EPA/635/R-17/486
IRIS Assessment Protocol
www.epa.gov/iris

Systematic Review Protocol for the IRIS Chloroform Assessment (Inhalation)

[CASRN 67-66-3]

January 2018

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

DISCLAIMER

This document is a **Preliminary Materials Draft**. This information is distributed solely for the purpose of pre-dissemination 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 clarity and science policy implications. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

CONTENTS

TABLES

FIGURES

ABBREVIATIONS

ADME	absorption, distribution, metabolism, or elimination
BMD	benchmark dose
BMDL	benchmark dose lower confidence limit
$BW^{3/4}$	body-weight scaling to the $^{3/4}$ power
BMDS	Benchmark Dose Software
CAA	Clean Air Act
CAS	Chemical Abstracts Service
CASRN	Chemical Abstracts Service Registry Number
CERCLA	Comprehensive Environmental Response, Compensation, and Liability Act
CI	confidence interval
COI	conflict of interest
EPA	Environmental Protection Agency
GLP	good laboratory practices
HAP	Hazardous Air Pollutant
GRADE	Grading of Recommendations Assessment, Development and Evaluation
HAWC	Health Assessment Workspace Collaborative
HEC	human equivalent concentration
HERO	Health and Environmental Research Online
IAP	IRIS Assessment Plan
IPCS	International Programme on Chemical Safety
IRIS	Integrated Risk Information System
ITER	International Toxicity Estimates for Risk
IUR	inhalation unit risk
LOAEL	lowest-observed- adverse- effect level
LOEL	lowest observed effect level
MeSH	Medical Subject Headings
MOA	mode of action
NCEA	National Center for Environmental Assessment
NMD	normalized mean difference
NOEL	no observed effect level
NTP	National Toxicology Program
NOAEL	no-observed- adverse- effect level
OAR	Office of Air and Radiation
OECD	Organization for Economic Co-operation and Development
OLEM	Office of Land and Emergency Management
ORD	Office of Research and Development
OSF	oral slope factor
PBPK	physiologically based pharmacokinetic
PECO	Populations, Comparators, Exposures, Outcomes
PK	pharmacokinetic
POD	point of departure
RfC	reference concentration
RfD	reference dose
ROBINS-I	Risk of Bias in Non-Randomized Studies of Interventions
SD	standard deviation
SEM	standard error of the mean
UF	uncertainty factor

AUTHORS | CONTRIBUTORS

Assessment Team

Ted Berner (Chemical Manager)	U.S. EPA/ORD/NCEA
Audrey Galizia	U.S. EPA/ORD/NCEA
Amanda Persad	U.S. EPA/ORD/NCEA
Michele Taylor	U.S. EPA/ORD/NCEA
Kristina Thayer	U.S. EPA/ORD/NCEA
Amina Wilkins	U.S. EPA/ORD/NCEA

Contributors and Production Team

Hillary Hollinger	HERO Librarian
Ryan Jones	HERO Director
Vicki Soto	Project Management Team
Dahnish Shams	Project Management Team
Maureen Johnson	NCEA Webmaster

Executive Direction

Tina Bahadori	NCEA Center Director
Mary Ross	NCEA Deputy Center Director
Emma Lavoie	NCEA Assistant Center Director for Scientific Support
Samantha Jones	NCEA Associate Director for Health (acting)
Kris Thayer	NCEA/IRIS Division Director
James Avery	NCEA/IRIS Deputy Director (acting)

1. INTRODUCTION

The Integrated Risk Information System (IRIS) Program is undertaking a reassessment of the health effects of chloroform via inhalation. IRIS assessments provide high quality, publicly available information on the toxicity of chemicals to which the public might be exposed. These assessments are not regulations, but provide a critical part of the scientific foundation for decisions made in EPA (Environmental Protection Agency) program and regional offices to protect public health.

Before beginning an assessment, the IRIS Program consults with EPA program and regional offices to define the scope of the assessment, including the nature of the hazard characterization needed, identification of the most important exposure pathways, and level of detail required to inform program and regional office decisions. Based on the scope, the IRIS Program undertakes problem formulation activities to frame the scientific questions that will be the focus of the assessment, which is conducted employing the principles of systematic review.

A draft assessment plan for chloroform was posted publicly and also presented at a September 27–28, 2017 Science Advisory Board Chemical Assessment Advisory Committee (SAB CAAC) public meeting to seek input from the scientific community and interested parties on the problem formulation components of the assessment plan. The draft assessment plan contains a summary of the IRIS Program's scoping and problem formulation conclusions; the objectives and specific aims of the assessment; draft Populations, Exposures, Comparators, and Outcomes (PECO) criteria; and identification of key areas of scientific complexity. The protocol then incorporates the elements of the assessment plan, but also presents more detailed methods for conducting the systematic review and dose-response analysis, including any adjustments made to the specific aims and PECO in response to public input into the assessment plan. While the IRIS Assessment Plan describes *what* the assessment plans to cover, chemical-specific protocols describe *how* the assessment will be conducted (see Figure 1). The IRIS Program posts assessment protocols on its website and considers public input while preparing the draft assessment. Major updates to the protocol (e.g., fundamental alterations to the PECO or addition of literature search results) will trigger release of a revised protocol document and an additional public comment period.

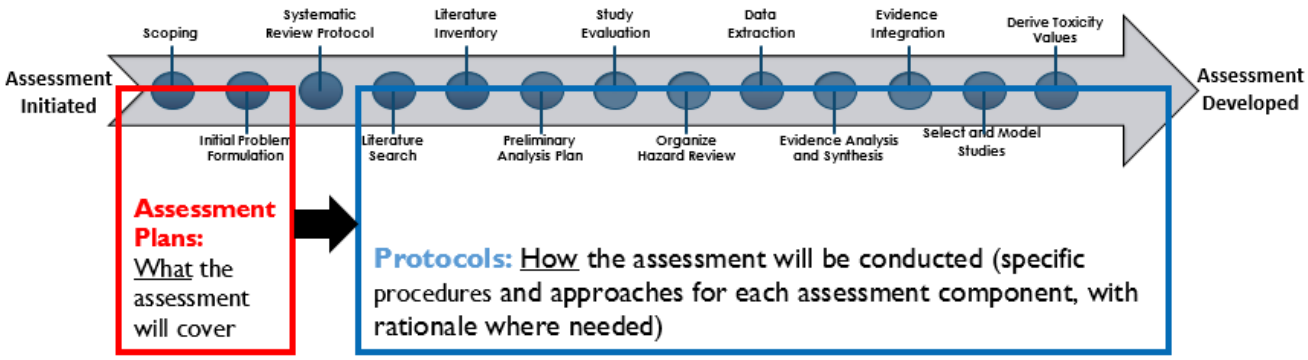

Figure 1. IRIS systematic review problem formulation and method documents.

1

2. SCOPING AND INITIAL PROBLEM FORMULATION SUMMARY

2.1. BACKGROUND

Chloroform (trihalomethane), or $CHCl_3$, is a colorless, volatile liquid at room temperature with a distinctive odor. Chloroform is nonflammable, slightly soluble in water, and readily miscible with most organic solvents. It was formerly used as an inhaled anesthetic during surgery until about 1950, but today, the primary use of chloroform is in industry and research labs, where it is typically used as a chemical intermediate and solvent, respectively. Because of its volatility, chloroform tends to escape from contaminated media (e.g., water or soil) into air. Therefore, humans are most commonly exposed environmentally to chloroform via inhalation (especially in indoor air) or through ingestion of chlorinated drinking water. Once inhaled or ingested, chloroform is rapidly absorbed and metabolized by cytochrome P450-dependent pathways. Metabolism occurs primarily in the liver, and to a lesser extent in the kidneys, and thus these organs tend to be the targets of chloroform toxicity.

