Materials Submitted to the National Research Council
Part I: Status of Implementation of Recommendations

U.S. Environmental Protection Agency

Integrated Risk Information System Program

January 30, 2013

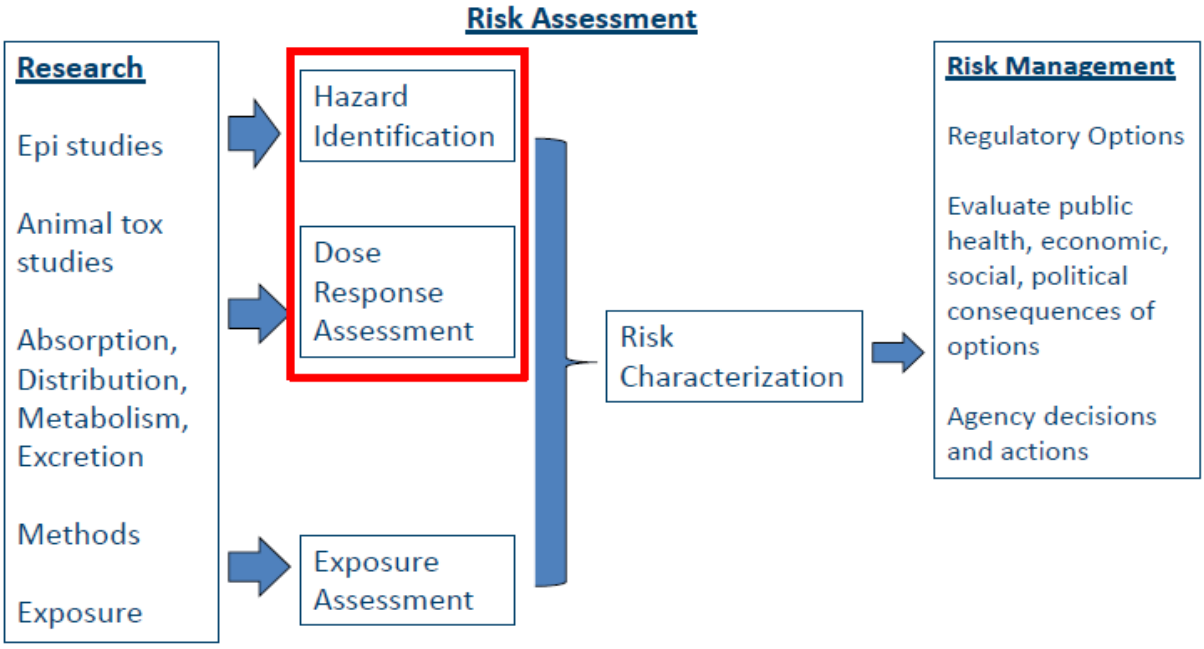
Table of Contents

I. Introduction

The U.S. Environmental Protection Agency's (EPA) Integrated Risk Information System (IRIS) Program develops human health assessments that provide health effects information on environmental chemicals to which the public may be exposed, providing a critical part of the scientific foundation for EPA's decisions to protect public health. In April 2011, the National Research Council (NRC), in their report *Review of the Environmental Protection Agency's Draft IRIS Assessment of Formaldehyde*, made several recommendations to EPA for improving IRIS assessments and the IRIS Program. The NRC's recommendations were focused on Step 1 of the IRIS process, the development of draft assessments. Consistent with the advice of the NRC, the IRIS Program is implementing these recommendations using a phased approach and is making the most extensive changes to assessments that are in the earlier stages of the IRIS process.

Background on IRIS

IRIS human health assessments contain information that can be used to support the first two steps (hazard identification and dose-response analysis) of the risk assessment paradigm. IRIS assessments are scientific reports that provide information on a chemical's hazards and, when supported by available data, quantitative toxicity values for cancer and noncancer health effects. IRIS assessments are not regulations, but they provide a critical part of the scientific foundation for decisions to protect public health across EPA's programs and regions under an array of environmental laws (e.g., Clean Air Act, Safe Drinking Water Act, Comprehensive Environmental Response, Compensation, and Liability Act, etc). EPA's program and regional offices combine IRIS assessments with specific exposure information for a chemical. This information is used by EPA, together with other considerations (e.g., statutory and legal requirements, cost/benefit information, technological feasibility, and economic factors), to characterize the public health risks of environmental chemical and make risk management decisions, including regulations, to protect public health. IRIS assessments are also a resource for risk assessors and environmental and health professionals from state and local governments and other countries. Figure 1 illustrates where IRIS assessments contribute information within the risk assessment and risk management paradigms.



[1] Adapted from the National Research Council risk assessment risk management paradigm (NRC 1983).

Figure 1. Risk Assessment Risk Management Paradigm (adapted from the National Research Council's paradigm, 1983). The red box shows the information included in IRIS assessments.

II. Charge to the NRC Expert Panel

In April 2012, EPA contracted with the NRC to conduct a comprehensive review of the IRIS assessment development process. The panel will review the IRIS process and the changes being made or planned by EPA and will recommend modifications or additional changes as appropriate to improve the process, and scientific and technical performance of the IRIS Program. The panel will focus on the development of IRIS assessments rather than the review process that follows draft development. In addition, the panel will review current methods for evidence-based reviews and recommend approaches for weighing scientific evidence for chemical hazard and dose-response assessments.

III. Overview of EPA's Implementation of NRC's Recommendations

EPA agrees with the NRC's 2011 recommendations for the development of IRIS assessments and plans to fully implement the recommendations consistent with the NRC panel's "Roadmap for Revision," which viewed the full implementation of their recommendations by the IRIS Program as a multi-year process. In response to the NRC's 2011 recommendations, the IRIS Program has made changes to streamline the assessment development process, improve transparency, and create efficiencies within the Program. The following sections outline the NRC's 2011 recommendations and provide an overview of how the IRIS Program is implementing the NRC's general and specific

recommendations. Further details regarding changes that have been made and will be made in response to the recommendations are provided in Appendices to this report.

In addition, chemical-specific examples demonstrating how the IRIS Program is currently implementing the NRC's 2011 recommendations have also been provided to the panel (see additional document provided, *Chemical-Specific Examples Demonstrating Implementation of NRC's 2011 Recommendations*). The examples cover literature search and screening, evaluation and display of individual studies, development of evidence tables, evidence integration, selecting studies for derivation of toxicity values, dose-response modeling output, and considerations for selecting organ/system-specific or overall toxicity values. The examples are not to be construed as final Agency conclusions and are provided for the sole purpose of demonstrating how the IRIS Program is implementing the NRC recommendations.

NRC's General Recommendations and Guidance

NRC Recommendations[1]:

- To enhance the clarity of the document, the draft IRIS assessment needs rigorous editing to reduce the volume of text substantially and address redundancies and inconsistencies. Long descriptions of particular studies should be replaced with informative evidence tables. When study details are appropriate, they could be provided in appendices.
- Chapter 1 needs to be expanded to describe more fully the methods of the assessment, including a description of search strategies used to identify studies with the exclusion and inclusion criteria articulated and a better description of the outcomes of the searches and clear descriptions of the weight-of-evidence approaches used for the various noncancer outcomes. The committee emphasizes that it is not recommending the addition of long descriptions of EPA guidelines to the introduction, but rather clear concise statements of criteria used to exclude, include, and advance studies for derivation of the RfCs and unit risk estimates.
- Elaborate an overall, documented, and quality-controlled process for IRIS assessments.
- Ensure standardization of review and evaluation approaches among contributors and teams of contributors; for example, include standard approaches for reviews of various types of studies to ensure uniformity.
- Assess disciplinary structure of teams needed to conduct the assessments.

Implementation:

➢ **New Document Structure**

Implemented

In their report, the NRC recommended that the IRIS Program enhance the clarity of the document, reduce the volume of text, and address redundancies and inconsistencies. To improve the clarity of IRIS assessments, the IRIS Program has revised the assessment template to substantially reduce the volume of text and address redundancies and inconsistencies in assessments. The new template provides sections for the literature search strategy, study selection and evaluation, and methods used to develop the assessment.

[1] National Research Council, 2011. Review of the Environmental Protection Agency's Draft IRIS Assessment of Formaldehyde.

The new document structure includes an *Executive Summary* in the beginning of each assessment which provides a concise summary of the major conclusions of the assessment. Additionally, a newly developed *Preamble* describes the methods used to develop the assessment. Each assessment will include information on the literature search strategy used to identify the evidence for consideration in developing the assessment, as well as the evaluation criteria and rationale used to make decisions about including or excluding studies in the assessment.

The main body of the IRIS assessment has been reorganized into two sections, *Hazard Identification* and *Dose-Response Analysis,* to better focus on the role of IRIS assessments in the risk assessment paradigm and to further reduce the volume of text and redundancies/inconsistencies. Information on assessments by other national and international health agencies, chemical and physical properties, toxicokinetics, and individual studies has been moved to appendices (which are provided as supplemental information) to improve the flow of the document.

In the *Hazard Identification* chapter of the new document template, the IRIS Program has developed subsections based on organ/system-specific hazards to systematically integrate the available evidence for a given chemical (i.e., epidemiology, toxicological, and mechanistic data). The assessment now uses evidence tables to present the key study findings that support how toxicological hazards are identified. In addition, exposure-response arrays are being used as visual tools to inform the hazard characterization. This chapter provides for a strengthened and more integrated and transparent discussion of the weight of the available evidence supporting hazard identification. The IRIS Program is also developing standardized study summary tables, which will be included in the supplemental information, to present more detailed study characteristics and summary information.

The *Dose-Response Analysis* section of the new document structure provides a section to explain the rationale used to select and advance studies for consideration in calculating toxicity values based on conclusions regarding the potential hazards associated with chemical exposure. Key data supporting the dose-response analysis are reported and the methodology and derivation of toxicity values are described. In addition, details of the dose-response analysis—including the data, models, methods, and software—are provided as supplemental information and described in sufficient detail to allow for independent replication and verification. The *Dose-Response Analysis* section also includes tables and figures showing candidate toxicity values for comparison across studies and endpoints. Finally, this section of the new document structure includes clear documentation of the conclusions and selection of the overall toxicity values.

The IRIS assessment template demonstrating the new document structure is provided in Appendix A.

- **IRIS Assessment Preamble**

Implemented

In their report, the NRC recommended that the IRIS Program expand Chapter 1 of IRIS assessments to "describe more fully the methods of the assessment, including a description of search strategies used to identify studies with the exclusion and inclusion criteria clearly articulated and a better

description of the outcomes of the searches and clear descriptions of the weight-of-evidence approaches used for the various noncancer outcomes."

In accordance with this recommendation, the IRIS Program has replaced the previous Chapter 1 of IRIS assessments with a section titled *Preamble to IRIS Toxicological Reviews* which describes the application of existing EPA guidance and the methods and criteria used in developing the assessments. The term "*Preamble*" is used to emphasize that these methods and criteria are being applied consistently across IRIS assessments. The new *Preamble* discusses the following topics:

- Scope of the IRIS Program;
- Process for developing and peer-reviewing IRIS assessments;
- Identifying and selecting pertinent studies;
- Evaluating the quality of individual studies;
- Evaluating the overall evidence of each effect;
- Selecting studies for derivation of toxicity values; and
- Deriving toxicity values.

For each of these topics, the *Preamble* summarizes and cites EPA guidance on methods used in the assessment. The *Preamble* was included in the draft IRIS assessments of ammonia and trimethylbenzenes when they were released for public comment in June 2012 and will be included in all new IRIS assessments.

The Preamble to IRIS Toxicological Reviews is included in Appendix B.

- **New Initiatives to Improve Overall Process, Quality Control, and Documentation**

In Progress

In their report, the NRC recommended that the IRIS Program "elaborate an overall, documented, and quality-controlled process for IRIS assessment" and "assess disciplinary structure of teams needed to conduct the assessments." In response to these recommendations, the IRIS Program has developed several new initiatives and enhanced existing processes. These initiatives help to ensure that standardized approaches are use throughout IRIS assessments and major science decisions are rigorously vetted.

IRIS assessments are developed by interdisciplinary teams of scientists (referred to as an "Assessment Team") internal to EPA. For each assessment, scientists with the necessary scientific backgrounds (e.g., neurotoxicology, epidemiology, developmental toxicology) are assigned to lead or assist in the development of the assessment. The expertise needed is chemical-specific and the personnel assigned to the assessment team are identified in the early stages of planning and document development.

Contractors may provide technical and analytical support to the chemical assessment teams during the development of assessments. This assistance may be provided in conducting literature searches and identifying pertinent studies; developing evidence tables and exposure-response arrays using studies identified and evaluated by the IRIS Program; and performing dose-response modeling (i.e., using EPA's benchmark dose modeling software [BMDS]). All materials provided by the contractor are evaluated in accordance with EPA policies regarding quality assurance and quality management, and specified in the contract. Contractor products are not incorporated into IRIS assessments without significant Agency scientist review. EPA is responsible for the content and conclusions within the assessments and all scientific and policy decisions are made by the Agency.

An example of instruction provided to contractors is available in Appendix C.

Additionally, discipline-specific workgroups within the IRIS Program assist the assessment teams in developing assessments. These workgroups coordinate across assessments to ensure consistency, solve cross-cutting issues, and advance scientific understanding that contributes to decision-making in IRIS assessments. The discipline-specific workgroups cover topics related to: statistics and dose-response analysis, physiologically-based pharmacokinetic modeling, and mechanistic data.

In late 2011, the IRIS Program developed a new initiative, Chemical Assessment Support Teams (CASTs), as a means of formalizing an internal process to provide continuing quality control in the development of IRIS assessments. This initiative uses a team approach to make judicious, consistent decisions during assessment development, to ensure that the necessary disciplinary expertise is available for assessment development and review, and to provide a forum for identifying and addressing key issues at each stage of the assessment. There are three CASTs and each team consists of four permanent core members: two senior scientists, a senior statistician, and a rapporteur (a staff scientist).

All on-going IRIS assessments have been distributed and assigned across the three CASTs. Each CAST meets periodically with the individual chemical assessment teams. In addition to meeting with each chemical assessment team, the CASTs convene as a group once a week to discuss issues that have surfaced in the chemical-specific CAST meetings from the previous week. Discussions at this meeting are relayed to scientists working on IRIS assessments during weekly meetings convened by the IRIS Program Director.

The CAST initiative:

- Provides a forum for problem solving;
- Ensures appropriate disciplinary structure of assessment teams;
- Pinpoints key issues early on in the assessment;
- Identifies overarching assessment issues that require Program-wide discussions;
- Increases objectivity in assessment decisions;
- Monitors progress in implementing NRC's 2011 recommendations;

- Assists in responding to Agency, interagency, external peer review, and public comments;
- Ensures consistency across assessments; and
- Serves as a mechanism for documenting and communicating decisions.

As noted above, the CASTs ensure documentation and communication of decisions. Documenting discussions and decisions from CAST meetings is the primary responsibility of the rapporteurs, who have developed a searchable database to capture comments received throughout assessment development and review as well as Agency decisions in response to these comments. This important information management tool, the *Comment Tracker Database*, allows for recording, reviewing, responding to, and analyzing comments and responses. The IRIS Program is currently testing the use of this database.

The CAST initiative is aimed at improving the quality and consistency of IRIS assessments as well as identifying overarching scientific issues to be addressed. This process facilitates communication across the organization and consistency across assessments to improve the overall efficiency of the IRIS Program.

The Comment Tracker Database is further described in Appendix D.

The IRIS Program also recognizes that it is important to understand the big picture in order to develop an assessment that is most informative and efficient for decision-makers. Having a clear understanding of the overarching environmental problems being addressed in the context of a chemical can help inform what an IRIS assessment will ultimately include. This concept was supported by the NRC in their 2009 report *Science and Decisions: Advancing Risk Assessment* when they recommended that EPA provide "greater attention on design in the formative stages of risk assessment." While the NRC was referring to the overall risk assessment paradigm, the spirit of the recommendation supports a scoping step before developing a hazard identification and dose-response assessment (i.e., IRIS assessment). Because of the importance of considering the scope of an IRIS assessment, the IRIS Program is developing a new initiative to include a "scoping" process as an early step in developing IRIS assessments. The scoping process involves consultation with clients in EPA's program and regional offices. This early consultation provides an opportunity to identify key questions for framing various analyses and helps ensure that the assessment meets the needs and critical timelines of Agency decision-makers.

The considerations for scoping during the development of IRIS assessments are further described in Appendix E.

The IRIS Program has recently initiated ways to improve stakeholder engagement to help ensure transparency and the use of the best available science in IRIS assessments. When IRIS toxicity values are combined with specific exposure information, government and private entities can use these values to help characterize the public health risks of chemical substances in various situations and support risk management decisions to protect public health. Environmental protection decisions can have potentially large impacts on the environment, human health, and the economy. Engaging with stakeholders can help facilitate the development of assessments and promote public

discussion of key scientific issues. Therefore, stakeholder and public scientific engagement is an important part of supporting the best decisions possible.

The IRIS Program considers a stakeholder to be any individual or group that participates in, has an impact on, or could be affected by products produced by the IRIS Program. Public and stakeholder engagement has always been an important part of the IRIS assessment development process. The May 2009 IRIS process provides multiple opportunities for engagement including: (1) public and stakeholder nomination of chemicals for assessment; (2) a public listening session for each draft assessment; (3) public review and comment of draft documents; (4) a public peer review process; and (5) two opportunities for review and comment on draft assessments by other EPA scientists, other Federal agencies, and the Executive Office of the President.

Recently, the IRIS Program convened two public meetings to engage with stakeholders. In November 2012, a public stakeholder meeting was held to discuss the IRIS Program in general. The meeting was intended to begin a series of dialogues between the IRIS Program and a broad and diverse group of stakeholders. The goals of the meeting were to: engage stakeholders in the IRIS process; listen to views and needs of IRIS users in an open and respectful environment; facilitate improvements to the IRIS process; and initiate an ongoing dialogue between the IRIS Program and stakeholders. In January 2013, a public stakeholder meeting, which focused on informing the plan for drafting a new IRIS assessment for inorganic arsenic, was convened. The meeting provided an opportunity for stakeholders to comment on their expectations for the IRIS assessment, the current state of scientific information that should be considered when developing the assessment, and the potential impacts of the completed assessment.

Another initiative involves the increased use of public peer consultation workshops to enhance the input from the scientific community as assessments are designed. Information regarding specific peer consultation workshops will be announced to the public in advance of the meetings. The goal of these workshops will vary. For example, the workshops may focus on the state-of-the-science for a particular chemical or provide a forum for discussion with experts about certain cross-cutting scientific issues that may impact the development of a scientifically complex assessment. One of the first of these peer consultation workshops will focus on mouse lung tumors as they relate to human cancer risk. This is an important issue for the IRIS assessments for naphthalene, styrene, and ethylbenzene.

The IRIS Program will also conduct public dialogue meetings to discuss the available chemical-specific data and the science issues for new IRIS assessments in the draft development stage. IRIS will share with the public the list of references and tables summarizing the key studies prior to the meeting.

NRC's Specific Recommendations and Guidance

The NRC made twenty-five specific recommendations in five broad categories:

- evidence identification,
- evidence evaluation,
- weight-of-evidence evaluation,
- selection of studies for derivation of toxicity values, and
- calculation of toxicity values

The IRIS Program has been working to improve the approaches for identifying and selecting pertinent studies; evaluating and displaying individual studies; strengthening and improving integration of evidence for hazard identification; and increasing transparency in dose-response analysis.

The IRIS Program recognized the value of providing specific information to its assessment teams and contractors in order to develop IRIS assessments that satisfy the needs of the NRC recommendations. In order to document these individual changes, the IRIS Program has compiled information into a working draft *Handbook for IRIS Assessment Development*. This document is intended to more clearly summarize the internal processes and evaluation steps used to develop IRIS assessments. The draft *Handbook* (which in its current form will be made publicly available) is a work in progress and currently does not fully discuss each step in the IRIS assessment development process. However, the draft *Handbook* contains important information that reflects the changes that have been implemented or will be implemented in response to the NRC recommendations. These changes are noted below and further described in the draft *Handbook for IRIS Assessment Development* in Appendix F.

Evidence Identification: Literature Collection and Collation Phase

NRC Recommendations:

- Select outcomes on the basis of available evidence and understanding of mode of action.
- Establish standard protocols for evidence identification.
- Develop a template for description of the search approach.
- Use a database, such as the Health and Environmental Research Online (HERO) database, to capture study information and relevant quantitative data.

Implementation:

➢ **Identifying and Selecting Pertinent Studies** In Progress

The IRIS Program is adopting the principles of systematic review in IRIS human health assessments with regard to providing an overview of methods and points to consider in the process of developing and documenting decisions. The focus of IRIS assessments is typically on the evidence of health effects (any kind of health effects) of a particular chemical. This is, by definition, a broad

topic. The systematic review process that has been developed and applied within the clinical medicine arena (evidence-based medicine) is generally applied to narrower, more focused questions. Nonetheless, the experiences within the clinical medicine field provide a strong foundation to draw upon. The IRIS Program is planning to convene a workshop in spring 2013 on this topic in order to have a public discussion of systematic review approaches that may be applicable to IRIS assessments.

An IRIS assessment is made up of multiple systematic reviews. The initial steps of the systematic review process formulate specific strategies to identify and select studies relating to each key question, evaluate study methods based on clearly defined criteria, and transparently document the process and its outcomes. Synthesizing and integrating data also falls under the purview of systematic review. Overall, this is an iterative process that identifies relevant scientific information needed to address key, assessment-specific questions.

The IRIS Program has improved the approach to identifying and selecting studies pertinent to IRIS assessments by adopting the principles of systematic review. One of the strengths of systematic review is its ability to identify relevant studies, published and unpublished, pertaining to the question of interest (e.g., what are the health effects of a chemical?). Additionally, by transparently presenting all decision points and the rationale for each decision, bias in study selection and evaluation is eliminated.

The new IRIS assessment document structure includes a detailed description of the literature search strategy and study selection process used to develop IRIS assessments. This section describes how the scientific literature was gathered, emphasizes how studies were selected to be included in the document, and, if applicable, explains the rationale for excluding potentially relevant studies from the assessment. This section of the new document structure is specific to each chemical assessment. It is designed to provide enough information that an independent literature search would be able to replicate the results of the literature search used by the IRIS Program in developing the assessment. In this section, a link to an external database (www.epa.gov/hero) that contains the references that were cited in the document, along with those that were considered for inclusion in the assessment but not cited is provided.

For more detailed information, see the "Identifying and Selecting Pertinent Studies: Literature Search and Screening" section in the draft Handbook for IRIS Assessment Development in Appendix F.

See also Section 3 ("Identifying and selecting pertinent studies") in the Preamble to IRIS Toxicological Reviews in Appendix B.

A chemical-specific example of the implementation of this recommendation is available as "EXAMPLE 1 – Literature Search and Screening" in the Chemical-specific Examples Demonstrating Implementation of NRC Recommendations document.

Evidence Evaluation: Hazard Identification

NRC Recommendations:

- All critical studies need to be thoroughly evaluated with standardized approaches that are clearly formulated and based on the type of research, for example, observational epidemiologic or animal bioassays. The findings of the reviews might be presented in tables to ensure transparency.
- Standardize the presentation of reviewed studies in tabular or graphic form to capture the key dimensions of study characteristics, weight of evidence, and utility as a basis for deriving reference values and unit risks.
- Standardized evidence tables for all health outcomes need to be developed. If there were appropriate tables, long text descriptions of studies could be moved to an appendix or deleted.
- Develop templates for evidence tables, forest plots, or other displays.
- Establish protocols for review of major types of studies, such as epidemiologic and bioassay.

Implementation:

➢ **Evaluating and Documenting the Quality of Individual Studies**

In Progress

The IRIS Program is improving the approach to evaluating and describing the strengths and weaknesses of critical studies and standardizing the documentation of this evaluation. This step in the systematic review process involves the evaluation of a variety of methodological features (e.g., study design, exposure measurement details, data analysis and presentation). The purpose of this step is generally not to eliminate studies, but rather to evaluate studies with respect to potential methodological considerations that could affect the interpretation of and relative confidence in the results. It is worth emphasizing that the systematic evaluation of the study described in this step is conducted at an early stage of assessment development (i.e., after identifying the relevant sources of primary data but before developing evidence tables and characterizing hazards associated with exposure to a chemical). The results of this systematic evaluation may inform decisions about which studies to use for hazard identification, considerations to keep in mind when interpreting the results of specific studies, and which studies to move forward for dose-response modeling for derivation of toxicity values.

For more detailed information, see "Study Quality Evaluation" and "Documentation of Study Quality Evaluations" in the Evaluation and Display of Individual Studies section in the draft Handbook for IRIS Assessment Development in Appendix F.

See also Section 4 ("Evaluating the quality of individual studies") in the Preamble to IRIS Toxicological Reviews in Appendix B.

A chemical-specific example of the implementation of this recommendation is available as "EXAMPLE 2 – Evaluation and Display of Individual Studies" in the Chemical-specific Examples Demonstrating Implementation of NRC Recommendations document.

➢ **Evidence Tables**

Implemented

The IRIS Program has developed templates for evidence tables to standardize the presentation of reviewed studies in IRIS assessments. Once a literature search has been conducted and the resulting database of studies has been evaluated, evidence tables are developed to present information from the collection of studies related to a specific outcome or endpoint of toxicity. The evidence tables include studies that have been judged adequate for hazard identification and display available study results, both positive and negative results. The studies that are considered to be most informative will depend on the extent and nature of the database for a given chemical, but may encompass a range of study designs and include epidemiology, toxicology, and, other toxicity data when appropriate.

For more detailed information, see "Reporting Study Results" in the Evaluation and Display of Individual Studies section in the draft Handbook for IRIS Assessment Development in Appendix F.

A chemical-specific example of the implementation of this recommendation is available as "EXAMPLE 3 – Evidence Tables" in the Chemical-specific Examples Demonstrating Implementation of NRC Recommendations document.

Weight-of-Evidence Evaluation: Integration of Evidence for Hazard Identification

NRC Recommendations:

- Strengthened, more integrative and more transparent discussions of weight of evidence are needed. The discussions would benefit from more rigorous and systematic coverage of the various determinants of weight of evidence, such as consistency.
- Review use of existing weight-of-evidence guidelines.
- Standardize approach to using weight-of-evidence guidelines.
- Conduct agency workshops on approaches to implementing weight-of-evidence guidelines.
- Develop uniform language to describe strength of evidence on noncancer effects.
- Expand and harmonize the approach for characterizing uncertainty and variability.
- To the extent possible, unify consideration of outcomes around common modes of action rather than considering multiple outcomes separately.

Implementation:

➢ **Integration of Evidence for Hazard Identification**

In Progress

The IRIS Program has strengthened and increased transparency in the weight-of-evidence for identifying hazards in IRIS assessments. Hazard identification involves the integration of evidence from human, animal, and mechanistic studies in order to draw conclusions about the hazards associated with exposure to a chemical. In general, IRIS assessments integrate evidence in the context of Hill (1965), which outlines aspects — such as consistency, strength, coherence, specificity, does-response, temporality, and biological plausibility — for consideration of causality

in epidemiologic investigations that were later modified by others and extended to experimental studies (U.S. EPA, 2005a).

All results, both positive and negative, of potentially relevant studies that have been evaluated for quality are considered (U.S. EPA, 2002) to answer the fundamental question: "Does exposure to chemical X cause hazard Y?" This requires a critical weighing of the available evidence (U.S. EPA, 2005a; 1994), but is not to be interpreted as a simple tallying of the number of positive and negative studies (U.S. EPA, 2002). Hazards are identified by an informed, expert evaluation and integration of the human, animal, and mechanistic evidence streams.

For more detailed information, see "Synthesis of Observational Epidemiology Evidence", "Synthesis of Animal Toxicology Evidence", and "Mechanistic Considerations in Elucidating Adverse Outcome Pathways" in the Evaluating the Overall Evidence of Each Effect section in the draft Handbook for IRIS Assessment Development in Appendix F.

See also Section 5 ("Evaluating the overall evidence of each effect") in the Preamble to IRIS Toxicological Reviews in Appendix B.

A chemical-specific example of the implementation of this recommendation is available as "EXAMPLE 4 – Evidence Integration" in the Chemical-specific Examples Demonstrating Implementation of NRC Recommendations document.

Currently, the IRIS Program is using existing guidelines that address these issues to inform assessments. In addition, the IRIS Program is taking a more systematic approach in analyzing the available human, animal, and mechanistic data is being used in IRIS assessments. In conducting this analysis and developing the synthesis, the IRIS Program evaluates the data for the:

- strength of the relationship between the exposure and response and the presence of a dose-response relationship;
- specificity of the response to chemical exposure and whether the exposure precedes the effect;
- consistency of the association between the chemical exposure and response; and
- biological plausibility of the response or effect and its relevance to humans.

The IRIS Program uses this weight of evidence approach to identify the potential hazards associated with chemical exposure.

The IRIS Program recognizes the benefit of adopting a formal weight-of-evidence framework that includes standardized classification of causality. In addition to the NRC task, in which the panel will review current methods for evidence-based reviews and recommend approaches for weighing scientific evidence for chemical hazard and dose-response assessments, the IRIS Program is planning to convene a workshop to discuss approaches to evidence integration. As part of this workshop, the various approaches that are currently in use will be acknowledged and compared for their strengths and limitations. The workshop will include scientists with expertise in the

classification of chemicals for various health effects. The workshop will be open to the public, and the details will be publicly announced.

The "Integration of Evidence Evaluation" section in the draft Handbook for IRIS Assessment Development in Appendix F is currently under development.

Selection of Studies for Derivation of Toxicity Values

NRC Recommendations:

- The rationales for the selection of the studies that are advanced for consideration in calculating the RfCs and unit risks need to be expanded. All candidate RfCs should be evaluated together with the aid of graphic displays that incorporate selected information on attributes relevant to the database.
- Establish clear guidelines for study selection.
- Balance strengths and weaknesses.
- Weigh human vs. experimental evidence.
- Determine whether combining estimates among studies is warranted.

Implementation:

➢ **Selection of Studies for Dose-Response Analysis**

Implemented

The IRIS Program has improved the process for selecting studies for derivation of toxicity values as well as increasing the transparency about this process by providing an improved discussion and rationale. Building on the individual study quality evaluations (described under *Evidence Evaluation: Hazard Identification* in this report) that identify strengths and weaknesses of individual studies, for each health effect for which there is credible evidence of hazard, a group of studies are identified and evaluated as part of the hazard identification. In evaluating these studies for selecting a subset to be considered for the derivation of toxicity values, the basic criterion is whether the quantitative exposure and response data are available to compute a point of departure (POD). The POD can be a no-observed-adverse-effect-level [NOAEL], lowest-observed-adverse-effect-level [LOAEL], or the benchmark dose/concentration lower confidence limit[BMDL/BMCL]).

Additional attributes (aspects of the study, data characteristics, and relevant considerations) pertinent to derivation of toxicity values are used as criteria to evaluate the subset of studies for dose-response analysis. Thus, the most relevant, informative studies are selected to move forward. The new document structure provides for transparent discussion of the studies identified for dose-response analysis.

For more detailed information, see "Selection of Studies for Derivation of Toxicity Values" in the Dose-Response Analysis section in the draft Handbook for IRIS Assessment Development in Appendix F.

See also Section 6 ("Selecting studies for dose-response analysis") in the Preamble to IRIS Toxicological Reviews in Appendix B.

A chemical-specific example of the implementation of this recommendation is available as "EXAMPLE 5 – Selecting Studies for Derivation of Toxicity Values" in the Chemical-specific Examples Demonstrating Implementation of NRC Recommendations document.

- **Considerations for Combining Data for Dose-Response Modeling** In Progress

The IRIS Program is now routinely considering whether combining data among studies is warranted for the derivation of toxicity values. For most IRIS assessments, the POD had been derived based on data from a single study dataset. This is because in most cases, datasets are often expected to be heterogeneous for biological or study design reasons.

However, there are cases where conducting dose-response modeling after combining data from multiple studies can be considered, resulting in a single POD based on multiple datasets. For instance, this may be useful to increase precision in the POD or to quantify the impact of specific sources of heterogeneity. The IRIS Program has developed considerations for combining data for dose-response modeling to be taken into account when performing dose-response analysis for an IRIS assessment.

In addition, multiple PODs or toxicity values can be combined (considering, for example, the highest quality studies, the most sensitive outcomes, or a clustering of values) to derive a single, overall toxicity value (or "meta-value"). For example, the IRIS assessment for trichloroethylene (TCE) identified multiple candidate RfDs that fell within a narrow dose range, and selected an overall RfD that reflected the midpoint among the similar candidate RfDs. This RfD is supported by multiple effects/studies and lead to a more robust (i.e., less sensitive to limitations of individual studies) (for more information: http://epa.gov/iris/subst/0199.htm, U.S. EPA, 2011).

For more detailed information, see "Considerations for Combining Data for Dose-Response Modeling" in the Dose-Response Analysis section in the draft Handbook for IRIS Assessment Development in Appendix F.

Calculation of Reference Values and Unit Risks

NRC Recommendations:

- Describe and justify assumptions and models used. This step includes review of dosimetry models and the implications of the models for uncertainty factors; determination of appropriate points of departure (such as benchmark dose, no-observed-adverse-effect level, and lowest observed-adverse-effect level), and assessment of the analyses that underlie the points of departure.
- Provide explanation of the risk-estimation modeling processes (for example, a statistical or biologic model fit to the data) that are used to develop a unit risk estimate.
- Provide adequate documentation for conclusions and estimation of reference values and unit risks. As noted by the committee throughout the present report, sufficient support for conclusions in the formaldehyde draft IRIS assessment is often lacking. Given that the development of specific IRIS assessments and their conclusions are of interest to many stakeholders, it is important that they provide sufficient references and supporting documentation for their conclusions. Detailed appendixes, which might be made available only electronically, should be provided, when appropriate.
- Assess the sensitivity of derived estimates to model assumptions and end points selected. This step should include appropriate tabular and graphic displays to illustrate the range of the estimates and the effect of uncertainty factors on the estimates.

Implementation:

➢ Conducting and Documenting Dose-Response Modeling and Deriving Toxicity Values

Implemented

IRIS assessments, in general, include dose-response analysis to derive toxicity values. In response to NRC recommendations, the IRIS Program has improved the quality control of the overall dose-response modeling process and increased transparency by documenting the approach for conducting dose-response modeling. Part of this documentation is achieved with the addition of considerations for selecting organ/system-specific and overall toxicity values, and a streamlined dose-response modeling output (both part of the new document structure). Additionally, tools and approaches to manage data and ensure quality (e.g., Data Management and Quality Control for Dose-Response Modeling) in dose-response analyses have been developed. The objectives are to minimize errors, maintain a transparent system for data management, automate tasks where possible, and maintain an archive of data and calculations used to develop assessments.

The IRIS Program has improved the documentation of dose-response modeling. *Preamble* Section 7 provides a description of the process for dose-response analysis. In addition, the text describing the dose-response analysis will include a description of how the toxicity values were derived and will cite EPA guidelines where appropriate.

For more detailed information, see "Data Management and Quality Control for Dose-Response Modeling," and "Considerations for Selecting Organ/System-Specific or Overall Toxicity Values" in the Dose-Response Analysis section in the draft Handbook for IRIS Assessment Development in Appendix F.

See also Section 7 ("Deriving toxicity values") in the Preamble to IRIS Toxicological Reviews in Appendix B.

Chemical-specific examples of the implementation of this recommendation are available as "EXAMPLE 6 – Dose-Response Modeling Output" and "EXAMPLE 7 – Considerations for Selecting Organ/System-Specific or Overall Toxicity Values" in the Chemical-specific Examples Demonstrating Implementation of NRC Recommendations document.

IV. Additional Initiatives

External Peer Review Enhancements

IRIS Peer Review Basics

Implemented

Rigorous, independent peer review is a cornerstone of IRIS assessments. Every IRIS assessment is reviewed by a group of internationally recognized experts in scientific disciplines relevant for the particular assessment. The peer review process used for IRIS assessments follows EPA guidance on peer review[2]. Most IRIS assessments are reviewed through contractor-organized or EPA's Science Advisory Board (SAB) peer reviews. All peer reviews, regardless of the reviewing body, involve a public comment period and public meeting (usually face-to-face). Following peer review, all revised IRIS assessments include an appendix describing how peer review and public comments were addressed.

Dedicated Chemical Assessment Advisory Committee

EPA's SAB has established a new standing committee, the Chemical Assessment Advisory Committee (CAAC), to review IRIS assessments. In the past, the SAB formed a new committee for each chemical assessment that the SAB reviewed. The new CAAC will provide the same high-level, transparent review as previous SAB reviews, but it will provide more continuous and overlapping membership for consistent advice.

The CAAC is comprised of 26 highly qualified scientists with a broad range of expertise relevant to human health assessment. The CAAC members will serve on panels reviewing individual IRIS assessments. Panels will be supplemented with added consultants who have expertise on the specific chemical substance or other areas of expertise needed to review the assessment. The CAAC review process is expected to be similar to how IRIS assessment reviews are currently conducted by the SAB and will include the following: the public will be invited to nominate peer reviewers for specific assessments; the proposed panels or pools of panelists will be posted for public comment; the proposed panelists will be screened by an Agency official for conflicts of interest; the final panel will be announced prior to the peer review phase.

