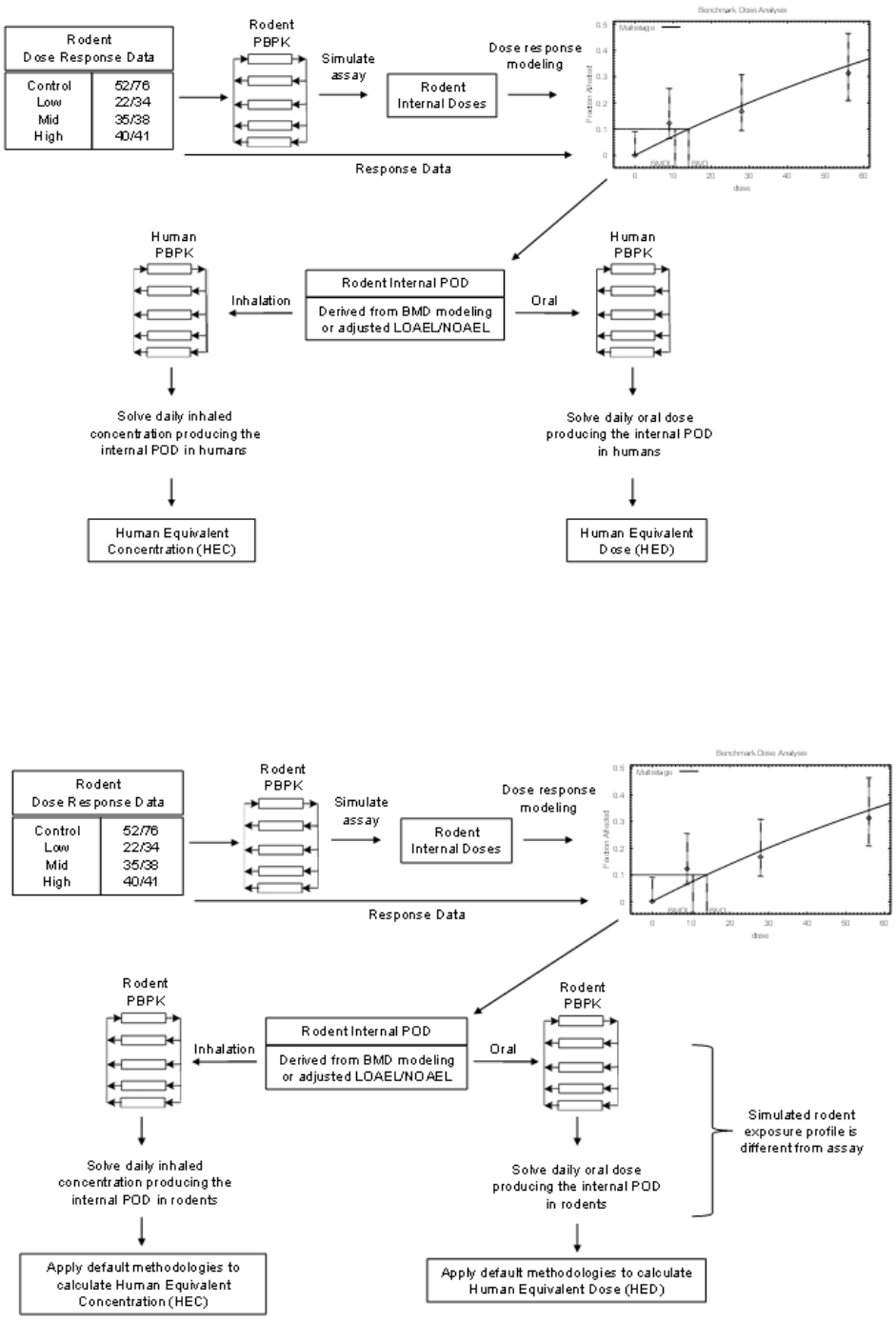


Figure 13-1. Process for deriving human equivalent exposures and performing route-to-route extrapolation using a rodent physiologically based pharmacokinetic (PBPK) model. (A) With a human PBPK model, and (B) in the absence of a human PBPK model.