An assessment of chloroform is currently available on the IRIS website and consists of (1) an inhalation assessment, (2) an oral assessment, and (3) a mode of action (MOA) analysis for cancer (https://cfpub.epa.gov/ncea/iris2/chemicalLanding.cfm?substance_nmbr=25). The inhalation assessment (posted in 1987) derived an inhalation unit risk (IUR) for chloroform of 2.3×10^{-5} per $\mu g/m^3$. This IUR was based on the incidence of liver tumors observed in an oral gavage study in mice that employed a route-to-route extrapolation without the use of a physiologically-based pharmacokinetic (PBPK) model.[1] This inhalation assessment did not include the derivation of a reference concentration (RfC) for chloroform. The oral assessment (posted in 2001) yielded a reference dose (RfD) for chloroform of 1×10^{-2} mg/kg-day based on liver effects in dogs. Also posted in 2001, the MOA analysis concluded that chloroform is likely carcinogenic to humans by all routes of exposure, but only under high-exposure conditions that lead to cytotoxicity and regenerative hyperplasia in susceptible tissues. Based on this MOA analysis, the RfD was determined to be protective with respect to cancer because, at the RfD, cytotoxicity—a key event in the MOA for cancer—was not observed. The inhalation assessment posted in 1987 was never updated to address the outdated route-to-route extrapolation approach employed or the more recent MOA analysis.

[1]Conducting a route-to-route extrapolation without the use of a PBPK model is no longer advocated by the EPA because of the potential inaccuracy of this methodology, especially when converting doses from the oral to the inhalation route of exposure.

As a result, the methodology used to derive the IUR posted in 1987 has two shortcomings: (1) it utilized a route-to-route extrapolation approach that did not employ a PBPK model, and (2) it incorporated a linear extrapolation approach for dose-response that implicitly assumes a risk of cancer at all nonzero exposures to chloroform (i.e., no threshold). The MOA analysis added in 2001, however, concluded that for cancer, chloroform exhibits a "threshold" by all routes of exposure. Thus, a chloroform dose exists that does not elicit cytotoxicity and presents no cancer risk. Therefore, the assumption underlying the IUR dose-response approach (linear extrapolation with no threshold) is inconsistent with the MOA analysis. These shortcomings, along with the absence of an RfC, present difficulties for EPA program offices and regions when trying to evaluate risks associated with chloroform exposure via inhalation. For example, the use of the IUR in establishing risk-based clean-up levels at several chloroform contaminated sites has been challenged by stakeholders. Thus, a specific need was identified to conduct a targeted update of the inhalation assessment for chloroform.

Exposure to chloroform from chlorinated drinking water is considered outside the scope of this assessment. Drinking water treated with chlorine typically contains chloroform, along with several other trihalomethanes, as well as a wide variety of other disinfection byproducts (U.S. EPA, 1994b). Chloroform is usually the predominant disinfection byproduct found in chlorinated drinking water, although some drinking water supplies subjected to high bromide levels can result in higher relative proportions of brominated disinfection byproduct species. Although exposure to chloroform in drinking water may result in inhalation of chloroform gas released from water into indoor air, epidemiological studies of disinfection byproducts are not considered pertinent to the current assessment because of unresolvable challenges in isolating any independent effects of chloroform because of co-exposures to other chemicals.

2.2. SCOPING SUMMARY

The chloroform inhalation assessment will be updated by deriving an RfC based on available inhalation data from human or animal studies and evaluating this RfC considering the MOA analysis posted in 2001 and addressing the inconsistency with the IUR. During scoping, the IRIS Program met with EPA program and regional offices that had an interest in an updated IRIS assessment for chloroform to discuss specific assessment needs. Table 1 provides a summary of input from this outreach. EPA's Office of Land and Emergency Management (OLEM), EPA's Office of Air and Radiation (OAR), and Region 4 expressed a specific need for an inhalation reference value for chloroform. Derivation of an RfC will address these program and regional office needs. In addition, the MOA analysis posted in 2001 will be used to determine whether this newly derived RfC is protective with respect to cancer, and if the IUR should be removed or updated. Finally, the derivation of the RfD, and the analysis that determined it was protective with respect to cancer, will

not be re-evaluated as part of this update to the chloroform assessment because EPA program and regional offices did not express a specific need for an updated RfD for chloroform.

2.3. PROBLEM FORMULATION

This assessment will consider all adverse effects elicited by inhalation exposure to chloroform for which data are available. After a preliminary review of the literature, the IRIS Program anticipates there will be fewer than 30 PECO-relevant studies, and the following health effects are likely to warrant inclusion in this assessment: nasal cavity effects, nervous system effects, liver and kidney effects, immune system effects, and reproductive or developmental effects.

Table 1. EPA program or regional offices interest in an updated chloroform assessment

Program or regional office	Oral	Inhalation	Statues/regulations	Anticipated uses/interest
OLEM		✓	CERCLA	Chloroform is listed as a hazardous substance under CERCLA. CERCLA authorizes EPA to conduct short- or long-term cleanups at Superfund sites and later recover cleanup costs from potentially responsible parties. Chloroform is commonly found at National Priorities List sites. Chloroform toxicological information developed for this assessment may be used to make risk determinations for response or remedial actions (e.g., short-term removals or long-term remedial response actions) at such sites.
Region 4[a]		✓		
OAR		✓	CAA	Chloroform is listed as a hazardous air pollutant (HAP) under Section 112 (42 U.S.C.§ 7412) of the CAA. Under CAA Section 112, 8 years after promulgation of standards requiring maximum achievable control technology, EPA must assess the remaining risk and revise the standards, if necessary. Chloroform toxicological information developed for this assessment may be used to inform these residual risk decisions.

CAA = Clean Air Act; CERCLA = Comprehensive Environmental Response, Compensation, and Liability Act; HAP = Hazardous Air Pollutant.

[a] Region 4 serves the states of Alabama, Florida, Georgia, Kentucky, Mississippi, North Carolina, South Carolina, Tennessee, and six Native American tribes.

2.4. ASSESSMENT APPROACH

The chloroform inhalation assessment will be updated by deriving an RfC based on available inhalation data in human or animal studies, and then evaluating this RfC considering the MOA analysis posted on the IRIS website in 2001. The results of this evaluation is anticipated to result in a new RfC that would replace the existing IUR from 1987.

2.5. KEY SCIENCE ISSUES

No specific key science issues have been identified outside of those described in the background and scoping summary.

3. OVERALL OBJECTIVES, SPECIFIC AIMS, AND POPULATIONS, COMPARATORS, EXPOSURES, OUTCOMES (PECO) CRITERIA

The overall objective of this assessment is to identify adverse health effects and characterize exposure-response relationships for these effects of chloroform to support development of toxicity values for this chemical. More specifically, the objective of this assessment is to derive an RfC for chloroform by using inhalation dose-response data from human or animal studies, without the need for route-to-route extrapolation. In addition, the MOA analysis for cancer for chloroform posted on the IRIS website in 2001 will be used to determine whether this newly derived RfC is protective with respect to cancer. This evaluation is anticipated to result in a new RfC that would replace the existing IUR from 1987. This assessment will use systematic review methods to evaluate the epidemiological and toxicological literature for chloroform. The evaluations conducted in this assessment will be consistent with relevant EPA guidance.[2]

3.1. SPECIFIC AIMS

- Identify epidemiological (i.e., human), toxicological (i.e., experimental animal), and physiologically-based pharmacokinetic (PBPK) model literature reporting effects of exposure to chloroform via inhalation as outlined in the PECO.

- Use an iterative prioritization approach to determine which mechanistic studies may be considered for evaluation and synthesis, primarily focusing on mechanistic studies that (1) present evidence to challenge the existing 2001 MOA analysis for cancer, and (2) could inform remaining questions following the synthesis of human and animal evidence for determining potential hazards other than cancer.