[2] U.S. EPA (2006) *Science Policy Council Peer Review Handbook - 3rd Edition, EPA document number EPA/100/B-06/002.*(http://www.epa.gov/peerreview/) and the EPA National Center for Environmental Assessment Policy and Procedures for Conducting IRIS Peer Reviews (2009, http://www.epa.gov/iris/pdfs/Policy_IRIS_Peer_Reviews.pdf).

V. Summary

EPA is committed to a strong, vital, and scientifically sound IRIS Program. Over the past two years, EPA has worked to strengthen and streamline the IRIS Program, improving transparency and creating efficiencies. Significant changes have been made in response to the NRC recommendations and further efforts are underway to fully implement the recommendations.

Appendix A – IRIS Toxicological Review Template

www.epa.gov/iris

Toxicological Review of [Chemical]

[CASRN X-X-X]

In Support of Summary Information on the Integrated Risk Information System (IRIS)

DATE

NOTICE

This document is an **[Agency Review, Interagency Science Consultation, Public Comment, External Review, or Final Agency/Interagency Science Discussion] draft.** This information is distributed solely for the purpose of pre-dissemination peer review under applicable information quality guidelines. It has not been formally disseminated by EPA. It does not represent and should not be construed to represent any Agency determination or policy. It is being circulated for review of its technical accuracy and science policy implications.

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

DISCLAIMER

This document is a preliminary draft for review purposes only. This information is distributed solely for the purpose of pre-dissemination peer review under applicable information quality guidelines. It has not been formally disseminated by EPA. It does not represent and should not be construed to represent any Agency determination or policy. Mention of trade names or commercial products does not constitute endorsement of recommendation for use.

CONTENTS

TABLES

FIGURES

ABBREVIATIONS

α2u-g	alpha2u-globulin
ACGIH	American Conference of Governmental Industrial Hygienists
AIC	Akaike's information criterion
ALD	approximate lethal dosage
ALT	alanine aminotransferase
AST	aspartate aminotransferase
atm	atmosphere
ATSDR	Agency for Toxic Substances and Disease Registry
BMD	benchmark dose
BMDL	benchmark dose lower confidence limit
BMDS	Benchmark Dose Software
BMR	benchmark response
BUN	blood urea nitrogen
BW	body weight
CA	chromosomal aberration
CASRN	Chemical Abstracts Service Registry Number
CBI	covalent binding index
CHO	Chinese hamster ovary (cell line cells)
CL	confidence limit
CNS	central nervous system
CPN	chronic progressive nephropathy
CYP450	cytochrome P450
DAF	dosimetric adjustment factor
DEN	diethylnitrosamine
DMSO	dimethylsulfoxide
DNA	deoxyribonucleic acid
EPA	Environmental Protection Agency
FDA	Food and Drug Administration
FEV_1	forced expiratory volume of 1 second
GD	gestation day
GDH	glutamate dehydrogenase
GGT	γ-glutamyl transferase
GSH	glutathione
GST	glutathione-S-transferase
Hb/g-A	animal blood:gas partition coefficient
Hb/g-H	human blood:gas partition coefficient
HEC	human equivalent concentration
HED	human equivalent dose
i.p.	intraperitoneal
IRIS	Integrated Risk Information System
IVF	in vitro fertilization
LC_{50}	median lethal concentration
LD_{50}	median lethal dose
LOAEL	lowest-observed-adverse-effect level
MN	micronuclei
MNPCE	micronucleated polychromatic erythrocyte
MTD	maximum tolerated dose
NAG	N-acetyl-β-D-glucosaminidase
NCEA	National Center for Environmental Assessment
NCI	National Cancer Institute
NOAEL	no-observed-adverse-effect level
NTP	National Toxicology Program
NZW	New Zealand White (rabbit breed)
OCT	ornithine carbamoyl transferase
ORD	Office of Research and Development
PBPK	physiologically based pharmacokinetic
PCNA	proliferating cell nuclear antigen
POD	point of departure
$POD_{[ADJ]}$	duration-adjusted POD
QSAR	quantitative structure-activity relationship
RDS	replicative DNA synthesis
RfC	inhalation reference concentration
RfD	oral reference dose
RGDR	regional gas dose ratio
RNA	ribonucleic acid
SAR	structure activity relationship
SCE	sister chromatid exchange
SD	standard deviation
SDH	sorbitol dehydrogenase
SE	standard error
SGOT	glutamic oxaloacetic transaminase, also known as AST
SGPT	glutamic pyruvic transaminase, also known as ALT
SSD	systemic scleroderma
TCA	trichloroacetic acid
TCE	trichloroethylene
TWA	time-weighted average
UF	uncertainty factor
UF_A	animal-to-human uncertainty factor
UF_H	human variation uncertainty factor
UF_L	LOAEL-to-NOAEL uncertain factor
UF_S	subchronic-to-chronic uncertainty factor
UF_D	database deficiencies uncertainty factor
U.S.	United States of America

AUTHORS | CONTRIBUTORS | REVIEWERS

Assessment Team

NAME (alphabetical order)	U.S. EPA/ORD/NCEA
NAME	Washington, DC
NAME(Chemical Manager)	
NAME	

Scientific Support Team

NAME (alphabetical order)	U.S. EPA/ORD/NCEA
NAME[3]	Washington, DC
NAME	
NAME	

Production Team

NAME (alphabetical order)	Agency, Office, Location
NAME	Agency, Office, Location
NAME	Agency, Office, Location
NAME	Agency, Office, Location

Contractor Support

NAME	Company, Location
NAME	
NAME	
NAME	Company, Location
NAME	
NAME	

Executive Direction

Kenneth Olden, Ph.D., Sc.D., L.H.D. (Center Director)	U.S. EPA/ORD/NCEA
John Vandenberg, Ph.D, (National Program Director, HHRA)	Washington, DC
Lynn Flowers, Ph.D., DABT (Associate Director for Health)	
Vincent Cogliano, Ph.D.[4] (IRIS Program Director—acting)	
Samantha Jones, Ph.D.[5] (IRIS Associate Director for Science)	
[Division Director]	
[Branch Chief]	

[3] Chemical Assessment Support Team (CAST) Member
[4] Chemical Assessment Support Team (CAST) Lead
[5] Chemical Assessment Support Team (CAST) Lead

Internal Review Team

NAME	Agency, Office, Location
NAME	Agency, Office, Location
NAME	Agency, Office, Location
NAME	Agency, Office, Location

Reviewers

This assessment was provided for review to scientists in EPA's program and regional offices. Comments were submitted by:

Office, Location
Office, Location
Office, Location
Office, Location

This assessment was provided for review to other federal agencies and Executive Offices of the President. Comments were submitted by:

AGENCY
AGENCY
AGENCY
AGENCY

This assessment was released for public comment on [month] [day], [year] and comments were due on [month] [day], [year]. The external peer-review comments are available on the IRIS Web site. A summary and EPA's disposition of the comments received from the independent external peer reviewers and from the public is included in Appendix [X] and is also available on the IRIS Web site. Comments were received from the following entities:

NAME	Affiliation, Location
NAME	Affiliation, Location
NAME	Affiliation, Location
NAME	Affiliation, Location

This assessment was peer reviewed by independent expert scientists external to EPA and a peer-review meeting was held on [month] [day], [year]. The external peer-review comments are available on the IRIS Web site. A summary and EPA's disposition of the comments received from the independent external peer reviewers and from the public is included in Appendix [X] and is also available on the IRIS Web site.

NAME	Affiliation, Location
NAME	Affiliation, Location
NAME	Affiliation, Location
NAME	Affiliation, Location

PREFACE

This Toxicological Review critically reviews the publicly available studies on [chemical] in order to identify its adverse health effects and to characterize exposure-response relationships. The assessment covers [....] It was prepared under the auspices of EPA's Integrated Risk Information System (IRIS) Program.

[Chemical] is listed as [....] [Why is EPA interested in this assessment? Is the chemical included on Agency lists (ex. HAPs, DWCL)?]

[If this is a reassessment...] This assessment updates a previous IRIS assessment of [chemical] that was developed in [year]. The previous assessment included [....]. New information has become available and this assessment reviews information on all health effects by all exposure routes. Organ/system-specific RfDs are calculated based on [applicable hazards, e.g., developmental, reproductive and immune system toxicity data]. These toxicity values may be useful for cumulative risk assessments that consider the combined effect of multiple agents acting on the same biological system.

This assessment was conducted in accordance with EPA guidance, which is cited and summarized in the *Preamble to IRIS Toxicological Reviews*. The findings of this assessment and related documents produced during its development are available on the IRIS Web site (http://www.epa.gov/iris). Appendices for chemical and physical properties, toxicokinetic information, and summaries of toxicology studies and other information are provided as *Supplemental Information* to this assessment.

For additional information about this assessment or for general questions regarding IRIS, please contact EPA's IRIS Hotline at 202-566-1676 (phone), 202-566-1749 (fax), or hotline.iris@epa.gov.

Assessments by Other National and International Health Agencies

Toxicity information on [chemical] has been evaluated by [....]. The results of these assessments are presented in Appendix A of the Supplemental Information. It is important to recognize that these assessments may have been prepared for different purposes and may utilize different methods, and that newer studies may be included in the IRIS assessment.

Chemical Properties and Uses

[Appendix B...]

PREAMBLE TO IRIS TOXICOLOGICAL REVIEWS

EXECUTIVE SUMMARY

Occurrence and Health Effects

[Placeholder for text]

Effects Other Than Cancer Observed Following Oral Exposure

Oral Reference Dose (RfD) for Effects Other Than Cancer

Table ES-1. Summary of reference dose (RfD) derivation

Critical effect	Point of departure*	UF	Chronic RfD

[* Conversion Factors and Assumptions—]

Confidence in the Chronic Oral RfD

Effects Other Than Cancer Observed Following Inhalation Exposure

Inhalation Reference Concentration (RfC) for Effects Other Than Cancer

Table ES-2. Summary of reference concentration (RfC) derivation

Critical effect	Point of departure*	UF	Chronic RfC

[* Conversion Factors and Assumptions—]

Confidence in the Chronic Inhalation RfC

Evidence for Human Carcinogenicity

Quantitative Estimate of Carcinogenic Risk From Oral Exposure

Quantitative Estimate of Carcinogenic Risk From Inhalation Exposure

Susceptible Populations and Lifestages

Key Issues Addressed in Assessment

LITERATURE SEARCH STRATEGY | STUDY SELECTION AND EVALUATION

1. HAZARD IDENTIFICATION

1.1. SYNTHESIS OF EVIDENCE

1.2. SUMMARY AND EVALUATION

2. DOSE-RESPONSE ANALYSIS

2.1. ORAL REFERENCE DOSE FOR EFFECTS OTHER THAN CANCER

The RfD (expressed in units of mg/kg-day) is defined as an estimate (with uncertainty spanning perhaps an order of magnitude) of a daily exposure to the human population (including sensitive subgroups) that is likely to be without an appreciable risk of deleterious effects during a lifetime. It can be derived from a no-observed-adverse-effect level (NOAEL), lowest-observed-adverse-effect level (LOAEL), or the 95% lower bound on the benchmark dose (BMDL), with uncertainty factors (UFs) generally applied to reflect limitations of the data used.

2.1.1. Identification of Studies and Effects for Dose-Response Analysis

Hazard A

Hazard B

Hazard C

2.1.2. Methods of Analysis

Table 2-1 summarizes the sequence of calculations leading to the derivation of a human-equivalent point of departure for each data set discussed above.

Table 2-1. Summary of derivation of points of departure following oral exposure

Endpoint and reference	Species/ sex	Model	BMR	BMD	BMDL	POD_{ADJ}	POD_{HED}
Hazard A (ex. DEVELOPMENTAL)							
Hazard B (ex. REPRODUCTIVE)							

2.1.3. Derivation of Candidate Values

Table 2-2 is a continuation of Table 2-1 and summarizes the application of uncertainty factors to each point of departure to derive candidate values for each data set. The candidate values presented in Table 2-2 are preliminary to the derivation of reference values in subsequent sections. The selection of uncertainty factors is based on EPA's *Review of the Reference Dose and Reference Concentration Processes* (U.S. EPA, 2002; Section 4.4.5) and is described in Section 7.6 of the *Preamble*. Figure 2-1 presents graphically the candidate values, uncertainty factors, and points of departure, with each bar corresponding to one data set described in Tables 2-1 and 2-2.

Table 2-2. Effects and corresponding derivation of candidate RfDs

Endpoint and reference	POD_{HED}	POD type	UF_A	UF_H	UF_L	UF_S	UF_D	Composite UF	Candidate value (mg/kg-day)
Hazard A (ex. DEVELOPMENTAL)									
Hazard B (ex. REPRODUCTIVE)									

[Insert rationale for the application of uncertainty factors. The value (e.g., 1, 3, or 10) of the uncertainty factor will depend upon the availability of data and what is known about the chemical…]

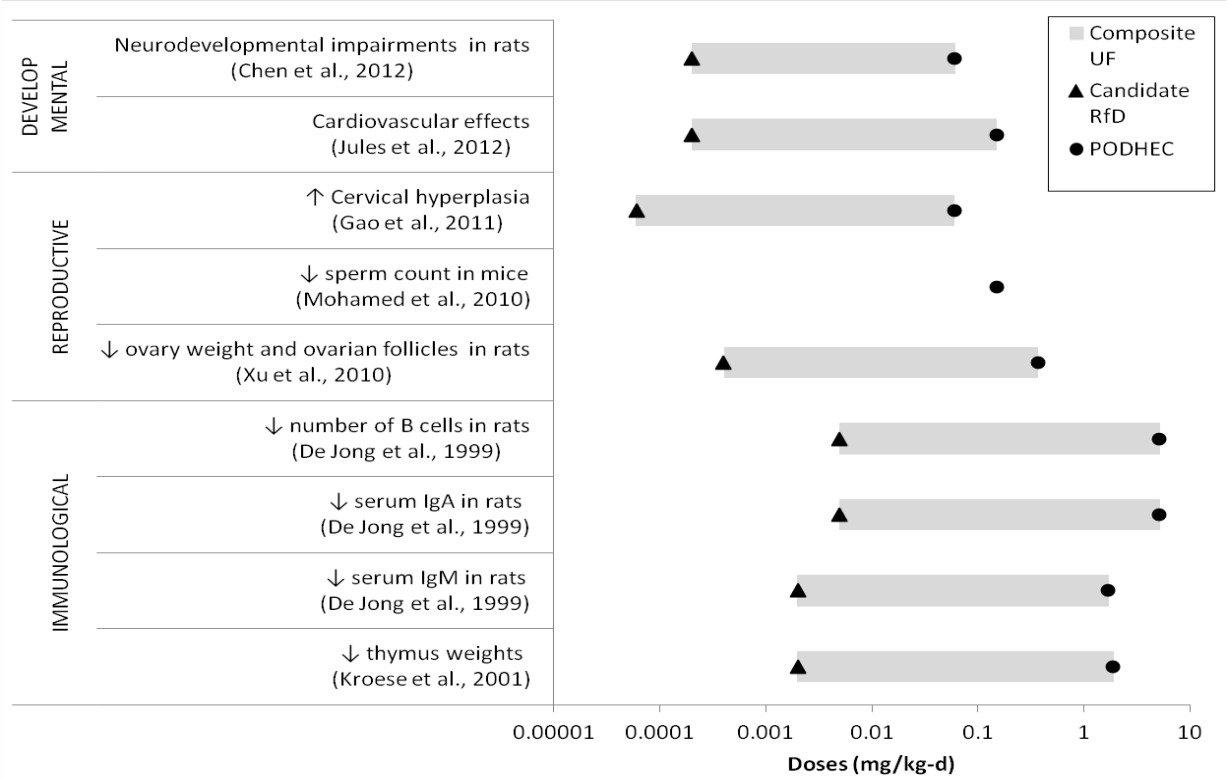


Figure 2-1. Candidate RfDs with corresponding POD and composite UF.
[Note: Data shown here are provided only for illustrative purposes]

2.1.4. Derivation of Organ/System-Specific Reference Doses

Table 2-3 distills the candidate values from Table 2-2 into a single value for each organ or system. These organ or system-specific reference values may be useful for subsequent cumulative risk assessments that consider the combined effect of multiple agents acting at a common site.

Hazard A

Hazard B

Hazard C

Table 2-3. Organ/system-specific RfDs and proposed overall RfD for [chemical]

Effect	Basis	RfD (mg/kg-day)	Exposure description	Confidence
Hazard A			Ex. chronic	
Hazard B			Ex. gestational	
Hazard C			Ex. subchronic	
Proposed overall RfD			**Ex. gestational**	

2.1.5. Selection of the Proposed Overall Reference Dose

2.1.6. Uncertainties in the Derivation of Reference Dose

2.1.7. Confidence Statement

A confidence level of high, medium, or low is assigned to the study used to derive the RfD, the overall database, and the RfD itself, as described in Section 4.3.9.2 of EPA's *Methods for Derivation of Inhalation Reference Concentrations and Application of Inhalation Dosimetry* (U.S. EPA, 1994).

2.1.8. Previous IRIS Assessment

2.2. INHALATION REFERENCE CONCENTRATION FOR EFFECTS OTHER THAN CANCER

The RfC (expressed in units of mg/m^3) is defined as an estimate (with uncertainty spanning perhaps an order of magnitude) of a continuous inhalation exposure to the human population (including sensitive subgroups) that is likely to be without an appreciable risk of deleterious effects during a lifetime. It can be derived from a NOAEL, LOAEL, or the 95% lower bound on the benchmark concentration (BMCL), with UFs generally applied to reflect limitations of the data used.

2.2.1. Identification of Studies and Effects for Dose-Response Analysis

Hazard A

Hazard B

Hazard C

2.2.2. Methods of Analysis

Table 2-4 summarizes the sequence of calculations leading to the derivation of a human-equivalent point of departure for each data set discussed above.

Table 2-4. Summary of derivation of points of departure following inhalation exposure

Endpoint and reference	Species/ sex	Model	BMR	BMC	BMCL	POD_{ADJ}	POD_{HED}
Hazard A (ex. DEVELOPMENTAL)							
Hazard B (ex. REPRODUCTIVE)							

2.2.3. Derivation of Candidate Values

Table 2-5 is a continuation of Table 2-4 and summarizes the application of uncertainty factors to each point of departure to derive a candidate values for each data set. The candidate values presented in Table 2-5 are for exploratory purposes only, and are preliminary to the derivation of reference values in subsequent sections. The selection of uncertainty factors was based on EPA's *Review of the Reference Dose and Reference Concentration Processes* (U.S. EPA, 2002; Section 4.4.5) and is described in Section 7.6 of the *Preamble*. Figure 2-2 graphically presents these candidate values, uncertainty factors, and points of departure with each bar corresponding to one data set described in Tables 2-4 and 2-5.

Table 2-5. Effects and corresponding derivation of candidate RfCs

Endpoint	POD_{HEC} ($\mu g/m^3$)	POD type	UF_A	UF_H	UF_L	UF_S	UF_D	Composite UF	Candidate value ($\mu g/m^3$)
Hazard A (ex. DEVELOPMENTAL)									
Hazard B (ex. REPRODUCTIVE)									

[Insert rationale for the application of uncertainty factors. The value (e.g., 1, 3, or 10) of the uncertainty factor will depend upon the availability of data and what is known about the chemical…]

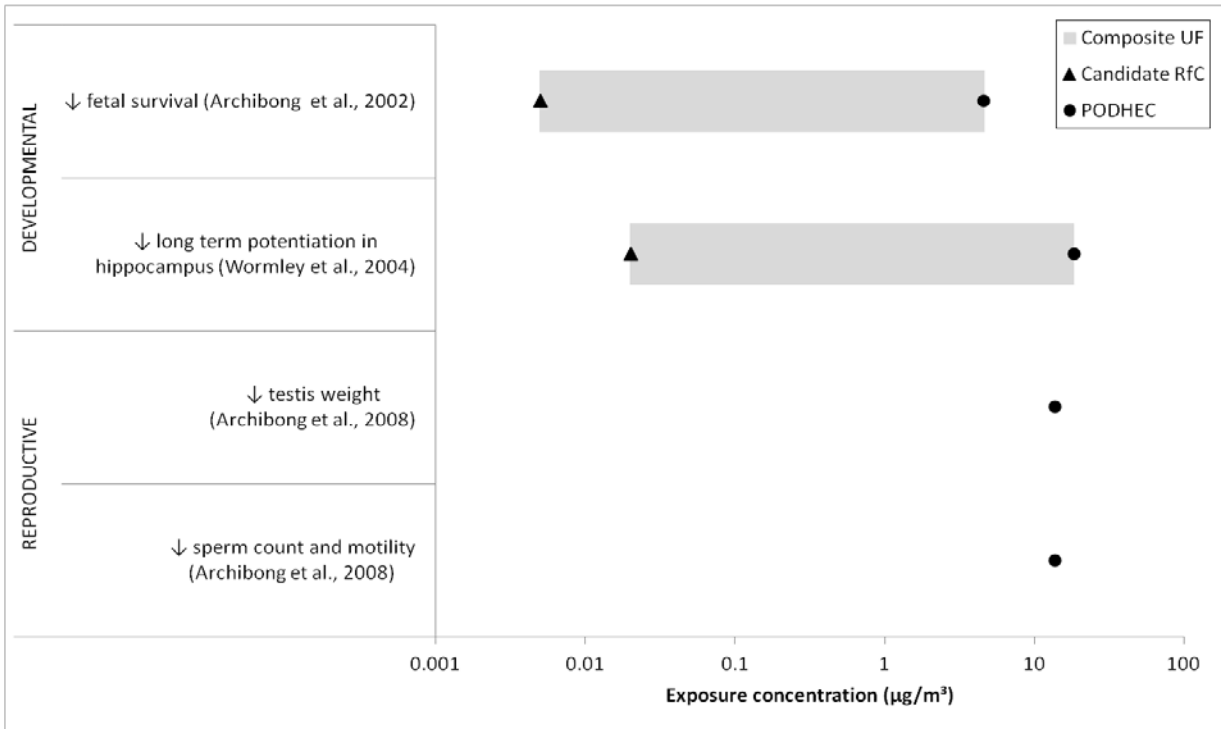


Figure 2-2. Candidate RfCs with corresponding POD and composite UF.
[Note: Data shown here are provided only for illustrative purposes.]

2.2.4. Derivation of Organ/System-Specific Reference Concentrations

Table 2-6 distills the candidate values from Table 2-5 into a single value for each organ or system. These organ or system-specific reference values may be useful for subsequent cumulative risk assessments that consider the combined effect of multiple agents acting at a common site.

Hazard A

Hazard B

Hazard C

Table 2-6. Organ/system-specific RfCs and proposed overall RfC for [chemical]

Effect	Basis	RfC (mg/m^3)	Exposure description	Confidence
Hazard A				
Hazard B				
Hazard C				
Proposed overall RfC				

2.2.5. Selection of the Proposed Overall Reference Concentration

2.2.6. Uncertainties in the Derivation of Reference Concentration

2.2.7. Confidence Statement

A confidence level of high, medium, or low is assigned to the study used to derive the RfC, the overall database, and the RfC itself, as described in Section 4.3.9.2 of EPA's *Methods for Derivation of Inhalation Reference Concentrations and Application of Inhalation Dosimetry* (U.S. EPA, 1994).

2.2.8. Previous IRIS Assessment

2.3. ORAL SLOPE FACTOR FOR CANCER

The carcinogenicity assessment provides information on the carcinogenic hazard potential of the substance in question, and quantitative estimates of risk from oral and inhalation exposure may be derived. Quantitative risk estimates may be derived from the application of a low-dose extrapolation procedure. If derived, the oral slope factor is a plausible upper bound on the estimate of risk per mg/kg-day of oral exposure. [Note: Similarly, an inhalation unit risk is a plausible upper bound on the estimate of risk per µg/m^3 air breathed.]

2.3.1. Analysis of Carcinogenicity Data

2.3.2. Dose-Response Analysis—Adjustments and Extrapolations Methods

2.3.3. Derivation of the Oral Slope Factor

2.3.4. Uncertainties in the Derivation of the Oral Slope Factor

2.3.5. Previous IRIS Assessment: Oral Slope Factor

2.4. INHALATION UNIT RISK FOR CANCER

The carcinogenicity assessment provides information on the carcinogenic hazard potential of the substance in question and quantitative estimates of risk from oral and inhalation exposure may be derived. Quantitative risk estimates may be derived from the application of a low-dose extrapolation procedure. If derived, the inhalation unit risk is a plausible upper bound on the estimate of risk per µg/m^3 air breathed.

2.4.1. Analysis of Carcinogenicity Data

2.4.2. Dose-Response Analysis—Adjustments and Extrapolations Methods

2.4.3. Inhalation Unit Risk Derivation

2.4.4. Uncertainties in the Derivation of the Inhalation Unit Risk

2.4.5. Previous IRIS Assessment: Inhalation Unit Risk

2.5. APPLICATION OF AGE-DEPENDENT ADJUSTMENT FACTORS

REFERENCES

Appendix B – Preamble to IRIS Toxicological Reviews

1. Scope of the IRIS Program

Soon after EPA was established in 1970, it was at the forefront of developing risk assessment as a science and applying it in decisions to protect human health and the environment. The Clean Air Act, for example, mandates that EPA provide "an ample margin of safety to protect public health"; the Safe Drinking Water Act, that "no adverse effects on the health of persons may reasonably be anticipated to occur, allowing an adequate margin of safety." Accordingly, EPA uses information on the adverse effects of chemicals and on exposure levels below which these effects are not anticipated to occur.

IRIS assessments critically review the publicly available studies to identify adverse health effects from long-term exposure to chemicals and to characterize exposure-response relationships. In terms set forth by the National Research Council (NRC, 1983), IRIS assessments cover the hazard identification and dose-response assessment steps of risk assessment, not the exposure assessment or risk characterization steps that are conducted by EPA's program and regional offices and by other federal, state, and local health agencies that evaluate risk in specific populations and exposure scenarios. IRIS assessments are distinct from and do not address political, economic, and technical considerations that influence the design and selection of risk management alternatives.

An IRIS assessment may cover a single chemical, a group of structurally or toxicologically related chemicals, or a complex mixture. Exceptions are chemicals currently used exclusively as pesticides, ionizing and non-ionizing radiation, and criteria air pollutants listed under section 108 of the Clean Air Act (carbon monoxide, lead, nitrogen oxides, ozone, particulate matter, and sulfur oxides).

Periodically, the IRIS Program asks other EPA programs and regions, other federal agencies, state health agencies, and the general public to nominate chemicals and mixtures for future assessment or reassessment. These agents may be found in air, water, soil, or sediment. Selection is based on program and regional office priorities and on availability of adequate information to evaluate the potential for adverse effects. The IRIS Program may assess other agents as an urgent public health need arises. IRIS also reassesses agents as significant new studies are published.

2. Process for developing and peer-reviewing IRIS assessments

The process for developing IRIS assessments (revised in May 2009) involves critical analysis of the pertinent studies, opportunities for public input, and multiple levels of scientific review. EPA revises draft assessments after each review, and external drafts and comments become part of the public record (U.S. EPA, 2009).

Step 1. Development of a draft Toxicological Review (generally about 11-1/2 months duration). The draft assessment considers all pertinent publicly available studies and applies consistent criteria to evaluate study quality, identify health effects, identify mechanistic events and pathways, integrate the evidence of causation for each effect, and derive toxicity values. A public dialogue meeting prior to the integration of evidence and derivation of toxicity values promotes public discussion of the literature search, evidence, and key issues.

Step 2. Internal review by scientists in EPA programs and regions (2 months). The draft assessment is revised to address comments from within EPA.

Step 3. Interagency science consultation with other federal agencies and the Executive Offices of the President (1-1/2 months). The draft assessment is revised to address the interagency comments. The science consultation draft, interagency comments, and EPA's response to major comments become part of the public record.

Step 4. Public review and comment, followed by external peer review (3-1/2 months or more, depending on the review process). EPA releases the draft assessment for public review and comment. Another public dialogue meeting provides an opportunity to discuss the assessment prior to peer review. EPA addresses the public comments and releases a draft for external peer review. The peer reviewers assess whether the evidence has been assembled and evaluated according to guidelines and whether the conclusions are justified by the evidence. The peer review meeting is open to the public and includes time for oral public comments. The peer review draft, peer review report, and written public comments become part of the public record.

Step 5. Revision of draft Toxicological Review and development of draft IRIS summary (2 months). The draft assessment is revised to reflect the peer review comments, public comments, and newly published studies that are critical to the conclusions of the assessment. The disposition of peer review comments and public comments becomes part of the public record.

Step 6. Final EPA review and interagency science discussion with other federal agencies and the Executive Offices of the President (1-1/2 months). The draft assessment and summary are revised to address EPA and interagency comments. The science discussion draft, written interagency comments, and EPA's response to major comments become part of the public record.

Step 7. Completion and posting (1 month). The Toxicological Review and IRIS summary are posted on the IRIS web site (http://www.epa.gov/iris/).

The remainder of this Preamble addresses step 1, the development of a draft Toxicological Review. IRIS assessments follow standard practices of evidence evaluation and peer review, many of which are discussed in EPA guidelines (U.S. EPA, 1986a, 1986b, 1991, 1996, 1998, 2000, 2005a, 2005b) and other methods (U.S. EPA, 1994, 2002, 2006a, 2006b, 2011, 2012a, 2012b). A practical draft *Handbook* is available for use by IRIS assessment teams (U.S. EPA, 2013). Transparent application of scientific judgment is of paramount importance. To provide a harmonized approach across IRIS assessments, this Preamble summarizes concepts from these guidelines and emphasizes principles of general applicability.

3. Identifying and selecting pertinent studies

3.1 Identifying studies

Before beginning an assessment, EPA conducts a comprehensive search of the primary scientific literature. The literature search follows standard practices and includes the PubMed and ToxNet databases of the National Library of Medicine, Web of Science, and other databases listed in EPA's HERO system (Health and Environmental Research Online, http://hero.epa.gov/). Searches for information on mechanisms of toxicity are inherently specialized and may include studies on other agents that act through related mechanisms.

Each assessment specifies the search strategies, keywords, and cut-off dates of its literature searches. EPA posts the results of the literature search on the IRIS web site and requests information from the public on additional studies and ongoing research.

EPA also considers studies received through the IRIS Submission Desk and studies (typically unpublished) submitted under the Toxic Substances Control Act or the Federal Insecticide, Fungicide, and Rodenticide Act. Material submitted as Confidential Business Information is considered only if it includes health and safety data that can be publicly released. If a study that may be critical to the conclusions of the assessment has not been peer-reviewed, EPA will have it peer-reviewed.

EPA also examines the toxicokinetics of the agent to identify other chemicals (for example, major metabolites of the agent) to include in the assessment if adequate information is available, in order to more fully explain the toxicity of the agent and to suggest dose metrics for subsequent modeling.

In assessments of chemical mixtures, mixture studies are preferred for their ability to reflect interactions among components. The literature search seeks, in decreasing order of preference (U.S. EPA, 1986a, 2000):

- Studies of the mixture being assessed.
- Studies of a sufficiently similar mixture. In evaluating similarity, the assessment considers the alteration of mixtures in the environment through partitioning and transformation.
- Studies of individual chemical components of the mixture, if there are not adequate studies of sufficiently similar mixtures.

3.2 Selecting pertinent epidemiologic studies

Study design is the key consideration for selecting pertinent epidemiologic studies from the results of the literature search.

- Cohort studies, case-control studies, and some population-based surveys (for example, NHANES) provide the strongest epidemiologic evidence, especially when they collect information about individual exposures and effects.
- Ecological studies (geographic correlation studies) relate exposures and effects by geographic area. They can provide strong evidence if there are large exposure contrasts between geographic areas, relatively little exposure variation within study areas, and population migration is limited.
- Case reports of high or accidental exposure lack definition of the population at risk and the expected number of cases. They can provide information about a rare effect or about the relevance of analogous results in animals.

The assessment briefly reviews ecological studies and case reports but reports details only if they suggest effects not identified by other studies.

3.3 Selecting pertinent experimental studies

Exposure route is a key design consideration for selecting pertinent experimental animal studies or human clinical studies.

- Studies of oral, inhalation, or dermal exposure involve passage through an absorption barrier and are considered most pertinent to human environmental exposure.
- Injection or implantation studies are often considered less pertinent but may provide valuable toxicokinetic or mechanistic information. They also may be useful for identifying effects in animals if deposition or absorption is problematic (for example, for particles and fibers).

Exposure duration is also a key design consideration for selecting pertinent experimental animal studies.

- Studies of effects from chronic exposure are most pertinent to lifetime human exposure.
- Studies of effects from less-than-chronic exposure are pertinent but less preferred for identifying effects from lifetime human exposure. Such studies may be indicative of effects from less-than-lifetime human exposure.

Short-duration studies involving animals or humans may provide toxicokinetic or mechanistic information.

For developmental toxicity and reproductive toxicity, irreversible effects may result from a brief exposure during a critical period of development. Accordingly, specialized study designs are used for these effects (U.S. EPA, 1991, 1996, 1998, 2006b).

4. Evaluating the quality of individual studies

After the subsets of pertinent epidemiologic and experimental studies have been selected from the literature searches, the assessment evaluates the quality of each individual study. This evaluation considers the design, methods, conduct, and documentation of each study, but not whether the results are positive, negative, or null. The

objective is to identify the stronger, more informative studies based on a uniform evaluation of quality characteristics across studies of similar design.

4.1 Evaluating the quality of epidemiologic studies

The assessment evaluates design and methodological aspects that can increase or decrease the weight given to each epidemiologic study in the overall evaluation (U.S. EPA, 1991, 1994, 1996, 1998, 2005a):

- Documentation of study design, methods, population characteristics, and results.
- Definition and selection of the study group and comparison group.
- Ascertainment of exposure to the chemical or mixture.
- Ascertainment of disease or health effect.
- Duration of exposure and follow-up and adequacy for assessing the occurrence of effects.
- Characterization of exposure during critical periods.
- Sample size and statistical power to detect anticipated effects.
- Participation rates and potential for selection bias as a result of the achieved participation rates.
- Measurement error (can lead to misclassification of exposure, health outcomes, and other factors) and other types of information bias.
- Potential confounding and other sources of bias addressed in the study design or in the analysis of results. The basis for consideration of confounding is a reasonable expectation that the confounder is related to both exposure and outcome and is sufficiently prevalent to result in bias.

For developmental toxicity, reproductive toxicity, neurotoxicity, and cancer there is further guidance on the nuances of evaluating epidemiologic studies of these effects (U.S. EPA, 1991, 1996, 1998, 2005a).

4.2 Evaluating the quality of experimental studies

The assessment evaluates design and methodological aspects that can increase or decrease the weight given to each experimental animal study, in-vitro study, or human clinical study (U.S. EPA, 1991, 1994, 1996, 1998, 2005a). Research involving human subjects is considered only if conducted according to ethical principles.

- Documentation of study design, animals or study population, methods, basic data, and results.
- Nature of the assay and validity for its intended purpose.
- Characterization of the nature and extent of impurities and contaminants of the administered chemical or mixture.
- Characterization of dose and dosing regimen (including age at exposure) and their adequacy to elicit adverse effects, including latent effects.
- Sample sizes and statistical power to detect dose-related differences or trends.
- Ascertainment of survival, vital signs, disease or effects, and cause of death.
- Control of other variables that could influence the occurrence of effects.

The assessment uses statistical tests to evaluate whether the observations may be due to chance. The standard for determining statistical significance of a response is a trend test or comparison of outcomes in the exposed groups against those of concurrent controls. In some situations, examination of historical control data from the same laboratory within a few years of the study may improve the analysis. For an uncommon effect that is not statistically significant compared with concurrent controls, historical controls may show that the effect is unlikely to be due to chance. For a response that appears significant against a concurrent control response that is unusual, historical controls may offer a different interpretation (U.S. EPA, 2005a).

For developmental toxicity, reproductive toxicity, neurotoxicity, and cancer there is further guidance on the nuances of evaluating experimental studies of these effects (U.S. EPA,

1991, 1996, 1998, 2005a). In multi-generation studies, agents that produce developmental effects at doses that are not toxic to the maternal animal are of special concern. Effects that occur at doses associated with mild maternal toxicity are not assumed to result only from maternal toxicity. Moreover, maternal effects may be reversible, while effects on the offspring may be permanent (U.S. EPA, 1991, 1998).

4.3 Reporting study results

The assessment uses evidence tables to present the design and key results of pertinent studies. There may be separate tables for each site of toxicity or type of study.

If a large number of studies observe the same effect, the assessment considers the study quality characteristics in this section to identify the strongest studies or types of study. The tables present details from these studies, and the assessment explains the reasons for not reporting details of other studies or groups of studies that do not add new information. Supplemental information provides references to all studies considered, including those not summarized in the tables.

The assessment discusses strengths and limitations that affect the interpretation of each study. If the interpretation of a study in the assessment differs from that of the study authors, the assessment discusses the basis for the difference.