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

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, 2005b), §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 may not be supported [e.g., (U.S. EPA, 1994), §4.1.2; (U.S. EPA, 2006a)].

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, 2005b), §3.1.1; (U.S. EPA, 1991), §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 $mg/kg^{3/4}$-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, 2005b), §3.1.3].

- Inhalation exposures are scaled using dosimetry models that apply species-specific physiologic and anatomic 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).

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

When evaluating animal data, EPA first considers toxicodynamic, or biologically based, models if any relevant to the assessment are available. Toxicodynamic modeling that incorporates data on biologic processes leading to an effect can be used to establish a POD and may reduce the extent of low-dose extrapolation needed for toxicity value derivation. Such models require sufficient data to ascertain the MOA and to support model parameters associated with its key events. Because different models may provide equivalent fits to the observed data but diverge substantially at lower exposure levels, critical biologic parameters should be measured from laboratory studies, not by model fitting. Confidence in the use of a toxicodynamic 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, 2005b), §3.2.2].

Because toxicodynamic 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 a biologically based model has been developed and judged to be useful in the low-dose range, extrapolation may use the fitted model below the observed range, if significant model uncertainty can be ruled out with reasonable confidence. Biologically informed, empirical dose-response modeling may also be used when biomarker data and MOA information can support it.

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 may be evaluated by fitting a variety of parametric and nonparametric dose-response models and then by applying a model-averaging procedure. Based on an explicit, case-specific evaluation of the uncertainties, a POD may be selected, or a decision may 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 lack of fit, the NOAEL (or absent that, the LOAEL) may then be used as the POD. Given that the hazard synthesis (see **Chapter 11**) supports the importance of these data 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].

13.2.3. 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, 2005b) suggest several approaches for characterizing total risk from 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).

13.2.4. 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; and
- Selection of a single toxicity value to represent each type of health effect.

The IRIS Assessment template for the dose-response modeling appendix of each chemical assessment currently documents BMDS-based modeling assumptions and conditions (including parameter constraints and parameters at boundaries) as well as model selection. This template is designed for animal studies and is adaptable for human data modeling.

13.3. 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, 2005b).

13.3.1. 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, 2005b). The result of linear extrapolation is described by the slope of the line from the response at the point of departure 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 may be used if there are multiple MOAs identified for the agent's carcinogenicity (U.S. EPA, 2005b). 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 origin 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 $\mu g/m^3$).

- 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, 2005c) provide more guidance on how and when to apply these ADAFs.

13.3.2. 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, 2005b), §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 13.2**), reference values are estimated by applying relevant adjustments to the PODs to account for five possible areas of uncertainty and variability: human variation, extrapolation from animals to humans, 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 is usually 10, 3, or 1, but different values based on chemical-specific information

may be applied if sufficient information exists in the chemical database. The assessment discusses the scientific bases for estimating these data-based adjustments and UFs.

- *Animal-to-human extrapolation*: If animal results are used to make inferences about humans, the toxicity value incorporates cross-species differences, which may arise from differences in toxicokinetics or toxicodynamics. If a biologically based model adjusts fully for toxicokinetic and toxicodynamic differences across species, this factor is not used. Otherwise, if the POD is standardized to equivalent human terms or is based on toxicokinetic or dosimetry modeling (U.S. EPA, 2014b, 2011a), a factor of $10^{1/2}$ (rounded to 3) is applied to account for the remaining uncertainty involving toxicokinetic and toxicodynamic differences.

- *Human variation*: The assessment accounts for variation in susceptibility across the human population and the possibility that the available data may 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 toxicodynamics or toxicokinetics is considered (U.S. EPA, 2014b).[22] Further, "when sufficient data are available, an intraspecies UF either less than or greater than 10× may be justified (U.S. EPA, 2002b). However, a reduction from the default (10) is only considered in cases when there are 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, 1996b), §4; (U.S. EPA, 1994), §4.3.9.1; (U.S. EPA, 1991), §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 may 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 13.2.2**).