- Conduct study evaluations (risk of bias and sensitivity) for individual epidemiological and toxicological studies. Studies considered uninformative will not be used for hazard identification or dose-response analysis. The suitability of identified PBPK models will also be evaluated.

- Extract data on relevant health outcomes from those epidemiological and toxicological studies considered most informative based on study evaluation.

[2]EPA guidance documents: http://www.epa.gov/iris/basic-information-about-integrated-risk-information-system#guidance/.

- Synthesize the evidence across studies, assessing similar health outcomes using a narrative approach or meta-analysis (if appropriate).

- For each health outcome, express confidence in conclusions from across studies (or subsets of studies) within human and animal evidence streams, evaluating each evidence stream (human and animal) separately.

- For each health outcome, integrate results across evidence streams (human and animal) to conclude whether a substance is hazardous to humans. Identify and discuss issues concerning potentially susceptible populations and lifestages. Biological support provided from mechanistic studies and non-mammalian model systems will be considered based on the iterative prioritization approach outlined in the PECO.

- Derive an RfC, as supported by the available data.

- After deriving an RfC, evaluate its protectiveness against cancer based on the 2001 MOA analysis. This evaluation is anticipated to result in a new RfC that would replace the existing IUR from 1987.

- Characterize uncertainties and identify key data gaps and research needs such as limitations of the evidence base, limitations of the systematic review, and relevance of dose and pharmacokinetic differences when extrapolating findings from higher dose animal studies to lower levels of human exposure.

3.2. POPULATIONS, COMPARATORS, EXPOSURES, OUTCOMES (PECO)

A PECO is used to focus the research question(s), search terms, and inclusion/exclusion criteria in a systematic review. The draft PECO for chloroform (see Table 2) was based on (1) nomination of the chemical for assessment, (2) discussions with scientists in EPA program and regional offices to determine the scope of the assessment that will best meet Agency needs, and (3) preliminary review of the health effects literature for chloroform (primarily reviews and authoritative health assessment documents) to identify the major health hazards associated with exposure to chloroform via inhalation and identify key areas of scientific complexity.

Table 2. Populations, comparators, exposures, outcomes (PECO) criteria for the chloroform assessment

PECO element	Evidence
Populations[a]	*Human*: Any population and lifestage (occupational or general population, including children and other sensitive populations). The following study designs will be considered most informative: controlled exposure, cohort, case-control, cross-sectional, and ecological. Note: Case reports and case series will be tracked during study screening, but are not the primary focus of this assessment. They may be retrieved for full-text review and subsequent evidence synthesis if no or few informative study designs are available. Case reports can also be used as supportive information to establish biologic plausibility for some target organs and health outcomes.
	Animal: Nonhuman mammalian animal species (whole organism) of any lifestage (including preconception, in utero, lactation, peripubertal, and adult stages).
Exposures	*Human*: Any exposure to chloroform, including occupational exposures, via inhalation. Exposures quantified by either actual exposure measurements or occupational exposure history are preferred. Studies of chloroform in the context of its use as an anesthetic gas will be excluded.
	Animal: Any exposure to chloroform via inhalation. Studies employing chronic exposures or short-term, developmental-only exposures will be considered the most informative. Studies involving exposures to mixtures will be included only if they include an arm with exposure to chloroform alone. Studies utilizing chloroform as an extraction solvent to isolate specific chemical constituents will be excluded.
	Studies describing physiologically-based pharmacokinetic (PBPK) models for chloroform will be included.
Comparators	*Human*: A comparison or referent population exposed to lower levels (or no exposure/exposure below detection limits) of chloroform, or exposed to chloroform for shorter periods of time.
	Animal: A concurrent control group exposed to vehicle-only treatment.
Outcomes	All health outcomes (both cancer and noncancer). In general, endpoints related to clinical diagnostic criteria, disease outcomes, histopathological examination, or other apical/phenotypic outcomes will be prioritized for evidence synthesis over outcomes such as biochemical measures. As discussed above, based on preliminary screening work, EPA anticipates that a systematic review for health effect categories other than those identified (i.e., nasal cavity effects, nervous system effects, liver and kidney effects, immunotoxic effects, and reproductive/developmental effects) will not be undertaken unless a significant amount of new evidence is found upon review of references during the comprehensive literature search.

[a] Evidence from in vitro, in silico, and other types of mechanistic studies will be prioritized based on likelihood to impact evidence synthesis conclusions for human health. For chloroform, mechanistic studies will only be considered for evaluation if they have the potential of impacting the existing 2001 MOA analysis, or are essential for answering questions identified during the human and animal evidence syntheses.

4. LITERATURE SEARCH AND SCREENING STRATEGIES

4.1. USE OF EXISTING ASSESSMENTS

A search for potentially relevant recent assessments was conducted. The following federal, state, and international organizations were searched: Agency for Toxic Substances and Disease Registry; National Institute for Occupational Safety and Health; National Toxicology Program; Occupational Safety and Health Administration; U.S. EPA, Office of Chemical Safety and Pollution Prevention; U.S. EPA, Office of Water; California EPA, Office of Environmental Health Hazard Assessment; New Jersey Department of Environmental Protection; Texas Commission on Environmental Quality; European Chemicals Agency; Environment Canada; Health Canada, International Agency for Research on Cancer; Netherlands National Institute for Public Health and the Environment; Public Health England; World Health Organization, and International Programme on Chemical Safety. A search of the International Toxicity Estimates for Risk database was also performed. The most recent chloroform assessment cited was a 2003 International Programme on Chemical Safety (IPCS) assessment, which supports the need for updated assessment.

4.2. LITERATURE SEARCH STRATEGIES

The literature search for this assessment focused on studies published since completion of the last literature search for chloroform conducted by the IRIS Program in January 2009 using EPA's Health and Environmental Research Online (HERO) database[3]. Health outcome studies identified from the January 2009 search were combined with the literature search results from the updated database search and screened for PECO relevance. The updated literature search focused only on the chemical name with no limitations on evidence streams (i.e., human, animal, in vitro, or in silico) or health outcomes. No language restrictions were applied. The detailed search strategy is presented in Appendix A. The databases listed below were searched using HERO for the date range of January 2009 through October 26, 2017:

- PubMed (National Library of Medicine)
- Web of Science (Thomson Reuters)
- ToxLine (National Library of Medicine)

[3]Health and Environmental Research Online: https://hero.epa.gov/hero/.

Additional relevant literature not found through database searching was identified through searching citations from key references. The literature search will be updated throughout draft development to identify literature published during preparation of the assessment. The last full literature search update will occur a few months before the planned release of the draft assessment for public comment.

The IRIS Program takes extra steps to ensure identification of pertinent studies by: encouraging the scientific community and the public to identify additional studies and ongoing research; searching for data submitted under the Toxic Substances Control Act or the Federal Insecticide, Fungicide, and Rodenticide Act; and considering late-breaking studies that would impact the credibility of the conclusions, even during the review process.[4] Studies identified after peer review begins will only be considered for inclusion if they are PECO relevant and fundamentally alter the assessment's conclusions.

4.3. UNPUBLISHED DATA

IRIS assessments include only publicly accessible, peer-reviewed information. However, it is possible that unpublished data directly relevant to the PECO may be identified during assessment development. In this case, if these data would likely make a substantial impact on assessment decisions or conclusions, EPA can conduct an external peer review of this information if the owners of the data are willing to have the study details and results made publicly accessible. This independent, contractor-led peer review would include an evaluation like what is done for the peer review of a journal publication. The contractor would identify and select two to three scientists knowledgeable in scientific disciplines relevant to the topic as potential peer reviewers. Persons invited to serve as peer reviewers would be screened for conflict of interest prior to confirming their service. In most instances, the peer review would be conducted by letter review. The study authors would be informed of the outcome of the peer review and given an opportunity to clarify issues or provide missing details. EPA would consider the peer review comments regarding the scientific and technical evaluation of the unpublished study in determining whether to include the study in its evaluation. The study and its related information, if used in the IRIS assessment, would become publicly available. In the assessment, EPA would acknowledge that the document underwent external peer review, and the names of the peer reviewers would be identified. Unpublished data from personal author communication can supplement a peer-reviewed study, if the information is made publicly available.