As a check on the selection and evaluation of pertinent studies, EPA asks peer reviewers to identify studies that were not adequately considered.

5. Evaluating the overall evidence of each effect

5.1 Concepts of causal inference

For each health effect, the assessment evaluates the evidence as a whole to determine whether it is reasonable to infer a causal association between exposure to the agent and the occurrence of the effect. This inference is based on information from pertinent human studies, animal studies, and mechanistic studies of adequate quality. Positive, negative, and null results are given weight according to study quality.

Causal inference involves scientific judgment, and the considerations are nuanced and complex. Several health agencies have developed frameworks for causal inference, among them the U.S. Surgeon General (DHEW, 1964; DHHS, 2004), the International Agency for Research on Cancer (2006), the Institute of Medicine (2008), and the U.S. Environmental Protection Agency (2005a, 2010). Although developed for different purposes, the frameworks are similar in nature and provide an established structure and language for causal inference. Each considers aspects of an association that suggest causation, discussed by Hill (1965) and elaborated by Rothman and Greenland (1998) (U.S. EPA, 1994, 2002, 2005a).

Strength of association: The finding of a large relative risk with narrow confidence intervals strongly suggests that an association is not due to chance, bias, or other factors. Modest relative risks, however, may reflect a small range of exposures, an agent of low potency, an increase in an effect that is common, exposure misclassification, or other sources of bias.

Consistency of association: An inference of causation is strengthened if elevated risks are observed in independent studies of different populations and exposure scenarios. Reproducibility of findings constitutes one of the strongest arguments for causation. Discordant results sometimes reflect differences in study design, exposure, or confounding factors.

Specificity of association: As originally intended, this refers to one cause associated with one effect. Current understanding that many agents cause multiple effects and many effects have multiple causes make this a less informative aspect of causation, unless the effect is rare or unlikely to have multiple causes.

Temporal relationship: A causal interpretation requires that exposure precede development of the effect.

Biologic gradient (exposure-response relationship): Exposure-response relationships strongly suggest causation. A monotonic increase is not the only pattern consistent with causation. The presence of an exposure-response gradient also weighs against bias and confounding as the source of an association.

Biologic plausibility: An inference of causation is strengthened by data demonstrating plausible biologic mechanisms, if available. Plausibility may reflect subjective prior beliefs if there is insufficient understanding of the biologic process involved.

Coherence: An inference of causation is strengthened by supportive results from animal experiments, toxicokinetic studies, and short-term tests. Coherence may also be found in other lines of evidence, such as changing disease patterns in the population.

"Natural experiments": A change in exposure that brings about a change in disease frequency provides strong evidence, as it tests the hypothesis of causation. An example would be an intervention to reduce exposure in the workplace or environment that is followed by a reduction of an adverse effect.

Analogy: Information on structural analogues or on chemicals that induce similar mechanistic events can provide insight into causation.

These considerations are consistent with guidelines for systematic reviews that evaluate the quality and weight of evidence. Confidence is increased if the magnitude of effect is large, if there is evidence of an exposure-response relationship, or if an association was observed and the plausible biases would tend to decrease the magnitude of the reported effect. Confidence is decreased for study limitations, inconsistency of results, indirectness of evidence, imprecision, or reporting bias (Guyatt et al., 2008a,b).

5.2 Evaluating evidence in humans

For each effect, the assessment evaluates the evidence from the epidemiologic studies as a whole. The objective is to determine whether a credible association has been observed and, if so, whether that association is consistent with causation. In doing this, the assessment explores alternative explanations (such as chance, bias, and confounding) and draws a conclusion about whether these alternatives can satisfactorily explain any observed association.

To make clear how much the epidemiologic evidence contributes to the overall weight of the evidence, the assessment may select a standard descriptor to characterize the epidemiologic evidence of association between exposure to the agent and occurrence of a health effect.

Sufficient epidemiologic evidence of an association consistent with causation: The evidence establishes a causal association for which alternative explanations such as chance, bias, and confounding can be ruled out with reasonable confidence.

Suggestive epidemiologic evidence of an association consistent with causation: The evidence suggests a causal association but chance, bias, or confounding cannot be ruled out as explaining the association.

Inadequate epidemiologic evidence to infer a causal association: The available studies do not permit a conclusion regarding the presence or absence of an association.

Epidemiologic evidence consistent with no causal association: Several adequate studies covering the full range of human exposures and considering susceptible populations, and for which alternative explanations such as bias and confounding can be ruled out, are mutually consistent in not finding an association.

5.3 Evaluating evidence in animals

For each effect, the assessment evaluates the evidence from the animal experiments as a whole to determine the extent to which they indicate a potential for effects in humans. Consistent results across various species and strains increase confidence that similar results would occur in humans. Several concepts discussed by Hill (1965) are pertinent to the weight of experimental results: consistency of response, dose-response relationships, strength of response, biologic plausibility, and coherence (U.S. EPA, 1994, 2002, 2005a).

In weighing evidence from multiple experiments, U.S. EPA (2005a) distinguishes

Conflicting evidence (that is, mixed positive and negative results in the same sex and strain using a similar study protocol) from ***Differing results*** (that is, positive results and negative results are in different sexes or strains or use different study protocols).

Negative or null results do not invalidate positive results in a different experimental system. EPA regards all as valid observations and looks to explain differing results using mechanistic information (for example, physiologic or metabolic differences across test systems) or methodological differences (for example, relative sensitivity of the tests, differences in dose levels, insufficient sample size, or timing of dosing or data collection).

It is well established that there are critical periods for some developmental and reproductive effects. Accordingly, the assessment determines whether critical periods have been adequately investigated (U.S. EPA, 1991, 1996, 1998, 2005a, 2005b, 2006b). Similarly, the assessment determines whether the database is adequate to evaluate other critical sites and effects.

In evaluating evidence of genetic toxicity:

- Demonstration of gene mutations, chromosome aberrations, or aneuploidy in humans or experimental mammals (*in vivo*) provides the strongest evidence.
- This is followed by positive results in lower organisms or in cultured cells (*in vitro*) or for other genetic events.
- Negative results carry less weight, partly because they cannot exclude the possibility of effects in other tissues (IARC, 2006).

For germ-cell mutagenicity, EPA has defined categories of evidence, ranging from positive results of human germ-cell mutagenicity to negative results for all effects of concern (U.S. EPA, 1986b).

5.4 Evaluating mechanistic data to identify adverse outcome pathways and modes of action

Mechanistic data can be useful in answering several questions.

- The biologic plausibility of a causal interpretation of human studies.
- The generalizability of animal studies to humans.
- The susceptibility of particular populations or lifestages.

The focus of the analysis is to describe, if possible, *adverse outcome pathways* that lead to a health effect. An adverse outcome pathway encompasses:

- *Toxicokinetic processes* of absorption, distribution, metabolism, and elimination that lead to the formation of an active agent and its presence at the site of initial biologic interaction.
- *Toxicodynamic processes* that lead to a health effect at this or another site (also known as a *mode of action*).

For each effect, the assessment discusses the available information on its *modes of action* and associated *key events* (*key events* being empirically observable, necessary precursor steps or biologic markers of such steps; *mode of action* being a series of key events involving interaction with cells, operational and anatomic changes, and resulting in disease). Pertinent information may also come from studies of metabolites or of compounds that are structurally similar or that act through similar mechanisms. Information on mode of action is not required for a conclusion that the agent is causally related to an effect (U.S. EPA, 2005a).

The assessment addresses several questions about each hypothesized mode of action (U.S. EPA, 2005a).

(1) **Is the hypothesized mode of action sufficiently supported in test animals?** Strong support for a key event being necessary to a mode of action can come from experimental challenge to the hypothesized mode of action, in which studies that

suppress a key event observe suppression of the effect. Support for a mode of action is meaningfully strengthened by consistent results in different experimental models, much more so than by replicate experiments in the same model. The assessment may consider various aspects of causation in addressing this question.

(2) **Is the hypothesized mode of action relevant to humans?** The assessment reviews the key events to identify critical similarities and differences between the test animals and humans. Site concordance is not assumed between animals and humans, though it may hold for certain effects or modes of action. Information suggesting quantitative differences in doses where effects would occur in animals or humans is considered in the dose-response analysis. Current levels of human exposure are not used to rule out human relevance, as IRIS assessments may be used in evaluating new or unforeseen circumstances that may entail higher exposures.

(3) **Which populations or lifestages can be particularly susceptible to the hypothesized mode of action?** The assessment reviews the key events to identify populations and lifestages that might be susceptible to their occurrence. Quantitative differences may result in separate toxicity values for susceptible populations or lifestages.

The assessment discusses the likelihood that an agent operates through multiple modes of action. An uneven level of support for different modes of action can reflect disproportionate resources spent investigating them (U.S. EPA, 2005a). It should be noted that in clinical reviews, the credibility of a series of studies is reduced if evidence is limited to studies funded by one interested sector (Guyatt et al., 2008b).

For cancer, the assessment evaluates evidence of a mutagenic mode of action to guide extrapolation to lower doses and consideration of susceptible lifestages. Key data include the ability of the agent or a metabolite to react with or bind to DNA, positive results in multiple test systems, or similar properties and structure-activity relationships to mutagenic carcinogens (U.S. EPA, 2005a),

5.5 Characterizing the overall weight of the evidence

After evaluating the human, animal, and mechanistic evidence pertinent to an effect, the assessment answers the question: Does the agent cause the adverse effect? (NRC, 1983, 2009). In doing this, the assessment develops a narrative that integrates the evidence pertinent to causation. To provide clarity and consistency, the narrative includes a standard hazard descriptor. For example, the following standard descriptors combine epidemiologic, experimental, and mechanistic evidence of carcinogenicity (U.S. EPA, 2005a).

Carcinogenic to humans: There is convincing epidemiologic evidence of a causal association (that is, there is reasonable confidence that the association cannot be fully explained by chance, bias, or confounding); or there is strong human evidence of cancer or its precursors, extensive animal evidence, identification of key precursor events in animals, and strong evidence that they are anticipated to occur in humans.

Likely to be carcinogenic to humans: The evidence demonstrates a potential hazard to humans but does not meet the criteria for *carcinogenic*. There may be a plausible association in humans, multiple positive results in animals, or a combination of human, animal, or other experimental evidence.

Suggestive evidence of carcinogenic potential: The evidence raises concern for effects in humans but is not sufficient for a stronger conclusion. This descriptor covers a range of evidence, from a positive result in the only available study to a single positive result in an extensive database that includes negative results in other species.

Inadequate information to assess carcinogenic potential: No other descriptors apply. *Conflicting evidence* can be classified as *inadequate information* if all positive results

are opposed by negative studies of equal quality in the same sex and strain. *Differing results*, however, can be classified as *suggestive evidence* or as *likely to be carcinogenic*.

Not likely to be carcinogenic to humans: There is robust evidence for concluding that there is no basis for concern. There may be no effects in both sexes of at least two appropriate animal species; positive animal results and strong, consistent evidence that each mode of action in animals does not operate in humans; or convincing evidence that effects are not likely by a particular exposure route or below a defined dose.

Multiple descriptors may be used if there is evidence that carcinogenic effects differ by dose range or exposure route (U.S. EPA, 2005a).

Another example of standard descriptors comes from EPA's Integrated Science Assessments, which evaluate causation for the effects of the criteria pollutants in ambient air (U.S. EPA, 2010).

Causal relationship: Sufficient evidence to conclude that there is a causal relationship. Observational studies cannot be explained by plausible alternatives, or they are supported by other lines of evidence, for example, animal studies or mechanistic information.

Likely to be a causal relationship: Sufficient evidence that a causal relationship is likely, but important uncertainties remain. For example, observational studies show an association but co-exposures are difficult to address or other lines of evidence are limited or inconsistent; or multiple animal studies from different laboratories demonstrate effects and there are limited or no human data.

Suggestive of a causal relationship: At least one high-quality epidemiologic study shows an association but other studies are inconsistent.

Inadequate to infer a causal relationship: The studies do not permit a conclusion regarding the presence or absence of an association.

Not likely to be a causal relationship: Several adequate studies, covering the full range of human exposure and considering susceptible populations, are mutually consistent in not showing an effect at any level of exposure.

EPA is investigating and may on a trial basis use these or other standard descriptors to characterize the overall weight of the evidence for effects other than cancer.

6. Selecting studies for derivation of toxicity values

For each effect where there is credible evidence of an association with the agent, the assessment derives toxicity values if there are suitable epidemiologic or experimental data. The decision to derive toxicity values may be linked to the hazard descriptor.

Dose-response analysis requires quantitative measures of dose and response. Then, other factors being equal (U.S. EPA, 1994, 2005a):

- Epidemiologic studies are preferred over animal studies, if quantitative measures of exposure are available and effects can be attributed to the agent.
- Among experimental animal models, those that respond most like humans are preferred, if the comparability of response can be determined.
- Studies by a route of human environmental exposure are preferred, although a validated toxicokinetic model can be used to extrapolate across exposure routes.
- Studies of longer exposure duration and follow-up are preferred, to minimize uncertainty about whether effects are representative of lifetime exposure.
- Studies with multiple exposure levels are preferred for their ability to provide information about the shape of the exposure-response curve.
- Studies with adequate power to detect effects at lower exposure levels are preferred, to minimize the extent of extrapolation to levels found in the environment.

Studies with non-monotonic exposure-response relationships are not necessarily excluded from the analysis. A diminished effect at higher

exposure levels may be satisfactorily explained by factors such as competing toxicity, saturation of absorption or metabolism, exposure misclassification, or selection bias.

If a large number of studies are suitable for dose-response analysis, the assessment considers the study characteristics in this section to focus on the most informative data. The assessment explains the reasons for not analyzing other groups of studies. As a check on the selection of studies for dose-response analysis, EPA asks peer reviewers to identify studies that were not adequately considered.

7. Deriving toxicity values

7.1 General framework for dose-response analysis

EPA uses a two-step approach that distinguishes analysis of the observed dose-response data from inferences about lower doses (U.S. EPA, 2005a).

Within the observed range, the preferred approach is to use modeling to incorporate a wide range of data into the analysis. The modeling yields a *point of departure* (an exposure level near the lower end of the observed range, without significant extrapolation to lower doses) (sections 7.2-7.3).

Extrapolation to lower doses considers what is known about the modes of action for each effect (sections 7.4-7.5). When response estimates at lower doses are not required, an alternative is to derive *reference values*, which are calculated by applying factors to the point of departure in order to account for sources of uncertainty and variability (section 7.6).

For a group of agents that induce an effect through a common mode of action, the dose-response analysis may derive a *relative potency factor* for each agent. A full dose-response analysis is conducted for one well-studied *index chemical* in the group, then the potencies of other members are expressed in relative terms based on relative toxic effects, relative absorption or metabolic rates, quantitative structure-activity relationships, or receptor binding characteristics (U.S. EPA, 2000, 2005a).

Increasingly, EPA is basing toxicity values on combined analyses of multiple data sets or multiple responses. EPA also considers multiple dose-response approaches when they can be supported by robust data.

7.2 Modeling dose to sites of biologic effects

The preferred approach for analysis of dose is toxicokinetic modeling because of its ability to incorporate a wide range of data. The preferred dose metric would refer to the active agent at the site of its biologic effect or to a close, reliable surrogate measure. The active agent may be the administered chemical or a metabolite. Confidence in the use of a toxicokinetic model depends on the robustness of its validation process and on the results of sensitivity analyses (U.S. EPA, 1994, 2005a, 2006a).

Because toxicokinetic modeling can require many parameters and more data than are typically 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 are standardized to a daily average over the duration of exposure. For chronic effects, daily exposures are averaged over the lifespan. Exposures during a critical period, however, are not averaged over a longer duration (U.S. EPA, 1991, 1996, 1998, 2005a).
- Doses are standardized to equivalent human terms to facilitate comparison of results from different species.
  - Oral doses are scaled allometrically using $mg/kg^{3/4}$-d 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, 2005a, 2011).
  - 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, 1994, 2012b).

It can be informative to convert doses across exposure routes. If this is done, the assessment describes the underlying data, algorithms, and assumptions (U.S. EPA, 2005a).

In the absence of study-specific data on, for example, intake rates or body weight, EPA has developed recommended values for use in dose-response analysis (U.S. EPA, 1988).

7.3 Modeling response in the range of observation

Toxicodynamic ("biologically based") modeling can incorporate data on biologic processes leading to an effect. Such models require sufficient data to ascertain a mode of action and to quantitatively support model parameters associated with its key events. Because different models may provide equivalent fits to the observed data but diverge substantially at lower doses, 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, 2005a).

Because toxicodynamic modeling can require many parameters and more knowledge and data than are typically available, EPA has developed a standard set of empirical ("curve-fitting") models (http://www.epa.gov/ncea/bmds/) that can be applied to typical data sets, including those that are nonlinear. EPA has also developed guidance on modeling dose-response data, assessing model fit, selecting suitable models, and reporting modeling results (U.S. EPA, 2012a). Additional judgment or alternative analyses are used when the procedure fails to yield reliable results, for example, if the fit is poor, modeling may be restricted to the lower doses, especially if there is competing toxicity at higher doses (U.S. EPA, 2005a).

Modeling is used to derive a point of departure (U.S. EPA, 2005a, 2012a). (See section 7.6 for alternatives if a point of departure cannot be derived by modeling.)

- When linear extrapolation is used, selection of a response level corresponding to the point of departure is not highly influential, so standard values near the low end of the observable range are generally used (for example, 10% extra risk for cancer bioassay data, 1% for epidemiologic data, lower for rare cancers).
- For nonlinear approaches, both statistical and biologic considerations are taken into account.
  - For dichotomous data, a response level of 10% extra risk is generally used for minimally adverse effects, 5% or lower for more severe effects.
  - For continuous data, a response level is ideally based on an established definition of biologic significance. In the absence of such definition, one control standard deviation from the control mean is often used for minimally adverse effects, one-half standard deviation for more severe effects.

The point of departure is the 95% lower bound on the dose associated with the selected response level.

7.4 Extrapolating to lower doses and response levels

The purpose of extrapolating to lower doses is to estimate responses at exposures below the observed data. Low-dose extrapolation is typically used for cancer data. Low-dose extrapolation considers what is known about modes of action (U.S. EPA, 2005a).

(1) If a biologically based model has been developed and validated for the agent, extrapolation may use the fitted model below the observed range if significant model uncertainty can be ruled out with reasonable confidence.
(2) Linear extrapolation is used if the dose-response curve is expected to have a linear

component below the point of departure. This includes:

- Agents or their metabolites that are DNA-reactive and have direct mutagenic activity.
- Agents or their metabolites for which human exposures or body burdens are near doses associated with key events leading to an effect.

Linear extrapolation is also used if there is an absence of sufficient information on modes of action.

The result of linear extrapolation is described by an *oral slope factor* or an *inhalation unit risk*, which is the slope of the dose-response curve at lower doses or concentrations, respectively.

(3) Nonlinear models are used for extrapolation if there are sufficient data to ascertain the mode of action and to conclude that it is not linear at lower doses, and the agent does not demonstrate mutagenic or other activity consistent with linearity at lower doses. If nonlinear extrapolation is appropriate but no model is developed, an alternative is to calculate reference values.

If linear extrapolation is used, 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. The assessment then derives or selects an overall slope factor and an overall unit risk for the agent, considering the various dose-response analyses, the study preferences discussed in section 6, and the possibility of basing a more robust result on multiple data sets.

7.5 Considering susceptible populations and lifestages

The assessment analyzes the available information on populations and lifestages that may be particularly susceptible to each effect. A tiered approach is used (U.S. EPA, 2005a).

(1) If an epidemiologic or experimental study reports quantitative results for a susceptible population or lifestage, these data are analyzed to derive separate toxicity values for susceptible individuals.

(2) If data on risk-related parameters allow comparison of the general population and susceptible individuals, these data are used to adjust the general-population toxicity values for application to susceptible individuals.

(3) In the absence of chemical-specific data, EPA has developed *age-dependent adjustment factors* for early-life exposure to potential carcinogens that have a mutagenic mode of action. There is evidence of early-life susceptibility to various carcinogenic agents, but most epidemiologic studies and cancer bioassays do not include early-life exposure. To address the potential for early-life susceptibility, EPA recommends (U.S. EPA, 2005b):

- 10-fold adjustment for exposures before age 2 years.
- 3-fold adjustment for exposures between ages 2 and 16 years.

7.6 Reference values and uncertainty factors

An *oral reference dose* or an *inhalation reference concentration* is an estimate of an exposure (including in susceptible subgroups) that is likely to be without an appreciable risk of adverse health effects over a lifetime (U.S. EPA, 2002). Reference values are typically calculated for effects other than cancer and for suspected carcinogens if a well characterized mode of action indicates that a necessary key event does not occur below a specific dose. Reference values provide no information about risks at higher exposure levels.

The assessment characterizes effects that form the basis for reference values as adverse, considered to be adverse, or a precursor to an adverse effect. For developmental toxicity, reproductive toxicity, and neurotoxicity there is guidance on adverse effects and their biologic markers (U.S. EPA, 1991, 1996, 1998).

To account for uncertainty and variability in the derivation of a lifetime human exposure where adverse effects are not anticipated to occur,

reference values are calculated by applying a series of *uncertainty factors* to the point of departure. If a point of departure cannot be derived by modeling, a no-observed-adverse-effect level or a lowest-observed-adverse-effect level is used instead. The assessment discusses scientific considerations involving several areas of variability or uncertainty.

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. A factor of 10 is generally used to account for this variation. This factor is reduced only if the point of departure is derived or adjusted specifically for susceptible individuals (not for a general population that includes both susceptible and non-susceptible individuals) (U.S. EPA, 1991, 1994, 1996, 1998, 2002).

Animal-to-human extrapolation. If animal results are used to make inferences about humans, the assessment adjusts for cross-species differences. These may arise from differences in toxicokinetics or toxicodynamics. Accordingly, if the point of departure is standardized to equivalent human terms or is based on toxicokinetic or dosimetry modeling, a factor of $10^{1/2}$ (rounded to 3) is applied to account for the remaining uncertainty involving toxicodynamic differences. If a biologically based model adjusts fully for toxicokinetic and toxicodynamic differences across species, this factor is not used. In most other cases, a factor of 10 is applied (U.S. EPA, 1991, 1994, 1996, 1998, 2002, 2011).

Adverse-effect level to no-observed-adverse-effect level. If a point of departure is based on a lowest-observed-adverse-effect level, the assessment must infer a dose where such effects are not expected. This can be a matter of great uncertainty, especially if there is no evidence available at lower doses. A factor of 10 is applied to account for the uncertainty in making this inference. 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, 1991, 1994, 1996, 1998, 2002).

Subchronic-to-chronic exposure. If a point of departure is based on subchronic studies, the assessment considers whether lifetime exposure could have effects at lower levels of exposure. A factor of 10 is applied to account for the uncertainty in using subchronic studies to make inferences about lifetime exposure. This factor may also be applied for developmental or reproductive effects if exposure covered less than the full critical period. A factor other than 10 may be used, depending on the duration of the studies and the nature of the response (U.S. EPA, 1994, 1998, 2002).

Incomplete database. If an incomplete database raises concern that further studies might identify a more sensitive effect, organ system, or lifestage, the assessment may apply a database uncertainty factor (U.S. EPA, 1991, 1994, 1996, 1998, 2002). The size of the factor depends on the nature of the database deficiency. For example, EPA typically follows the suggestion that a factor of 10 be applied if both a prenatal toxicity study and a two-generation reproduction study are missing and a factor of $10^{1/2}$ if either is missing (U.S. EPA, 2002).

In this way, the assessment derives candidate values for each suitable data set and effect that is credibly associated with the agent. These results are arrayed, using common dose metrics, to show where effects occur across a range of exposures (U.S. EPA, 1994).

The assessment derives or selects an *organ- or system-specific reference value* for each organ or system affected by the agent. The assessment explains the rationale for each organ/system-specific reference value (based on, for example, the highest quality studies, the most sensitive outcome, or a clustering of values). By providing these organ/system-specific reference 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 (U.S. EPA, 2002; NRC, 2009).

The assessment then selects an overall reference dose and an overall reference concentration for the agent to represent lifetime human exposure levels where effects are not anticipated to occur. This is generally the most sensitive organ/system-specific reference value, though consideration of study quality and confidence in each value may lead to a different selection.

7.7 Confidence and uncertainty in the reference values

The assessment selects a standard descriptor to characterize the level of confidence in each reference value, based on the likelihood that the value would change with further testing. Confidence in reference values is based on quality of the studies used and completeness of the database, with more weight given to the latter. The level of confidence is increased for reference values based on human data supported by animal data (U.S. EPA, 1994).

High confidence: The reference value is not likely to change with further testing, except for mechanistic studies that might affect the interpretation of prior test results.

Medium confidence: This is a matter of judgment, between high and low confidence.

Low confidence: The reference value is especially vulnerable to change with further testing.

These criteria are consistent with guidelines for systematic reviews that evaluate the quality of evidence. These also focus on whether further research would be likely to change confidence in the estimate of effect (Guyatt et al., 2008a).

All assessments discuss the significant uncertainties encountered in the analysis. EPA provides guidance on characterization of uncertainty (U.S. EPA, 2005a). For example, the discussion distinguishes model uncertainty (lack of knowledge about the most appropriate experimental or analytic model) and parameter uncertainty (lack of knowledge about the parameters of a model). Assessments also discuss human variation (interpersonal differences in biologic susceptibility or in exposures that modify the effects of the agent).

References

DHEW (1964) *Smoking and Health: Report of the Advisory Committee to the Surgeon General of the Public Health Service*. Public Health Service Pub. No. 1103. http://profiles.nlm.nih.gov/ps/access/NNBBMQ.pdf.

DHHS (2004) *The Health Consequences of Smoking: A Report of the Surgeon General*. http://www.cdc.gov/tobacco/data_statistics/sgr/2004/index.htm.

Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, Schünemann HJ (2008a) GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. British Medical Journal 336: 924-926, http://www.bmj.com/content/336/7650/924.full.

Guyatt GH, Oxman AD, Kunz R, Vist GE, Falck-Ytter Y, Schünemann HJ (2008b) GRADE: what is "quality of evidence" and why is it important to clinicians? British Medical Journal 336: 995-998, http://www.bmj.com/content/336/7651/995.full.

Hill AB (1965) The environment and disease: association or causation? Proceedings of the Royal Society of Medicine 58: 295-300. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1898525/.

IARC (2006) Preamble to the IARC Monographs. http://monographs.iarc.fr/.

IOM (2008) *Improving the Presumptive Disability Decision-Making Process for Veterans*. Washington: National Academy Press. http://www.nap.edu/catalog.php?record_id=11908.

NRC (1983) *Risk Assessment in the Federal Government: Managing the Process*. Washington: National Academy Press. http://www.nap.edu/catalog.php?record_id=366.

NRC (2009) *Science and Decisions: Advancing Risk Assessment*. Washington: National Academy Press. http://www.nap.edu/catalog.php?record_id=12209.

Rothman KJ, Greenland S (1998) Modern Epidemiology. Philadelphia: Lippincott Williams and Wilkins.

U.S. EPA (1986a) Guidelines for the Health Risk Assessment of Chemical Mixtures. EPA/630/R-98/002.

U.S. EPA (1986b) Guidelines for Mutagenicity Risk Assessment. EPA/630/R-98/003.

U.S. EPA (1988) Recommendations for and Documentation of Biological Values for Use in Risk Assessment. EPA/600/6-87/008.

U.S. EPA (1991) Guidelines for Developmental Toxicity Risk Assessment. EPA/600/FR-91/001.

U.S. EPA (1994) Methods for Derivation of Inhalation Reference Concentrations and Application of Inhalation Dosimetry. EPA/600/8-90/066F.

U.S. EPA (1996) Guidelines for Reproductive Toxicity Risk Assessment. EPA/630/R-96/009.

U.S. EPA (1998) Guidelines for Neurotoxicity Risk Assessment. EPA/630/R-95/001F.

U.S. EPA (2000) Supplementary Guidance for Conducting Health Risk Assessment of Chemical Mixtures. EPA/630/R-00/002.

U.S. EPA (2002) A Review of the Reference Dose and Reference Concentration Processes. EPA/630/P-02/002F.

U.S. EPA (2005a) Guidelines for Carcinogen Risk Assessment. EPA/630/P-03/001F.

U.S. EPA (2005b) Supplemental Guidance for Assessing Susceptibility from Early-Life Exposure to Carcinogens. EPA/630/R-03/003F.

U.S. EPA (2006a) Approaches for the Application of Physiologically Based Pharmacokinetic (PBPK) Models and Supporting Data in Risk Assessment. EPA/600/R-05/043F.

U.S. EPA (2006b) A Framework for Assessing Health Risks of Environmental Exposures to Children. EPA/600/R-05/093F.

U.S. EPA (2009) IRIS Process. http://www.epa.gov/iris/process.htm.

U.S. EPA (2010) Integrated Science Assessment for Carbon Monoxide. EPA/600/R-09/019F.

U.S. EPA (2011) Recommended Use of Body Weight$^{3/4}$ as the Default Method in Derivation of the Oral Reference Dose. EPA/100/R11/0001.

U.S. EPA (2012a) Benchmark Dose Technical Guidance. EPA/100/R-12/001.

U.S. EPA (2012b) Advances in Inhalation Gas Dosimetry for Derivation of a Reference Concentration (RfC) and Use in Risk Assessment. EPA/600/R-12/044.

U.S. EPA (2013) draft Handbook for IRIS Assessment Development. http://www.epa.gov/iris/xxx.

January 2013

Appendix C – Example of IRIS Program Direction to Contractors

CONDUCTING DOSE-RESPONSE MODELING

NOTE: This section addresses only dose-response modeling of animal bioassay data from standard experimental designs. Analysis of animal data from complex experimental designs (for example, repeated measurements on the same test subjects) and analysis of epidemiological studies require specialized methods that are documented on a case by case basis.

The IRIS Program's Statistics Workgroup (SWG) has provided instructions for conducting the majority of dose-response modeling of animal bioassay data using BMDS software. This was written to provide unambiguous instructions to contractors regarding what the IRIS Program requires as an initial analysis before reviewing results. Analysis is often iterative, because review of results may suggest corrections or additional analyses, or prompt a second QA review of data. This draft cannot provide completely detailed advice; modelers should consult a statistician familiar with the relevant methods for detailed advice and for a review of work.

This section summarizes the process of assembling data for dose-response modeling from animal bioassay toxicology studies and then conducting and reporting those analyses. It is intended for use by the IRIS Program. It can also be included with a Statement of Work (Performance Work Statement) to provide detailed directions to a contractor and can also serve as a review guide. . The objective is to reduce errors, streamline operations, and ensure that analyses and model selection are done consistently for each assessment.

The section is intended to describe a common core of best practices, but is not necessarily exhaustive and may be modified in the future. In addition, Chemical Managers and data analysts within EPA may modify these practices to suit the particular needs of an assessment or the features of a particular data set. This section assumes that the user has a good working knowledge of EPA's Benchmark Dose Software (BMDS).

The selection and justification of a final point of departure (POD) is the responsibility of EPA and not to be delegated to a contractor, although contractors may be asked to make recommendations on a wide range of topics and to draft related justifications. This section is intended to be applied in conjunction with detailed guidance, including EPA's *Benchmark Dose Technical Guidance* (EPA/100/R-12/001, June 2012).

Any potential problems or questions that the analyst or modeler identifies should be brought to the attention of the chemical manager as soon as possible. For example, there may be questions about which data sets should be used for dose-response modeling (i.e., suitability, study quality, biological significance, and significance of the response to dose), problems with data quality or missing data, or difficulties encountered while modeling the data. Decisions to include or exclude studies for DRA are significant and should be reviewed by the chemical manager or by EPA staff designated by the chemical manager. The IRIS Program's Statistics Workgroup (SWG) can provide advice on implementing this section and on complicated or non-routine statistical analysis issues.

A. Preparing Data for Dose-Response Modeling

The steps in this section, especially steps 1 and 2, will be most efficiently accomplished during preparation of data for hazard evaluation. When a number of studies need to be compared, use of common units and allowance for different dosing regimens as described below will facilitate the exercise.

1. **For continuous response data, verify quantities described as either standard deviations or standard errors.** BMDS requires standard deviation (S), which characterizes variability in responses among individual animals. Some investigators report the standard error of the group mean, which is $S/\sqrt{n}$. Some studies may report confidence limits. To calculate S from these data, you will need to know the confidence level and method used. Have a statistician review the report and your calculations. Document your assumptions and calculations.[6]

2. **Convert doses to standard units.**

   a. Standard units for oral and inhalation exposures are mg/kg-day and mg/m^3, respectively. If the original study does not provide these data, the data analyst will need to apply the necessary assumptions (especially those regarding body weight and food or water consumption) and must document the necessary calculations.

Best Practices for Data Management to Support Modeling

All data should be proofed against the original cited source before being used in any evaluations, comparisons, or analyses. This should be done using double inspection after data entry or by double entry followed by machine comparison.

Permanently document all calculations and conversions, using a database or spreadsheet to record the individual terms, used to do calculations and to produce the adjusted dose or concentration.

   b. For oral exposures in food or water, exposure is treated as if continuous over 24 hours, 7 days per week. The source publication will usually report the oral intake of the chemical in mg/kg-day. If not, it will be necessary to calculate this from study-specific data on intake and body weights or to make inferences specific to the species, strain, and sex (e.g., U.S. EPA, 1988). For oral exposures by gavage, the entire dose is assumed to be distributed across 24 hours (in effect, assuming 24-hour dosing), so the most common adjustment is for days/week of dosing (typically 5 days per week).

   c. One may need to obtain animal body weights applicable to the dose-response data. Use data from the study, if available. If body-weight data are presented only in figures, numerical data can be recovered by digitizing the figures. Average body weight should be calculated across the period of exposure, for the purpose of calculating exposure per unit body weight.

[6] BMDS provides a "Transformation" option to convert SE to SD.

If there are no study-specific data, there are sources of default body weights for adults of each sex for various strains of rodents (e.g., U.S. EPA, 1988).

d. For the purpose of animal to human extrapolation of oral exposures, the IRIS Program uses a standard weight for each species, regardless of the availability of study-specific information, as recommended by Agency guidance (U.S. EPA, 2011).

e. For inhalation exposures, RfCs are typically expressed in mg/m^3 by multiplying the POD in ppm by [(Molecular Weight)/24.45]. This conversion can be made before or after dose-response modeling. RfCs may be extrapolated to humans using RfC methodology (U.S. EPA, 1994). This conversion can also be made before dose-response modeling, or after if there were no substantive differences in breathing rate with exposure level.

f. The foregoing may not apply if you are using internal dose metrics from a PBPK model. Instructions for such analyses are beyond the scope of this draft.

3. Make standard dose adjustments to account for intermittent and less-than-lifetime exposures. Report the individual terms and final multiplier employed in the adjustments, in the data summary table in the modeling appendix of the assessment.

a. This section may not apply if you are using internal dose metrics from a PBPK model. The PBPK dose metrics may partially or fully account for intermittent exposure. The dose-response analyst must communicate with the PBPK analyst regarding adjustments to dose metrics.

b. <u>Cancer bioassays:</u>

*dose * (hours/day)/24 * (days/week)/7 * (weeks exposed / weeks on study)*

If sacrifice time is less than the standard 'lifetime' (104 weeks for rats and mice), also multiply by $[(weeks\ on\ study\ before\ sacrifice\ /\ 104)^3]$ (see, e.g., Portier et al., 1986).

Example: 8-week-old mice were exposed by inhalation for 6 hours per day, 5 days per week, for 78 weeks, and were sacrificed at 91 weeks after starting exposure; the adjusted exposure concentration is (nominal exposure concentration) (6/24) * (5/7) * (78/91) * $[(91\ wk / 104\ wk)^3]$.

c. <u>Noncancer endpoints (in cancer bioassays or other studies):</u>

*dose * (hours/day)/24 * (days/week)/7 * (weeks exposed / weeks on study)*

If exposure lasted until the final sacrifice, the last term has no effect on the conversion. Unlike with cancer bioassays, exposures are not extrapolated to longer durations without chemical-specific data.

d. <u>Time-varying exposures</u>: These dosing regimens require calculation of a time-weighted average dose or concentration, before applying the factors above.

Example: a study applied one dose, D1, for weeks 1-12 and another, slightly different dose, D2, for weeks 13-78 (and nothing thereafter to week 91). The time-weighted average dose is (12 weeks/91 weeks)*D1 + (66 weeks/91 weeks)*D2 + (13 weeks/91 weeks)*0).

4. If survival rates differ substantially among dose groups for <u>cancer</u> bioassays

a. **If data are available for individual animals on time of death:** Assemble data on individual times of death and tumor incidence for use in time-to-tumor modeling (described below).

b. **If only grouped data are available:** Estimate the number of animals at risk in each dose group using the number alive at 52 weeks (if the first tumor was observed later than 52 weeks) or the number alive at the week when the first tumor was observed. Note the use of an adjusted number at risk when reporting the DRA. This adjustment is not as effective as using individual animal data for survival adjustment, so some bias in estimates is to be expected. The results of DRA must be qualified by noting the possible inaccuracy (bias) caused by incomplete survival adjustment.

c. If no survival adjustment is possible, results of DRA must be qualified by noting the possible inaccuracy (bias) caused by lack of a survival adjustment.