- *LOAEL to NOAEL*: If a POD is based on a LOAEL or a BMDL associated with an adverse effect level (see **Section 13.1**), 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. 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 may be used depending on the magnitude and nature of the response and the shape of the dose-response curve (U.S. EPA, 2002b, 1998, 1996b, 1994, 1991).

- *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 when using subchronic studies to make inferences about lifetime exposure. A factor other than 10 may be used, depending on the duration of the studies and the nature of the response (U.S. EPA,

[22]Examples of adjusting the toxicokinetic 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).

2002b, 1998, 1994). This factor may also be applied, albeit rarely, for developmental or reproductive effects if exposure covered less than the full critical period.

- 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 may apply a database UF (U.S. EPA, 2002b, 1998, 1996b, 1994, 1991). The size of the factor depends on the nature of the database deficiency. For example, the 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 $10^{1/2}$ (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 **Chapter 12**.).

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 13-2**.

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 may be distinguished from the UF considerations outlined above.

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 13-2. Example summary of candidate toxicity values (for RfD derivation). Candidate values for three effects (nervous system, kidney/urogenital system, and male reproductive system).

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.

13.4. CHARACTERIZING UNCERTAINTY AND CONFIDENCE IN TOXICITY VALUES

13.4.1. 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 may 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 particular data set. It may 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, 2005b), **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 may 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. Supplemental cancer guidelines (U.S. EPA, 2005c) 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 may 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 may 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 (Shao and Gift, 2013; Shao, 2012; Wheeler and Bailer, 2009), 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).

13.4.2. Characterizing Confidence

In assessments in 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 are provided in *Methods for Derivation of Inhalation Reference Concentrations and Application of Inhalation Dosimetry* (U.S. EPA, 1994). Briefly, the confidence ranking (*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. 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 may 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/studies used in dose-response analysis may 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 may factor more heavily in that final determination given assessment- or endpoint-specific situations. In other words, confidence in the database may be the predominating factor in the overall confidence in one risk estimate, whereas the quantification of the POD may be the most important factor in the confidence for another risk estimate.

13.5. SELECTING FINAL TOXICITY VALUES

13.5.1. 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 12**). 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 toxicokinetics 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 13.3**) may 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 may 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 12.2**; the composite approach may 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].)

13.5.2. Overall Toxicity Values

The selection of overall toxicity values for noncancer and cancer effects involves the study preferences discussed in **Chapter 12**, 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 may 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)] may be sufficient for both communicating uncertainty and selecting a final toxicity value.

REFERENCES

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; Skakkebæk, 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. (2018). DRAFT Guidance for the preparation of toxicological profiles. Department of Health and Human Services. https://www.atsdr.cdc.gov/toxprofiles/guidance/profile_development_guidance.pdf

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

Bogen, KT. (1990). Uncertainty in environmental health risk assessment (Environment - Problems and solutions). New York, NY: Garland Publishing.

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.

Brazma, A; Hingamp, P; Quackenbush, J; Sherlock, G; Spellman, P; Stoeckert, C; Aach, J; Ansorge, W; Ball, CA; Causton, HC; Gaasterland, T; Glenisson, P; Holstege, FC; Kim, IF; Markowitz, V; Matese, JC; Parkinson, H; Robinson, A; Sarkans, U; Schulze-Kremer, S; Stewart, J; Taylor, R; Vilo, J; Vingron, M. (2001). Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat Genet 29: 365-371. http://dx.doi.org/10.1038/ng1201-365

Chiu, WA; Guyton, KZ; Martin, MT; Reif, DM; Rusyn, I. (2018). Use of high-throughput in vitro toxicity screening data in cancer hazard evaluations by IARC Monograph Working Groups. ALTEX 35: 51-64. http://dx.doi.org/10.14573/altex.1703231

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

Dean, JL; Zhao, QJ; Lambert, JC; Hawkins, BS; Thomas, RS; Wesselkamper, SC. (2017). Editor's Highlight: Application of Gene Set Enrichment Analysis for Identification of Chemically Induced, Biologically Relevant Transcriptomic Networks and Potential Utilization in Human Health Risk Assessment. Toxicol Sci 157: 85-99. http://dx.doi.org/10.1093/toxsci/kfx021

Dickersin, K. (1990). The existence of publication bias and risk factors for its occurrence. JAMA 263: 1385-1389.