[4]IRIS "stopping rules": https://www.epa.gov/sites/production/files/2014-06/documents/iris_stoppingrules.pdf.

4.4. LITERATURE SCREENING PROCESS

The PECO was used to determine inclusion or exclusion criteria for references that served as primary sources of health effects data for chloroform. In addition, the exclusion criteria noted below were applied:

- Records that did not contain original data, such as reviews, editorials, or commentaries; and
- Study materials that have not been peer reviewed (e.g., conference abstracts, theses/dissertations, working papers from research groups or committees, and white papers).

The reference lists from these excluded records and materials were reviewed to identify PECO-relevant studies that may have been missed during database searching.

Studies were screened for inclusion using a structured form based on the PECO in DistillerSR (Evidence Partners; https://www.evidencepartners.com/products/distillersr-systematic-review-software/). Following a pilot phase to calibrate screening guidance, two screeners independently conducted a title and abstract screen of the search results to identify records that appear to meet the PECO. Records that were not excluded based on the title and abstract screen advanced to full-text review. For citations with no abstract, articles were screened based on all or some of the following: title relevance, number of pages (articles two pages in length or less may be assumed to be conference reports, editorials, or letters), and relevant PubMed Medical Subject Headings (MeSH; e.g., a study might not be considered further if there are no human health or biology related MeSH terms). Screening conflicts during title and abstract review were resolved by discussion among the primary screeners with consultation by a third reviewer or technical advisor (if needed) to resolve any remaining disagreements. Assessments of non-English studies were accomplished by translating these studies using native language speakers at the EPA, EPA contractors, or Google Translate, and then reviewing them for PECO relevance. Other informative studies not directly applicable to PECO (e.g., absorption, distribution, metabolism, or elimination [ADME] or exposure characteristics) were tracked during the screening process and tagged as supporting information. Conflict resolution was not required during the screening process to identify other informative studies (i.e., tagging by a single screener is sufficient to identify the study as containing potentially relevant information).

Full-text copies of potentially relevant records identified from title and abstract screening were retrieved, stored in the HERO database, and again independently assessed by two screeners to confirm eligibility according to the PECO. Screening conflicts following full-text review were resolved by discussion among the primary screeners with consultation by a third reviewer or technical advisor (as needed) to resolve any remaining disagreements.

The included and excluded studies, identified by applying the PECO during this two-step screening process, are posted on the project page for this assessment in HERO (hero.epa.gov) and "tagged" with appropriate category descriptors. Release of the PECO-screened literature in the

protocol (or protocol update) for public comment provides an opportunity for stakeholders to identify any missing studies, which, if identified, will be screened as outlined above for adherence to the PECO.

4.4.1. Multiple Publications of the Same Data

If there are multiple publications using the same or overlapping data, all publications will be included, with one selected for use as the primary study; the others will be considered as secondary publications with annotation indicating their relationship to the primary record during data abstraction. For epidemiology studies, the primary publication will generally be the one with the longest follow-up, the largest number of cases, or the most recent publication date. For animal studies, the primary publication will typically be the one with the longest duration of exposure, or with the outcome(s) most informative to the PECO. EPA will include relevant data from all publications, although if the same outcome is reported in more than one publication, the duplicate data will be excluded.

4.5. LITERATURE SURVEYS AND SUMMARY-LEVEL INVENTORIES

During title/abstract or full-text screening, studies were categorized (or "tagged") based on features such as evidence stream (human, animal, in vitro, or in silico), route of administration, health outcome(s) and/or endpoint measure(s), or type of mechanistic information (in vitro, PBPK, ADME, etc.). These literature inventories facilitate subsequent understanding of the extent of the evidence for primary PECO-relevant studies, as well as for studies that may be considered in the assessment as supporting material (e.g., mechanistic information, including alternative model systems, epidemiological or animal toxicology studies assessing routes of administration other than inhalation, mixture studies, case reports of chloroform poisoning, use of chloroform as an anesthetic, studies of chloroform produced as a byproduct from its use to disinfect drinking water, or exposure to chloroform from swimming pools).

Mechanistic studies that were tagged preliminarily during title/abstract screening as "Supplemental material" will be sorted according to hazard categories or types of mechanistic outcomes/pathways. Here, the objective of tagging is to create an inventory of studies for potential later consideration (e.g., by relevance to the research question[s] for each potential hazard) to support analyses of related data. These studies will then be surveyed to assess whether any new literature suggests a re-analysis is warranted of the 2001 MOA conclusions that for cancer chloroform exhibits a "threshold" by all routes of exposure. In addition, they will be screened following the human and animal evidence syntheses to identify studies that may address specific outstanding questions that are likely to have a substantial impact on the assessment conclusions. The inventory also facilitates generation and evaluation of hypothesized mechanistic pathways, and quantification of specific biological processes (i.e., ADME and PBPK data).

4.6. TRACKING STUDY ELIGIBILITY AND REPORTING THE FLOW OF INFORMATION

The main reason for exclusion at the full-text-review stage was annotated and reported in a literature flow diagram (see Figure 2). Categories for exclusion included the following: (1) not relevant to PECO; (2) review, commentary, or letter with no original data; (3) conference abstract or thesis (and the criteria for including unpublished data, described above, were not met); or (4) unable to obtain full-text.