5. If survival rates differ substantially among dose groups for <u>noncancer</u> effects: The foregoing methods have not been used in IRIS assessments for noncancer effects, but they could be. If there was severe early mortality (either different or similar among dose groups), this should be called to the attention of the chemical manager (and noted in the assessment, if the data are used).

6. Sort/organize endpoints by type of health effect (i.e., different target systems). Seek the Chemical Manager's advice on this. An example list:

- kidney
- liver
- other organs
- body weight
- neurological
- immunological (includes thymus, adrenal)
- respiratory tract
- reproductive
- developmental (when summarized by dam; excluding nested data)

Physiologically-Based Pharmacokinetic (PBPK) models

PBPK models relate external exposures, also referred to as applied exposure, to internal measures of concentration or exposure. Preferably a PBPK model will describe the concentration of the active toxicant (often a metabolite of the material whose risk is being assessed) in the target tissue. PBPK models may also describe early (precursor) molecular interactions, such as binding to a receptor, inhibition of an enzyme, or formation of DNA adducts or cross-links. Even if a PBPK model only describes the blood concentration of a toxicant, or rate of metabolic formation of the toxicant (in the liver), these measures of exposure or dose are closer to and hence presumed to be

more predictive of a toxicant response than the applied dose. There are a number of potential advantages to using a PBPK model:

- Since the internal dose metric is mechanistically closer to the toxic response, subsequent dose-response modeling (BMD modeling in particular) should more accurately interpolate among the dose-response data and better characterize uncertainty in that relationship.
- If a PBPK model is calibrated for a route of exposure for which toxicity data are not available (e.g., toxicity data are only available for oral exposure but the PBPK model is calibrated for both oral and inhalation exposure), then the model can also be used for route-to-route extrapolation.
- Given PBPK models which are calibrated for both a test animal species and humans should allow for a more accurate prediction of the human equivalent (exposure) concentration (HEC) or dose (HED), since they use chemical- and species-specific data to match the exposure associated with a toxicological point-of-departure in the animal with the corresponding exposure in humans.
- Also, in conjunction with Monte Carlo sampling, if the distributions of PBPK parameters are defined for a human population, then the model can be used to obtain a data-derived uncertainty factor for human variability in pharmacokinetics (UF_H,PK).

However, it must be first noted that these benefits are dependent first on the PBPK model's ability to predict a metric which is in fact mechanistically closer to the effect of concern. If toxicity is caused by a metabolite for which the model has not been calibrated, or occurs in a portal-of-entry tissue for which PK data are lacking, then other internal metrics that the model does predict may not be more predictive of toxicity than applied dose. And because PBPK models are used to predict internal doses under exposure conditions (e.g., over chronic periods) and at concentrations for which direct data may not be available, there is higher uncertainty in the prediction than in directly measured or known exposure data. To assure that these issues of applicability and certainty are adequately addressed, PBPK models should be subject to careful scientific and quality reviews before they are used. It is not possible to be as specific and proscriptive about when and how a PBPK model should be used as is the case for the statistical dose-response modeling, because PBPK modeling has not reached the level of maturity that exists for statistics, but also because the models are much more compound-specific. Each model tends to contain chemical-specific aspects, and each PK data set presents unique challenges for modeling. Because PBPK modeling therefore involves a significant use of scientific judgment, it is more important for models to be reviewed by a team of experts, a primary role of the Pharmacokinetics Work Group (PKWG). Having multiple modelers consider a given model (or set of models) provides input from multiple experts with a variety of backgrounds and perspectives.

The following items should be considered in reviewing a PBPK model and considering how to apply it.

(i) Both the science and the model code must be evaluated. The science -- model structure, equation forms, and hypotheses that these represent -- will be described or at least implied in the corresponding publication or report. But a QA of the model code is also essential, to

assure that it accurately represents the science as described, correctly converts units, and uses the correct parameters (again as listed in a paper or report). If a model cannot describe all the data with a consistent set of parameters, or ones which vary in a predictable or measurable way (e.g., respiration rate can be varied to match measured values), then that calls into question its ability to predict kinetics for exposure situations (bioassays) for which calibration data are not available.

(ii) The degree of mechanistic complexity should be evaluated against the available data. This is a matter of professional judgment, specific rules can't be laid out, but a more complex model may represent hypotheses that have not been adequately tested and hence is not necessarily considered the best for use in the assessment. Likewise a metric which is closer to a toxic endpoint (e.g., binding to a receptor in the brain) may have a higher degree of uncertainty due to limited calibration/validation data than one which is intermediately close (e.g., blood concentration).

(iii) Use of PBPK models for animal test species to replace applied dose with internal dose metrics can improve the quality of the statistical modeling if it takes out (really explains) some of the exposure-response nonlinearity. Therefore this approach is suggested, though it is not necessary. Using PBPK-predicted internal doses in the dose-response modeling makes it harder to update an assessment if the PBPK model is changed or there is an update in the analyzed dose-response data, so conducting the BMD modeling using applied doses is an acceptable alternative. Using PBPK-derived internal doses should be most considered, though, when the statistical BMD models have trouble fitting dose-response data.

(iv) When toxicity data are available for multiple exposure routes (i.e., inhalation, oral, or dermal), ideally a PBPK model can explain apparent route sensitivity differences. Specifically, the dose-response may appear discrepant when water ingestion/inhalation rates are used, but become aligned when an internal metric is used. This is mentioned in the current draft as a possible means of combining data sets and it is suggested that this possibility be evaluated where possible. However the metric will depend on how well the model captures any portal-of-entry/first-pass effects. Predictions for oral exposure of some compounds have been found to depend strongly on the assumed drinking water pattern. Oral ingestion is typically not continuous, and a bolus exposure can saturate metabolic processes when a continuous exposure to the same total dose would not. So some thought should be put into providing realistic oral ingestion patterns. The PKWG can provide guidance on this.

(v) When there is a background or endogenous level for a material being considered, some adjustment should be made to account for the fact that animals or people with no exogenous exposure will still have some internal dose. A simple approach that may be considered is to simply subtract the background levels from PK data and then calibrate a PBPK model by fitting the resulting background-subtracted data. However this approach implicitly assumes that the background is additive and that it is constant. Since PBPK models are particularly useful when the PK are nonlinear, the first assumption in particular may be inconsistent with the model and background subtraction can also distort the apparent linearity or non-linearity in dose-response data. Incorporating a background term

into a PBPK model may make the model somewhat more complex, but it will allow for a more accurate and transparent description and analysis of the dose-response relationship.

B. Conducting Dose-Response Modeling

In general, follow the advice on dose-response modeling in EPA's *Benchmark Dose Technical Guidance* (EPA/100/R-12/001, June 2012), aka BMD-TG. The instructions that follow assume data from 'chronic' studies, and rely mainly on use of EPA's Benchmark Dose Software (http://www.epa.gov/ncea/bmds/). After the data and endpoints are selected (previous section), these general principles apply:

- identify important or unusual statistical issues
- model all orders of multistage and polynomial models, up to and including the number of dose groups minus one.
- select a best-fitting model using model selection procedures at BMD-TG §2.3.9
- the decision flow chart at BMD-TG §2.5 is a useful guide
- IRIS assessment modeling usually aims at getting a lower confidence limit (BMDL) for the benchmark dose (BMD). For that end, a profile-likelihood interval or a Bayesian interval is preferred to a Wald interval because the latter is less accurate for BMD. BMDS uses profile-likelihood (BMD-TG 2.3.8). Most commercial statistical software (e.g., SAS) will report Wald intervals. Profile intervals can be obtained using custom programs in SAS, R, and other software.
- The POD may be below the lowest non-zero dose. Confidence in the inference will depend upon the degree of extrapolation, BMD/BMDL ratio, the type of model, the model fit, and what is known about the chemical and endpoint, including the probable MOA.

Advice to consider when initial modeling attempts are unsuccessful or the results are highly uncertain (e.g., poor model fit, large confidence intervals, large differences between models).

- Non-standard approaches (e.g., adding parameter constraints) or additional models (not included in BMDS) might then be considered (consult a statistician).
- Lack of fit might be evident for high doses (especially if the response decreases at the high dose, which can be owed to mortality or to a change in the response pattern or mechanisms at high doses). In some cases, the high dose group(s) may be omitted (BMD-TG §2.3.6) after an unsuccessful attempt at modeling.
- Model fit might be improved by using internal dose measures based on toxicokinetic modeling, if available.
- If adequately-fitting models differ greatly in BMDLs, and an adequately-fitting model has a very large ratio of BMD/BMDL, the data do not permit accurate estimation of a BMD and the data may be judged "not amenable to modeling." These facts should be reported. An alternative is to use LOAEL or NOAEL as a basis for the POD. The response and its confidence interval should always be reported with the LOAEL or NOAEL.
- other advice on improving model fit is given at BMD-TG §2.3.6

General advice concerning hypothesis tests:

- Calculating a LOAEL or NOAEL from bioassay data (in most cases this is available from the source publication): seek advice from a statistician. Various methods are available, depending on the properties of the data. Individual observations will be needed unless the distributional properties of the data are well known from other studies. In general, the method should account for multiple comparisons.
- Trend tests: seek advice from a statistician. The SWG is developing more detailed advice. There are various methods, having different strengths and limitations.

C. Modeling Cancer Endpoints (Single Tumor Sites)

Three approaches may be available: (1) a biologically based dose-response model, (2) time-to-tumor modeling, using survival times for individual animals, (3) modeling cumulative incidence of cancer in dose groups.

A biologically based model could be used if one is available and deemed appropriate. This approach requires specialized knowledge and custom software, and is not discussed further.

If dose groups differed substantially in survival, and if data for individual animals are available, time-to-tumor modeling or survival-adjusted quantal modeling is appropriate (see below).

Incidence of cancers, grouped by dose, may be modeled using BMDS, as described further below. This approach currently accounts for the majority of cancer dose-response modeling.

Modeling cancer incidence:

- If dose groups differ substantially in survival, adjustments to the number of animals at risk (below) will be needed
- Adenomas and carcinomas are combined (i.e., counting the number of animals with either adenomas or carcinomas) when they arise from the same cell type and the adenomas are believed to progress to carcinomas (U.S. EPA, 2005a)[7]
- When low-dose linearity is expected, current IRIS practice is to use the "cancer model" in BMDS (Gehlhaus et al. 2011)[8]. The decision to use this approach must be made by EPA.
- Apply cancer (multistage) models from the highest order model and all lower order models down to first order. E.g., given 4 dose groups, fit models of 3rd, 2nd and 1st order. Selection among these models is usually based upon minimum AIC (Akaike Information Criterion)

[7] U.S. EPA Guidelines for Carcinogen Risk Assessment (2005a), Section **2.2.2.1.2.** ***Statistical considerations***, p.2-19, states: "Statistical analysis of a long-term study should be performed for each tumor type separately. The incidence of benign and malignant lesions of the same cell type, usually within a single tissue or organ, are considered separately but may be combined when scientifically defensible (McConnell et al., 1986)."

[8] The BMDS "cancer model" is a multistage model with non-negative coefficients, and it reports the "cancer slope factor" or potency, i.e., the ratio BMR/BMDL.

- If the "cancer model" does not fit adequately—i.e., if p is not greater than 0.05 (BMD-TG Sections 2.3.5 and 2.3.9)— fit other BMDS models and select the best-fit model, as for noncancer.
- If low-dose non-linearity is expected based upon mode-of-action information (U.S. EPA, 2005a), fit the full suite of BMDS dichotomous models to the relevant precursor effect data and then select the best-fit model as for noncancer. The decision to use this approach must be made by EPA.

D. Combined Cancer Risk (multiple tumor types in a single animal study)

When there is increased risk from multiple tumor types (sites), when the tumor types can be assumed to be approximately independent, and when no single type substantially dominates risk (e.g., by over 10-fold), then composite ("total") cancer risk estimates are derived by combining risk across tumor types (NRC 1994). Evaluating composite cancer risk by modeling data for the total incidence for all cancers (incidence of "tumor-bearing animals") is not preferred (NRC 1994).

BMDS provides a 'multitumor' option for modeling composite risk. This does not refer to multiple tumors in one animal. It refers to multiple types of tumors. For this model, incidence is measured by the number of animals exhibiting a type of tumor, not by counting numbers of a certain tumor in each animal.

The BMDS 'multitumor' option requires the use of a single, common dose metric. Occasionally there is a need to estimate composite risk using different dose metrics for different tumors (e.g., best-fitting cancer models are based on different dose metrics; e.g., one cancer is modeled using a BMDS quantal model and another is modeled using a time to tumor model). For these and other special circumstances, consult the IRIS Program's SWG for alternative methods for estimating composite risk.

E. Time-to-Tumor Analysis and Survival Adjustment

Studies should be reviewed to identify those that may benefit from a survival adjusted analysis. At present, we have no set criterion for the survival difference among dose groups that would trigger such an analysis. However, if the incidence appears to plateau (at less than 100% of animals) or decreases at a high dose, and survival is also lower in higher dose groups, these methods should be used (either to augment or to replace standard BMDS modeling of cumulative incidence).

Two approaches are available for conducting a survival-adjusted analysis, (a) using a time-to-tumor model, and (b) calculating a survival-adjusted number at risk and then using a standard BMDS quantal model. Both require data for each animal on time of death, and both assume that tumors are observed 'incidental' to the cause of death. In most cases, cause of death is not available for most animals. If there is a need to analyze data in which death of each animal can be attributed to the tumor of interest (vs. some other cause), consult a statistician.

1. Time-to-Tumor Modeling

- Use the "MSW" (multistage Weibull) program (at www.epa.gov/ncea/bmds) for time-to-tumor analysis. A guide to using the program is available at the BMDS web site, as is an external review document. Report any failures of the MSW program to solve the BMDL.
- Evaluate estimates for all model orders between 1 and the number of dose groups minus one. Select a model based on minimum AIC (Akaike Information Criterion).
- These programs do not report a goodness of fit statistic. Evaluating goodness of fit for models based on interval-censored or current-status data is still an open issue (Lawless, 2002), but recent literature may point to solutions.[9]

2. Survival-Adjusted Number at Risk (Poly-3 Method)

- Calculate a survival-adjusted number at risk using the poly-3 method[10]
- Use the observed incidence and adjusted N in BMDS quantal models
- This method gave BMDs and BMDLs similar to those from the MSW model in a limited number of comparisons.[11]

F. Modeling Noncancer Endpoints

This section describes analysis of data from animal bioassays in which animals are randomized to treatments and a single response is of interest. The randomization may involve restricted or stratified schemes that ensure similarity among dose groups in animal weights or other attributes. However, no repeated measures or cluster sampling is involved. An average or other summary statistic of repeated measures might be used, but its standard deviation must be correctly estimated.

1. Fit all 'standard' BMDS models

a. For continuous responses, apply the power, polynomial, Hill, and exponential models in BMDS. All orders of polynomial model (up to order equal to the number of dose groups-1) should be evaluated

b. For dichotomous responses, apply[12] the gamma, logistic, log-logistic, probit, log-probit, multistage, and Weibull models[13]. All orders of multistage model (up to order equal to the number of dose groups-1) should be evaluated

[9] References are available on request.

[10] See Piegorsch, W.W., and A.J. Bailer, 1997, Statistics for Environmental Biology and Toxicology, London: Chapman & Hall. Section 6.3.2, pp. 235-236. SWG can provide a template spreadsheet for poly-3 calculation or an R program that makes the calculation along with a survival-adjusted trend test. Note that use of a weighting method to produce "adjusted" numbers at risk (i.e., poly-3, causes the dose-response modeling assumption of binomially distributed observations no longer to be exact. Software is also provided at http://www.jstatsoft.org/v16/i07, "A Computational Tool for Testing Dose-related Trend Using an Age-adjusted Bootstrap-based Poly-k Test", by H. Moon et al. (2006) J. Statist. Software 16(7), 14 pages.

[11] It is not clear that either method should always be preferred. There may be a variance-bias trade-off (greater accuracy and lower precision for MSW) such that neither is uniformly 'best'

[12] The "quantal-linear" and "quantal-quadratic" models are special cases of other models in BMDS; there is no need to use them routinely. Those models should only be applied in the unusual case that a mechanistic hypothesis (based on independent evidence and reasoning) supports one of these models.

2. Identify a best-fitting model. Apply model selection procedures and identify a best-fitting model following EPA's Benchmark Dose Technical Guidance §2.3.9

G. Modeling Clustered Developmental Data

[Specific advice for this type of data is being developed by SWG]

H. Other Types of Data

Other forms of data are not handled in BMDS and will require collaboration with a statistician to ensure correct analysis and interpretation. Some of these are listed below and briefly described.

1. Multivariate Response Data

Data collected as multivariate measurements on animals are usually reported and analysed singly as univariate measures. There can be advantages to joint analysis, including a more accurate estimation of risk and BMD (eg, Catalano et al., 1993, 1997; Dunson, 2000; Krewski and Zhu, 1995; Najita et al., 2009; Ryan, 1992).

2. Repeated Measures Data

a. Toxico-Diffusion Model for Time-Dependent Neurobehavioral Data

The BMDS web site provides a toxico-diffusion model implemented in R (e.g., Zhu et al., 2005a, b). The model and program are based on published work and are intended for modeling dose-response for time-dependent (i.e. repeated measures) neurobehavioral data from neurotoxicity experiments.

b. Growth Models

There is a large literature on growth models with repeated measures. Mixed effects models (for example using NLMIXED procedure in SAS) are suitable for such analyses. Those who need to apply such models should consult a statistician and pertinent textbooks and monographs. Applicability of the toxico-diffusion model (above) to such data has not been evaluated.

3. Categorical Response Data

The BMDS web site provides the "CatReg" model. This software or related methods have been applied to risk estimation and meta-analysis in a variety of cases (eg, Brown and Strickland, 2003; Krewski et al., 2010a, b; Simpson et al., 1996; Guth et al., 1997).

Modeling of categorical data is a well developed topic encompassing many statistical models and methods described in a number of textbooks. It is unwise to proceed with modeling categorical data without the assistance of a toxicologist and a statistician familiar with these methods. Users may consult examples of categorical modeling applied to risk estimation (Teuschler et al, 1999).

4. Concentration-Time Data

[13] As of December, 2012, there remain some occasional convergence difficulties with the dichotomous Hill model. If it is plausible that response might plateau at < 100%, this model should be applied, but some trial and error (initializing parameter values) may be needed occasionally to achieve convergence.

Typically these data involve responses after short-duration exposures (i.e., acute responses). The BMDS web site provides software implementing the Ten Berge CT model. It is unwise to proceed with modeling such data without the assistance of a statistician familiar with these methods. Users may consult examples from the literature as a starting point (Brown and Foureman, 2005).

References

Brown, K.G. and G.L. Foureman, 2005, Regul. Toxicol. Pharmacol. 43: 45-54.

Brown, K.G. and J.A. Strickland. 2003, Regul. Toxicol. Pharmacol. 37:305-317.

Catalano, P.J., 1997, Statistics in Medicine 16:883-900.

Catalano, P.J. et al, 1993, Teratology 47: 281-290.

Dunson, D.B., 2000, Risk analysis 20:429-437.

Guth et al., 1997, Risk Analysis 17: 321-332.

Krewski, D. and Y. Zhu, 1995, Risk Analysis 15:29-39.

Krewski et al., 2010, J. Toxicol. Environ. Health Part A 73:187-207.

Krewski et al., 2010, J. Toxicol. Environ. Health Part A 73:208-216.

Lawless, J.F. 2002. Statistical Models and Methods for Lifetime Data. 2nd ed. Wiley-Interscience.

NRC (National Research Council). 1994. Science and Judgment in Risk Assessment. Washington, DC: National Academy Press [Chapter 11, Appendix I-1, Appendix I-2]

Najita, J.S., et al., 2009, Applied Statistics 58:555-573

Ryan, L., 1992, Biometrics 48:163-174.

Simpson, D.G. et al., 1996, J. Agric. Biol. Environ. Stat. 1:354-376.

Teuschler, L.K. et al., 1999, Regul. Toxicol. Pharmacol. 30 (Supplement):S19-S26.

U.S. EPA. (1988) Recommendations for and Documentation of Biological Values for Use in Risk Assessment. U.S. Environmental Protection Agency, Washington, D.C., EPA/600/6-87/008 (NTIS PB88179874). http://cfpub.epa.gov/ncea/cfm/recordisplay.cfm?deid=34855.

U.S. EPA Guidelines for Carcinogen Risk Assessment (2005). U.S. Environmental Protection Agency, Washington, DC, EPA/630/P-03/001F, 2005.

Zhu, Y., 2005, Environmetrics 16:603-617.

Zhu, Y., et al., 2005, Regulatory Toxicology and Pharmacology 41: 190-201 and 240-255;

Appendix D – Information Management Tool: Comment Tracker Database

During 2012, the IRIS Program as part of the CAST initiative identified a need for better documentation and communication of decisions. To address this need, an information management tool that allowed for recording, reviewing, responding to, and analyzing comments and responses was determined to be of value to the IRIS Program. To this end, two databases were developed to (1) facilitate the analysis of, and response to, comments received during the course of developing an assessment and review, and (2) allow comparison of comments and recommendations as well as Agency responses made in multiple assessments.

The first database consists of a MS Access 2007 database with a form designed to streamline data entry (Figure C-1). The form collects the following information fields on a given comment:

Database ID #	Overarching Issues*
Charge Question ID (*if relevant*)	Reviewer Agreement with EPA*
Verbatim Charge Question (*if relevant*)	Assessment Team Response/Level of Effort*
Reviewer	Revisions to Toxicological Review
Topic*	Response to Comment Appendix Location (Pg # and Charge Question)
Stage at which Comment was Received*	Official Response to Comment
Verbatim Reviewer Comment	Individual Addressing Comment
Summary of Reviewer Points/Recommendations	Completion Date
Major Comment*	Type of Review*

*Fields contain a limited number of options to facilitate comparison across chemicals.

Some fields contain a limited number of options for the user to choose from; for example, the topic field allows the user to link comments with various sections of the assessment, or with broader topics that may not be limited to a specific place in the document. The form also includes a means of navigating the list of records as well as adding records for individuals with less experience with MS Access. Additionally, pre-defined templates have been created that generate reports for different purposes (e.g., project management for the chemical manager, CAST review of comments received in peer review). Further templates will be developed as needed. Finally, the database contains a query function (see Figures C-2 and C-3), allowing the user to look for specific comments, or patterns of comments using different selection criteria. For example, a user can determine if certain comments are repeated at multiple points during assessment development, or restrict a search on a given term to a specific stage or type of review (e.g., limit the search to peer review comments only).

The Comment Tracker Database provides a number of benefits to the IRIS Program, including:

- **The database serves as a quality-control tool.** With the flexibility to track public comments received during the life of assessment development as well as formal comments received during peer review, the database ensures that comments received from the public or external peer reviewers are adequately considered.

- **The database facilitates project management.** Chemical managers will now have a tool for efficiently assigning initial responsibility for addressing comments to assessment team members and clearly defining roles and responsibilities of team members on an assessment.
- **The database simplifies management review.** For example, as part of the post-peer review CAST process, assessment teams provide reports to their CAST teams detailing their response to comments, including estimations of feasibility and level of effort required to address a comment. Comments that would require a significant level of effort to address can quickly be identified and decisions made in consultation with management on resource allocation to resolve a scientific issue raised by a reviewer/commenter.
- **The database promotes a deeper review of comments.** The ability to sort, limit, and query the full text in the database encourages team members to look for broader issues raised by commenters. For example, the recurrence of a comment at multiple stages of review, even if previously addressed by the assessment team, may indicate areas in a document where further attention is warranted, or where the clarity of a section of the document could be improved. Alternatively, comments that are positive towards a particular analysis or issue presented in an assessment may be instructive to other assessment teams or for team members in their work on other assessments.

The IRIS Program does not anticipate that using the database will significantly alter the length of time it takes to complete an assessment. Entering information into the database is unlikely to take longer than current methods used to report comments and Agency responses. After an assessment is finalized and posted, further modifications to that database will be restricted. A copy of the database will remain with the files specific to that assessment, while another copy will be reserved for use in the cross-chemical database.

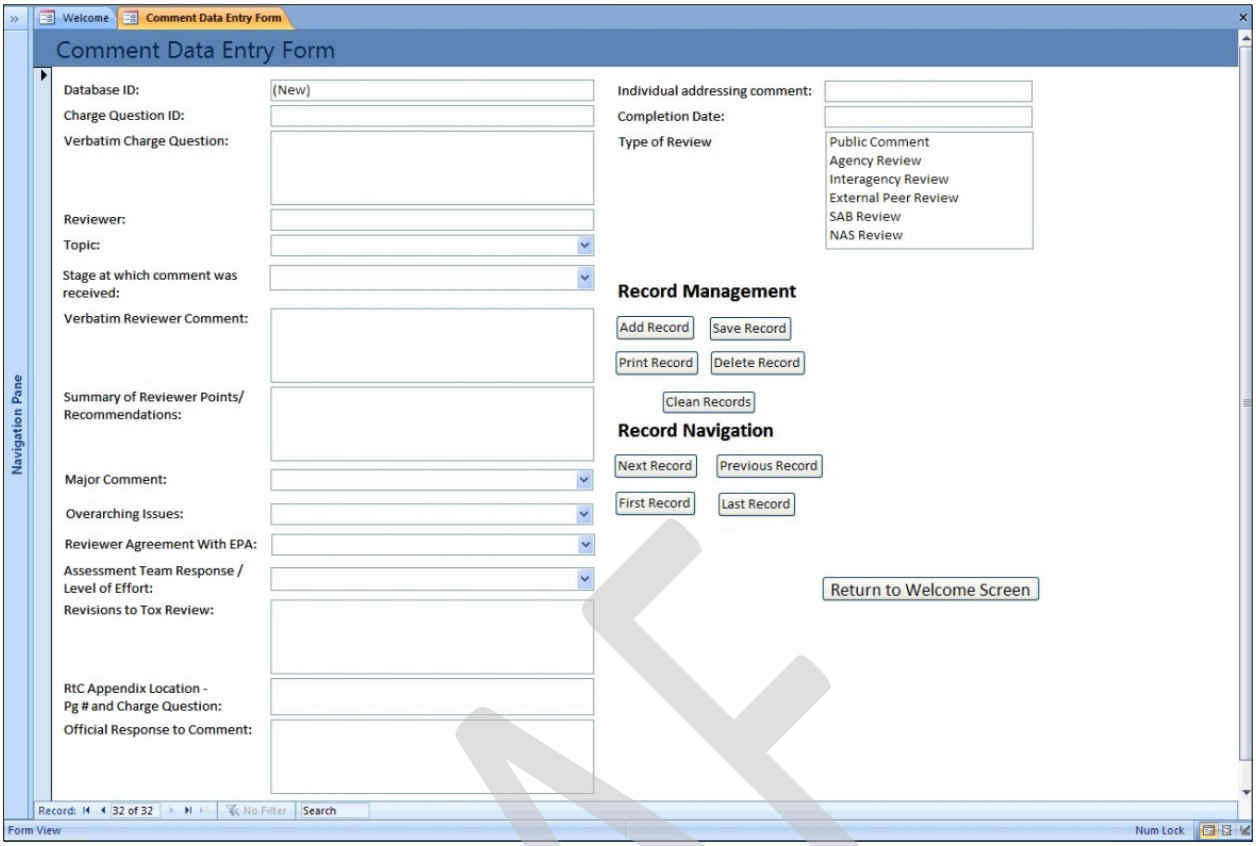


Figure C-1. Screen capture of Comment Data Entry Form. Drop-down menus contain pre-defined lists of content to facilitate management review and searchability of the database.

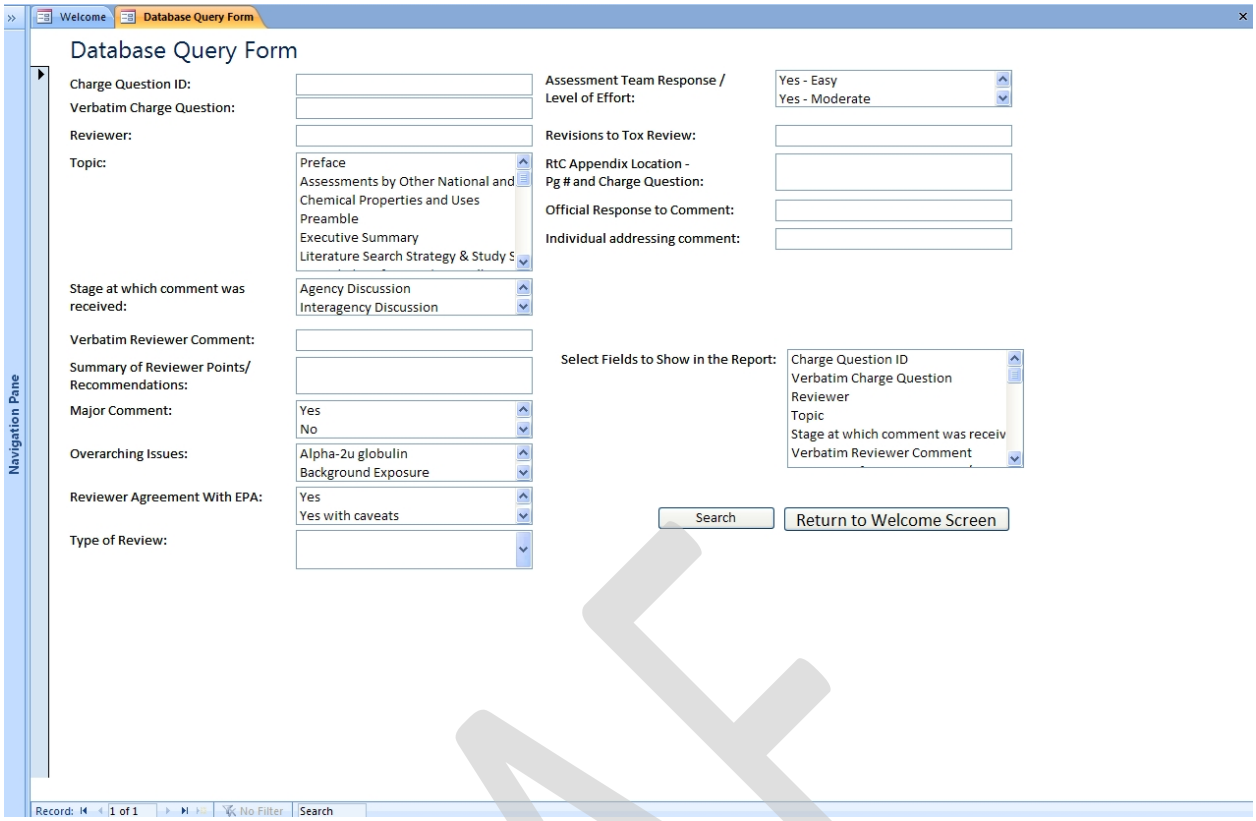


Figure C-2. Screen capture of the Database Query Form. Fields in the query form are largely the same as in the Comment Data Entry Form to allow the user to define any search parameters they may find useful.

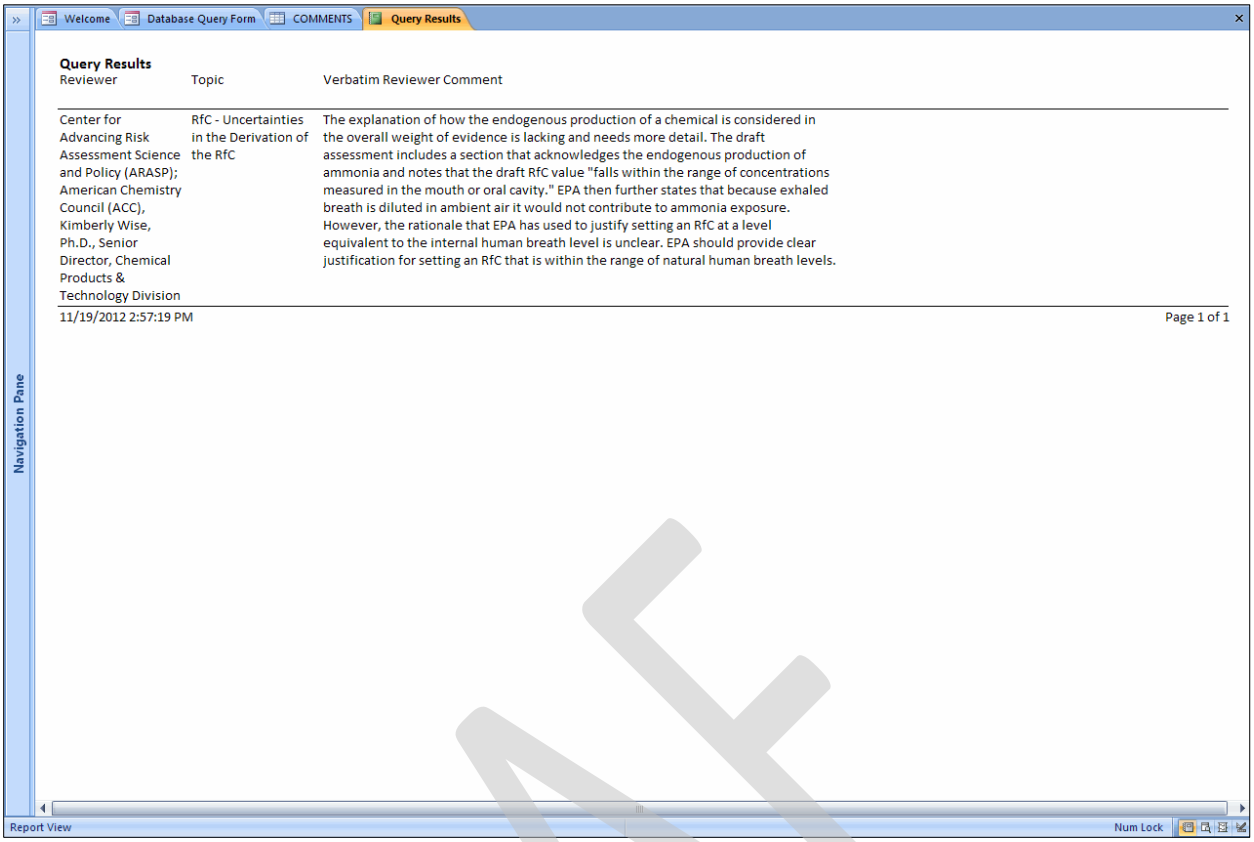


Figure C-3. Screen capture of a database query result, where the user searched the public comments on the draft ammonia assessment for use of the term "endogenous." The search returned one result, with the user generating a report with the Reviewer, Topic (i.e., section of the toxicological review associated with the comment), and Verbatim Reviewer Comment fields.

Cross-Chemical Comparison Database

Along with the chemical-specific comment tracker databases, the IRIS Program is developing an additional database tool to query multiple chemicals. Also developed in MS Access, the query tool is similar to the query tool used in the Comment Tracker Database; however this query form includes an additional field for the user to select which chemical(s) they wish to query (Figure C-4).

The Cross-Chemical Comparison Database allows the user to search for comments on a specific topic across chemical assessments. For example, if a chemical manager wanted to find all of the peer review comments related to a specific mode of action, a simple search would identify all of the relevant comments across all of the assessments in the database. The search results are based on information entered at the time the comment was added to a database. Users can conduct a full-text search within multiple fields, including the verbatim reviewer comment, Agency response to comment, and other fields where the responses can be variable dependent on user input. Figure C-5 shows an example output selecting from public comments received on the draft IRIS assessments for ammonia and trimethylbenzenes.

There are several benefits to the cross-chemical database, some of which overlap with the chemical-specific database. Similar to the chemical-specific databases, the cross-chemical database

is expected to aid staff during assessment development as well as management review. For example, the cross-chemical database will:

- **Help staff identify how scientific issues have been addressed in different assessments.** Early in draft development, chemical managers and assessment teams identify major scientific issues that will need to be addressed in the assessment. The cross-chemical database will allow staff to identify how these issues were addressed in other chemical assessments, as well as how the public and/or peer reviewers responded to EPA's analysis of that scientific issue. Understanding how a scientific issue was considered during peer review or public comment can highlight important points for chemical managers and assessment teams to consider early in draft development. Recognizing and considering these issues earlier in draft development help reduce the time needed to complete an assessment.
- **Simplify post-hoc analysis of comments across chemicals.** A searchable database will allow management and staff to review comments received across assessments and identify when issues are being raised consistently by single or multiple reviewers. These comments may point to cross-cutting issues that warrant further attention.
- **Allow for real-time searches during the review and public comment process.** During any of the review or public comment steps (Agency/Interagency, public comment, peer review), the database will allow scientists to identify, in real time, comments received on a topic at different steps in the IRIS process and look for consistency or inconsistencies. For example, if a public comment contradicts advice received during a peer review meeting, that inconsistency can be brought up during the meeting to ensure that peer reviewers are aware of previous arguments, and if needed, have further discussion on the issue.