Eastmond, DA. (2017). Recommendations for the evaluation of complex genetic toxicity data sets when assessing carcinogenic risks to humans. Environ Mol Mutagen 58: 380-385. http://dx.doi.org/10.1002/em.22078

Eastmond, DA; Hartwig, A; Anderson, D; Anwar, WA; Cimino, MC; Dobrev, I; Douglas, GR; Nohmi, T; Phillips, DH; Vickers, C. (2009). Mutagenicity testing for chemical risk assessment: Update of the WHO/IPCS harmonized scheme. Mutagenesis 24: 341-349. http://dx.doi.org/10.1093/mutage/gep014

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

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/cgiwgcommuniation_i.pdf

Farmahin, R; Williams, A; Kuo, B; Chepelev, NL; Thomas, RS; Barton-Maclaren, TS; Curran, IH; Nong, A; Wade, MG; Yauk, CL. (2017). Recommended approaches in the application of toxicogenomics to derive points of departure for chemical risk assessment. Arch Toxicol 91: 2045-2065. http://dx.doi.org/10.1007/s00204-016-1886-5

Fedak, KM; Bernal, A; Capshaw, ZA; Gross, S. (2015). Applying the Bradford Hill criteria in the 21st century: how data integration has changed causal inference in molecular epidemiology. Emerg Themes Epidemiol 12: 14. http://dx.doi.org/10.1186/s12982-015-0037-4

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

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

Guyatt, G; Oxman, AD; Akl, EA; Kunz, R; Vist, G; Brozek, J; Norris, S; Falck-Ytter, Y; Glasziou, P; DeBeer, H; Jaeschke, R; Rind, D; Meerpohl, J; Dahm, P; Schünemann, HJ. (2011a). GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. J Clin Epidemiol 64: 383-394. http://dx.doi.org/10.1016/j.jclinepi.2010.04.026

Guyatt, GH; Oxman, AD; Montori, V; Vist, G; Kunz, R; Brozek, J; Alonso-Coello, P; Djulbegovic, B; Atkins, D; Falck-Ytter, Y; Williams, JW, Jr; Meerpohl, J; Norris, SL; Akl, EA; Schünemann, HJ. (2011b). GRADE guidelines: 5. Rating the quality of evidence-publication bias. J Clin Epidemiol 64: 1277-1282. http://dx.doi.org/10.1016/j.jclinepi.2011.01.011

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

Higgins, J; Green, S. (2011a). Cochrane handbook for systematic reviews of interventions. Version 5.1.0: The Cochrane Collaboration, 2011. http://handbook.cochrane.org

Higgins, JPT; Green, S. (2011b). Cochrane handbook for systematic reviews of interventions. Version 5.1.0 (Updated March 2011): The Cochrane Collaboration. Retrieved from http://handbook.cochrane.org/

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. (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. (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. (2016). SWIFT-Review: a text-mining workbench for systematic review. Syst Rev 5: 87. http://dx.doi.org/10.1186/s13643-016-0263-z

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

IARC. (2006). Preamble to the IARC Monographs (amended January 2006). http://monographs.iarc.fr/ENG/Preamble/index.php

Ioannidis, JPA; Munafò, MR; Fusar-Poli, P; Nosek, BA; David, SP. (2014). Publication and other reporting biases in cognitive sciences: detection, prevalence, and prevention. Trends Cogn Sci 18: 235-241. http://dx.doi.org/10.1016/j.tics.2014.02.010

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

IPCS (International Programme for 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

Judson, RS; Houck, KA; Kavlock, RJ; Knudsen, TB; Martin, MT; Mortensen, HM; Reif, DM; Rotroff, DM; Shah, I; Richard, AM; Dix, DJ. (2010). In vitro screening of environmental chemicals for targeted testing prioritization: The ToxCast project. Environ Health Perspect 118: 485-492. http://dx.doi.org/10.1289/ehp.0901392

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.