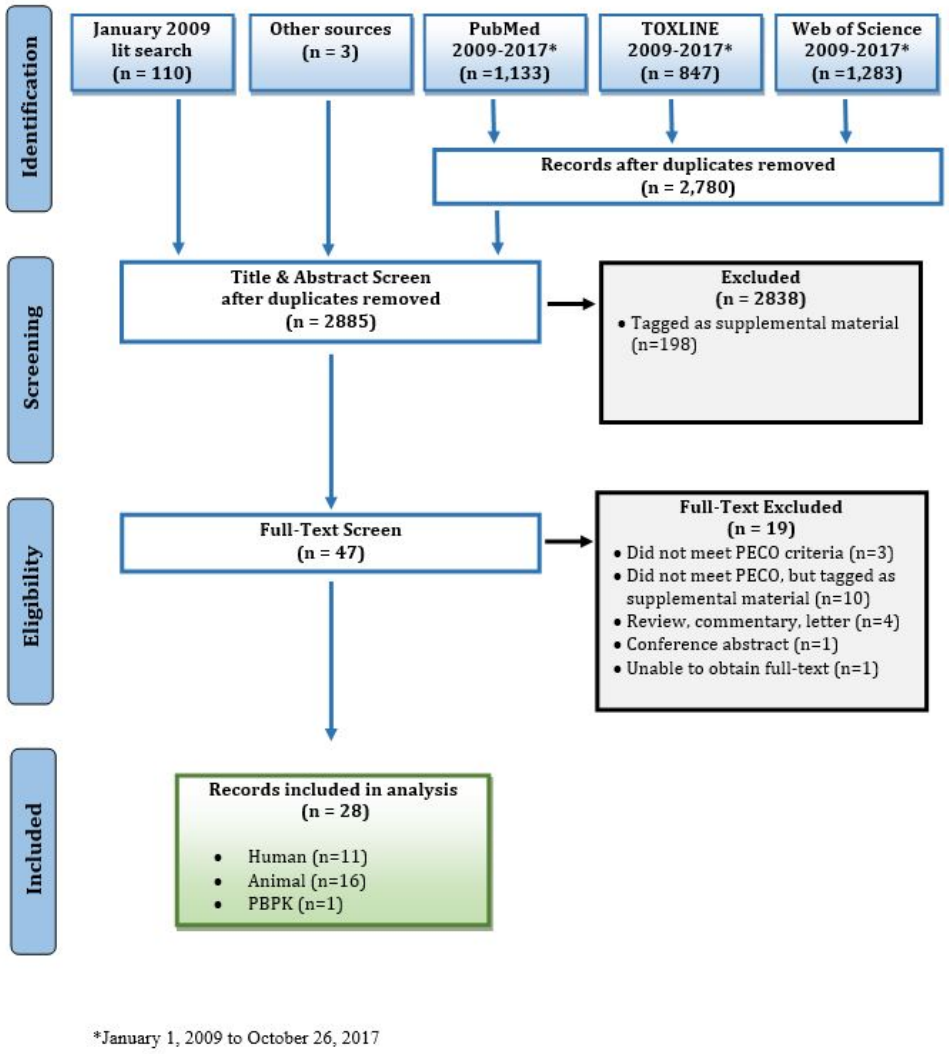

Figure 2. Study flow selection diagram.

5. REFINED ANALYSIS PLAN

The evidence base for this assessment was relatively small and public comments on the assessment plan did not suggest a change was warranted to the specific aims or PECO, thus no refined analysis plan was needed (i.e., all PECO-relevant studies will be considered in the assessment).

6. STUDY EVALUATION (REPORTING, RISK OF BIAS, AND SENSITIVITY) STRATEGY

IRIS assessments evaluate each study's methods using uniform approaches for each group of similar studies so that subsequent syntheses can weigh study results on their merits. Key concerns for the review of epidemiology and animal toxicology studies are potential bias (factors that affect the magnitude or direction of an effect) and insensitivity (factors that limit the ability of a study to detect a true effect). The domains reviewed during study evaluation for epidemiology and animal toxicology studies are shown in Table 3. Epidemiological or animal studies tagged as supplemental material during screening do not undergo study evaluation, unless they have a prominent role in the assessment conclusions.

Table 3. Study evaluation domains

Epidemiology studies	Animal toxicology studies
Exposure measurement	Reporting quality
Outcome ascertainment	Selection or performance bias
Participant selection	Confounding/variable control
Confounding	Reporting or attrition bias
Analysis	Exposure methods sensitivity
Selective reporting	Outcome measures and results display
Sensitivity	

Study evaluation considerations are specific to each study design, health effect, and agent. The study evaluations emphasize attempts to discern the expected magnitude of any identified limitations (focusing on potential sources of bias or insensitivity that could substantively change a result), considering also the expected direction of the bias. Low sensitivity is a bias towards the null. The study evaluations result in an overall judgment regarding confidence (i.e., in the reliability of the results) in the study (or a specific analysis in a study).

6.1. STUDY EVALUATION OVERVIEW

The general approach (described in this section) for evaluating epidemiology and animal toxicology studies is the same, but the specifics of applying the approach differ and are thus described separately in subsequent sections (see Sections 6.2 and 6.3).

Subject-matter experts will evaluate each group of studies to identify characteristics that bear on the informativeness (i.e., validity and sensitivity) of the results. For carcinogenicity, neurotoxicity, reproductive toxicity, and developmental toxicity, EPA guidance for study evaluation is available (U.S. EPA, 2005a, 1998, 1996, 1991). Outcome-specific study evaluations will be

conducted with at least two reviewers independently assessing each study, with inclusion of a pilot phase to assess and refine the evaluation process, comparison of decisions and reaching consensus between reviewers, and when necessary, resolution of differences by discussion between the reviewers, the chemical assessment team, or technical experts. As subject-matter experts examine a group of studies, additional chemical-specific knowledge or methodologic concerns may emerge and a second pass may become necessary. Refinements to the study evaluation process made during the pilot phase and subsequent implementation will be acknowledged as updates to the protocol.

For studies that examine more than one health outcome, the evaluation process will be performed separately for each outcome, because the utility of a study can vary for different outcomes. If a study examines multiple endpoints or measures for the same outcome,[5] evaluation may be performed at that more granular level, if appropriate, but these measures may still be grouped in the analysis plan or for evidence synthesis. For each study[6] (specifically, an outcome in an individual study), in each evaluation domain, reviewers will reach a consensus judgment of *Good, Adequate, Poor, Not reported,* or *Critically deficient*. It is important to stress that these evaluations are performed in the context of the study's utility for identifying individual hazards. While limitations specific to the usability of the study for dose-response analysis are useful to note (to inform those later decisions), they do not contribute to the study confidence classifications. These five categories are applied to each evaluation domain for each study as follows:

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

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

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

- *Not reported* indicates that the information necessary to evaluate the domain question was not available in the study. Generally, this term carries the same functional interpretation as *Poor* for the purposes of the study confidence classification (described below). Depending on the number and severity of other limitations identified in the study, it may or may not be worth reaching out to the study authors for this missing information (see discussion below).

[5]Note: "outcome" will be used throughout these methods; this term can also apply to an endpoint or measure within a larger outcome.

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

- *Critically deficient* reflects a judgment that the study conduct relating to the evaluation domain question introduced a serious flaw that is the primary driver of any observed effect(s) or makes the study uninterpretable. Studies for which domains were judged to be critically deficient will not be used for hazard identification or dose-response analysis without exceptional justification (e.g., it is the only study of its kind and may highlight possible research gaps).

Once the evaluation domains have been considered, the identified strengths and limitations will be combined to reach a study confidence classification of *High, Medium,* or *Low* confidence, or *Uninformative* for a specific health outcome. This classification will be based on the reviewer judgments across the evaluation domains, and will include consideration of the likely impact of the noted deficiencies in bias and sensitivity, or inadequate reporting, on the results. The classifications, which reflect a consensus judgment between reviewers, are defined as follows:

- *High* confidence: No notable deficiencies or concerns were identified; the potential for bias is unlikely or minimal, and the study used sensitive methodologies. In general, although classifications are not decided by "scoring," *High* confidence studies would reflect judgments of *Good* across all or most evaluation domains.

- *Medium* confidence: Possible deficiencies or concerns were noted, but the limitations are unlikely to be of a notable degree. Generally, *Medium* confidence studies will include *Adequate* or *Good* judgments across most domains, with the impact of any identified limitation not being judged as severe.

- *Low* confidence: Deficiencies or concerns were noted, and the potential for substantive bias or inadequate sensitivity could have a significant impact on the study results or their interpretation. Typically, *Low* confidence studies would have a *Poor* evaluation for one or more domains (unless the impact of the limitations on the results is judged as unlikely to be severe). Generally, *Low* confidence results will be given less weight compared to *High* or *Medium* confidence results during evidence synthesis and integration (see Section 10.1, Tables 10 and 11), and are generally not used for either hazard identification or dose-response analysis unless they are the only studies available.

- *Uninformative*: Serious flaw(s) make the study results unusable for informing hazard identification. Studies with *Critically deficient* judgements in any evaluation domain will almost always be classified as *Uninformative* (see explanation above). Studies with multiple *Poor* judgments across domains may also be considered *Uninformative*, particularly when there is a robust database of studies on the outcome(s) of interest or when the impact of the limitations is viewed as severe. *Uninformative* studies will not be considered further in the synthesis and integration of evidence or for dose response.

Authors will be queried to obtain missing critical information, in particular, questions about relationships among variables, missing data, or additional analyses that could address potential limitations. The decision on whether to seek missing information includes consideration of what additional information would be useful, specifically with respect to any information that could result in a re-evaluation of the classification of the domains, and subsequently the overall confidence in the study. Outreach to study authors will be documented and considered

unsuccessful if researchers do not respond within a reasonable amount of time to multiple e-mail or phone requests.