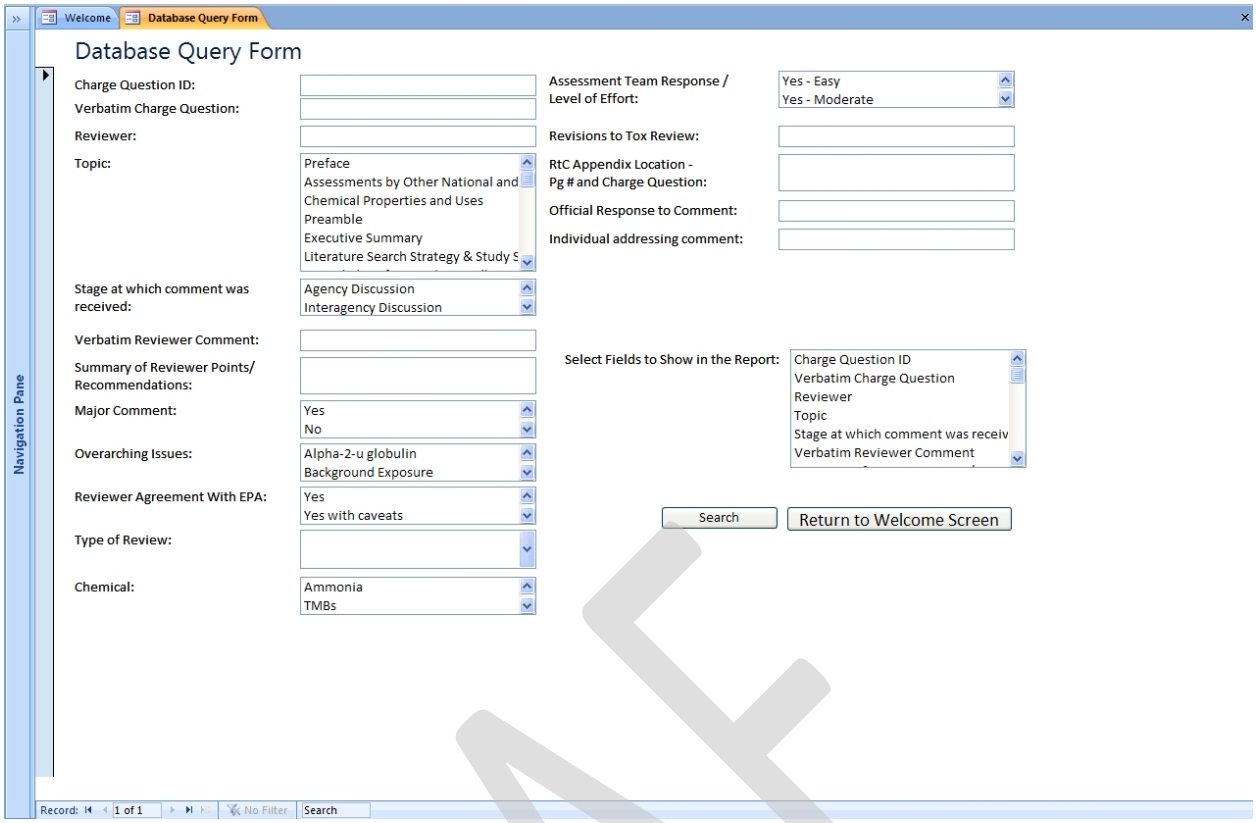


Figure C-4. Database Query Form for multiple chemical comparisons. The query form is identical to the query form in the Comment Tracker Database except for the inclusion of a field labeled "Chemical," which allows the user to select specific chemicals for comparison.

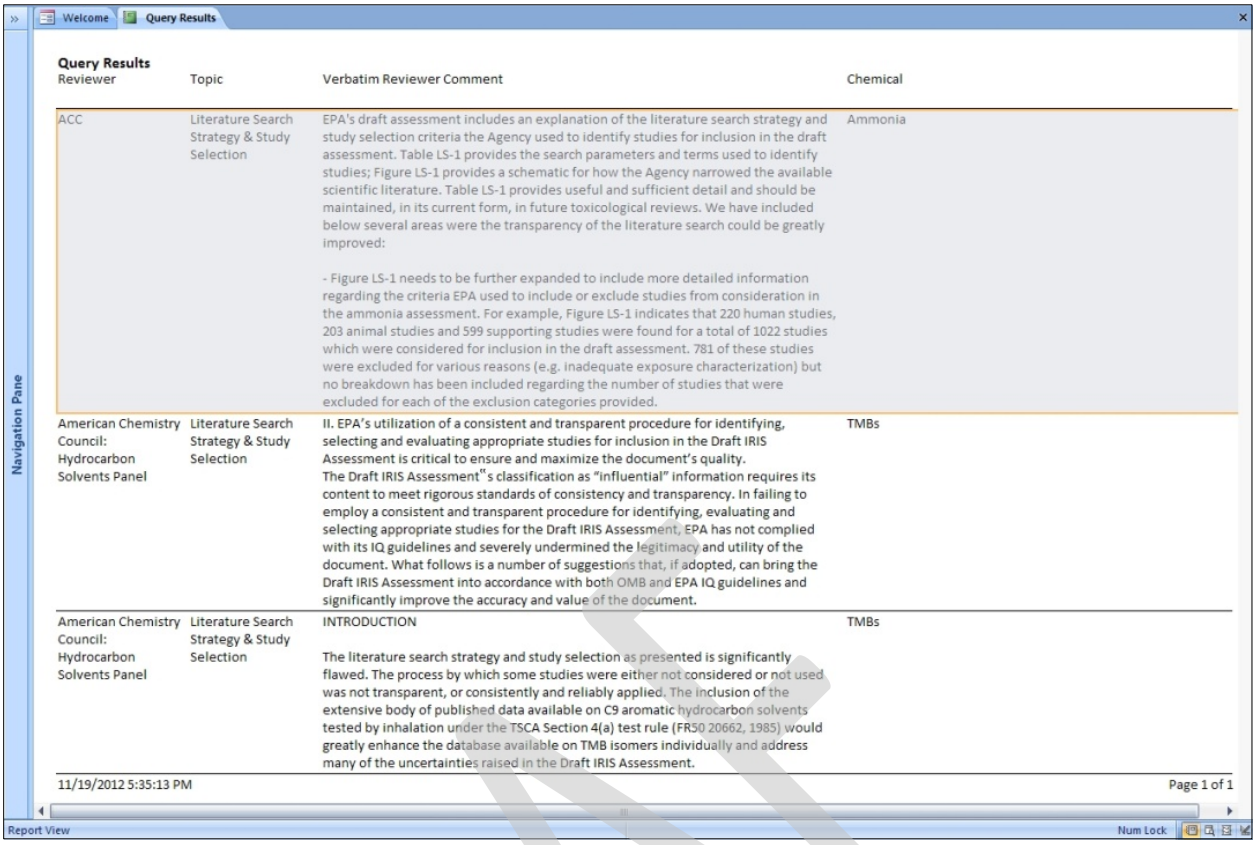


Figure C-5. Results of a multi-chemical query on the ammonia and trimethylbenzenes databases (public comments only). The user searched for the term "transparen*" in the verbatim text field and limited results to comments associated with the Literature Search Strategy and Study Selection topic/section of the toxicological review, and the query returned information on the Reviewer, Topic, Verbatim Comment, and Chemical. Use of the wildcard "*" insured that derivatives including "transparency" and "transparent" were both captured in the search. The search returned three comments, two for the trimethylbenzenes draft assessment and one for the draft ammonia assessment; all the comments were made by the American Chemistry Council.

Summary

The databases described above are currently being tested with assessments of varying size and complexity. Wider implementation of the databases is planned in January 2013. Once in place, the databases will serve as an important information management tool for assessment development as well as documentation and quality control of IRIS assessments.

Appendix E – Scoping to Inform the Development of IRIS Assessments

The following provides considerations for scoping to inform assessment development. These considerations are not to be construed as a standard process for scoping an IRIS assessment as the process will likely evolve as the IRIS Program gains experience in this area.

Purpose

The primary purpose of the scoping process is to understand the needs of clients in EPA's program and regional offices with regards to a chemical or group of chemicals addressed by an IRIS assessment. The scoping process builds upon information developed during the process of identifying chemicals for the IRIS agenda which helps to outline clear objectives for the IRIS assessment by defining and clarifying hazard identification and dose-response needs. During the scoping phase, more detailed questions are asked to seek greater understanding of the specific needs of the client offices.

Process

Scoping involves gathering specific information from EPA's program and regional offices (through either a meeting or other communications) before beginning an IRIS assessment. This allows client offices to identify their needs by explaining the environmental issues they need to address and what type of information is needed in a hazard identification and dose-response assessment to inform the decisions they will need to make. This exchange helps the IRIS Program understand the types of information needed and the level of detail necessary to address client needs. Understanding the clients' timelines is also important and may be factored into decisions about the scope of the IRIS assessment. The following provides examples of the types of questions that might be asked during the scoping process. It is important to note that these questions focus on the "what" rather than the "how" of developing an IRIS assessment.

- What are the environmental issues and the types of decisions that will have to be made? If there is more than one client interested in the IRIS assessment, do their decisions have different scopes?
- What risk assessment activities, if any, have been carried out for this chemical by EPA's program and regional offices, and other State and Federal stakeholders? Are there experts within these organizations with whom the IRIS Program can consult?
- What routes of exposure are of most *a priori* concern (e.g., oral, dermal, inhalation)?
- What form(s) of the compound are most relevant for the clients' needs (e.g., elemental forms or certain compounds for metals)? What ionic forms are of concern? In addition, should the IRIS assessment consider the effect of counter ions on the toxicity of the ion in question (e.g., NO_3^- in NH_4NO_3 or Br_2^- in $PbBr_2$)?

- Are there particular concerns regarding dose-response issues related to bioavailability of the compound? How stable is the chemical in physiological media? (If it readily decomposes, decomposition products may need to be considered in the IRIS assessment.)
- Do other chemicals routinely co-occur or are co-released with the compound(s) under consideration? If so, what are they? Where there is typically co-exposure, will the potential regulatory decision address only one chemical in the mixture, or the mixture as a whole? Would it be useful to develop a single IRIS assessment for the group of chemicals? For example: a single chemical (TCDD) versus all dioxin-like compounds.
- Does the decision-making or potential regulatory action pertain to a group of chemicals (such as substitutes for each other)?
- Are specific lifestages and windows of exposure of particular concern (e.g., children, geriatric, in utero, perinatal, lactation)?
- What are the typical durations of exposures of concern (e.g., acute, short-term, subchronic, and chronic)? Do exposure levels from scenarios of concern fluctuate significantly over time (which might impact the importance of short-term values to evaluate peak levels)?
- Are there urgent or time-sensitive decision-making needs faced by EPA's program or regional offices? (This question will help the IRIS Program address assessment deadlines by weighing priorities among IRIS assessments and by possibly re-aligning, if necessary, the scope of the IRIS assessment.)
- Would a conclusion on hazard identification without, or prior to, dose-response analysis be useful?
- Is it critical to have dose-response information that enables cost-benefit analysis, with some estimates of changes in health impacts between decision-making options?

Example of scoping: IRIS Assessment of Inorganic Arsenic

IRIS implemented a scoping process to inform the development of the inorganic arsenic assessment. Scoping meetings were held with EPA programs and regions as well as with interested stakeholders from the public and other federal offices. These meetings will inform the final planning and scoping statement for the inorganic arsenic assessment.

The following factors were discussed and identified by EPA clients at the arsenic planning and scoping meetings:

Hazard Identification Needs:

- Consideration of cancer and noncancer endpoints due to inorganic arsenic exposure.
- Inclusion of inhalation route in addition to oral.
- Consideration of exposure through occupational uses.
- Consideration of metabolites and oxidation state.
- Consideration of sensitive populations and lifestages: in particular children, and in utero and perinatal exposure. Evaluation of genetic and epigenetic factors affecting susceptibility, and impact of non-chemical stressors.

Dose-Response Needs:

- Impact of measures of exposure, bioavailability, and arsenic speciation.
- Risk at exposure to naturally occurring levels of inorganic arsenic.
- Estimation of risk beyond a reference concentration.
- Dose-response analyses that facilitate cost-benefit analyses.
- Impact of uncertainties in the dose-response analysis.
- Transparent presentation of choices made in the assessment and the supporting rationale.

- Is there a particular subpopulation already known to be of greater concern so that it is especially important to understand that particular susceptibility?
- Do the EPA decisions involve occupational risks (which may be at ranges above what are typical environmental exposures)?
- Is this a decision for which uncertainty and variability assessments might be particularly important? What kind of uncertainty-variability information will best inform the decision-making (taking into consideration the resource and time-intensive aspect of certain extensive uncertainty and variability analyses)?

Outcome

The outcome of the scoping process is a statement that outlines the focus of the assessment, the nature of the hazard characterization needed, and a clear indication of issues that are beyond the scope of the IRIS assessment.

Conclusion

The scoping process is an evolving tool. Although some level of scoping takes place for every IRIS assessment, the IRIS Program is just beginning to implement it as an early step in developing an assessment. As the Program gains more experience in this area, standard procedures may be developed. The IRIS Program needs to develop institutional experience and knowledge with the planning and scoping process for several assessments before formulating standard procedures. While face-to-face meetings may be necessary for some chemicals, email or other virtual consultation may be sufficient in other cases.

Appendix F – Draft Handbook for IRIS Assessment Development

The draft *Handbook for IRIS Assessment Development* provides information to IRIS assessment teams regarding internal processes and evaluation steps used in the development of IRIS assessments; however, the draft *Handbook* is a work in progress and currently does not fully discuss each step in the IRIS assessment development process. The draft *Handbook* is designed to provide the chemical assessment team with instructions and considerations involved in conducting a literature search; screening for relevance to identify and select pertinent studies; evaluating and documenting the quality of individual studies; reporting individual study results; synthesizing and integrating evidence for epidemiological, toxicological, and mechanistic data; selecting studies for derivation of toxicity values; considering combining data for dose-response modeling; managing data for dose-response modeling; and selecting an organ/system-specific or overall toxicity value.

Components that remain missing from the working draft are integrating across evidence (epidemiological, toxicological, and mechanistic data) to identify hazards and transition to dose-response analysis; conducting dose-response modeling; extrapolating to lower doses and response levels; considering susceptible populations and lifestages; developing candidate toxicity values; characterizing confidence and uncertainty in toxicity values; and selecting final toxicity values.

Identifying and Selecting Pertinent Studies: Literature Search and Screening

The focus of IRIS assessments is typically on the evidence of all types of health effects of a particular chemical, and their exposure-response relationships. This is, by definition, a broad topic. The systematic review process that has been developed and applied within the clinical medicine arena (evidence-based medicine) is generally applied to narrower, more focused questions. Nonetheless, the experiences within the clinical medicine arena provide a strong foundation for a similar endeavor in IRIS assessments. (Some useful references describing systematic review within clinical medicine are described at the end of this section.)

Systematic review, as applied in IRIS health assessments, is an iterative process that identifies relevant scientific information needed to address key, assessment-specific questions. The initial steps of the systematic review process formulate specific strategies to identify and select studies relating to each key question, evaluate study methods based on clearly defined criteria, and transparently document the systematic review process and its outcomes. The systematic review process must be conducted in a way that protects from bias in study selection and evaluation by transparently presenting all decision points and the rationale for each decision.

This section of the draft *Handbook* provides a discussion of the principles, overview of methods, and points to consider as you go through the process of developing and documenting a systematic review within the context of IRIS health assessments. It is not meant to be a "cookbook" or a checklist of procedures. The topics covered are the first two steps in the systematic review process: literature search and screening for relevance. Other steps involving evaluating the quality of individual studies and evaluating and synthesizing data across multiple studies will be covered in subsequent sections of the draft *Handbook*.

Throughout this draft *Handbook*, the examples provided are meant to be generalizable and are presented in a simplified form, such as would be expected for health assessments with a small literature database. Chemical-specific examples that have been drafted or completed can illustrate how these approaches may be envisioned for more complex datasets.

STEP 1: LITERATURE SEARCH

The strength of a systematic review of research rests on its ability to identify relevant studies, both published and unpublished, pertaining to the question of interest (e.g., health effects of a chemical). All search strategies balance competing needs for "sensitivity" (i.e., the ability to identify all potentially relevant studies) and specificity (i.e., the ability to avoid identification of non-relevant studies), using a process that is both manageable and reproducible. The efficiency of this process depends on optimizing the approaches used in initial searching and screening steps.

The goal of the search strategy is to identify full reports of **primary studies** (i.e., original sources of data) pertaining to the key question(s). These studies can be published papers or unpublished reports, but need to provide sufficient detail to allow evaluation of the study methods. The initial search strategy "casts a wide net"; subsequent steps in the process are used to screen and exclude articles that are not relevant, and to sort the relevant studies into categories (e.g., experimental studies in animals, observational studies in humans) for further evaluation.

In addition to the process described below, the IRIS Program takes other steps to identify relevant studies that may have been missed by the formal search strategy. The IRIS Program invites public review of the results of the literature search as one of the early steps in the development of the assessment. In addition, more targeted requests for review, for example by investigators active in the field of research, may also be a useful way to find studies that were not otherwise identified. Studies that are identified through this process can be included in the evaluation phase described in Study Quality Evaluation. It is also important to try to identify why the studies were missed in the initial literature search (e.g., the chemical was not part of the indexing terms used for that particular publication; publication not in any of the databases searched) and consider if any modifications to the search strategy are warranted.

The following sections discuss the key steps in the literature search process: selecting databases, selecting search terms, augmentation of a database search, documenting the search strategy, and updating the literature search.

1A. Selecting Databases

Systematic Reviews Conducted For IRIS Assessments Should Include Several (Types of) Databases, Including Sources of Unpublished Studies

The IRIS assessment team is responsible for devising and executing the literature search and screening strategy, with assistance from EPA resources (e.g., HERO staff) and from contractors as needed. If contractors are used, it is essential that EPA expertise guides the search process.

A search of PubMed is one way to start the process, to gain familiarity with the subject matter, but by itself is not sufficient. Table F-1 describes the databases that serve as the foundation for IRIS assessments. **PubMed**, **Web of Science**, and **Toxline** are overlapping databases of journals focusing on medical and life science, science and social science, and toxicology literature, respectively. These three databases are the core sources the IRIS Program uses for published studies. Other databases may be useful for specific chemicals or questions, so IRIS is not limited only to these core sources.

Another source of primary studies is the bioassays conducted by the National Toxicology Program (NTP). Although some of these reports may be found through the searches of the databases described above, it is also useful to directly search the NTP web site. This will assure you that you have found reports that have not yet been published. (All of the NTP reports have undergone external peer review, regardless of publication status.)

The category of unpublished studies can be quite broad, and could vary from anecdotal reports to standardized animal bioassays conducted under established protocols by reputable laboratories. These unpublished data are sometimes referred to as "gray literature." Under the federal Toxic Substances Control Act (TSCA), companies that manufacture, process or commercially distribute a chemical are required to submit to EPA results of chemical testing and health and safety studies. The **Toxic Substances Control Act Test Submissions (TSCATS)** database is a repository of unpublished studies submitted to EPA under TSCA. Submissions from 1985 to 2004 can be found through TOXLINE; subsequent submissions can be found through an EPA web site (see

TSCATS2, Table F-1). 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. Reports in this database should be included in the identification and evaluation process.

Another type of "gray literature" is conference proceedings and abstracts. IRIS assessments generally do not include these types of publications as primary research literature because the level of detail is insufficient to evaluate the methods. This group of references is kept as a separate category to facilitate its evaluation, but in general a study that is only available in abstract form would not be included in an IRIS assessment.

Chemicals used as pesticides must be registered for used in the U.S. by EPA's Office of Pesticide Programs (OPP). To support registration determinations, EPA requires more than 100 different scientific studies and tests from applicants, including toxicology studies. Searches of OPP databases should be performed for chemicals that are also used as pesticides. OPP's two main data bases are PRISM Documentum and the Integrated Hazard Assessment Database (IHAD). PRISM Documentum contains toxicology studies for all pesticides, nearly all of which are Good Laboratory Practices (GLP) guideline studies. IHAD contains summaries of the studies and reviews (Data Evaluation Records, DERs) which describe the studies and rank them for study quality and guideline compliance. Access to PRISM Documentum and IHAD is limited to EPA employees with FIFRA confidential business information access authorization, which requires one hour of online training.

The primary options for conducting searches are 1) using the HERO interface to selected databases, 2) directly searching databases, downloading citations into EndNote for review (and eventual import into HERO), and 3) supervising the search process conducted by contractors. It is possible that a combination of approaches will need to be used.

When using HERO for the search process, you will need access to the LitSearch, LitCiter, and LitTagger functions. It is important to test the search string in each of the selected databases, select the "no pdfs" option for this initial search, and include "tags" for each database as part of initial project page set-up. These tags can be used to track the source(s) for each citation identified in the search.

When doing direct searching of databases, you will need to take additional steps to eliminate duplicate references after combining the results of more than one database. (HERO does this automatically through the PMID number, but EndNote uses an algorithm that is more prone to errors based on differences in punctuation or capitalization). Also, there is no easy method to keep track of the source(s) of each individual citation. However, this option may provide greater flexibility in searching than the HERO interface provides.

When working with contractors, you will need to review each of the key decisions (i.e., selection of databases, search strings, additional sources). You do not want to simply receive a product; rather you want to be part of the process that creates the product by providing oversight and technical direction in accordance with the procedures specified in the contract and task order.

Table F-1. Description of Core Databases For Primary Literature

Database	Description	Notes
PubMed[a]	Approximately 5,600 medical, biology, and other life sciences journals (through MEDLINE), most back to 1966. www.pubmed.com	Uses Medical Subjects Headings (MeSH) terms
Web of Science[a]	12,000 science and social science journals, back to 1970. Also includes conference abstracts. Maintained by Thompson Reuters. http://apps.webofknowledge.com	Can also do "forward" searching, i.e., searching for publications that cite a specified reference.
TOXLINE[a]	Toxicology journals, including developmental and reproductive toxicology (DART), technical reports and research projects, and archival collections; back to 1965 (a few citations dating back to the 1940's); run by NLM. http://toxnet.nlm.nih.gov/cgi-bin/sis/htmlgen?TOXLINE	CASRN and synonyms
TSCATS	**Toxic Substances Control Act Test Submissions** Unpublished, studies 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 FYI (voluntary documents). TSCATS is included in the TOXLINE database via HERO from 1985 through 2004; for submissions after 2004, use TSCATS2 at: http://yosemite.epa.gov/oppts/epatscat8.nsf/ReportSearch?OpenForm Or for recent 8E and FYI submissions, search: http://www.epa.gov/oppt/tsca8e/pubs/8eandfyisubmissions.html	Chemical name; CASRN Section 8(e) submissions most relevant
Office of Pesticide Programs (EPA)		
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/ o OPP Applications (under popular sites in green box on left) o e-Registration Workflow (Documentum Login)	
Integrated Hazard Assessment Database (IHAD)	Contains Data Evaluation Records (DERs; reviews of toxicology study reports), memoranda, cancer reports, metabolism reports, etc. for all of OPP. Accessible through Lotus NOTES database to any EPA employee with FIFRA confidential business information access authorization	

[a] Accessible through HERO

1B. Selecting Search Terms

Consult a Reference Librarian Early and Often When Developing and Refining Your Search Strategy

IRIS assessments are not limited to a single, narrowly defined study question. The focus of IRIS assessments is typically the evidence of toxicity or health effect (of any kind) of a particular chemical. The search strategy thus would generally begin by selecting the appropriate forms of the chemical name, CAS number, and if relevant, major metabolite(s).

The process of selecting the search terms should be done in close collaboration between the EPA assessment team and a reference librarian, either with HERO or with a contractor working on the assessment. Both of these resources offer extensive experience with database searching and information management. Correctly using limits in the search strategy, and correctly constructing complex structure using AND, OR, and NOT terms requires a high level of training and experience.

For some chemicals, the initial search using the chemical terms will yield an easily manageable reference base (e.g., < 1000 citations). In other situations, it may be necessary to refine the search strategy. For example, if the chemical is used as a positive control in certain assays, or used as an extraction solvent, you may find yourself in the position of looking for the few relevant studies among thousands of citations. This situation presents challenges from the standpoint of efficiency, as well as accuracy, because even with the various computer-based systems to facilitate screening titles and abstracts, reviewer fatigue (and subsequent error) is possible.

For very large databases, the searching and screening processes may also be improved by developing a series of secondary searches, each focused on a particular question (i.e., reproductive toxicity, cancer, pulmonary function). These more focused searches would result in smaller collections of citations that can more easily be reviewed by people with the appropriate scientific background.

There are situations in which expansion of general search terms within a category of effects is warranted. Review articles and other key documents should be consulted for information about specific types of effects that are of particular concern for the chemical under study. For example, for some chemicals, focusing the immune-related effects on allergic sensitization may make sense. For others, autoimmune effects may be most relevant. Male infertility may be a primary endpoint of interest within the category of reproductive effects for one chemical, and ovulatory disorders may be of more interest for another chemical. In each of these cases, more targeted development of search strings may be warranted.

Examples of secondary search terms used in conjunction with the chemical name terms to focus a search are shown in Table F-2. The first set is an example of terms that could be used to focus a search on pulmonary effects. The second set is an example of terms that could be used to exclude studies that use a chemical in certain types of assays (in this case, formaldehyde), but which are not studies of the effects of that chemical. NOTE: Be very careful when using "NOT" as a Boolean operator. An abstract that contains the term will be removed, even if other parts of the study are relevant to the primary focus of the search.

It is important to evaluate the set of studies that are excluded to see if the exclusion process was overly broad. For example, if this secondary filter eliminated 10,000 out of 12,000 initial references, a random sample of the excluded articles (small enough to be manageable, e.g., but large

enough to be representative, 100-200) should be selected and manually reviewed (based on titles and abstracts) to determine if the "error rate" is acceptable; further refinements or additional manual review may be necessary.

The Literature Search Is Often an Iterative Process: What You Learn From Reviewing the Results Should Feed Back Into Ways to Expand or Refine the Strategy

Table F-2. Illustrative Example of Strategies to Improve Literature Search Results

Chemical Terms (include synonyms and relevant metabolites)	Purpose of Additional Search	Example Secondary Terms For Refinement of Search
Beryllium and beryllium compounds	Focus on pulmonary toxicity	"Chemical" AND (alveolar OR BAL OR brochoalveolar OR "carbon monoxide" OR "chest pain" OR "chest tightness" OR cough OR crackles OR DL_{CO} OR dyspnea OR FVC OR "pulmonary edema" OR FEV OR fibrosis OR granuloma* OR hypoxemia OR pneumon* OR pulmon* OR spirometry OR "radiographic X-ray")
Formaldehyde	Exclude use of formaldehyde as a fixative	"Chemical" NOT ("formaldehyde fixation" OR "formalin fixation" OR "formalin fixed" OR "formaldehyde fixed")

* indicates wild-card search term; search will include all permutations of the word with the specified backbone (e.g., *toxic* will include neurotoxic, toxicity, immunotoxicant, etc.)

1C. Augmentation of a Database Search

Do Not Rely Solely On the Initial Computer-Based Search of Databases to Identify Relevant Studies

Some publications will be missed with even the best-designed algorithm-based search strategy. Publications can be missed because they are not indexed correctly, or because the relevant data in the paper are not mentioned in the abstract. Articles published before 1965 are likely to be missed because of the coverage of the primary databases used in the search process. In addition, many older papers (e.g., published before 1970) do not include an abstract and so are difficult to evaluate in the initial screening process.

The team responsible for the IRIS assessment should identify key studies (review articles, other comprehensive documents, and articles with primary (i.e., original) data) that will serve as the basis for additional searches. It is useful to include reviews from different time periods, because earlier review papers may have more descriptive information about earlier studies.

Additional search strategies that can be employed through a database (e.g., Web of Science) include "forward searching", and "backward searching" based on articles identified as key studies. Forward searching identifies articles that cite the key study, and backward searching identifies articles that the key study cites. Using the forward and backward search options through Web of

Science does not eliminate the need for additional review, particularly of earlier time periods, given the coverage limitations of the database.

Regulatory resources and other web sites for information pertaining to a chemical of interest should be checked for additional resources.

As an additional check on the completeness of your search, you can send the list of identified relevant studies (i.e., studies that pass the screening step and are moved to Study Quality Evaluation) to researchers who are currently or were previously active in the specified area of research. They may know of other studies, including "file drawer" unpublished studies.

When using the "forward" or "backward" searching strategies through a database, the resulting set of references will need to undergo the screening for relevance step described in the next section. Other strategies will result in a more refined set of additional references that can bypass that step. For example, when reviewing the references in a discussion section of a paper, you may find six citations for similar studies, one of which is not already included in your search results. That one reference would be added to your literature search, but would not need to be further screened for relevance.

1D. Documenting the Search Strategy

Tell the Story of Your Strategy, Preferably in Such a Way that Someone Else Can Reproduce It

Accurate documentation of the search strategy is an essential component of the systematic review process. You want to be able to provide the reader with the information needed to understand what you did. Documentation of database searches should include, at a minimum, the database(s) and date range covered by the search, search terms used and the fields (e.g., title, abstract) to which they were applied, and dates the searches were performed (Table F-3). It is also useful to keep a log of the sources and approaches you used to augment the initial database search (Table F-4). In addition to these details, information pertaining to the context of your search strategy (not what you did but why you did it), such as why you focused the search in particular ways and other ways in which the search strategy evolved, should be included in the text describing your literature search.

Table F-3. Example Worksheet Summarizing the Database Search Process (Note: this is a research aid; this is not expected to be included in the finalized assessment)

Database	Set #	Terms	Hits
PubMed Date range Search date	1A	CHEMICAL TERMS; ADDITIONAL TERMS	
Web of Science Date range Search date	1B		
ToxNet Date range Search date	1C		
Other Database Date range Search date	1D		
Merged Reference Set	1	(duplicates eliminated through electronic screen)	

Table F-4. Summary of Additional Search Strategies

System Used	Selected Key Reference(s)	Date	Additional References Identified
Web of Science, forward search	Review study: Yuko et al., 2000 Review article: Smith et al., 2010 Primary study: Kim et al., 2006		N, citation(s) N, citation(s) N, citation(s)
Manual search of cited references	Primary study: Kim et al., 2006 Review article: Drew et al., 1966		N, citation(s) N, citation(s)

1E. Updating the Literature Search

Establish a System to Regularly Update the Literature Search for IRIS Assessments

IRIS assessments can take over 2 years to complete. The length of this process necessitates the establishment of a system to regularly update the literature search. The team responsible for the IRIS assessment is responsible for this updating process. An "alert" system can be set up through the core literature databases for automatic notification of new citations based on a designated search string. The frequency of the updates depends on personal preferences and the relative amount of research activity for the chemical under review, ranging from weekly for very large and active research areas, to bi-monthly for chemicals with relatively little active research. A cut-off date can be established for various steps. For example, although the updating process would continue throughout the development of the assessment, only studies identified up to a certain point (e.g., 45 days) before the literature search results are posted for public review would be

included. An additional cut-off date would be used for subsequent steps, such as finalization of the external peer review draft. Notably, after a certain step of the process (e.g., external peer review), additional studies will only be added if it is expected that they will substantially change the conclusions of the assessment; thus, a full literature search update will no longer be necessary.

Review of General Principles: Literature Search

- Use your initial search strategy to "cast a wide net"
- Include databases of published papers (PubMed, Web of Science, Toxline) and unpublished studies (TSCATS, TSCATS2; OPP documents for pesticides)
- Consult a reference librarian (through EPA resources or contractor resources) to develop search terms
- If using HERO, include "tags" for each database in initial project page set-up
- Augment database search with additional sources (e.g., lists of references from reviews and primary studies); solicited review of the identified studies by knowledgeable investigators may yield additional references (published and unpublished)
- Update literature search regularly (at least bi-monthly), with specified cut-off dates for study inclusion before key steps
- Keep a record of database(s), dates, search terms, results (n citations) and augmentation strategies
- It may be necessary to update your search criteria based on discoveries made at later stages of the systematic review process, such as evidence synthesis and integration. Be flexible to change, but be sure these criteria are documented and applied systematically

STEP 2: SCREENING FOR RELEVANCE

This Step Addresses the Following Question:
Is This Study Relevant to the Question of Interest?

It is highly likely that many, or even most, of the studies identified using your search strategy are not useful because they do not address the question of interest (i.e., the health effects of a chemical, or for more focused searches, a particular type of health effect or specified mode of action). This result should not be viewed as a deficiency of the search strategy process, but rather is expected given your goal of "casting a wide net." Your next step is to undergo a systematic review of each of the citations to determine its relevance; neither the quality of the study, nor the results are considered in this step. Depending on just how wide a net you ended up casting, this process could be somewhat akin to finding a needle in a haystack. In an effort to efficiently identify the non-relevant studies, this screening step is broken down into two sequential stages, title and abstract screening followed by full text screening. In some situations, a three-stage process may be more efficient, with an initial screen based on title, followed by screening based on abstract, followed by

full text screening. There is not a "right" or 'wrong" choice; however, whichever you choose, be sure to document the process you use.

The articles identified as relevant are then organized into topic-specific bins to aid you in performing the next step of evaluating the quality of individual studies.

2A. Review Process

This step in the review process is based on review of the title and abstract, and in some cases, the full text of the article, and should be conducted by two reviewers. If a contractor is used for this step, one of the reviewers should be an EPA staff member. It is meant to be a limited review of each citation as a way to relatively quickly exclude the large portion of citations that are not relevant.

There are numerous reasons a study may not be relevant to the subject matter of interest. Some reasons are common to all chemicals, and some may be tailored based on the specifics of the chemical. In some cases it is not possible to determine if the paper contains relevant data based on the information contained in the title and abstract; these citations should be set aside for additional perusal. Examples of the decisions that may be made about a citation, and reasons for these decisions, are shown in Table F-5, and discussed below.

Table F-5. Examples of Decisions Made Regarding Relevancy of Citation to Research Question (e.g., health effects of Chemical X)

Decision	Reasons
1. Exclude from consideration	• Duplicate • Abstract only (full report not available) • Not relevant - define categories as appropriate, for example: - study that uses chemical in sample preparation or assay - study that uses chemical as a positive control - study of effects on ecosystems
2. Not a primary data source of health effects data, but keep as additional resource	• Review articles, meta-analyses, editorials, risk assessments (use as source of additional references, discussion of key issues) • Articles describing development of measurement methods or exposure levels • Absorption, distribution, metabolism, and excretion studies • other (to be specified)
3. Further review needed	• No abstract • Language other than English • Case reports • Not enough information in title and abstract to determine relevancy • other (to be specified)
4. Move to full text screening	• Seems to be relevant to question of health effects of Chemical X

Decision Category 1. Exclude from consideration

Excluded Studies

- This category should contain a set of studies that two reviewers agree can be eliminated from further review; these studies are not included in the assessment.
- The goal of this step is to eliminate studies that do not contain original research that addresses the relevant question; neither the quality of the methods nor the details of the results should be considered in assessing the question of relevance at this stage.

This category will be relatively large. It will contain duplicates that were not caught electronically after the merging of the citations from the different databases (i.e., differences in punctuation or capitalization style may result in two citations being counted as two separate entries rather than as identical entries; when encountered, these types of duplicates need to be manually eliminated from the database). It also includes studies available only in abstract form (i.e., presented as a poster or presentation at a conference, but never published). As mentioned previously, this abstract-only group is treated as a separate category. In general, abstracts do not present enough information to allow evaluation of the study details. The assessment team can review this category and decide the appropriate level of effort to be used to pursue these studies; it may be possible to contact the study author(s), particularly if it is a relatively recent study, to obtain additional information. The decision to seek additional information should be consistently applied across all studies within the database that are similarly lacking information and should not be based on criteria that rely on the direction or magnitude of the study results.

This category will also include the studies that are "not relevant" – that is, studies that do not address the question of the health effects of the chemical of interest. Some of these types of "not relevant" studies are shown in Table F-5. You will find many other reasons that a study does not pass the "relevancy" test. The process of sorting through the database is facilitated by development of a list of common types of "not relevant" studies you are likely to encounter. The IRIS assessment team is responsible for developing this list, drawing upon the experiences of the reference librarians at EPA or a contractor you may be working with. You can do this by reviewing the results of other chemicals that have undergone this type of review (i.e., what are the categories that have been used previously?). Since each database may present unique issues, it may also be useful, particularly for chemicals with large databases, to systematically review a sample of the citations to develop a set of relevancy-exclusion categories – i.e., sort by year, take 5-10 citations per year, review titles/abstracts to get a sense of what this database includes, and use this information to develop the categories or questions that can be used to more easily sort through the entire database. You may initially have a large "miscellaneous reasons" category; this category can be examined and organized further as part of the review process.

There may be multiple reasons that a study can be considered "not relevant"; the reviewers should agree that the reason is correctly applied to each of the studies included in that category, but it is not necessary to count a study within every category that applies to it. It may be possible to create a hierarchy of categories, with those that are likely to be most easily determined (e.g., duplicates, reviews) placed first, to facilitate the review process. The two reviewers need to assure they have the same interpretation of the meaning of each category. For large databases especially, this may involve working through selected batches of 50-100 citations as "training" exercises. New categories should be documented with a written description of its definition, with enough detail that someone else could read it and determine that it was correctly applied. If a hierarchy of categories is used for the review process, this, too, should be documented.