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

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

Lynch, HN; Goodman, JE; Tabony, JA; Rhomberg, LR. (2016). Systematic comparison of study quality criteria. Regul Toxicol Pharmacol 76: 187-198. http://dx.doi.org/10.1016/j.yrtph.2015.12.017

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.

Matthews, GA; Dumville, JC; Hewitt, CE; Torgerson, DJ. (2011). Retrospective cohort study highlighted outcome reporting bias in UK publicly funded trials [Review]. J Clin Epidemiol 64: 1317-1324. http://dx.doi.org/10.1016/j.jclinepi.2011.03.013

McConnell, ER; Bell, SM; Cote, I; Wang, RL; Perkins, EJ; Garcia-Reyero, N; Gong, P; Burgoon, LD. (2014). Systematic Omics Analysis Review (SOAR) tool to support risk assessment. PLoS ONE 9: e110379. http://dx.doi.org/10.1371/journal.pone.0110379

Miake-Lye, IM; 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

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.

Morgan, RL; Thayer, KA; Bero, L; Bruce, N; Falck-Ytter, Y; Ghersi, D; Guyatt, G; Hooijmans, C; Langendam, M; Mandrioli, D; Mustafa, RA; Rehfuess, EA; Rooney, AA; Shea, B; Silbergeld, EK; Sutton, P; Wolfe, MS; Woodruff, TJ; Verbeek, JH; Holloway, AC; Santesso, N; Schünemann, HJ. (2016). GRADE: Assessing the quality of evidence in environmental and occupational health. Environ Int 92-93: 611-616. http://dx.doi.org/10.1016/j.envint.2016.01.004

NASEM.(National Academies of Science Engineering and Medicine) (2018). Progress toward transforming the Integrated Risk Information System (IRIS) program. A 2018 evaluation. Washington, DC: The National Academies Press. http://dx.doi.org/10.17226/25086

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 for 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

NRC (National Research Council). (1994). Science and judgment in risk assessment. Washington, DC: The National Academies Press. http://dx.doi.org/10.17226/2125

NRC (National Research Council). (2009). Science and decisions: Advancing risk assessment. Washington, DC: The National Academies Press. http://dx.doi.org/10.17226/12209

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

NTP (National Toxicology Program). (2015). Handbook for conducting a literature-based health assessment using OHAT approach for systematic review and evidence integration. U.S. Dept. of Health and Human Services, National Toxicology Program. https://ntp.niehs.nih.gov/ntp/ohat/pubs/handbookjan2015_508.pdf

NTP (National Toxicology Program). (2018). National Toxicology Program approach to genomic dose-response modeling. (NTP Research Report No. 5). https://ntp.niehs.nih.gov/ntp/results/pubs/rr/reports/rr05_508.pdf

OECD (Organisation for Economic Cooperation and Development). (2018). Guidance document on good in vitro method practices. Paris, France. http://dx.doi.org/10.1787/9789264304796-en

Oshiro, WM; Beasley, TE; McDaniel, KL; Taylor, MM; Evansky, P; Moser, VC; Gilbert, ME; Bushnell, PJ. (2014). Selective cognitive deficits in adult rats after prenatal exposure to inhaled ethanol. Neurotoxicol Teratol 45: 44-58. http://dx.doi.org/10.1016/j.ntt.2014.07.001

Parekh-Bhurke, S; Kwok, CS; Pang, C; Hooper, L; Loke, YK; Ryder, JJ; Sutton, AJ; Hing, CB; Harvey, I; Song, F. (2011). Uptake of methods to deal with publication bias in systematic reviews has increased over time, but there is still much scope for improvement. J Clin Epidemiol 64: 349-357. http://dx.doi.org/10.1016/j.jclinepi.2010.04.022