Study evaluation determinations reached by each reviewer and the consensus judgment between reviewers will be recorded in Health Assessment Workspace Collaborative (HAWC), a free and open source web-based application.[7] Study evaluation results housed in HAWC will be made available with the release of the draft assessment for peer review.

6.2. EPIDEMIOLOGY STUDY EVALUATION

Evaluation of epidemiology studies to assess bias and study sensitivity will be conducted for the following domains: exposure measures, outcome ascertainment, participant selection, potential confounding, analysis, selective reporting of results, and study sensitivity (see Table 4). Bias can result in false positives or false negatives, while study sensitivity is typically concerned with identifying the latter.

The principles and framework used for evaluating epidemiology studies are based on the Cochrane Risk of Bias in Non-randomized Studies of Interventions [ROBINS-I; (Sterne et al., 2016)], but modified to address environmental and occupational exposures. The underlying philosophy of ROBINS-I is to describe attributes of an "ideal" study with respect to each of the evaluation domains (e.g., exposure measurement, outcome classification, etc.). Core and prompting questions are used to collect information to guide evaluation of each domain.

Core and prompting questions for each domain are presented in Table 5. Core questions represent key concepts while the prompting questions help the reviewer focus on relevant details under each key domain. Criteria for responding to core and prompting questions will be refined during a pilot phase with engagement from topic specific experts, especially to reflect exposure- and outcome-specific considerations. The types of information that may be the focus of those criteria are listed in Table 4.

[7]HAWC: A Modular Web-Based Interface to Facilitate Development of Human Health Assessments of Chemicals. https://hawcproject.org/portal/.

Table 4. Information relevant to evaluation domains for epidemiology studies

Domain	Example information
Exposure measurement	Source(s) of exposure (e.g., consumer products, occupational, an industrial accident) and source(s) of exposure data, blinding to outcome, level of detail for job history data, when measurements were taken, type of biomarker(s), assay information, reliability data from repeat measures studies, and validation studies.
Outcome ascertainment	Source of outcome (effect) measure, blinding to exposure status or level, how measured/classified, incident vs. prevalent disease, evidence from validation studies, and prevalence (or distribution summary statistics for continuous measures).
Participant selection	Study design, where and when was the study conducted, and who was included? Recruitment process, exclusion and inclusion criteria, type of controls, total eligible, comparison between participants and nonparticipants (or followed and not followed), and final analysis group. Does the study include potential susceptible populations[8] or lifestages?
Confounding	Background research on key confounders for specific populations or settings; participant characteristic data, by group; strategy/approach for consideration of potential confounding; strength of associations between exposure and potential confounders and between potential confounders and outcome; and degree of exposure to the confounder in the population.
Analysis	Extent (and if applicable, treatment) of missing data for exposure, outcome, and confounders, approach to modeling, classification of exposure and outcome variables (continuous vs. categorical), testing of assumptions, sample size for specific analyses, and relevant sensitivity analyses.
Selective reporting	Are results presented with adequate detail for all the endpoints and exposure measures of interest in the context of the PECO? Are results presented for the full sample as well as for specified subgroups? Were stratified analyses (effect modification) motivated by a specific hypothesis?
Sensitivity	What are the ages of participants (e.g., not too young in studies of pubertal development)? What is the length of follow-up (for outcomes with long latency periods)? Choice of referent group, the exposure range, and level of exposure contrast between groups is critical (i.e., the extent to which the "unexposed group" is truly unexposed, and the prevalence of exposure in the group designated as "exposed").

1

[8]Various terms have been used to characterize populations that may be at increased risk of developing health effects from exposure to environmental chemicals, including "susceptible," "vulnerable," and "sensitive." Further, these terms have been inconsistently defined across the scientific literature. This protocol adopts the following definitions for these terms provided by Hines et al. (2010):

> "Susceptibility is defined as a capacity characterized by biological (intrinsic) factors that can modify the effect of a specific exposure, leading to higher health risk at a given relevant exposure level. The term sensitivity is used to describe the capacity for higher risk due to the combined effect of susceptibility (biological factors) and differences in exposure. Vulnerability incorporates the concepts of susceptibility and sensitivity, as well as additional factors that include social and cultural parameters (e.g., socio-economic status and location of residence) that can contribute to an increased health risk."

The term susceptibility is used in this protocol to describe populations at increased risk, focusing on biological (intrinsic) factors that can modify the effect of a specific exposure.

Table 5. Questions to guide the development of criteria for each domain in epidemiology studies

Core question	Prompting questions	Follow-up questions
Exposure measurement Does the exposure measure reliably distinguish between levels of exposure in a time window considered most relevant for a causal effect with respect to the development of the outcome?	For all: • Does the exposure measure capture the variability in exposure among the participants, considering intensity, frequency, and duration of exposure? • Does the exposure measure reflect a relevant time window? If not, can the relationship between measures in this time window and the relevant time window be estimated reliably? • Was the exposure measurement likely to be affected by knowledge of the outcome? • Was the exposure measurement likely to be affected by the presence of the outcome (i.e., reverse causality)? For case-control studies of occupational exposures: • Is exposure based on a comprehensive job history describing tasks, setting, time period, and use of specific materials? For biomarkers of exposure, general population: • Is a standard assay used? What are the intra- and interassay coefficients of variation? Is the assay likely to be affected by contamination? Are values less than the limit of detection dealt with adequately? • What exposure time period is reflected by the biomarker? If the half-life is short, what is the correlation between serial measurements of exposure?	Is the degree of exposure misclassification likely to vary by exposure level? If the correlation between exposure measurements is moderate, is there an adequate statistical approach to ameliorate variability in measurements? If there is a concern about the potential for bias, what is the predicted direction or distortion of the bias on the effect estimate (if there is enough information)?

Table 5. Questions to guide the development of criteria for each domain in epidemiology studies (continued)

Core question	Prompting questions	Follow-up questions
Outcome ascertainment Does the outcome measure reliably distinguish the presence or absence (or degree of severity) of the outcome?	For all: • Is outcome ascertainment likely to be affected by knowledge of, or presence of, exposure (e.g., consider access to health care, if based on self-reported history of diagnosis)? For case-control studies: • Is the comparison group without the outcome (e.g., controls in a case-control study) based on objective criteria with little or no likelihood of inclusion of people with the outcome? For mortality measures: • How well does cause of death data reflect occurrence of the outcome in an individual? How well do mortality data reflect incidence of the outcome? For diagnosis of outcome measures: • Is diagnosis based on standard clinical criteria? If based on self-report of diagnosis, what is the validity of this measure? For laboratory-based measures (e.g., hormone levels): • Is a standard assay used? Does the assay have an acceptable level of interassay variability? Is the sensitivity of the assay appropriate for the outcome measure in this study population?	Is there a concern that any outcome misclassification is nondifferential, differential, or both? What is the predicted direction or distortion of the bias on the effect estimate (if there is enough information)?
Participant selection Is there evidence that selection into or out of the study (or analysis sample) was jointly related to exposure and to outcome?	For longitudinal cohort: • Did participants volunteer for the cohort based on knowledge of exposure and/or preclinical disease symptoms? Was entry into the cohort or continuation in the cohort related to exposure and outcome? For occupational cohort: • Did entry into the cohort begin with the start of the exposure? • Was follow-up or outcome assessment incomplete, and if so, was follow-up related to both exposure and outcome status?	Were differences in participant enrollment and follow-up evaluated to assess bias? If there is a concern about the potential for bias, what is the predicted direction or distortion of the bias on the effect estimate (if there is enough information)? Were appropriate analyses performed to address changing exposures over time in relation to symptoms?