One strategy for accomplishing this task is to have one member do the initial screening and sorting of the database, with the second member responsible for checking the accuracy of each of the resulting group (i.e., assuring that the reason for exclusion applies to each study in this group). A final step is resolution of the differences or discrepancies that are found. This approach allows for each study to be reviewed using two different frameworks: one asking "Does this study belong in the "not relevant" category, and if so why?" and the other asking "Is it true that [the specific lack of relevance category] applies to this study?" As with all of the other steps in the systematic review process, be sure to document the procedure(s) you use.

What you end up with in this category is a set of studies that two reviewers agree does not need to be considered further. If there are cases where the reviewers do not reach resolution and you are unsure of a study's relevance, set it aside for further review.

The main options for conducting the literature screening step are 1) "tagging" through HERO, 2) sorting sets of citations in EndNote (with eventual importation into HERO), and 3) supervising the screening process conducted by contractors. As discussed in the literature searching step, when working with contractors, you want to take an active role in decision making and quality assurance.

The most common approach for "tagging" in HERO is through use of the "LitTagger" function and EndNote. It is possible to directly "tag" citations in HERO, but that option does not work well for more than a minimal number of citations at a time. The "tags" used to represent the different exclusion categories should be specified when setting up the initial project page; modifications can be made but will need to be requested through the HERO librarians. When working with the downloaded HERO citations through EndNote, you will need to save the file to your desktop or file system if you want to complete the tagging process in multiple sessions. The HERO team is working on enhancements to the HERO database that will, in the future, allow you to complete the tagging over multiple sessions directly in HERO.

If you conducted your initial searches through the individual databases, rather than through HERO, you can use the EndNote grouping function for the screening process. After the database is uploaded into HERO, you can use the EndNote groupings to generate lists of HERO IDs for each of the exclusion categories. This process can be somewhat cumbersome for long lists, so you may need to ask for help from a HERO librarian. In brief, the list is generated by changing style to "HERO ID", selecting the group of references, right clicking

the mouse, and copying into the project page; a preface of 'hero.' then needs to be added to each number.

Decision Category 2. Not a primary data source of health effects data: keep as additional resource

Additional Resources

- This category includes reviews and types of studies that can serve as a useful additional source of potentially relevant primary articles, and studies that provide background information that could be useful in evaluating the health effects literature.

Review articles may address the question of the health effects of a chemical, but they are not considered relevant in that they are not a source of original data (i.e., a "primary" article). These types of studies, including meta-analyses, should be eliminated from consideration of primary data, but should be kept as an additional resource. For example, earlier reviews may contain information about studies that were not obtained in your search strategy because of limitations of the online databases and changes in indexing terms over time. In addition, as noted previously, reference lists of review articles also serve as a good source to augment your algorithm-based search strategy ("backward searching"). Reviews can also provide background information on issues you will need to consider as you evaluate the literature. Finally, some review articles do contain primary data (i.e., updates of previously reported data by the review authors), so additional review of the paper to specifically look for new primary data should also be conducted by one of the members of the screening review team.

Depending on the database, there may be other sets of studies that do not contain primary data pertaining to the toxicity of the chemical, but do contain background information that may be useful. For example, you may find studies describing the sensitivity or specificity of a particular type of effect measure, or studies of exposure levels in the general population or in different types of occupational settings. These studies can also be set aside as additional resources to draw upon in your evaluation of the primary studies. Absorption, distribution, metabolism, and excretion (ADME) studies are other examples of studies that you want to retain for use in the assessment.

The options for documenting this category are the same as those discussed for the first group of exclusion categories, and will most likely involve working through HERO and EndNote, or through a contractor.

Decision Category 3. Possible further review

Possible Further Review

- This category includes sets of studies that may be relatively easy to define, but difficult to process.

Some of the groups of studies in this category may be the easiest sets of studies to create. That is, it is relatively easy to select studies with no abstract and case reports through standard database search and sorting capabilities (e.g., through EndNote). The reference librarian(s) working with you on your search strategy can help with this sorting process.

Once organized, however, another question that must be tackled is what is to be done with them. For example, review of the set of case reports can give you an idea of the types of effects that have been seen, but only a limited number of these citations would need to be included if more extensive epidemiological studies examining the types of effects described in the case reports are available.

The category of citations with no abstract can be a relatively large group, and can include letters to the editor, commentaries, older studies (e.g., published before the 1970's), and some "brief reports" or "brief communications" found in some journals. It is often very difficult to determine the type of article, or the topic of the article, solely from the title. In some cases, the title may be sufficient to allow you to move it to another category (e.g., review article). Neither decision at either end of the spectrum (i.e., exclude them all or spend the time and resources to obtain, translate if necessary, and review them all) is likely to be an optimal decision.

The category of non-English language studies can also be relatively large, and it can be difficult to determine relevancy based on the limited information available. Often a relatively easily-obtained translation (such as through Google Translate) of the title and abstract will be enough to determine if an article belongs in one of the "not relevant" categories. Another approach is to use a "forward search" for papers which cite the foreign language studies; this can provide a better sense of the information that each contains. Decisions to translate the full text of articles that appear to be primary data sources of health effects data (and which are not also published as an English-language report) need to consider characteristics of the database, and available resources.

As the IRIS Program gains more experience with this process, more definitive advice may be developed as to how to proceed (i.e., whether attempts are made to obtain the complete publication, and to translate it if necessary). At this time, however, it is up to the assessment team to review each of these batches of citations to develop a decision making process that works given the scope of what is found for a given chemical. As has been noted previously, information that is available about the magnitude or direction of effects seen in a given study (e.g., from the phrasing of the title) should not be used in the decision regarding how to proceed. That is, you do not want to decide to translate a study because it looks like it has "positive" or "negative" results; rather, your decision should be based on your perception of the likelihood that the citation contains primary data pertaining to the research question (i.e., the health effects of a specific chemical). The options for

documenting this category are the same as those discussed previously, and will most likely involve working through HERO and EndNote, or through a contractor.

Decision Category 4. Move to full text screening

Move to Full Text Screening

- This category consists of the set of articles that appears to contain primary data pertaining to chemical toxicity; there is enough information available in the title and abstract to warrant further review.

The initial screening process should leave you with a much smaller set of studies than you started with. This smaller set of studies that will be subject of additional review through examination of the full text. During this process, it is likely that for some, the additional perusal of the full article will result in the realization that the study does not, in fact, belong in the group of "relevant" studies either because of one or more of the reasons used to define Decision Category 1 (Excluded Studies) or because of some other reason. In these situations, the citation should be "tagged" into the appropriate exclusion category.

Steps 1 and 2 of the literature search process, literature search and screening for relevance, are summarized in Figure F-1.

Review of General Principles: Screening for Relevance

- Focus is on this question: Does the study provide primary data relevant to the question of health effects of exposure to the chemical?
- Quality of the study is not considered in this stage of the review
- Based on title, abstract, and in some cases, full text
- Document screening process
- Two reviewers for screening process
- Some sets of articles will need to be put aside for additional decision-making (i.e., should the full article be obtained?)
- If using HERO, include "tags" for each exclusion category in initial project page set-up

Search Topic: Sources of Primary Data on Health Effects of Chemical X

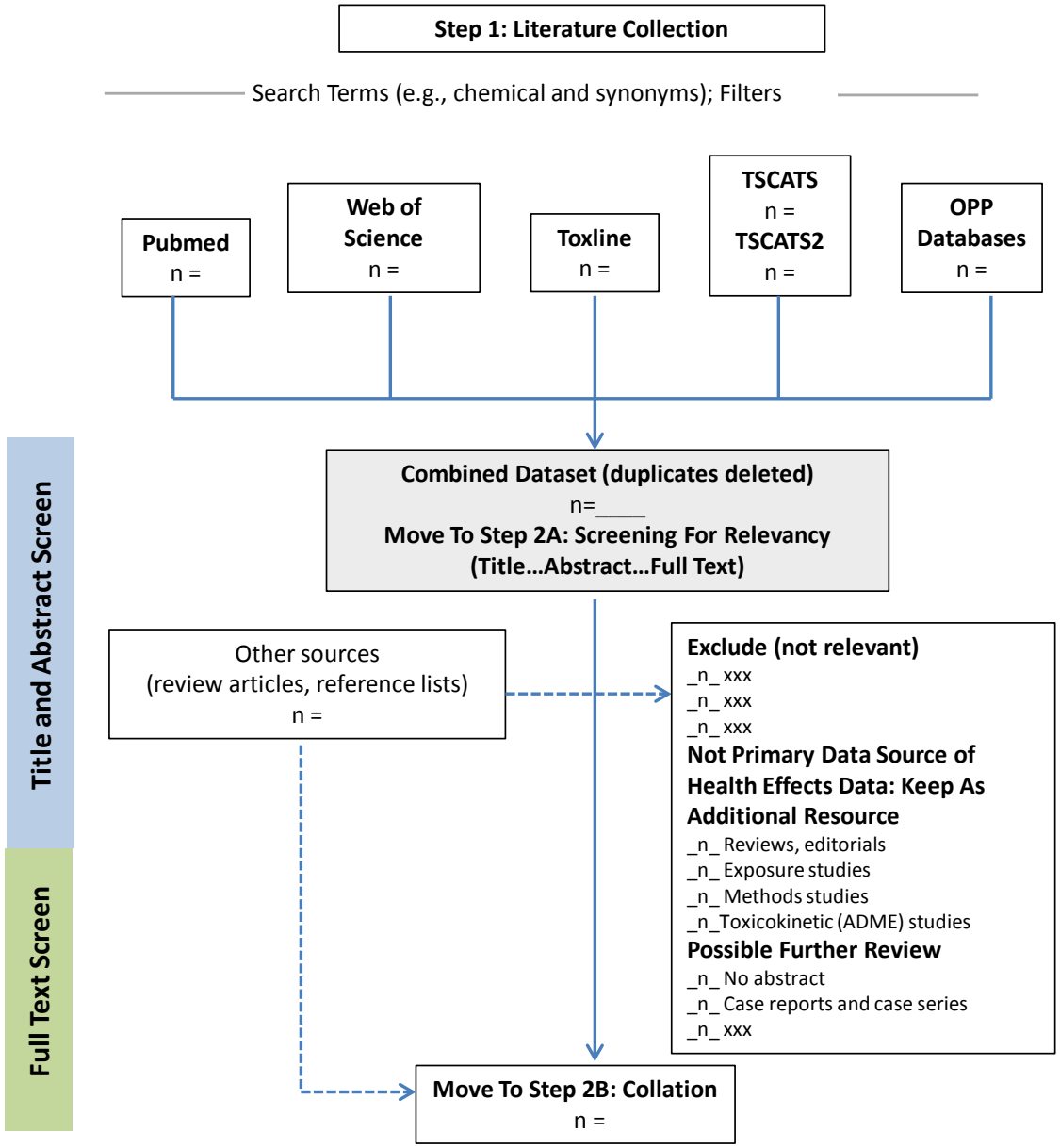


Figure F-1. Summary of Literature Collection and Screening for Relevance

1

2B. Collation (Sorting)

You should now be left with a (relatively) manageable number of citations that have a (relatively) high likelihood of providing primary data pertaining to the question of the health effects of a chemical. This collection of studies could include acute exposure animal experiments, two-year bioassays, experimental chamber studies in humans, observational epidemiology studies, in vitro studies, and many other types of designs. A basic organizational structure for the database is needed to facilitate the evaluation of this collection of studies.

The optimal organization structure will depend on many factors including the number of citations and breadth of topics and designs it includes. The assessment team should peruse the database to get a sense of the specific question(s) addressed by the available studies and to make a determination as to the optimal approach to sorting the studies. The actual process of sorting the database can be done by a contractor or by an EPA staff member. The goal is to create groups of studies that are of the same "type," such that specific evaluation tools (described in the Study Quality Evaluation section) can be applied. In general, the most likely divisions will be studies of the chemical's toxicity in humans, the chemical's toxicity in animals, and mode or mechanism of action (including in vitro studies). For large databases, however, additional categories or subdivisions within these categories may be needed. Figure F-2 provides an example of a sorting structure that may be useful for human and animal studies. For certain databases, it may be necessary to provide a greater level of detail than that presented in Figures F-1 and F-2.

Review of General Principles: Collation

- End result is references organized into categories that make sense for next step (study quality evaluation)

Search Topic: Sources of Primary Data on Health Effects of Chemical X

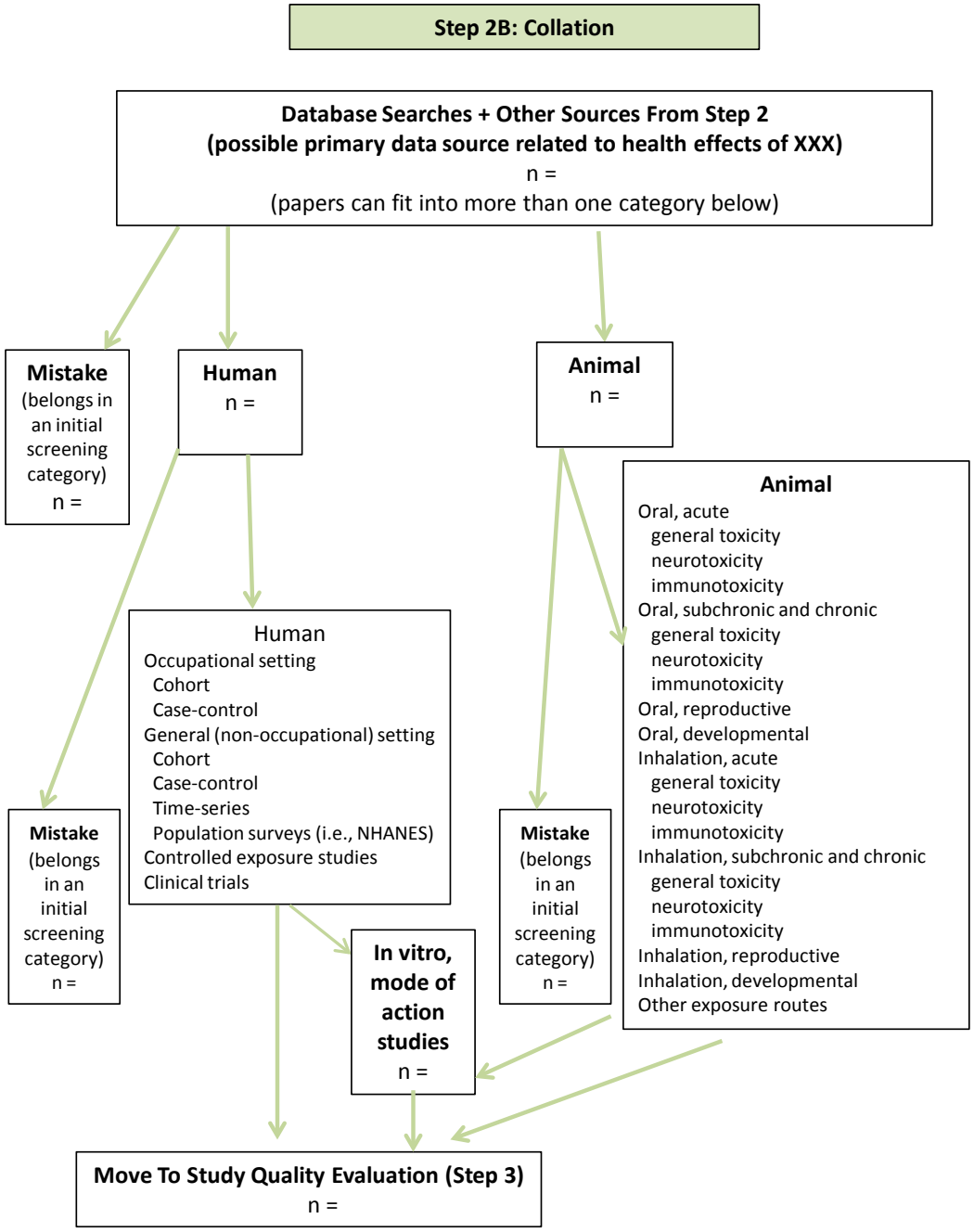


Figure F-2. Example of Collation, Human and Animal Studies

1
2

Systematic Review References

Briss PA, Zaza S, Pappaioanou M, et. al. (2000) *Developing an Evidence-Based Guide to Community Preventive Services-Methods*. Am J Prev Med. 18(1S).

- [An example of a codified system for evaluating specific aspects of the design and execution of individual studies, in conjunction with the pattern of results seen across studies, for the purpose of evaluating the evidence pertaining to effectiveness of a type of intervention.]

Higgins JPT and Green S, eds. (2008) *Cochrane Handbook for Systematic Review of Interventions*. West Sussex, England: John Wiley & Sons, Ltd.

- [A guide to the content and methods of systematic reviews, as developed and applied in evidence-based medicine.]

Eden J, Levit L, Berg A, Morton S., eds. (2011) *Finding What Works in Health Care: Standards for Systematic Reviews*. Washington, DC: National Academies Press,

- [The Institute of Medicine's report on standards for systematic reviews for comparative effectiveness evaluation for clinical practice.]

Evaluation and Display of Individual Studies

STUDY QUALITY EVALUATION

Study Quality Evaluation: Overview

- Be inclusive: it is better to include a study and evaluate effects of potential limitations than to exclude a study and eliminate any information the study could have provided
- Evaluate studies BEFORE developing evidence tables
- Series of focused questions; applied systematically to all primary data studies identified as relevant in the screening steps
- Evaluation is endpoint-specific; a given study evaluating several endpoints may have different strengths and limitations with respect to each endpoint

Study "quality," as defined herein, is a broad term encompassing interpretations regarding a variety of methodological features (e.g., study design, exposure measurement details, study execution, data analysis and presentation). The purpose of this step in the systematic review process is not to eliminate studies, but rather to evaluate studies with respect to potential methodological considerations that could affect the interpretation of or confidence in the results. For larger databases, in particular, this evaluation can provide a transparent means to convey your assessment of a study's methodological strengths and limitations, and thus your ability to rely on the results. The results of this systematic evaluation may also inform decisions about which studies to move forward for dose-response modeling for derivation of toxicity values.

The systematic evaluation described in this step should be conducted at an early stage of assessment development, i.e., after identifying the relevant sources of primary data but before developing evidence tables and characterizing hazards associated with exposure to a chemical. All studies identified as relevant from the literature screening process should be evaluated. Even if a deficiency in an aspect of the study is obvious, it can be useful to complete the evaluation of all of the component questions so that a full record of the evaluation can be maintained.

Examination of specific methodological features of each study can be accomplished by applying a series of focused questions. A useful starting point for generating these assessment and endpoint specific questions would be to consider the examples provided in Tables F-6 and F-7 for observational epidemiology and animal toxicology studies, respectively. Documentation of the important methodological features of a study may be an iterative process, requiring modification of an initial set of questions, as specific features of the chemical, endpoint(s), or study design(s) are discovered. It is essential that these focused questions be applied uniformly to all studies evaluated. This will allow for a comparison of the considered studies that is both systematic in design and independent of the study results. Ideally, two reviewers would independently identify the relevant methodological details, and then compare their results and interpretations and resolve any differences.

For studies that examine more than one endpoint or outcome, the evaluation process should be endpoint-specific, as the utility of a study may vary for the different endpoints.

The methods section of the paper will generally provide the majority of information needed for this evaluation (except, of course, for considerations relating to the level of detail of the reported results). In some cases, however, study details may be presented elsewhere in the manuscript or report, such as the introduction or discussion sections. Identification of some study details may require additional investigation, for example, by consulting other publications describing the study or studies on the reliability of an assay, or by contacting the study authors. In general, study quality evaluation should be independent of considerations regarding the direction or magnitude of the study's results.

It is useful to check the citation in one of the primary databases (e.g., PubMed) to see if there is any linked material, such as an erratum, supplementary or appendix material, letter to the editor (and authors' reply) regarding the citation, or companion study. This kind of preliminary work can prevent significant heartburn and headaches in subsequent steps.

It is useful to record the pertinent methodological features in an easy to read form (e.g., a tabular format) so that these study details can be easily reviewed. Because observational epidemiology and animal toxicology studies have fundamental differences, the documentation and evaluation of these studies will differ.

There may be situations, most commonly when extensive literature databases exist for a given chemical and effect, in which an individual study or sets of studies can be excluded from further consideration. For example, acute animal toxicology studies may be excluded when abundant subchronic and chronic exposure studies examining similar endpoints are available.

The following discussion of study quality evaluation is focused on evaluation of observational epidemiology, animal toxicology, and human controlled exposure studies. This approach could also be adapted for the evaluation of in vitro studies and other types of studies relevant to mechanisms of action.

Study Quality Evaluation: Logistics

- Methods section of the study should provide most of the information you need; study quality evaluation should be independent of considerations regarding the direction or magnitude of the study's results
- Look for errata, supplemental files, and other material linked to the primary data citation for additional information about the study
- Published correspondence (e.g., letters to the editor, editorials) may provide additional background information on important methodological features.
- Ideally, use two *independent* reviewers, with procedures for disagreements to be reviewed and resolved

Evaluation of Observational Epidemiology Studies

The process of study evaluation is akin to detective work. You need to investigate specific study features that directly affect the interpretation of the experimental results, including:

- exposure measures (reliability, validity, probability and level of exposure in different situations or settings)

- outcome measures (reliability, validity, prevalence in different populations, disease course, relation between survival and access to health care or other socioeconomic factors)
- confounders (strong risk factors for the outcome that are also known to be strongly associated with the exposure within the study)

These investigations may require "mini-reviews" and consultation with experts in different fields. Without this background understanding, you may not be able to accurately evaluate the studies.

Exposure assessment is especially important in the environmental or occupational arena. The ability to correctly classify "exposed" and "unexposed", estimate quantitative measures of exposure, and the range of exposure encompassed in the study is a key difference between observational epidemiology and randomized clinical trials in which "exposure" (e.g., "intention to treat" or type of treatment) may be less subject to measurement error and the exposure contrast is less variable between studies.

As noted above, an inclusive approach is generally recommended: that is, it is better to include a study in this systematic evaluation and examine the impact of potential limitations, rather than exclude a study and thus lose any information it could have provided. For epidemiology studies, to the extent possible, you want to assess not just the "risk of bias," but also the likelihood, direction, and magnitude of bias.

The study characteristics that inform the evaluation of observational epidemiology studies are summarized in Table F-6. The first feature, the type of study design, provides a framework for the subsequent evaluation; that is, the specific questions and issues will vary depending on the type of study. The other features encompass aspects of the study populations, exposure measures, outcome (effect) measures, and the analysis and presentation of results. Although in general your evaluation is based on the information provided about study methods, review of some of the results is needed, for example within the context of the evaluation of confounding, since confounding depends on the strength of various relationships (i.e., between the exposure and the potential confounder and between the potential confounder and the outcome).

A structured form may be useful for recording the key features needed to evaluate a study. An example form is shown in Figure F-3; details of such a form will need to be modified based on the specifics of the chemical, exposure scenarios, and effect measures under study.

Study Quality Evaluation: Observational Epidemiology Studies

- As noted in the overview, the evaluation process is inclusive in nature, is conducted BEFORE developing evidence tables, uses a series of systematically applied, focused questions, and is end-point specific
- Do your detective work ahead of time: investigate exposure measures, effect measures, and confounders for the chemical-effect under review
- To the extent possible, assess likelihood, direction, and magnitude of bias

Table F-6. General Considerations for Evaluation of Features of Epidemiology Studies

Feature	Example Questions or Details	Useful Information
Study design	Major types, based on approach to sample selection: cohort, case-control, nested case-control, population-based survey (e.g., NHANES), times series, case-crossover	Study methods
Study population; target population; setting	Where and when was the study conducted? What is the source(s) of exposure (environmental media, consumer products, occupational, an industrial accident, or other)? What was the recruitment process? How was eligibility determined? Does the study provide information on potential vulnerable or susceptible groups? Address: Potential generalizability of study results, potential for selection bias, potential to address effect modification	Geographic area, site (occupational, etc.), time period. Age and sex distribution, other details as needed (may include race/ethnicity, socioeconomic status); recruitment process; exclusion and inclusion criteria
Participation rate; follow-up	Did rates vary by exposure (or disease) status? Were there differences between individuals who did and did not participate, or who were or were not lost to follow-up? Is it known (or possible) that participation (or loss) is related **both** to exposure **and** disease status? Is there evidence of "healthy worker" or "healthy worker survivor" effect? Are differences likely to impact the observed associations (and if so, how)? Address: Potential for selection bias	Total eligible; participation at each stage and for final analysis group; loss to follow-up, denominators used to make these calculations; length of follow-up
Comparability (exposed and non-exposed; cases and controls)	How were potential differences between groups addressed in the study design (e.g. randomization, restriction, matching) and/or analysis (e.g. stratification, multivariate methods)? How were variables associated with exposure and with outcome, or which alter the association between exposure and outcome, addressed in the study? Address: potential for confounding and effect modification	"Table 1" type participant characteristic data, by group; approach to consideration of potential confounding (if applicable); strength of associations between exposure and potential confounders and between potential confounders and outcome
Exposure measures (procedure, range)	Are exposure estimates qualitative, semi-quantitative or quantitative? How well does the exposure protocol correctly classify or rank participants with respect to exposure? What is the likelihood of systematic (differential) error? What is the likelihood of random (non-differential) error? Does the protocol adequately characterize exposure during the relevant time window? What exposure range is spanned in this study? Address: potential for exposure misclassification (either non-differential or differential).	Describe, i.e., type of biomarker(s), occupational history, lifetime consumption, evidence from validation studies, variability within and between exposure groups
Outcome measures	What is source of outcome (effect) measure? How well do the outcome(s) measures correctly classify participants with respect to the outcome? What is the likelihood of systematic (differential) error? What is the likelihood of random (non-differential) error? Address: potential for outcome misclassification (either non-differential or differential).	Describe (i.e., source, how measured/classified, incident versus prevalent disease), evidence from validation studies
Data Presentation and Statistical Analysis	Is the analysis appropriate for the data and the study question? Are aspects of the data (i.e., non-normal distributions, correlation structure) adequately accounted for? Is the rationale for inclusion of variables in a model clear and logical? Are results presented with adequate detail? Is the study population of adequate size and composition to detect a true association (of a relevant effect size) between exposure and outcome? Were stratified analyses (effect modified) motivated by a specific hypothesis? Address: ability to interpret and level of confidence in results	How groups are compared (may include t-tests, ANOVA, regression models, etc.); what results are presented in text, tables, and figures; n exposed cases (case-control studies) or N cases among exposed (cohort studies).

Figure F-3. Example Worksheet for Recording Methodological Details of Observational Epidemiology Studies (Note: this is a research aid; this is not expected to be included in the finalized assessment)

Reference (primary)		
Additional reference(s)		
Study Design ____ cohort ____ case-control ____ nested case-control ____ population-based survey (e.g., NHANES) ____ times series ____ case-crossover _____ controlled exposure ____ other (describe)		
Setting	Describe, i.e., geographic area, worksite, clinic; time period….	
Study Population	Excluded if…..	Descriptive Statistics (e.g., median, range, etc – what is reported will vary among studies)
Age Sex Other details as relevant (e.g., socioeconomic status) Duration of exposure		
Participant Recruitment	Describe process Evidence that knowledge of exposure and diseases status affected participation?	
Participation Rates / Follow-up (separate data for cases and controls, exposed and non-exposed if provided)	Total eligible: Participated. any part (describe): Participated, all parts: Loss to follow-up: Length of follow-up:	Evidence of differential participation?
Comparability of Groups	Comparability between exposed and non-exposed; cases and controls ("Table 1" type sociodemographic data)	
Exposure measurement protocol	Describe, i.e., type of biomarker(s), occupational history, lifetime consumption…	
Biomarker(s) details	Sample collection (time of day, fasting?)	Assay (e.g., coefficient of variation, limit of detection, proportion < limit of detection), blinded to outcome status?)
Outcome measurement protocol	Describe (i.e., source, how measured/classified, incident versus prevalent disease, etc)	
Medical records		
Likely confounders?	Variables strongly associated both with exposure and with effect? What is strength of associations in this study? How addressed?	
Analysis and presentation of results	Approach used; assumptions made regarding distributions or shapes? Standard error or confidence intervals presented for effect estimates (or could be computed)?	
Statistical power considerations *may differ for different effects*	n exposed cases (case-control studies) or n cases among exposed (cohort studies)	
Reviewer Comments	*May include summary evaluation of likelihood, direction, and magnitude of bias. May include usefulness and feasibility of re-analysis.*	

Evaluation of Animal Toxicology Studies

In contrast to observational epidemiology studies, animal toxicology studies seek, by their very nature, to control exposure and environmental conditions. Considerations relevant to the evaluation of toxicology studies include exposure and endpoint methodology, as well as control of potentially confounding variables.

Table F-7 provides a list of questions relating to study features that should be considered when evaluating in vivo animal toxicology studies, namely: exposure, test animals, study design, toxicity endpoints, data presentation and statistics, as well as the reporting of this information. These questions are based on previous approaches for evaluating toxicological data (e.g., Klimisch et al., 1997; U.S. EPA 2002, 1994). These study features reflect aspects of an experiment that have been placed into modular components to assist in the analysis and transparent documentation of decisions; however, there is some overlap among the study features and it may be useful to reorganize some of the study features or clarifying questions for a given chemical or research question. For each study feature, the table provides a primary question that an assessor should try to answer using expert judgment. The example clarifying questions are included to provide direction and suggest ways to evaluate and document the study evidence that underlies these decisions. By no means are these questions comprehensive and, for most assessments and endpoints, some of these questions will not be applicable. For example, determining whether maternal toxicity was considered by study authors will only apply to evaluations of developmental and reproductive toxicology studies.

The purpose of the questions in Table F-7 is not to exclude studies from consideration. Rather, these questions are intended to help identify and characterize features of a given study that, together, can provide a picture of how well that study informs the specific endpoint in question. These evaluations should not preclude toxicologists from looking for patterns across studies on a given endpoint, even if all of the identified studies do a relatively "poor" job at analyzing the endpoint in question.

Additionally, not all clarifying questions or considerations are of equal importance. Although the relative importance of specific criteria may vary by endpoint, chemical, or database, in general the criteria in bolded text represent some of the more important questions to examine. Evaluations of exposure quality, study design, and toxicity endpoints will generally require the greatest effort. Exposure quality refers to the characterization of the animals' interaction with the test article, which should be specific to the chemical of interest and tightly regulated by the study director. Exposure quality is a particular concern for inhalation toxicology studies because of the inherent complexity in generating and characterizing test atmospheres. Study design refers specifically to the setup of the experiment. It includes consideration of items such as the length of exposure, the distribution of test animals into dosing groups, and the timing of endpoint(s) evaluation. Endpoint evaluation refers to the specific methods used to assess the hazard in question, including whether the protocols used to evaluate the endpoints were appropriate and complete, as well as whether said protocols are subject to modification by factors present in the study other than the chemical of interest. It is important to reiterate that all of these features should be evaluated without consideration of the magnitude or direction of the reported study results. Finally, any decisions made during the evaluation of a given study should be applied consistently throughout the database of studies on that particular endpoint.

Study Quality Evaluation: Animal Toxicology Studies

- As noted in the overview, the evaluation process is inclusive in nature, is conducted BEFORE developing evidence tables, uses a series of systematically applied, focused questions, and is end-point specific (a study may be very useful for one type of endpoint, but not for another)

Table F-7. General Considerations for Evaluation of Features of Animal Toxicology Studies

Feature	Primary Question	Example Clarifying Questions/Considerations
Exposure Quality	Are the exposures well designed and tightly controlled?	*General Attributes* ▪ **How well was the test article identified and characterized?** Are co-exposures expected as a result of test article composition? ▪ **Was there a vehicle control group?**
		Inhalation Studies ▪ **For generation and measurement of the test article, how accurate and appropriate were the methods employed?** Were analytical concentrations in the test animals' breathing zone reported (i.e., not just target or nominal concentrations)? For aerosol studies, were the mass median aerodynamic diameter (MMAD) and geometric standard deviation (GSD) reported? ▪ **Was a dynamic chamber used?** Static chambers are not recommended.
		Oral Studies ▪ Diet/Water: Could accurate doses be determined (e.g., was consumption measured)? Are there any expected or reported issues related to stability, homogeneity, or palatability of the test substance? ▪ Gavage: To what extent would toxicokinetic differences due to bolus dosing be expected to influence the results?
Test Animals	Are the test animals appropriate for evaluating the specified effect(s)?	▪ **How well are the control and exposed test animals matched in aspects other than exposure?** Was information available to evaluate potential effects such as systemic or maternal toxicity that could confound interpretation of the endpoint of interest? Were there any notable issues regarding animal housing or food and water consumption? ▪ **Based on what is known about the endpoint in question, how well do the species, strain, sex, age, and/ or number of test animals examined inform this evaluation?**
Study Design	Is the study design appropriate for the test article and the evaluated effect(s)?	▪ **How well do the timing, frequency, and duration of exposures inform the effect(s) measured?** For example, are critical windows of development encompassed by the exposures when assessing developmental toxicity? Were multiple exposure groups tested? ▪ **If the results are expected to be subject to confounding by factors introduced as a result of selection bias, were efforts made to protect against this (e.g., control for potential litter bias in developmental studies; randomization of treatment groups)?** ▪ **How well do the timing and/or frequency of the endpoint evaluation(s) inform the measured effect(s)? For example, is the latency between exposure and testing expected to influence the level of confidence in the results?** ▪ **Was the study conducted under Good Laboratory Practices (GLP)?** ▪ How well does the study conform to established guidelines (e.g., EPA, OECD)? Was it designed to specifically test the endpoint(s) in question? ▪ Did the study include other experimental conditions or procedures (e.g., surgery) that may influence the results of the toxicity endpoint(s) in question? If so, were appropriate control groups (e.g., sham) included in the study design?
Endpoint Evaluation	Are the protocols used for evaluating the endpoint(s) reliable and specific?	▪ **How well do the procedures used to evaluate the endpoint(s) in question conform to established protocols?** If novel or uncommon, are the approaches biologically sound? ▪ **What is the level of specificity of the protocols used?** Did they include control experiments to discern effect-specific contributions (e.g., learning and memory) from nonspecific contributions (e.g., from motor activity) to the output measure (e.g., escape latency) ▪ **How sensitive are the protocols for a given endpoint?** ▪ As appropriate, were steps taken to minimize potential experimenter bias (e.g., blinding) and sampling bias (e.g., evaluation of multiple tissue sections/ organ)?
Data Presentation	Do the results provided allow	▪ **Are the statistical methods and comparisons appropriate and transparent?** If not, is sufficient information available for the IRIS Program to conduct its own analyses?

Table F-7. General Considerations for Evaluation of Features of Animal Toxicology Studies

Feature	Primary Question	Example Clarifying Questions/Considerations
and Analysis	one to accurately identify the direction and magnitude of the observed effect?	▪ Are there any notable issues regarding presentation of the results? For example, if data were pooled (e.g., pooled exposure groups; pooled sexes) and this is expected to influence interpretation of the results for a given endpoint, are the reasons justified? ▪ Did the study report an unexpectedly high/low level of within-study variability and/or variation from historical measures that was not addressed?
Reporting	Are the methods and results well documented?	▪ **Are all aspects of the study described in sufficient detail such that it can be evaluated across the five study features presented above? Are any critical descriptions missing?** ▪ Are group sizes and results reported quantitatively for each exposure group, time-point, and endpoint indicated as examined?

Criteria in **bolded text** represent the more important considerations.

Additional information on study protocols (e.g., guidelines developed by EPA and the Organization for Economic Co-operation and Development [OECD]) that may prove helpful in evaluating study features can be found in the annotated reference list. Consult EPA and OECD guidelines for recommendations on the design and interpretation of toxicology experiments. Remember, however, that these are not intended to be comprehensive protocols designed to provide in-depth analyses of all endpoints of toxicity; thus, they are not to be used as the "end-all, be-all" references for evaluations regarding study quality across every study or endpoint.

Study Quality Evaluation: Animal Toxicology Studies

- Because all aspects of a toxicology study should be controlled, it is expected that the exposure causes the outcome. Anything that makes you question this is likely a study limitation.
- Decisions made during the evaluation should be applied in a consistent manner throughout a given database of studies

Evaluation of Human Controlled-Exposure Studies

Human controlled-exposure studies combine aspects of observational epidemiology studies and animal toxicology studies. Examples of human controlled-exposure studies include randomized controlled trials, randomized intervention studies, and chamber studies. The main distinguishing feature of controlled-exposure studies relative to observational epidemiology studies is that the exposure is determined by the investigator (similar to an animal toxicology study). Therefore, many of the considerations relevant to evaluating animal toxicology studies in Table F-7, and in particular considerations related to exposure, apply to the evaluation of human controlled-exposure studies. Many of the same study features and considerations outlined for observational studies, in particular those related to study population, are also relevant for controlled exposure studies (see Table F-6). It is also important to consider the informed consent and other human subjects research ethics procedures undertaken in these studies, relative to the ethical standards prevailing at the time the research was conducted.