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

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

Rooney, AA; Cooper, GS; Jahnke, GD; Lam, J; Morgan, RL; Boyles, AL; Ratcliffe, JM; Kraft, AD; Schünemann, HJ; Schwingl, P; Walker, TD; Thayer, KA; Lunn, RM. (2016). How credible are the study results? Evaluating and applying internal validity tools to literature-based assessments of environmental health hazards. Environ Int 92-93: 617-629. http://dx.doi.org/10.1016/j.envint.2016.01.005

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

Salami, K; Alkayed, K. (2013). Publication bias in pediatric hematology and oncology: Analysis of abstracts presented at the annual meeting of the American Society of Pediatric Hematology and Oncology. Pediatr Hematol Oncol 30: 165-169. http://dx.doi.org/10.3109/08880018.2013.774078

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; Hill, S; Guyatt, G; Akl, EA; Ahmed, F. (2011). The GRADE approach and Bradford Hill's criteria for causation. J Epidemiol Community Health 65: 392-395. http://dx.doi.org/10.1136/jech.2010.119933

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

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; Cogliano, 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

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; Jüni, 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. Br Med J 355: i4919.

Sterne, JAC; Smith, GD; Cox, DR. (2001). Sifting the evidence -- what's wrong with significance tests? Br Med J 322: 226-231.

Stiteler, 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.

Thomas, RS; Philbert, MA; Auerbach, SS; Wetmore, BA; Devito, MJ; Cote, I; Rowlands, JC; Whelan, MP; Hays, SM; Andersen, ME; Meek, ME; Reiter, LW; Lambert, JC; Clewell, HJ, 3rd; Stephens, ML; Zhao, QJ; Wesselkamper, SC; Flowers, L; Carney, EW; Pastoor, TP; Petersen, DD; Yauk, CL; Nong, A. (2013). Incorporating new technologies into toxicity testing and risk assessment: Moving from 21st century vision to a data-driven framework. Toxicol Sci 136: 4-18. http://dx.doi.org/10.1093/toxsci/kft178

Thomas, RS; Waters, MD. (2016). Chapter 5. Transcriptomic dose-response analysis for mode of action and risk assessment. In MD Waters; RS Thomas (Eds.), Toxicogenomics in predictive carcinogenicity (pp. 154-184). Cambridge, England: Royal Society of Chemistry. http://dx.doi.org/10.1039/9781782624059-00154

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. EPA (U.S. Environmental Protection Agency). (1988). Recommendations for and documentation of biological values for use in risk assessment [EPA Report] (pp. 1-395). (EPA/600/6-87/008). Cincinnati, OH: U.S. Environmental Protection Agency, Office of Research and Development, Office of Health and Environmental Assessment. http://cfpub.epa.gov/ncea/cfm/recordisplay.cfm?deid=34855

U.S. EPA. (U.S. Environmental Protection Agency) (1991). Guidelines for developmental toxicity risk assessment (pp. 1-71). (EPA/600/FR-91/001). Washington, DC: U.S. Environmental Protection Agency, Risk Assessment Forum. http://cfpub.epa.gov/ncea/cfm/recordisplay.cfm?deid=23162

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 Research and Development, Office of Health and Environmental Assessment, Environmental Criteria and Assessment Office. https://cfpub.epa.gov/ncea/risk/recordisplay.cfm?deid=71993&CFID=51174829&CFTOKEN=25006317

U.S. EPA (U.S. Environmental Protection Agency). (1996a). Guidelines for reproductive toxicity risk assessment. Fed Reg 61: 56274-56322.

U.S. EPA (U.S. Environmental Protection Agency). (1996b). Guidelines for reproductive toxicity risk assessment (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/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/630/P-02/002F). Washington, DC: U.S. Environmental Protection Agency, Risk Assessment Forum. 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. 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, IRIS. http://nepis.epa.gov/exe/ZyPURL.cgi?Dockey=P1006CK9.txt

U.S. EPA (U.S. Environmental Protection Agency). (2005a). Guidance on selecting age groups for monitoring and assessing childhood exposures to environmental contaminants. (EPA/630/P-03/003F). Washington, DC. https://nepis.epa.gov/Exe/ZyPURL.cgi?Dockey=2000D2JZ.txt