Table 5. Questions to guide the development of criteria for each domain in epidemiology studies (continued)

Core question	Prompting questions	Follow-up questions
	• Could exposure produce symptoms that would result in a change in work assignment/work status ("healthy worker survivor effect")? For case-control study: • Were controls representative of population and time periods from which cases were drawn? • Are hospital controls selected from a group whose reason for admission is independent of exposure? • Could recruitment strategies, eligibility criteria, or participation rates result in differential participation relating to both disease and exposure? For population-based survey: • Was recruitment based on advertisement to people with knowledge of exposure, outcome, and hypothesis?	Is there a comparison of participants and nonparticipants to address whether differential selection is likely?
<u>Confounding</u> Is confounding of the effect of the exposure likely?	Is confounding adequately addressed by considerations in… a. … participant selection (matching or restriction)? b. … accurate information on potential confounders, and statistical adjustment procedures? c. … lack of association between confounder and outcome, or confounder and exposure in the study? d. … information from other sources? Is the assessment of confounders based on a thoughtful review of published literature, potential relationships (e.g., as can be gained through directed acyclic graphing), and minimizing potential overcontrol (e.g., inclusion of a variable on the pathway between exposure and outcome)?	If there is a concern about the potential for bias, what is the predicted direction or distortion of the bias on the effect estimate (if there is enough information)?

Table 5. Questions to guide the development of criteria for each domain in epidemiology studies (continued)

Core question	Prompting questions	Follow-up questions
Analysis Does the analysis strategy and presentation convey the necessary familiarity with the data and assumptions?	• Are missing outcome, exposure, and covariate data recognized, and if necessary, accounted for in the analysis? • Does the analysis appropriately consider variable distributions and modeling assumptions? • Does the analysis appropriately consider subgroups of interest (e.g., based on variability in exposure level, duration, or susceptibility)? • Is an appropriate analysis used for the study design? • Is effect modification considered, based on considerations developed a priori? • Does the study include additional analyses addressing potential biases or limitations (i.e., sensitivity analyses)?	If there is a concern about the potential for bias, what is the predicted direction or distortion of the bias on the effect estimate (if there is enough information)?
Sensitivity Is there a concern that sensitivity of the study is not adequate to detect an effect?	• Is the exposure range adequate? • Was the appropriate population included? • Was the length of follow-up adequate? Is the time/age of outcome ascertainment optimal given the interval of exposure and the health outcome? • Are there other aspects related to risk of bias or otherwise that raise concerns about sensitivity?	
Selective reporting Is there reason to be concerned about selective reporting?	• Are the results needed for the analysis (based on *a priori* specification) presented? If not, can these results be obtained? • Are only statistically significant results presented?	

6.3. ANIMAL STUDY EVALUATION

Using the process for the evaluation of individual epidemiology studies described above, the evaluation of animal studies to assess risk of bias and sensitivity will be conducted for the following domains: reporting quality, selection or performance bias, confounding/variable control, reporting or attrition bias, exposure methods sensitivity, and outcome measures and results display (see Table 6).

Table 6. Considerations to evaluate domains from animal toxicology studies

Domain	Metric	Considerations
Reporting quality	Reporting of information necessary for study evaluation	Key information necessary for study evaluation (study would be deemed *critically deficient* if not reported[a]): • Species; test article description; levels and duration of exposure; endpoints investigated; and qualitative or quantitative results. Important information that should also be reported is listed below. The brackets contain secondary information that would ideally be reported, and based on the needs of a given assessment, may be considered important, or key, information. • *Test animal*—strain; sex; source (e.g., vendor); husbandry procedures (e.g., housing, feed, mating); baseline health [e.g., colony monitoring procedures]; age and/or body weight at start of study. • *Exposure methods*—test article source; description of vehicle control; route of administration; methods of administration (e.g., gavage or exposure chamber); information on stability; purity; analytical verification methods. • *Experimental design*—periodicity of exposure; animal age/lifestage during exposure and at endpoint evaluation(s); timing of endpoint evaluation[s] [e.g., latency between exposure and testing]. • *Endpoint evaluations*—procedural details to understand how endpoints were measured; procedural controls, including information on positive and negative controls; related details (e.g., biological matrix or specific region of tissue/organ evaluated); information on other manipulations (e.g., surgery or cotreatment). • *Results presentation*—presentation of findings for all endpoints of interest that were investigated; information on variability; experimental units assessed; sample size; statistical procedures; (related details—e.g., maternal toxicity in developmental studies; handling of early mortality in long-term bioassays). [a] Although such decisions should be made on an assessment-specific basis, if this information is not reported, it is generally not useful to reach out to the study authors. However, for other missing study details that might change study confidence conclusions, if such details were available, efforts should be made to contact the study authors. *Note*: Studies adhering to GLP (good laboratory practices) or to testing guidelines established by (inter)national agencies are assumed to be of good reporting quality.

1

Table 6. Considerations to evaluate domains from animal toxicology studies (continued)

Domain	Metric	Considerations
Selection or performance bias	Allocation to experimental groups	Ideally, experimental units are randomly assigned, with each animal or litter having an equal chance of being assigned to any experimental group, including controls, and allocation procedures are sufficiently described. Less ideally, but generally adequate or good, are studies indicating normalization of experimental groups prior to exposure, for example according to body weight or litter, but without indication of randomization. The least preferred situation is studies with no indication of how groups were assigned.
	Blinding of investigators, particularly during outcome assessment	Good studies will conceal the treatment groups from the researchers conducting the endpoint evaluations (and, in rare but ideal situations, from all research personnel and technicians). Concerns regarding blinding may be attenuated when outcome measures are more objective (e.g., as is the case of obtaining organ weights) or measurement is automated using computer driven systems (e.g., as is the case in many behavioral assessments).
Confounding/variable control	Control for variables across experimental groups	In a good study, outside of the (chemical) exposure of interest, all variables will be controlled for and consistent across experimental groups. Concern regarding additional variables, introduced intentionally or unintentionally, may be mitigated by knowledge or inferences regarding the likelihood and extent to which the variable can influence the endpoint(s) of interest. A very important example to consider is whether the exposure was sufficiently controlled to attribute the effects of exposure to the compound of interest alone. Generally, well-conducted exposures will not have any evidence of coexposures and will include experimental controls that minimize the potential for confounding (e.g., use of a suitable vehicle control). Other examples of variables that may be uncontrolled or inconsistent across experimental groups include protective or toxic factors that could mask or exacerbate effects, diet composition, or surgical procedures (e.g., ovariectomy).
Reporting or attrition bias	Lack of selective data reporting and unaccounted loss of animals	In a good study, information is reported on all prespecified outcomes and comparisons for all animals, across treatment groups and scheduled sacrifices. Aspects to consider include whether all study animals were accounted for in the results (if not, are explanations, such as death while on study and adjustments, provided), and whether expected comparisons or certain groups were excluded from the analyses. In some studies, the outcomes evaluated must be inferred (e.g., a suite of standard measures in a guideline study). *Note*: This metric does not address whether quantitative data were reported, nor considers statistical test methods.