DOCUMENTATION OF STUDY QUALITY EVALUATIONS

The method for documenting information on study features that inform study quality may vary depending on the size and characteristics of the epidemiology or toxicology database. For example, if only a small number of epidemiology studies are available, it may be sufficient to summarize methodologic details in a single table. For chemicals with a small number of animal toxicology studies of generally uniform study design and quality, it may be sufficient to describe the information relevant to evaluation of study quality in the text. For data-rich chemicals with a large number of epidemiology or toxicology studies, however, more detailed documentation in tables is recommended to allow the user to see at once the number and type of studies available, and the level of information available from each. Database tools have been developed for organization and management of this type of information (e.g., through LitCiter Lite or DistillerSR software), but additional testing and refinement is needed to establish their usefulness for IRIS assessments.

Options for displaying relevant study quality information from epidemiology and animal toxicology studies in tables are described in the sections that follow. These tables could be included in an appendix of an IRIS assessment. These tables serve to 1) document all the studies that were considered; 2) provide the means to identify and track how informative a given study was throughout the assessment process; and 3) document why some studies were not further considered in the assessment.

Once the study information has been recorded and evaluated, it may be useful to sort studies into "tiers" according to the level of information they provided. The considerations and judgments used to "tier" studies should be clearly and transparently documented.

Documentation of Observational Epidemiology Study Evaluations

Table F-8 is an example of a summary display of relevant information for observational epidemiology studies. The shading of specific cells represents those features for which a specific limitation was noted.

In some situations, the collection of studies may be divided based on the likelihood and the types of limitations or biases identified in the evaluation of study quality. A study in the top quality tier would typically use an appropriate study design, have high-quality measures of exposure and outcome, and use adequate methods to analyze and present results. These studies would be given the greatest consideration within the context of hazard identification. Studies of lower quality are limited in one or more other ways. Because the type(s) of limitation(s) noted in a study can influence the direction of bias, it may be important to further classify this group based on the type(s) of limitations identified. For example, one group may consist of studies where the limitation(s) are likely to result in an attenuated effect measure, such as studies for which the major limitation is a substantial amount of non-differential exposure misclassification. Another group may include studies with different types of limitations for which it is difficult to determine the likely direction (or likelihood) of bias. Another group may include studies where the limitation(s) were considered to be likely to result in observation of a spurious association, such as a study that did not control for a known risk factor for the disease that was also strongly related to the exposure in the study.

Table F-8. Evaluation of Observational Epidemiology Studies of Chemical X.

Reference, Setting and Design	Participants, Selection, Follow-up	Comparability	Exposure Measure and Range	Outcome Measure	Consideration of Likely Confounding	Analysis and Presentation of Results (Estimate and Variability)	Sample Size; Power	Evaluation of Major Limitations
Lee et al., 1995 US (New York) chemical X production plant (cohort)	All men, age at baseline not reported; duration ≥ 12 months (mean 2.2 years), worked at plant 1960 – 1972 - plant operations began in 1945. Follow-up through 1990, 2% loss to follow-up, mean follow-up time 32 years	External (state mortality rates) referent; age and time-period matched (5 year groupings). Healthy worker effect seen for CVD (SMR 0.7) and all cancers (SMR 0.9). Internal referent: "no" exposure group	Exposure based on job records and personal/air monitoring; cumulative exposure calculated based on summations across all jobs (duration times average exposure)	Mortality (death certificates, ICD-8 and 9, underlying and contributing causes of death)	External comparison: use of age and time-period matched mortality rates.	SMR and 95% CI	Brain cancer: 4 obs cases	Low statistical power; not an inception cohort (had to "survive" to 1960 to be included)
Johnson et al., 1996 US (24 states) (case-control)	All deaths 1984-1986. Controls (died of causes other than cancer; frequency matched by age, sex, state and race)	Matching procedures for cases and controls	Death certificate occupation data; job exposure matrix developed to assess 11 chemical exposures	Mortality (death certificates, ICD-9), underlying cause of death)	Sex-specific odds ratios adjusted for marital status, race, socioeconomic status (3-levels), age at death	OR and 95% CI	10,540 cases, 42,160 controls	Non-differential exposure misclassification likely, particularly for women (lower quality occupation data for women)

Documentation of Animal Toxicology Study Evaluations

Study quality evaluation requires an analysis and documentation of the six categories of study features described above. An example tabular documentation of study quality features for animal toxicology studies is provided in Table F-9. Because the delivery of exposures in inhalation toxicology studies is complex, it may be advisable to develop a separate table (as shown in Table F-9a) that documents exposure quality in greater detail; the overall characterization of the exposure quality can be characterized in terms such as "robust" and "marginal." The quality of the exposure characterization is then incorporated into the broader evaluation of the other 5 previously described study features for each study (Table F-9b). It should be noted that this example was derived for an evaluation of a large and complex dataset; more simplified documentation is likely to be adequate for other types of datasets. Additional separate quality tables could be developed for other exposure routes or for other specific study features requiring more in-depth analyses (e.g., endpoint evaluation of neurotoxicity and respiratory pathology).

As previously described for epidemiology studies, a "tiering" system may be appropriate for categorizing animal toxicology studies according to aspects of study design, methods, and execution.

Review of General Principles: Study Quality Evaluation and Documentation

- To the extent possible, evaluation of a study is independent of consideration of the direction or magnitude of the study's results
- The goal is not necessarily to eliminate studies, but rather to understand potential limitations that would affect the interpretation of the results
- Record pertinent study details: what do you need to know about how the study was designed and conducted?
- "Tiering" can be useful to allow an easier flow of discussion during evidence synthesis and can transparently inform weight-of-evidence considerations
- Document judgments made regarding basis for "tiering" of studies

Table F-9. Example of a tabular documentation of study evaluation for a large dataset. This example includes important issues regarding inhalation exposure quality, and is broken into: (a) an evaluation of inhalation exposure quality; and (b) incorporation of this exposure quality analysis into a larger evaluation with the other 5 study features. The data are generalized and the endpoint is not specified. The results of this evaluation could be used to document an expert judgment that "Smith et al., 1984" is likely to be a more informative study (and "Gray et al., 2012," less) evaluating the endpoint in question.

(a) Evaluation of **inhalation exposure** quality

Reference (Species)	Test Article Characterization	Generation Method	Analytical Method	Analytical Concentrations	MMAD (GSD)	Chamber Type	Vehicle Control
Robust Exposure Characterization			Meet a robust standard for exposure quality				
Smith et al. (1984) (Monkey)	Test article (99%) solution in water	Bubble generator	Infrared spectrophotometry	Reported	Not applicable	Dynamic whole- body	Not needed
Marginal Exposure Characterization			Studies that meet a marginal standard for exposure quality. Key exposure data are missing				
Jones et al. (1986) (Mouse)	Solid test article (98.5%) ***Co-exposure likely***	Thermal depolymerization	Chromotropic acid	Reported	1.3 (1.7)	Dynamic nose-only	**No**
Poor Exposure Characterization			Studies that may be inadequate for exposure response but which may support other studies in informing hazard				
Gray et al. (2012) (Rat)	**Not reported**	**Not reported**	**Not reported**	**Not reported**	Not applicable	**Static**	**No**

Study deficiencies noted in **bolded text**.

(b) Evaluation of **all features** of animal toxicology study quality

Reference (Species)	Exposure Quality[a]	Test Subjects	Study Design	Toxicity Endpoints	Data and Statistics	Reporting
Smith et al. (1984) (Monkey)	++	++ Note: N=20	++ Note: 102 wk study	++	Not applicable	++
Jones et al. (1986) (Mouse)	+ co-exposure likely	+ N=5; variable ages at onset of exposure across groups	++ Note: 13 wk study	Potential sampling bias; No observer blinding indicated; protocols incompletely reported	+ data represents pooled sexes	+ Surgical procedures not reported
Gray et al. (2012) (Rat)	Test article and exposure methods not specified	Bacterial infection noted in animal colony; N= 3 litters; males only; overt maternal toxicity	No randomization across litters into treatment groups; testing during exposure expected to confound results; acute exposure	++	Not applicable	Results data not reported

[a] Summary results from inhalation exposure quality analysis in " Table F-9a." Criteria for the 6 categories developed based on the chemical and hazard type in question. In this example: gray box = examination of relevant study details identified potential limitations that could influence interpretations of the study's results; '+' = criteria not completely met or potential issues identified, but unlikely to directly affect study interpretation; ++ = criteria determined to be completely met. Text accompanying summary table would explain key study details informing these determinations.

Study Quality References

Klimisch HJ, Andreae M, Tillmann U. (1997) A systematic approach for evaluating the quality of experimental toxicological and ecotoxicological data. Regul Toxicol Pharmacol 25(1):1-5.

- [Presents an approach to systematically evaluating the quality of animal toxicology data and their use in hazard and risk assessment.]

NTP (National Toxicology Program). (2012) Protocol: evaluation of cancer studies in experimental animals. U.S. Department of Health and Human Services. Available online at www.ntp.niehs.nih.gov/NTP/roc/thirteenth/Protocols/PCP_animalcancer_508.pdf

- [Presents the protocol for cancer assessment of animal studies for the NTP's Report on Carcinogens Monograph on pentachlorophenol. Appendix C is a particularly useful section: Assessment of the quality of the individual animal cancer studies. Various study performance elements are described as they pertain to evaluating study quality.]

OECD guidelines and guidance documents are the standard for toxicology study quality. Available online at http://www.oecd-ilibrary.org/environment/oecd-guidelines-for-the-testing-of-chemicals-section-4-health-effects_20745788

U.S. EPA (Environmental Protection Agency). (1993) Reference dose (RfD): description and use in health risk assessments. Available online at http://www.epa.gov/iris/rfd.htm

- [Describes the EPA's principal approach to and rationale for assessing risk for health effects other than cancer and gene mutations from chronic chemical exposure. Section 1.3.1.1.6 (Quality of the study) provides an overview of the types of factors that are generally considered while making determinations pertaining to study quality.]

U.S. EPA (Environmental Protection Agency). (1994) Methods for derivation of inhalation reference concentrations (RfCs) and application of inhalation dosimetry. Environmental Criteria and Assessment Office, Research Triangle Park, NC; EPA/600/8-90/066F. Available online at http://cfpub.epa.gov/ncea/cfm/recordisplay.cfm?deid=71993.

- [Discusses criteria and information to be considered in selecting key studies for RfC derivation. Appendix F is a particularly useful section: Criteria for Assessing the Quality of Individual Animal Toxicity Studies.]

U.S. EPA (Environmental Protection Agency). (1998) Guidelines for neurotoxicity risk assessment. Risk Assessment Forum, Washington, DC; EPA/630/R-95/001F. Available online at http://www.epa.gov/raf/publications/pdfs/NEUROTOX.PDF

- [Summarizes the procedures that EPA uses in evaluating the potential for agents to cause neurotoxicity.]

U.S. EPA (Environmental Protection Agency). (1996) Guidelines for reproductive toxicity risk assessment. Risk Assessment Forum, Washington, DC; EPA/630/R-96/009. Available online at http://www.epa.gov/raf/publications/pdfs/REPRO51.PDF

- [Summarizes the procedures that EPA uses in evaluating the potential for agents to cause reproductive toxicity.]

U.S. EPA (Environmental Protection Agency). (1991) Guidelines for Developmental Toxicity Risk Assessment. Risk Assessment Forum, Washington, DC; EPA/600/FR-91/001. Available online at http://www.epa.gov/raf/publications/pdfs/DEVTOX.PDF

- [Summarizes the procedures that EPA uses in evaluating the potential for agents to cause developmental toxicity.]

U.S. EPA (Environmental Protection Agency). (2005a) Guidelines for Carcinogen Risk Assessment. Risk Assessment Forum, Washington, DC; EPA/630/P-03/001F. Available online at http://www.epa.gov/raf/publications/pdfs/CANCER_GUIDELINES_FINAL_3-25-05.PDF

- [Summarizes the procedures that EPA uses in evaluating the potential for agents to cause cancer.]

U.S. FDA (Food and Drug Administration). (1982) Toxicological Principles for the Safety Assessment of Food Ingredients (also known as Redbook 2000). Bureau of Foods (now Center for Food Safety and Applied Nutrition), Washington, DC. Available online at www.fda.gov/food/guidancecomplianceregulatoryinformation/guidancedocuments/foodingredientsandpackaging/redbook/default.htm

- [The U.S. Food and Drug Administration published this as guidance to industry and other stakeholders regarding toxicological information submitted to its Center for Food Safety and Applied Nutrition. The Redbook is an alternative resource wherein EPA scientists may find recommendations that are useful in evaluating animal toxicology studies.]

WHO (World Health Organization). (2012) Guidance for Immunotoxicity Risk Assessment for Chemicals. Harmonization Project Document No. 10. Available online at www.who.int/ipcs/methods/harmonization/areas/guidance_immunotoxicity.pdf

- [A comprehensive immunotoxicity resource that includes useful information on aspects of evaluating immunotoxicity studies in humans and animals.]

[Specific guidance on inhalation testing and reporting can be found in OECD Guidance Document 39 (GD 39): http://search.oecd.org/officialdocuments/displaydocumentpdf/?cote=env/jm/mono(2009)28&doclanguage=en.]

REPORTING STUDY RESULTS

Once a literature search has been conducted and the resulting database of primary (i.e., original research) studies have been evaluated with respect to strengths and limitations, the next step is to display the results in a form that facilitates perusal, review, and synthesis. Most applications to date have been with tabular display of results; however, presentations should not be limited to this type of display and in some situations, particularly for large collections of data, a graphical or some other type of figure may be a better choice.

Evidence Tables

Evidence tables present information from the collection of studies related to a specific outcome or endpoint of toxicity; for example, an evidence table for liver toxicity may include studies which evaluated changes in liver enzyme levels or diagnosis of liver disease (epidemiology studies), or increased liver weight or histopathological effects (animal studies). Included in the table are the studies which have been judged adequate for hazard identification following the principles outlined in Step 3 (see 4.3, *Reporting Study Results* of the IRIS *Preamble*). Evidence tables display findings of informative studies evaluating a relevant exposure scenario (taking into consideration route, timing, and dose). A key point is that evidence tables display the available study results, and are not restricted to those which observed 'statistically significant' or 'positive' associations.

The studies considered to be informative will depend on the extent and nature of the database for a given chemical, but may encompass a range of study designs and include epidemiology, toxicology, and, other toxicity data when appropriate. Consequently, evidence tables may be organized differently, when compared across assessments, depending on the data at hand; for example, it may make sense to organize studies by route and duration of exposure, or by specific endpoints within a toxicity type. If the database is extensive, the evidence tables may be organized into two or more tiers based on the relevance and quality of the studies applied in the hazard determination. Below are templates for evidence tables summarizing findings of observational epidemiology and animal toxicology studies; as noted, these should not be considered fixed structures, and may be adapted to best suit the database for a given chemical.

Human Evidence Tables:

Table F-10. Template for Reporting Results From Observational Epidemiology Study

Reference and Study Design	Results
Outcome	
[Reference] (location) Study design, time period, description of study population (including sample size), Exposure assessment and estimates Outcome measure Related references (i.e., earlier publications of a cohort with exposure measurement details)	Prevalence of outcome (if applicable, i.e., cohort studies) Prevalence of exposure (if applicable, i.e., case-control studies) Effect estimates (and variability measure (e.g., Beta and standard error, odds ratio and 95% confidence interval) Include results from analysis of exposure as a continuous measure and a categorical measure [if applicable)]

When constructing the evidence tables for human studies the following should be considered:

Study Design

- Order of the presentation of data in 'Study Design' column is flexible but must be consistent throughout tables in document.
- Study size may be the overall number of participants, or preferably, the number in each exposure or outcome group.
- Description of comparison groups may include population from which they were selected and prevalence of important potential confounders relevant to the endpoint of concern (e.g., % male, mean age, % smokers).
- Exposure estimate format will vary according to study; it is helpful to have some measure of both average (such as median) and upper end (such as 90th percentile) in each comparison group (such as exposed and unexposed, or cases and controls).
- If multiple dose metrics are provided (for example, both 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. It may be helpful to convert exposure metrics in order to compare results between studies; if so, provide the conversion calculation as a table footnote.

Results

- If few or no quantitative results are reported, a qualitative description of results may be provided using brief sentences or phrases.
- The effect measure(s) reported will depend on study design; for example, a mortality follow-up study may present standardized mortality ratios, while a case-control study may present odds ratios. Other examples include β coefficients from a regression model, risk ratios, or hazard ratios.

- As noted above, positive and negative results should be displayed regardless of 'statistical significance.' If available, however, there should be some indication of the variability in the result (such as a 95% confidence interval).
- Include whether or how potential confounding variables were considered or adjusted for in the analysis.

Animal Evidence Tables:

Table F-11. Template Option 1 for Reporting Results From Animal Toxicology Studies

Reference and Study Design	Results					
	Descriptor of Effect					
Reference	[effect] (*percent change compared to control*)					
species, strain, n /sex/group	M	0	5	10	15	20
doses (converted doses)		-	3%	7%	20%*	40%*
exposure route and details	F	0	6	12	17	23
age and duration of exposure		-	3%	7%	20%*	40%*
Species, strain, n /sex/group,	M	0	5	10	15	20
(describe the chamber type;		-	3%	7%	20%*	40%*
e.g., dynamic nose-only)	F	0	6	12	17	23
Exposure regimen (e.g., 6		-	3%	7%	20%*	40%*
h/day, 5 days/wk for 13 weeks)						
Test article (substance used to generate the atmosphere)						
Analytical concentrations (in mg/m^3; do not report target or nominal concentrations)						
MMAD (GSD): (aerosol only)						
Other critical information (e.g., sections of nasal turbinates examined)						

Percent change compared to control = (treated value – control value) ÷ control value x 100
* Statistically significant (p<0.05) based on analysis by study authors

Table F-12. Template Option 2 for Reporting Results From Animal Toxicology Studies

Reference and Study Design	Results			
	Descriptor of Effect			
Reference	[effect] (percent *change compared to control*)			
species, strain, n /sex/group	Male		Female	
doses (converted doses)	0	-	0	-
exposure route and details	5	3%	6	3%
age and duration of exposure	10	7%	12	7%
	15	20%*	17	20%*
	20	40%*	23	40%*

*Statistically significant (p<0.05) based on analysis of data conducted by study authors.
Percentage change compared to control = (treated value – control value) ÷ control value × 100.

When constructing the evidence tables for animal studies, the following should be considered:

Study Design

- The organization of the information in 'Study Design' column is flexible (i.e., species, duration, route) but must be consistent throughout tables in document.
- Details about species and number of test subjects should be presented as species, strain, n/sex/group.
- In the study design column, report administered doses, as specified in the study, and converted doses (when necessary) in mg/kg-d or mg/m^3. Do not adjust for intermittent dosing. In the results column, report converted doses only.
- Present average doses administered (converted from applied doses, using appropriate factors) [e.g., 0, 1.0, 2.5, 3.9 mg/kg-d]. Do not use 'or' or 'and' before last dose (i.e., not 0, 1.0, 2.5, and 3.9 mg/kg-d). Use at least two significant figures, except when presenting whole numbers (e.g., 0, 2.5, 5, 10, 112, 1,024).
- If converted doses are different in males and females present as: 0, 1, 2, 3 mg/kg-d in males; 0, 4, 5, 6 mg/kg-d in females.
- Provide the dose conversion calculation as a table footnote; note if authors reported the dose conversion.

Results

- If a study reports an effect but does not provide quantitative data, a qualitative description of the observed result must be provided, as a brief sentence or phrase. For example, "treatment-related histopathological changes were not observed".
- For continuous data, report the percent change compared to control (generally, round to whole percent unless one decimal point is needed).
- Provide the percent change formula as a table footnote: "(Treatment Mean – Control Mean)/ Control Mean". Decreases calculated in this manner will have negative signs to sufficiently describe the direction of the change in effect. Do not create confusion by including descriptions such as "Decrease in…" or "Increase in…" before the numerical value.
- For quantal data, present incidence and number at risk (e.g., 0/20, 5/20, etc), percent (if needed), and/or percent relative to control (if needed)
- Provide information for results that were not statistically significant but demonstrated an increase or decrease that was biologically relevant
- In specifying the effect, simply name the effect (e.g., Rotorod latency). Do not qualify the 'effect' with descriptions such as "change in..," "increase in…," or "decrease in…" (e.g. "liver weight," not "increased liver weight".) Also, do not use arrows to describe the direction of change in the effect observed.
- For statistical tests comparing treatment groups to control, remember to:
  - state when statistical analysis designations are based on analysis conducted by study authors;
  - document in study design cell when the study did not report statistical comparison to control; and/or
  - State in a table footnote when statistical tests are performed by EPA.

Evaluating the Overall Evidence of Each Effect

Hazard identification involves the integration of evidence from human, animal, and mechanistic studies in order to draw conclusions about the hazards associated with exposure to a chemical. In general, evidence is integrated in the context of Hill (1965), which outlines aspects — such as consistency, strength, coherence, specificity, does-response, temporality, and biological plausibility — for consideration of causality in epidemiologic investigations that were later modified by others and extended to experimental studies (U.S. EPA, 2005a).

All results, both positive and negative, of potentially relevant studies that have been evaluated for quality are considered (U.S. EPA, 2002). This requires a critical weighing of the available evidence (U.S. EPA, 2005a; 1994), but is not to be interpreted as a simple tallying of the number of positive and negative studies (U.S. EPA, 2002). Hazards are identified by an informed and expert evaluation and integration of the evidence. The sections that follow discuss evidence integration for human, animal, and mechanistic data with the ultimate goal of integrating across these evidence streams to answer the fundamental question of: **Does exposure to chemical X cause hazard Y?**

SYNTHESIS OF OBSERVATIONAL EPIDEMIOLOGY EVIDENCE

Focus of this section

Studies in humans may include epidemiologic studies, case studies, and, more rarely, controlled human exposure studies. While all of these study types may be included in an IRIS assessment, epidemiology studies are the predominant source of human evidence for most IRIS assessments. Therefore, this section is focused on the synthesis of evidence from epidemiology studies.

Evaluation of epidemiologic evidence

The synthesis of epidemiologic evidence and conclusions regarding summary descriptors focuses on whether and to what degree the collective evidence supports a conclusion that there is an *association* between the exposure and a health outcome. That is, the goal is to answer the question, "Is there evidence to conclude that an association or lack of an association exists between an exposure and a health outcome, for which reasonable alternative explanations (e.g., reverse causation, chance, bias, or confounding) are judged to be unlikely?"

The IRIS *Preamble* describes the framework for weighing the evidence from epidemiologic studies. The *Preamble* states that, "for each effect, the assessment evaluates the evidence from the epidemiologic studies as a whole to determine the extent to which any observed associations may be causal." While the *Preamble* refers to the concept of causality here, the evaluation of available studies involving humans constitutes one line of evidence in the process of drawing an overall conclusion regarding causality. In the context of an IRIS hazard evaluation, determinations of causality involve consideration of the weight of evidence from all available sources, including human, animal and mode of action (MOA) studies. Although a causal conclusion can be based on human evidence alone, evidence from animal and MOA studies can add weight to a less robust set of studies in humans.

Most epidemiologic studies used for risk assessment are non-experimental in design, in that the investigator generally does not control exposures or intervene with the study population. Broadly, epidemiologic studies are observational in nature and test specific hypotheses and evaluate associations between exposures and health outcomes. These analytical studies fall into several categories: e.g., cross-sectional, cohort and case-control studies. Each study design can make an important contribution to an overall conclusion regarding an association, although any particular design will have a specific interpretation with regard to individual aspects of the weight of evidence evaluation. For example, a cross-sectional study may be less informative regarding the temporal relationship between exposure and a health outcome, but it can be highly informative about an association if the health response is immediate, rather than delayed. Case studies involving one or a small number of affected individuals highlight potential toxicity of an exposure but are the least informative in an overall evaluation of association. While controlled human exposure studies, like clinical trials, offer advantages because of their experimental design, they may be less informative for hazard evaluations focused on long-term low level exposures, or health outcomes that occur many years after an exposure occurred. Properly interpreted, all types of study designs may contribute to the weight of evidence concerning an association. The process of weighing the evidence from human studies builds on the conclusions regarding the quality of individual studies. Each study, including both those that do and do not show an association between exposure and health outcome, is evaluated for study quality and considered as part of the weight of evidence evaluation.

Aspects suggesting causality

This section discusses "aspects"[14] of an association that suggest causality, drawn from Hill (1965), elaborated by Rothman and Greenland (1998), and referred to in other risk assessment documents such as those developed by the Environmental Protection Agency (U.S. EPA, 1994, 2002, 2005a), U.S. Surgeon General (DHHS, 2004;) and the Committee on Evaluation of the Presumptive Disability Decision-Making Process for Veterans (Samet and Bodurow, 2008). The 1964 Surgeon General's report on tobacco smoking discussed criteria for the evaluation of epidemiologic studies, focusing on consistency, strength, specificity, temporal relationship, and coherence (HEW, 1964). These aspects of causality are briefly described in the *Preamble*, and in more detail here.

First, greater *strength of association* lends greater confidence that the association is not due to chance or bias. However, while an association may be of small magnitude (due to factors such as low potency or a low level of exposure in the study population), a widespread exposure could lead to a significant public health burden, as seen for air pollution and risk of cardiovascular disease. 'Strength' encompasses not only magnitude of the association, but statistical confidence in effect measure estimates. Higher precision, as reflected by narrow confidence bounds or smaller standard errors, also adds confidence in the observed association.

Second, *consistency* of the association across studies is another important weight of evidence consideration. Observing an association in different study types, study populations, and exposure scenarios makes it less likely that the association is due to confounding or other factors specific to a given study, or is confined to a specific susceptible population. Characterizations of

[14] The "aspects" described by Sir Austin Bradford Hill (Hill, 1965) have become, in the subsequent literature, more commonly described as "criteria." The original term "aspects" is used here to avoid confusion with "criteria" as it is used, with different meaning, in the Clean Air Act.

consistency should distinguish between heterogeneity of findings which may be explained (e.g., due to differences in populations, exposure measures, ranges of exposures, potential co-exposures, and other factors specific to the exposure and health outcomes under evaluation) and unexplained variability suggesting potentially spurious findings (White et al., in press). For example, one would not necessarily expect to find identical results of exposure to an endocrine disruptor among those exposed prenatally versus during adulthood. This difference in timing of exposure is an expected source of heterogeneity in findings, rather than a signal that the findings are 'inconsistent'. In addition, a group of studies should not be characterized as 'inconsistent' if the results are not all statistically significant, or if effect measures are of different magnitudes, but are predominantly negative, null or positive.

The third aspect of *specificity* refers to one (or a few) causes for one health outcome. This aspect draws on Koch's postulates for infectious causes of disease, but may be less relevant in other contexts. For example, many environmental exposures may have carcinogenic action, but all contribute to a single health outcome. Conversely, a single exposure may be linked to a range of health outcomes. Thus, specificity may lend greater confidence in an association when it exists, but should not detract from an association if it does not.

Temporality is generally agreed to be the only aspect which is necessary for an association to be causal. That is, the exposure must precede the health outcome. In terms of epidemiologic studies, temporality is often cited as a main weakness of cross-sectional study designs. However, in evaluating a body of evidence, other study designs which do inform temporality can lend strength to the group of studies as a whole.

The *biologic gradient* or *exposure-response relationship* is another aspect which lends confidence to an observed association. Observing incremental changes in the risk of a health outcome which correspond to incremental changes in the exposure of interest, is a powerful argument against a spurious association, since that would necessitate a third (uncontrolled) factor which changes in the same manner (direction and magnitude) as the exposure of interest. Although this aspect is sometimes interpreted to imply that a monotonic relationship is required, the true exposure-response curve may indeed be non-linear. In evaluating a body of epidemiologic studies, it may be that any one study only includes a portion of the range of exposure. Piecing together evidence from multiple studies may yield a fuller understanding of the response and the shape of the exposure-response curve over the full range of exposures. Similarly, an observed lack of response in any one study does not imply a lack of an association between exposure and a health outcome. This may be due to exposure misclassification, or the exposures in a study were below some threshold for response, or that the range of exposures was too narrow to differentiate between groups (White et al., in press).

The next group of aspects comprises biologic plausibility, coherence and analogy. These were originally separate (related) aspects as laid out by Bradford-Hill, but more recently are seen as variations of a common theme. Biologic plausibility, coherence and analogy are addressed when weighing the totality of evidence including human, animal and mode of action. Generally, the association between exposure and a health outcome should be consistent with (or not violate) known scientific principles or other existing information from epidemiology, toxicology, clinical medicine, or other disciplines. A difficulty in applying these aspects is the reliance on current information, or the 'state of the science.' Associations in the epidemiologic literature may be

observed well in advance of experiments being performed or insight into mechanism or mode of action, but confidence that an association exists is strengthened by these aspects.

The final aspect is the existence of *natural experiments*, occurring when environmental conditions change in such a way as to mimic a controlled experiment or randomized trial—such as a change in workplace standards which reduces occupational exposure, or change in medication use with the introduction, or withdrawal, of a drug from the market. When such a change in the exposure is followed by changes in the risk of a health outcome of interest, this provides greater confidence that an association exists.

As discussed in the U.S. EPA Integrated Science Assessments for particulate matter (U.S. EPA 2009) and carbon monoxide (U.S. EPA 2010), although these aforementioned aspects provide a framework for assessing the evidence, they do not lend themselves to being considered in terms of simple formulae or fixed rules of evidence leading to conclusions about causality (Hill, 1965). For example, one cannot simply count the number of studies reporting statistically significant results or statistically nonsignificant results and reach credible conclusions about the relative weight of the evidence and the likelihood of causality (U.S. EPA 2009, 2010). Rather, these aspects are taken into account with the goal of producing an objective appraisal of the evidence, which includes weighing alternative explanations. In addition, it is important to note that the aspects of causality cannot be used as a checklist, but rather are used as a guide to help determine the weight of the evidence for inferring causality. (U.S. EPA 2009, 2010) In particular, not meeting one or more of the aspects does not preclude a determination of causality [(U.S. EPA 2009, 2010), see discussion in (CDC, 2004)]. Scientific judgment is needed to evaluate individual study quality and to weight the overall body of evidence.

<u>Evaluation of potential alternative explanation of observed epidemiologic associations</u>

In evaluating epidemiologic studies, consideration of many study design factors and issues must be taken into account to properly inform their interpretation and determine whether observed associations are likely to represent the truth or if there are reasonable alternative explanations (e.g. biases or other threats to internal validity). Such alternative explanations include "reverse causality" where the health outcome precedes exposures, chance, bias (selection bias and information bias) and confounding, and these alternatives are carefully considered in the evaluation of the aspects of causality and of the evidence as a whole.

As noted earlier, a logical time sequence (temporality) is an essential aspect of causality and ensures that "reverse causation" is unlikely. Chance can always be a potential explanation for the results in any collection of studies but is less likely as more studies are accrued that have similar observations across different settings, study designs and populations.

A further key consideration is evaluation of the potential effects of selection bias which may occur when study groups (exposed and unexposed, cases and controls) are not sufficiently comparable. Selection bias may alter epidemiologic findings when participation or follow-up rates are related to the probability of exposure and to the outcome of interest. For example, effect estimates that are based on a comparison of exposed workers to a general population (e.g., standardized mortality ratios) may be affected by a selection bias called the healthy-worker effect, because the baseline health of workers is typically better than the baseline health of the population as a whole. This type of selection bias could obscure a truly larger effect of toxicant exposure in analyses based on "external" comparisons with mortality in the general population. Although this

type of bias would not influence analyses using "internal" or matched comparison groups, other types of healthy worker effect bias should also be considered for these types of studies. Selection bias can lead to either an overestimate or underestimate of risk, and the potential direction and size of the bias must be considered when deciding whether individual studies are given more weight or less weight for a hazard evaluation. Studies where selection bias is less of a concern are typically given more weight.

Another key consideration is evaluation of the potential effects of measurement error which can lead to information bias. One example is the uncertainty associated with using surrogate exposure metrics to represent the actual exposure of an individual or population. This exposure measurement error can be an important contributor to variability in epidemiologic study results. Exposure measurement error can lead to misclassification (a type of information bias) that can under- or over-estimate epidemiologic associations between exposures and health outcomes, distort exposure-response relationships and widen confidence intervals around effect estimates (i.e. decrease precision). There are several components that contribute to exposure measurement error in epidemiologic studies, including the difference between true and measured concentrations and the use of average population exposure rather than individual exposure estimates. The importance of exposure misclassification varies with study design and is dependent on the spatial and temporal aspects of the available data. For a given set of epidemiologic studies informing a hazard evaluation, results from studies with more accurate exposure estimates (minimizing exposure misclassification) are given more weight, barring other serious design limitations (e.g., selection bias). Generally, exposure misclassification, when nondifferential, results in a bias toward the null and this is a potential explanation for relatively small effect estimates or for variability in results across studies with different degrees of exposure misclassification.

Confounding is a type of bias that leads to "… a confusion of effects. Specifically, the apparent effect of the exposure of interest is distorted because the effect of an extraneous factor is mistaken for, or mixed with, the actual exposure effect (which may be null)" (Rothman and Greenland, 1998). A confounder is a common cause of both the exposure and the health outcome—thus, it is associated with both the exposure and the health outcome, but is not an intermediary between the two. For example, confounding can occur between correlated toxicants (such as pesticides used in a mixture) that are also associated with the same health outcome. Knowledge of the broader literature on risk factors for the health outcome is important. Scientific judgment is needed to evaluate the likely sources and extent of confounding, together with consideration of how well the existing constellation of study designs, results, and analyses address this potential threat to inferential validity. The ability to statistically adjust for confounding in an epidemiologic study is dependent on the ability to identify and measure potential confounders. Consistency in reported effect estimates across multiple studies, conducted in various settings using different populations or exposures, can increase confidence that unmeasured confounding is an unlikely alternative explanation for the observed associations. Such consistency also reduces the likelihood of chance as an alternative explanation through the accumulation of a larger body of similar evidence, as noted above. The observations of exposure-response trends across different studies similarly reduce the likelihood that chance, bias, or confounding can explain the observed association. Studies in which confounding is a minimal concern are typically given more weight.

<u>Summary descriptors of epidemiologic evidence</u>

The considerations described above are consistent with guidelines for systematic reviews that evaluate the quality and strength of evidence. Confidence that a true association between exposure and a health outcome exists is increased if the effect estimates across multiple studies are judged to be consistent or when apparent inconsistencies may be explainable due to differences in study designs, populations studied, exposure concentration or timing, and/or issues of potential confounding, information bias and selection bias. Confidence is also increased if there is evidence of an exposure-response relationship or when the magnitude of effects is considered sufficient to conclude that a role of residual bias is negligible. Greater weight is given to the aspects of consistency, strength of association, temporality, and biologic gradient (exposure-response relationship) when assessing the epidemiologic evidence.

To make clear how much the epidemiologic evidence contributes to the overall weight of the evidence, the assessment may include a descriptor such as "*Sufficient epidemiologic evidence of an association consistent with causation*", "*Suggestive epidemiologic evidence of an association consistent with causation", "Inadequate epidemiologic evidence to infer a causal association*", or "*Epidemiologic evidence consistent with no association"* to characterize the epidemiologic evidence of each outcome. While each epidemiologic database is distinct and requires specific judgments be made on the relative merits of those studies, some examples of the constellation of the aspects of an association that suggest causality are provided below.

"*Sufficient epidemiologic evidence of an association consistent with causation*"

This descriptor is appropriate when the epidemiologic evidence is sufficient to establish an association between exposure and a health outcome for which reasonable alternative explanations, such as confounding, information bias and selection bias, are judged to be unlikely. Evidence of a consistent finding of an association between exposure and a health outcome along with evidence of an exposure-response relationship contribute considerable weight toward evidence of an association. Such evidence is increased when the association is relatively strong but may not necessarily be diminished when the observed associations are small in magnitude. Likewise, evidence of a coherent temporal relationship allowing for disease latency (where applicable) adds weight to this conclusion but the absence of such information does not necessarily detract from the conclusion.

"*Suggestive epidemiologic evidence of an association consistent with causation*"

This descriptor is appropriate when the epidemiologic evidence is suggestive of a causal association between exposure and a health outcome, but where there is less certainty that alternative explanations such as selection bias, information bias, and confounding, have been addressed. This descriptor covers a spectrum of evidence associated with varying levels of concern for a health outcome. Depending on the extent of the database, additional studies may or may not provide further insights.

An example of an aggregation of suggestive epidemiologic evidence might include apparent unexplained inconsistency of risks across studies with varying strength of the association but multiple studies reporting exposure-response relationships and a coherent temporal relationship allowing for disease latency. Another example of a constellation of suggestive epidemiologic evidence might include repeated observations of increases in risk across studies, especially for high exposures, but only a relatively modest overall strength of the association and limited evidence of an exposure-response relationship from one or more high quality studies.

"Inadequate epidemiologic evidence to infer a causal association"

This descriptor is appropriate when the epidemiologic evidence is judged inadequate for describing an association. An example of inadequate epidemiologic evidence might include explained heterogeneity of the observed increases in risk across studies with the majority of studies having relatively poor quality exposure assessment methodology and reporting null results contrasted with a single large high quality study with clear evidence of an exposure-response relationship.