U.S. EPA (U.S. Environmental Protection Agency). (2005b). Guidelines for carcinogen risk assessment [EPA Report]. (EPA/630/P-03/001B). 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). (2005c). 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://www3.epa.gov/airtoxics/childrens_supplement_final.pdf

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 (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). (2007). Interim guidance for microarray-based assays: Data submission, quality, analysis, management, and training considerations (External review draft). Science Policy Council. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.111.1389&rep=rep1&type=pdf

U.S. EPA (U.S. Environmental Protection Agency). (2009). An approach to using toxicogenomic data in U.S. EPA human health risk assessments: A dibutyl phthalate case study (Final Report) [EPA Report]. (EPA/600/R-09/028F). Washington, DC. http://cfpub.epa.gov/ncea/cfm/recordisplay.cfm?deid=213405

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 (pp. 1-50). (EPA/100/R-11/0001). Washington, DC: U.S. Environmental Protection Agency, Risk Assessment Forum, Office of the Science Advisor. 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 (CASRN 79-01-6) in support of summary information on the Integrated Risk Information System (IRIS) [EPA Report]. (EPA/635/R-09/011F). Washington, DC. https://cfpub.epa.gov/ncea/iris/iris_documents/documents/toxreviews/0199tr/0199tr.pdf

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 (pp. 1-140). (EPA/600/R-12/044). Washington, DC.

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/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). Washington, DC: 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). (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 (U.S. Environmental Protection Agency). (2017). Guidance to assist interested persons in developing and submitting draft risk evaluations under the Toxic Substances Control Act. (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). (2018a). Application of systematic review in TSCA risk evaluations. (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 (U.S. Environmental Protection Agency). (2018b). 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.

Vandenberg, LN; Colborn, T; Hayes, TB; Heindel, JJ; Jacobs, DR; Lee, DH; Shioda, T; Soto, AM; vom Saal, FS; Welshons, WV; Zoeller, RT; Myers, JP. (2012). Hormones and endocrine-disrupting chemicals: Low-dose effects and nonmonotonic dose responses [Review]. Endocr Rev 33: 378-455. http://dx.doi.org/10.1210/er.2011-1050

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/rtph.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-Reyero, 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-Reyero, 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

Wambaugh, JF; Hughes, MF; Ring, CL; MacMillan, DK; Ford, J; Fennell, TR; Black, SR; Snyder, RW; Sipes, NS; Wetmore, BA; Westerhout, J; Setzer, RW; Pearce, RG; Simmons, JE; Thomas, RS. (2018). Evaluating in vitro-in vivo extrapolation of toxicokinetics. Toxicol Sci 163: 152-169. http://dx.doi.org/10.1093/toxsci/kfy020

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

Wetmore, BA; Allen, B; Clewell, HJ, III; Parker, T; Wambaugh, JF; Almond, LM; Sochaski, MA; Thomas, RS. (2014). Incorporating population variability and susceptible subpopulations into dosimetry for high-throughput toxicity testing. Toxicol Sci 142: 210-224. http://dx.doi.org/10.1093/toxsci/kfu169

Wetmore, BA; Wambaugh, JF; Allen, B; Ferguson, SS; Sochaski, MA; Setzer, RW; Houck, KA; Strope, CL; Cantwell, K; Judson, RS; LeCluyse, E; Clewell, HJ; Thomas, RS; Andersen, ME. (2015). Incorporating high-throughput exposure predictions with dosimetry-adjusted in vitro bioactivity to inform chemical toxicity testing. Toxicol Sci 148: 121-136. http://dx.doi.org/10.1093/toxsci/kfv171

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

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/IPCS (World Health Organization/ International Programme for 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

WHO/IPCS (World Health Organization/ International Programme for 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

Wigle, DT; Lanphear, BP. (2005). Human health risks from low-level environmental exposures: No apparent safety thresholds. PLoS Med 2: 1232-1234.

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

Ziliak, ST. (2011). Matrixx v. Siracusano and Student v. Fisher. Significance 8: 131-134.