Additional high quality studies generally would be expected to provide further insights. Additional supportive evidence demonstrating exposure-response relationships might lend more confidence that associations reported in epidemiologic studies are not due to alternative explanations, while additional evidence from other high quality studies showing a lack of exposure-response or that previous findings may be due to confounding might tip the balance in another direction.

"Epidemiologic evidence consistent with no association"

This descriptor is appropriate when the available data are considered robust for deciding that there is no basis for human hazard concern. An example of evidence suggestive of no association would include a consistent pattern of results indicating a lack of an association across a large number of studies that had adequate statistical power spanning different exposure patterns and exposure ranges, including high exposure levels, and evidence of the absence of exposure-response relationships even at high exposure levels.

SYNTHESIS OF ANIMAL TOXICOLOGY EVIDENCE

In IRIS assessments, human data are generally preferred for hazard identification because these data are more relevant in the assessment of toxicity to human health and avoid the uncertainty associated with potential interspecies differences when using animal data. However, many chemical databases contain little or no human data; thus, IRIS assessments frequently rely on available animal data in order to determine potential chemical hazards. In the absence of human data, well-conducted animal toxicology studies can support the identification of hazards. Animal data are used under the assumption that toxicity is conserved across species, in that effects observed in animals would be expected to occur in humans (U.S. EPA, 1998c; 1996; 1991). This section discusses how to approach synthesis of evidence from animal toxicology studies and focuses on whether and to what degree the collective evidence supports a conclusion that there is an association between chemical exposure and an effect.

In contrast to observational epidemiology studies that do not control exposures or intervene with the study population, experimental animal toxicology studies are designed to control exposure and environmental conditions. These studies permit the use of study design to control the number and composition (age, gender, species) of test subjects, the levels of doses tested, and the measurement of specific responses. Use of a designed study typically leads to more meaningful statistical conclusions than an uncontrolled observational study where additional confounding factors must also be considered for their impact on the conclusions. Thus, the observed responses in animals are expected to be due to chemical exposure. However, dose-response relationships observed in animal toxicology studies are often at much higher doses than would be anticipated for humans.

Animal toxicology studies fall into two broad categories: (1) general toxicology studies designed to evaluate a comprehensive array of endpoints following varying durations of exposure (e.g., chronic, sub-chronic, short-term, or acute) or (2) toxicology studies designed to evaluate a specific type of toxicity: e.g., neurotoxicity, immunotoxicity, reproductive toxicity, and developmental toxicity.

As noted in the *Preamble* and described in the Synthesis of Epidemiology Studies section, several aspects of causality discussed by Hill (1965) are pertinent to the interpretation of animal evidence: consistency of response, exposure-response relationship, strength of response, specificity of response, biological plausibility and coherence, and temporality (U.S. EPA, 2005a, 2002, 1994). These considerations, as they relate to synthesizing animal toxicology evidence in animals, are further described below.

Principles and Considerations for Writing a Synthesis of Animal Evidence

NOTE: In general these considerations apply to both human and animal data; however, for purposes of providing a simple example, this section is focused on animal evidence.

For each health effect, the evidence from animal experiments is evaluated to determine the extent to which this evidence indicates a potential for effects in humans. The starting points for a synthesis of the data for a given health effect (e.g., hepatic, immune system, cancer) are the following: (1) the evidence table(s) as developed in the Reporting Study Results subsection in the Evaluation and Display of Individual Studies section, (2) the actual papers or reports captured in the evidence tables, (3) information on study quality as documented in the Evaluation and Display of Individual Studies section, and (4) other information not summarized in the evidence table that contributes to the evidence of an association between exposure to the chemical and the given health effect. This other information could include short-term and acute experimental animal studies, and data from studies using routes other than oral, inhalation, or dermal. Keep in mind that while the evidence tables provide a useful framework for starting the evaluation, they are not sufficient for completing the synthesis. You will need the additional content provided in the papers and reports themselves to prepare the synthesis.

In general, the evidence table and accompanying synthesis text should be complementary, and provide a comprehensive and critical evaluation of the animal evidence. The synthesis should not be a text version of the information contained in the evidence table. For example, the synthesis text should not repeat study design details provided in the evidence tables, but should discuss strengths and limitations with studies (identified and documented in the Evaluation and Display of Individual Studies section) that would influence interpretation of the study results. To the extent possible, the information in the evidence table and text should be presented in the same order. However, it is important to remember that the synthesis should be a discussion from the perspective of the evidence for particular effects *across* studies, not by study, with the caveat that you will generally discuss evidence following oral, inhalation, or dermal exposures of chronic durations (i.e., more relevant to estimating potential toxicity to humans following chronic exposure to chemicals) prior to describing results in shorter duration studies. However, developmental and reproductive toxicology studies may provide pertinent evidence resulting from short-term exposures during a critical period of development.

There is no formula for writing a synthesis of the animal evidence. The approach for organizing the information will depend on the nature and extent of the literature for a given

chemical and health effect. Potentially relevant studies have been evaluated for quality (as identified and documented in the Evaluation and Display of Individual Studies section), and studies of higher quality are given more weight than those of low quality. This is mirrored in the development of evidence tables, which also capture findings from the most pertinent and higher quality studies without consideration of the presence or absence of an effect. All results, both positive and negative, are considered (U.S. EPA, 2002) and discussed.

In comparing and contrasting results across studies, evidence evaluations of study quality are further considered and discussed due to potential impacts on the interpretation of results (e.g., may explain differences in the results for a given endpoint). For example, could differences in test article preparation or delivery vehicle between studies account for differences in the reported results? Was the study adequately powered to identify an effect associated with a chemical exposure? Could co-exposures alter the response in one study versus another? Issues with study power, design, or conduct may limit the ability to draw conclusions about chemical-related effects when a single study is considered in isolation; when considered in the totality of studies that examine a given health effect these flawed studies can still add qualitative evidence for an effect associated with chemical exposure. Additionally, historical background levels of effects (if available) should be considered. While comparisons to concurrent controls are preferred when identifying effects, the use of appropriate historical control data may be informative when a particular effect is rare

The write-up should address the **consistency (including any lack of consistency) of the results** across studies. Consistent results across species, strains, sexes, life stages, routes of exposure, and exposure regimens and durations increases confidence that similar results would occur in humans. While consistency across higher quality studies is preferred, consistency of an effect across studies of varying quality and statistical power may provide qualitative information about a given effect. Inconsistency of effects among studies and/or species that cannot be explained by differences in timing and/or magnitude of exposure or toxicokinetics/metabolism can decrease confidence. As discussed in the *Preamble* (Section 5.2), distinguishing between ***conflicting evidence*** (that is, mixed positive and negative results in the same sex and strain using a similar study protocol) and ***differing results*** (that is, positive results and negative results in different sexes or strains or using different study protocols) is also important. Ask yourself why valid results are inconsistent and include information that could reconcile the differences in your evaluation. For example, did the "negative" study use an exposure range that was too low (e.g., were the highest exposures in the "negative" study similar to the range that produced no exposure-related response in the "positive" study)? Where you have a positive and negative study for a specific endpoint (e.g., neurotoxicity), investigate whether the negative study was adequately designed to look for that endpoint. Can differences in response be explained by differences in toxicokinetics across species? Refer to Agency guidance, where available, for additional information on evaluating specific health effects.

Your discussion should also indicate whether effects showed an **exposure-response relationship**, i.e., whether the incidence and/or intensity of response changes in an orderly manner as a function of exposure. Note, however, that the exposure-response relationship need not be monotonic. U-shaped (or inverted U-shaped) exposure-response functions are not uncommon in toxicology. In addition, information on the **strength (or magnitude) of the response** (in general terms) and the **exposure range where effects are first observed** should be provided. Confidence

in an association between a chemical exposure and a given health effect is increased when an exposure-response relationship is demonstrated and when the magnitude of effect is large. Also, be precise in characterizing the strength of the association between chemical exposure and effect. For example, the word "demonstrates" indicates a relatively strong association and should be used with caution. Words such as "suggests" or "indicates" are appropriately used when the evidence for an association is not as strong (e.g., an association based on a small number of studies or less consistent results).

If related effects in a target organ are observed (e.g., changes in serum enzymes that are markers of liver damage, increased liver weight, and liver histopathology), it is worthwhile to note the **coherence** of these related effects as well as **characterize the exposure ranges at which these effects** were observed. Coherence of the exposure ranges for related effects strengthens the biological plausibility for a given effect as well as provides a more complete picture of the toxicity associated with exposure to a chemical. For example, changes in liver enzymes are likely to occur at earlier time points and/or at lower exposures than histopathologic changes of the liver.

Another criterion that is important in interpreting data is the **temporal relationship** between exposure and effect. Temporality is generally assumed in animal toxicology studies since the exposure precedes measurement of effects. However, temporal considerations also need to be evaluated in the context of the observed effects. That is, the exposure should precede the effect at an interval that is consistent with what is known about the toxicokinetics and mode of action of the chemical. It may be the case, however, that higher exposures produce a shorter latency to effect than do lower exposures. Additionally, exposure-response relationships may vary due to temporal considerations. For example, if a study's dose groups result in premature mortality at higher doses, you may not observe an effect with increased frequency (or severity) at the higher doses because the animal died prior to the time needed to develop an effect with a long latency. Similarly, when considering cancer effects, the number of adenomas may decrease with increasing dose as they progress into larger tumors or progress to carcinomas. In some cases, initial effects may disappear as the pathogenesis of a lesion evolves or resolves (and early effects may not even be observed depending on the doses used and the resulting exposure-response relationship). How different studies illustrate the development of a lesion, weaving in considerations of the exposure-response relationship and temporality can increase or decrease the biological plausibility of a given effect.

Biological plausibility and coherence are also evaluated; although these aspects are best considered when integrating evidence across human, animal, and mechanistic evidence streams. Several types of information should be considered (e.g., toxicokinetics/metabolism, similarity of effects, exposure-response relationships, mode-of-action, and temporal relationships) when determining the likelihood of the occurrence of effects in humans based on observations in animals. All of this information must be weighed in light of the known heterogeneity of the human population versus the relatively inbred status of laboratory animals used in toxicology studies and housed under carefully controlled environmental conditions (U.S. EPA, 2002). These concepts are more fully described in the discussion of overall integration of evidence.

Additionally, in writing a synthesis of animal evidence, there are a couple of other considerations. Keep in mind that the evaluation is not a study by study summary of the literature. When discussing "significance," be clear as to whether you mean statistical or biological

significance[15]. Biological and statistical significance are both considered when making a judgment about the adversity of an observed effect. Where possible, emphasize biological significance over statistical significance, or be clear that biological significance is not well understood for a particular endpoint.

A Practical Example for Synthesizing Animal Toxicology Evidence

The following text example highlights some key considerations in writing synthesized text for a specific endpoint. Compare the text in "Draft 1" with the revisions made in "Draft 2."

> **Draft 1:** In several studies in rats and mice, decreased sperm count, motility, and production, and an increase in morphologically abnormal sperm have been observed. Decreased epididymal sperm count (approximately 50% at 1 mg/kg-day) and sperm motility (approximately 20% at 1 mg/kg-day) were observed in mice exposed by gavage to doses ≥ 1 mg/kg-day for 42 days prior to mating to unexposed females (Mohamed et al., 2010). This study also demonstrated transgenerational impacts on sperm parameters, as these endpoints were also decreased in the F1 and F2 generations produced from treated F0 males. Decreased epididymal sperm count (25%) and a slight increase in abnormal sperm morphology were observed in rats treated with 5 mg/kg-day benzo[a]pyrene by gavage for 84 days (Chen et al., 2001). A decrease in sperm motility (approximately 30%) and an apparent (but not statistically significant) decrease in epididymal sperm count (approximately 15%) were also observed in rats treated by gavage at 0.01 mg/kg-day for 90 days (Chung et al., 2011).
>
> Similar effects on sperm parameters have been observed in short term oral studies and inhalational studies. Significantly decreased sperm count, number of motile sperm, and daily sperm production (~40% decrease from control in each parameter) were observed following 10 days of gavage exposure to 50 mg/kg-day benzo[a]pyrene in rats (Arafa et al., 2009). In addition, decrements in sperm parameters (specifically sperm motility, sperm count, and percent morphologically normal sperm) were observed following inhalation exposure to benzo[a]pyrene in rats for 60 days to 75 µg/m3 (Archibong et al., 2008; Ramesh et al., 2008). In addition, decreased sperm motility, but not sperm count, was found to be decreased in rats exposed by inhalation to benzo[a]pyrene for 10 days at ≥ 75 µg/m3 (Inyang et al., 2003).

> **Draft 2:** In several studies in rats and mice, decreased sperm count, motility, and production, and an increase in morphologically abnormal sperm have been reported. Alterations in these sperm parameters have been observed in different strains of rats and mice and across different study designs and routes of exposure.
>
> Decreases in epididymal sperm (25 to 50% compared to controls) counts have been observed in SD rats and C57BL6 mice treated with 1- 5 mg/kg-day benzo[a]pyrene by oral exposure for 42 or 90 days (Chen et al., 2011; Mohamed et al., 2010).

[15] Biological significance is the determination that the observed effect (a biochemical change, a functional impairment, or a pathological lesion) is likely to impair the performance or reduce the ability of an individual to function or to respond to additional challenge from the agent (U.S. EPA, 2002).

Additionally, a 15% decrease in epididymal sperm count was observed at a dose two magnitudes lower in Sprague Dawley rats exposed to benzo[a]pyrene for 90 days (Chung et al., 2011). However, confidence in this study is limited as authors dosed animals with 0.001, 0.01, and 0.1 mg/kg-day benzo[a]pyrene but only reported on sperm parameters at the mid-dose. A short term study in mice and a subchronic inhalation study in rats lend support for the endpoint of decreased sperm count (Arafa et al., 2009; Archibong et al., 2008; Ramesh et al., 2008). Significantly decreased sperm count and daily sperm production (~40% decrease from control in each parameter) were observed following 10 days of gavage exposure to 50 mg/kg-day benzo[a]pyrene in mice (Arafa et al., 2009). In addition, decrements in sperm count were observed in rats following inhalation exposure to 75 μg/m^3 benzo[a]pyrene for 60 days (Archibong et al., 2008; Ramesh et al., 2008).

In addition to effects on sperm count, both oral and inhalation exposure of rodents to benzo[a]pyrene has been shown to lead to decreased epididymal sperm motility and altered morphology. Decreased motility of 20-30% compared to controls was observed in benzo[a]pyrene-exposed C57BL6 mice (≥ 1mg/kg-day) and SD rats (0.01 mg/kg-day) (Chung et al., Mohamed et al., 2010). The effective doses spanned two degrees of magnitude; however, as noted above, confidence in the study observing effects at 0.01 mg/kg-day benzo[a]pyrene (Chung et al., 2011) is limited by poor reporting. A short term oral study in mice also reported significantly decreased number of motile sperm (~40% decrease) following 10 days of gavage exposure to 50 mg/kg-day benzo[a]pyrene in mice (Arafa et al., 2009). In addition, decreased sperm motility was observed following inhalation exposure to 75 μg/m^3 benzo[a]pyrene in rats for 60 days (Archibong et al., 2008; Ramesh et al., 2008) and ≥ 75 μg/m^3 for 10 days (Inyang et al., 2003). Abnormal sperm morphology was observed in Sprague Dawley rats treated with 5 mg/kg-day benzo[a]pyrene by gavage for 84 days (Chen et al., 2001) and in rats exposed to 75 μg/m^3 benzo[a]pyrene by inhalation for 60 days (Archibong et al., 2008; Ramesh et al., 2008).

Note the following elements of the Draft 2 compared to Draft 1:

- Rather than providing the results of each study in a sentence (as in Draft 1), the Draft 2 text pulls together studies on the same effect (decreased sperm count, decreased motility, altered morphology) to provide a more integrated analysis of the results from multiple studies simultaneously.
- Information on the magnitude of the effect (or range of magnitudes of effect) is provided.
- Studies are generally organized by duration, with longer duration studies described first (within the limits of the available data).
- Study quality considerations are included in the discussion; in one instance, confidence in the findings is limited because the authors only reported effects in one mid-dose group.

While no single example could capture all the elements of a synthesized summary of evidence of animal toxicity, the above text highlights major considerations when writing up your synthesis of the animal toxicology data.

MECHANISTIC CONSIDERATIONS IN ELUCIDATING ADVERSE OUTCOME PATHWAYS

Mechanistic data contribute to the hazard evaluation of empirical evidence from human and animal studies by informing the following:

- The biological plausibility of a causal interpretation in humans
- The biological plausibility that animal experimental data is generalizable to humans
- The susceptibility of certain populations or lifestages

Evaluating mechanistic considerations is a critical part of weighing the evidence for hazard identification. The focus of this evaluation is on adverse outcome pathways (AOPs) that encompass both:

1) the toxicokinetic processes of absorption, distribution, metabolism, and excretion (ADME) that lead to the formation of the active agent and carry it through its distribution to the target cell, and
2) the toxicodynamic processes in the mode(s) of action (MOA[s]), leading to the adverse outcome.

While not a prescribed process—the database for every chemical and endpoint will be unique—the following steps may be informative in conducting the evaluation.

For each endpoint, the evaluation of AOPs begins by identifying:

- Information that may help identify the toxic moiety and the target site, and how the toxic agent is delivered to that site. Note that the target site at which the initial biological interaction occurs is not necessarily the site of the adverse effect.
- Information that may help identify key events in the hypothesized MOA(s).

This information may include both experimental and observational evidence specific to the chemical and endpoint, as well as additional evidence such as:

- Information on compounds that are similar in structure, function, and/or metabolism
- Information on how the chemical may disrupt normal biological processes or interacts with background aging or disease processes
- Interactions with other chemicals and/or mixtures
- Factors affecting biological susceptibility

Using this evidence, AOPs are described as sequences or networks of steps, from exposure to the chemical, formation of the active agent and delivery to the target site, and the key events leading to the adverse outcome.

Based on the evaluation of the available information, one or more of the following determinations may be possible:

- Whether there is sufficient information available to specify AOP hypotheses with respect to (1) and (2), above. In many cases the answer will be “no” to one or both of these due to lack of data.
- Whether the ADME and/or MOA data add to the biological plausibility of the hazard being evaluated.
- Whether a hypothesized AOP(s) is(are) sufficiently supported.

- Whether observed animal responses are generalizable to humans (i.e., human relevancy).
- Whether differences are anticipated in responses among humans, including susceptible subpopulations and lifestage-specific sensitivities.

While these determinations are qualitative in scope, the evaluation of AOP(s) should flag important quantitative information that may be carried over to dose-response analysis. These may include:

- Dosimetry for route-to-route extrapolation.
- Quantitative inter- or intraspecies differences in dosimetry.
- Quantitative inter- or intraspecies differences in response susceptibility.
- The shape of the dose-response relationship below the POD that may inform the choice of linear or non-linear extrapolation.

INTEGRATION OF EVIDENCE EVALUATION

[Note: EPA is investigating the use of standard descriptors to characterize the overall weight of the evidence for effects other than cancer. The NRC will hold a workshop in March 2013 on this topic, and EPA will follow up with a workshop to further develop this topic. In the meantime, the Preamble cites descriptors from EPA's 2005 Cancer Guidelines and, for effects other than cancer, the descriptors from EPA's Integrated Science Assessments.]

Dose-Response Analysis

SELECTING STUDIES FOR DERIVATION OF TOXICITY VALUES

For each health effect for which there is credible evidence of hazard, a group of studies has been identified and evaluated as part of the hazard identification (See Section on Evaluating the Overall Evidence of Each Effect). Once these studies have been identified, the basic criterion for selecting a subset for the derivation of toxicity values is whether the quantitative exposure and response data are available to compute a NOAEL, LOAEL or benchmark dose/concentration. When there are many studies, the assessment may focus on those that are more pertinent or of higher quality.

The relative merits of deriving toxicity values for each endpoint will depend on the size of the relevant database and various preferences as stated in the IRIS *Preamble* Section 6 as well as specific considerations appropriate for each chemical and health endpoint. All studies of sufficient quality (as evaluated in the Section on Evaluation and Display of Individual Studies) with data suitable for deriving toxicity values are considered (see the Benchmark Dose Technical Guidance sections 2.1.3, 2.1.5; U.S. EPA, 2012). All aspects of study quality evaluation for Hazard Identification are also important for dose-response (Tables F-6 and F-7). This section discusses **additional** considerations that are specific to dose-response analysis.

Table F-13. Attributes used to evaluate studies for derivation of toxicity values.

Aspects of Study	Data Characteristic	Considerations
Species studied	Human studies	Human data are preferred to reduce interspecies extrapolation uncertainties.
	Animal studies	Animal data are considered as supporting studies when adequate human studies are available, and as principal studies when adequate human studies are not available. Results from experiments using mammalian laboratory animals are favored over those conducted using non-mammalian species.
Relevance of exposure paradigm	Exposure route	Studies by a route of human environmental exposure are preferred, although 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. There are exceptions, such as when a susceptible population or life stage is more sensitive in a particular time window (e.g., developmental exposure).

Table F-13. Attributes used to evaluate studies for derivation of toxicity values.

Aspects of Study	Data Characteristic	Considerations
	Exposure levels	Studies with multiple exposure levels are preferred to the extent that they provide information about the shape of the exposure-response relationship (BMDTG 2.1.1).
Potential selection bias	Representativeness of the study sample to the target population and the potential for selection to be based jointly on both exposure status and disease status	In both cohort studies and case-control studies, higher participation rates are preferred. In cohort studies, higher follow-up rates are preferred. With lower participation (or follow-up rates), evidence for or against the potential for **differential** selection (e.g., greater participation of diseased among exposed compared with non-exposed) should be considered.
Potential confounding	A confounder is a common cause of both the exposure and the health outcome—thus, it is associated with both the exposure and the health outcome, but is not an intermediary between the two.	Studies with a design (e.g., matching procedures) or analysis (e.g., procedures for statistical adjustment) that adequately address the relevant sources of potential confounding for a given outcome are preferred.
Measurement of exposure	Standardized exposure assessment tools; validity and reliability	Studies are preferred that evaluate exposure during a biologically relevant time window for the outcome of interest, using higher quality exposure assessment methods that reduce measurement error. Measurement of exposure at the level of the individual is preferable to group-level exposures. Measurements of exposure should not be influenced by knowledge of health outcome status.
Measurement of health outcome	Standardized outcome assessment methods: validity and reliability	Studies that evaluate outcomes using generally accepted, standardized tools (e.g., disease classification systems, neuropsychological evaluation questionnaires) are preferred. Measurement or assignment of the outcome should not be influenced by knowledge of exposure status.
Power and precision	Numbers of test subjects and doses; experimental design	Preference is given to studies using designs reasonably expected to have power to detect responses of suitable magnitude.[3] This does not mean that studies with substantial responses but low power would be ignored, but they should be interpreted in light of a confidence interval or variance for the response.

Table F-13. Attributes used to evaluate studies for derivation of toxicity values.

Aspects of Study	Data Characteristic	Considerations

NOTES:

1 USEPA (2002), A Review of the Reference Dose and Reference Concentration Processes, EPA/630/P-02/002F (page 4-11).

2 Eliminating studies for which responses were not statistically significant will lead to bias toward larger effects. However, responses can be evaluated and weighted using standard errors or confidence intervals for the responses during hazard evaluation.[3] A judgment about endpoint and study 'sensitivity' or protectiveness can be made after dose-response modeling[5], in light of the range of candidate RfVs[5] and their precision and quality.

3 Power is an attribute of the design and population parameters; it cannot be inferred post-hoc using data from one experiment (Hoenig & Heisey, 2001, The American Statistician 55:19-24). Power is an ensemble property (based on a concept of repeatedly sampling a population) and is not a property of an individual study.

CONSIDERATIONS FOR COMBINING DATA FOR DOSE-RESPONSE MODELING

For most IRIS assessments, each POD has been derived based on data from a single study dataset. This is because in most cases, datasets are often **expected** to be heterogeneous for biological or study design reasons. Sources of potential heterogeneity include:

- Laboratory procedures used
- Population, species, and/or strain studied
- Sex
- Route of exposure

However, there are cases where one may consider conducting dose-response modeling after combining data from multiple studies, resulting in a single POD based on multiple datasets. For instance, this may be useful to increase precision in the POD or to quantify the impact of specific sources of heterogeneity.

Note that deriving toxicity values based on combining the results of multiple datasets **subsequent** to dose-response modeling of **each** dataset is discussed separately (see section 7.6 of *Preamble*, and associated draft *Handbook* text).

Examples of preliminary considerations as to whether are **potentially** suitable to derive a POD based on combining multiple datasets include the following:

a. *Sufficient quality for deriving PODs (see section 6 of Preamble, and associated draft Handbook text).* Note that statistical precision should not be a quality consideration for this question, as it can be automatically accounted by statistical weighting. Indeed, one of the reasons for considering combining datasets may be to increase overall precision.

b. *A common endpoint of concern reported.* Note that here "common endpoint" refers to the same specific outcome measurement, not just a common target site.

c. *A common measure of dose available.* PBPK models may be useful for estimating a common (internal) dose measure, particularly across routes of exposure.

d. *Comparable durations, given the nature of the endpoint.* Note that this may include exposure duration as well as observation duration (e.g., follow-up for cancer epidemiology).
e. *Evidence for homogeneous responses to dose*. Species and sexes often differ in response to dose, so convincing evidence would be needed to consider combining. A hypothesis test of no difference would not be convincing unless it has high power to detect a difference that matters (e.g., of the same magnitude as standard error of the mean).
f. *There is no one study that is clearly preferred.*

If potentially suitable datasets are available, then a statistician needs to be consulted to evaluate in more detail whether the datasets are appropriate for combining, and if so, what modeling approaches are appropriate to employ. Specific criteria for such evaluations will depend on the design of the underlying studies and the sources of potential heterogeneity.

CONDUCTING DOSE-REPONSE MODELING

[*Note: EPA has guidance addressing this topic. The draft Handbook will eventually contain more detailed information that summarizes the implementation of EPA's guidance in IRIS assessments.*]

DATA MANAGEMENT AND QUALITY CONTROL FOR DOSE-RESPONSE MODELING

The IRIS Program has developed tools and approaches to manage data and ensure quality in dose-response analyses. The objectives, described in more detail below, are to minimize errors, maintain a transparent system for data management, automate tasks, where possible, and maintain an archive of data and calculations used to develop assessments.

A. Objectives

1. Minimize Errors

Data (and metadata) should be entered into a database as early as possible in the process, verified, and "locked" to prevent accidental changes. Verification should be done either by double inspection or by double entry followed by machine comparison. Data should be entered once, before use in evidence tables (which require computations), and a subset of the same data will then be moved forward for calculating PODs (using dose-response analyses or tabulation of LOAELs and NOAELs) and for use in exposure-response arrays. Work after initial data entry and quality assurance (QA) should not involve any cut and paste operations. No calculations will be made except those recorded and retained transparently in the database. Later data entry or revision will be subject to the same QA process. Initial QA and later changes will be recorded and identified as to person responsible for data entry and data QA.

2. Transparent From Source to Result

Data entry is only one source of errors. Conversions and calculations, even simple ones can introduce errors; tracing such errors can be time-consuming.

The objective is to have data entered as reported by the source and then verified. Subsequent conversions and other calculations will be made transparently in a database. For example, if the source reported inhalation concentrations in ppm and

exposures of 6 hours/day, 5 days/week, for 78 weeks, then for cancers, (a) an average exposure would be calculated for a standard rodent lifetime as ppm × (6/24) × (5/7) × $(78/104)^3$, and (b) ppm would be converted to mg/m^3 using molecular weight and other quantities in a "dosimetry tool".

Also, the database should cite the source (reference) and the page(s) or table(s) or figure(s) from which the endpoint data were extracted.

This approach will enable ready verification (and quick revision when necessary).

3. Automate Tasks While Reducing Errors

Data should be maintained in a modern database management system (dbms) designed specifically for IRIS assessments to handle the types of endpoint data typically required. The dbms allows users to prepare data for dose-response analysis (including choice of BMRs), to execute analyses using Benchmark Dose Software (BMDS), and to marshal and organize the results for review and model selection. The dbms allows users to prepare custom reports and Microsoft (MS) Word tables and reports required for IRIS assessments in the new streamlined formats. The dbms allows users to prepare exposure-response arrays (figures) and import these into MS Word reports.

The dbms automates data processing, runs BMDS models, delivers results quickly, and requires minimal human intervention (after initial setup and QA of modeling choices and BMRs).

4. Accessibility: Retain and Archive Working Files and Data

Data in a dbms is easy to review and update. It is simple to re-run modeling for selected endpoints or an entire set of endpoints. Metadata should identify sources and data within sources unambiguously.

The dbms allows saving working files and data in a project "folder" when a project is suspended or completed, making it relatively simple to renew work or revise previous work. Project files can be shared among staff members working on a project (with locking of files in current use). Completed project files will serve as an archive to document work, including modeling decisions.

B. Current Database Management System and Excel Tools

The IRIS Program currently uses several tools for data management. These software tools are still being refined to satisfy all of the objectives outlined above. These software tools have been applied to several IRIS assessments, including dioxin, arsenic, and several other chemicals.

1. "BMDS Wizard"

The BMDS Wizard is an MS Excel-based tool that was designed to facilitate benchmark dose modeling when developing IRIS assessments. It handles one endpoint at a time in each MS Excel workbook. It expedites setting up BMDS modeling and automates running the various BMDS models. It includes forms for data and some other information about the study.

The BMDS Wizard expedites setting up modeling choices. A user selects models from a menu and the Wizard creates a table showing options for each of these models. One type of model can be selected several times, each for a different BMR value. The user

then reviews the options and BMRs in the table before asking the Wizard to automatically create the BMDS session file and BMDS model option files in a folder specific to the data set. This expedites setting up a BMDS run of multiple models; the table layout ensures better QA of modeling, option choices, and BMRs.

The Wizard then runs the chosen models and collects results in a single worksheet. It also reports a number of warnings and flags that can be used to review models and make a final selection of one model. Pop-ups reveal BMDS dose-response plots and tables of estimates and residuals. The warnings and flags are based on an included "logic" worksheet that can be modified by the user. The default logic worksheet includes model selection criteria recommended in EPA's Benchmark Dose Technical Guidance (U.S. EPA, 2012). The results worksheet makes it easy to compare models side by side and to document (in a comment column) any special reasons for rejecting and accepting models and any unusual situations.

Wizard also allows the user to request MS Word tables and plots for selected models, providing a well-organized summary of modeling results and model selection criteria. MS Word templates are provided for this purpose, and these can be modified readily as IRIS streamlined reporting requirements evolve.

2. "Dragon"

Dragon is a custom database management system (dbms) built in MS Access. Dragon can be used for all data pertaining to an IRIS assessment. Dragon is still being improved using feedback from IRIS users. Dragon works with several other software tools: BMDS Wizard, Dosimetry Tool, and Exposure-Response Array software.

Dragon allows for: data entry; QA, review; data transformations; dosimetry calculations (using the Dosimetry Tool); BMDS modeling for selected data (using the Wizard) and collection of modeling results and model selection decisions; creating a variety of reports and MS Word tables; and generating "skeleton" chapters and appendices, with modeling results, for an IRIS assessment. It is also designed to make MS Word reports, suitable as study summary tables and evidence tables. A summary of Dragon features follows:

- Data are entered and viewed using forms.
- QA/QC of data entry is integrated into the tool.
- Data entry: Study quality, dose-response data, design and other metadata.
- Intermediate results: Dose conversions (dosimetry tool), BMD modeling (Wizard).
- Review of results: model and endpoint selection by user.
- Final results: Summary tables, Figures, MS Word reports (Tox Review and Appendix).
- Customized for IRIS assessment requirements.
- Easy to set up model runs: creates BMDS session and option files for the model run.
- Organizes results for review: one model per line, all stats; flags problems.
- Identifies best model(s) using established criteria (user-modifiable).
- Writes IRIS assessment tables (Ch. 2 *Dose-Response Analysis* and modeling appendices in supplemental information of the IRIS assessment), Exposure-Response Arrays.
- Greatly reduces time to complete and report dose-response analyses.
- Will incorporate controlled nomenclature for endpoints.
- Flexible import and export capability, allowing data exchange with other software.

DRAGON stores the following information:

- Chemical-specific information—including name, molecular weight, etc.
- Study-specific information—including citation, HERO ID, study quality, etc.
- Dose-Protocol information—including species/strain/sex, route of exposure, dosing protocol, etc.
- Dose information—including doses from PBPK modeling and dosimetric conversions.
- Endpoint information—including NOAEL/LOAEL and statistical significance of each dose-group.
- BMD information—including output from the BMDS Wizard.

3. Dosimetry Tool

A Dosimetry Tool was developed using MS Excel to overcome several challenges to reliably making dosimetric conversions:

- Simple calculations, but numerous values to keep track of
  - Default values are in multiple guidance documents
  - Equations are specific to endpoint and study type
- Transparency and documentation
  - Calculation methods vary by author
  - Reporting format is not consistent
- Missing study data
  - Sources of body weight or food and water consumption are difficult to locate.

Dosimetry Tool Capabilities:

- Makes dose conversions, consistently and transparently, that are easy to document. The tool determines the correct formulas and defaults based on user inputs
- Used for Provisional Peer Reviewed Toxicity Values (PPRTVs)
  - Used by multiple study authors in ~100 PPRTV assessments
  - Sent as a compact deliverable showing all the inputs and results
- Consistency across multiple authors
  - Everyone uses the same equations, defaults, and reporting formats
- Easy QA of inputs
  - Summary tables show all input data and defaults
  - Equations show step-by-step calculations
- All conversions for an assessment can be saved in one workbook
  - Formatted tables stand alone as supporting documentation
- Will make default conversions (oral and inhalation)
- For more complex dosimetry calculations, converted doses can be entered into the tool (for reproductive studies, etc.)
- Can use study-specific information on body weight, inhalation rate, food and water consumption rates.

EXTRAPOLATION TO LOWER DOSES AND RESPONSE LEVELS

[*Note: EPA has guidance addressing this topic. The draft Handbook will eventually contain more detailed information that summarizes the implementation of EPA's guidance in IRIS assessments.*]

CONSIDERING SUSCEPTIBLE POPULATIONS AND LIFESTAGES

[*Note: EPA has guidance addressing this topic. The draft Handbook will eventually contain more detailed information that summarizes the implementation of EPA's guidance in IRIS assessments.*]

DEVELOPING CANDIDATE TOXICITY VALUES

[*Note: EPA has guidance addressing this topic. The draft Handbook will eventually contain more detailed information that summarizes the implementation of EPA's guidance in IRIS assessments.*]

CONSIDERATIONS FOR SELECTING ORGAN/SYSTEM-SPECIFIC OR OVERALL TOXICITY VALUES

The assessment derives or selects an organ/system-specific toxicity value for each organ or system affected by the agent. The assessment explains the rationale for each organ/system-specific toxicity value (for example, based on the highest quality studies, based on the most sensitive outcome, or based on a clustering of values). 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 (U.S. EPA, 2002).

Given multiple candidate toxicity values for a particular organ or system, each candidate value should be evaluated with respect to the multiple considerations:

- **Strength of evidence of hazard for the health effect or endpoint.** All other considerations being equal, effects and endpoints with stronger evidence of a causal relationship are preferred.
- **Attributes previously 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 Section for Selecting Studies for Derivation of Toxicity Values). All other considerations being equal, studies of higher quality when evaluated according to these attributes are preferred.
- **Basis of the POD.** All other considerations being equal, a modeled benchmark dose (BMD) 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 dose.
- **Other uncertainties in dose-response modeling.** These include the uncertainty in the BMD (e.g., reflected in the BMD/BMDL ratio) and uncertainty due to poor model fit.
- **Uncertainties due to other extrapolations.** All other considerations being equal, toxicity values for which other extrapolations are less uncertain are preferred. Note that the size of the composite uncertainty factor may **not** be a good indication of the remaining uncertainty, because some "uncertainty factors" overlap with aspects are already addressed separately above (e.g., study population/species, use of a LOAEL as opposed to a NOAEL). Therefore,

to avoid double-counting, the remaining uncertainties that are discussed should be explicitly enumerated.

Based on the results 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.
- Based on deriving a "composite" value supported by multiple candidate toxicity values that protects against toxicity in the given organ or system. One should carefully document how the supporting candidate toxicity values are selected and how the composite value is derived.

The assessment then selects an overall reference dose and an overall reference concentration for the agent to represent lifetime human exposure levels where effects are not anticipated to occur. This is generally the most sensitive organ/system-specific toxicity value, though consideration of study quality and confidence in each value may lead to a different selection.

CHARACTERIZING CONFIDENCE AND UNCERTAINTY IN THE TOXICITY VALUES

[*Note: EPA has guidance addressing this topic. The draft Handbook will eventually contain more detailed information that summarizes the implementation of EPA's guidance in IRIS assessments.*]

SELECTING FINAL TOXICITY VALUES

[*Note: EPA has guidance addressing this topic. The draft Handbook will eventually contain more detailed information that summarizes the implementation of EPA's guidance in IRIS assessments.*]