Table 6. Considerations to evaluate domains from animal toxicology studies (continued)

Domain	Metric	Considerations
Exposure methods sensitivity	Characterization of the exposure to the compound of interest	Consider whether there are notable issues that raise doubt about the reliability of the exposure levels, or of exposure to the compound of interest. Depending on the chemical being assessed, this may include considering factors such as the stability and composition (e.g., purity; isomeric composition) of the test article; exposure generation and analytic verification methods (including whether the tested levels and spacing between exposure groups is resolvable using current methods); and details of exposure methods (e.g., inhalation chamber type; gavage volume). In some cases, exposure biomarkers in blood, urine, or tissues of treated animals can mitigate concerns regarding inaccurate dosing (dependent on the validity of the biomarker for the chemical of interest). *Note*: While this identifies uncertainties in dose-response, it is typically not a valid reason for exclusion from hazard identification.
	Utility of the exposure design for the endpoint of interest	Based on the known or presumed biological progression of the outcomes being evaluated, consider whether there are notable concerns regarding the timing, frequency, or duration of exposure. For example, better developmental studies will cover the critical window of exposure (if studies have determined the critical window for the specific outcome) or the largest developmental interval (if studies have not defined the critical window for the specific outcome), while better studies for assessing cancer or other chronic outcomes will be of longer duration. Studies that expose animals infrequently or sporadically, or, conversely, on a continuous basis (which, depending on the exposure level, can impact food/water consumption, sleep cycles, or pregnancy/maternal care), might introduce additional complications.

Table 6. Considerations to evaluate domains from animal toxicology studies (continued)

Domain	Metric	Considerations
Outcome measures and results display	Sensitivity and specificity of the endpoint evaluations	Consider whether there are notable concerns about aspects of the procedures for, or the timing of, the endpoint evaluations. Based on the endpoint evaluation protocol used for the endpoints of interest, specific considerations will typically include: • Concerns regarding the sensitivity for evaluating the endpoint(s) of interest (i.e., assays can differ dramatically in terms of their ability to detect effects), and/or timing of treatment and assessment (i.e., the age of animals at assessment can be critical to the appropriateness and sensitivity of the evaluation). This includes both overestimates or underestimates of the true effect, as well as a much higher (or lower) probability for detecting the effect(s) being assessed. • Concerns regarding the specificity and validity of the protocols. This includes the use of appropriate protocol controls to rule out nonspecific effects, which can often be inferred from established guidelines or historical assay data. It may be considered useful for insensitive, complex, or novel protocols to include positive and/or negative controls. • Concerns regarding adequate sampling. This includes both the experimental unit (e.g., litter; animal) and endpoint (e.g., number of slides evaluated). This is typically inferred from historical knowledge of the assay or comparable assays. *Note*: Human relevance of the endpoint is not addressed during study evaluation; for under sampling without blinding (e.g., sampling bias), this will typically lead to gross overestimates of effect; sample size is generally not a reason for exclusion. Rather, human relevance of the endpoint is considered either during developing the PECO (endpoints not considered relevant to humans would not be included) or during evidence integration (Section 10).
	Usability and transparency of the presented data	Consider whether the results are analyzed or presented in a way that limits concerns regarding the reliability of the findings. Items that will typically be important to consider include: • Concern that the level of detail provided does not allow for an informed interpretation of the results (e.g., authors' conclusions without quantitative data; discussing neoplasms without distinguishing between benign and malignant tumors; not presenting variability). • Concern that the way in which the data were analyzed, compared, or presented is inappropriate or misleading. Examples include: failing to control for litter effects (e.g., when presenting pup data rather than the preferred litter data); pooling results from males and females or across lesion types; failing to address observed or presumed toxicity (e.g., in assessed animals; in dams) when exposure levels are known or expected to be highly toxic; incomplete presentation of the data (e.g., presenting continuous data as dichotomized); or non-preferred display of results

Table 6. Considerations to evaluate domains from animal toxicology studies (continued)

Domain	Metric	Considerations
		(e.g., using a different readout than is expected for that assay). The evaluator should support how or why, and to what extent, this might mislead interpretations. *Note*: Concerns regarding the statistical methods applied are not addressed during study evaluation, but should be flagged for review by a statistician. Missing information related to this metric should typically be requested from study authors.
Other	(Optional)	Example 1: Control for other threats to internal validity: This exceptional metric might be used to consider animal husbandry concerns, reports of pre-dosing toxicity or infection, etc. Example 2: Lack of concern for sensitivity of the animal model. This exceptional metric should be used only when there is demonstrated evidence of differences in model (e.g., species, sex, strain) sensitivity.

General Note: The rationale for judgments should be documented clearly and consistently. In addition, for metrics other than reporting quality, it is important to document and consider the overall confidence determination the level of concern raised by any identified limitations. This should, to the extent possible, reflect an interpretation of the potential influence on the results (including the direction and/or magnitude of influence) that limitation might provide and be conducted on a per outcome basis. For a given assessment, evaluators should establish and document a priori criteria for judging the information described within each metric, to the extent possible.

1

7. DATA EXTRACTION OF STUDY METHODS AND RESULTS

Data extraction and content management will be carried out using HAWC. Data extraction elements that may be collected from epidemiological and animal studies are listed in Appendix B. Choices about what data to extract will be guided by determining the elements that contribute to analyses that inform the synthesis of evidence. The content of the data extraction may be revised following the identification of the studies included in the review as part of a pilot phase to assess the data extraction workflow. Not all studies relevant to the initial PECO will go through data extraction. Studies evaluated as being "not informative" will not be considered further and, therefore, will not be considered for data extraction. In addition, outcomes that are determined to be less relevant during PECO refinement may not go through data extraction, or may have only minimal data extraction. The same may be true for low confidence studies if sufficient medium and high confidence studies are available.

The data extraction results for included studies will be presented in the assessment and available for download from HAWC in Excel format when the assessment is publicly released. [NOTE: The following browsers are fully supported for accessing HAWC: Google Chrome (preferred), Mozilla Foxfire, and Apple Safari. There are errors in functionality when viewed with Internet Explorer.] Data extraction will be performed by one member of the evaluation team and independently checked by another member. Any discrepancies in data extraction will be resolved by discussion or consultation with a third member of the evaluation team. Once data have been verified, they will be "locked" to prevent accidental changes. Digital rulers, such as WebPlotDigitizer (http://arohatgi.info/WebPlotDigitizer/), will be used to extract numerical information from figures.

As previously described, routine attempts will be made to obtain missing information from epidemiologic and animal studies, if that information is considered influential during study evaluations (see Section 6) or if needed to conduct a meta-analysis (e.g., missing group sizes or missing variance descriptors, such as standard deviations or confidence intervals). Missing data from individual mechanistic (e.g., in vitro) studies will generally not be sought. Outreach to study authors will be considered unsuccessful if they do not respond to email or phone requests after multiple attempts.

7.1. STANDARDIZING REPORTING OF EFFECT SIZES

Results from outcome measures will be transformed, when possible, to a common metric to help assess dissimilar but related outcomes measured with different scales. These considerations

are essential for meta-analysis, but also facilitate systematic evaluation and hazard identification when meta-analysis is not feasible or otherwise not necessary for an assessment. The following considerations outline issues in selecting the most appropriate common metric for a collection of related endpoints (Vesterinen et al., 2014):

Common metrics for continuous outcomes:

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

- *Percent control response (or normalized mean difference [NMD]*). This approach is commonly recommended. Percent control group calculations are based on means. Standard deviation (or standard error) values presented in the studies for these normalized effect sizes can also be estimated if sufficient information has been provided. Typically, effect sizes fall between −100% and +100%. Note that some outcomes reported as percentages, such as mean percentage of affected offspring per litter, can lead to distorted effect sizes when further characterized as percentage change from control. Such measures are better expressed as absolute difference in means, or even better, transformed to incidences using approaches for event or incidence data (see below).

- *Standardized mean difference*. The NMD approach above is relevant to ratio scales, but sometimes it is not possible to infer what a "normal" animal would score, such as when data for animals without lesions are not available. In these circumstances, standardized mean differences can be used. The difference in group means is divided by a measure of the pooled variance to convert all outcome measures to a standardized scale with units of standard deviations. This approach can also be applied to data for which different measurement scales are reported for the same outcome measure (e.g., different measures of lesion size such as infarct volume and infarct area).

Common metrics for event or incidence data:

- Percent change from control. This metric is analogous to the continuous data case above.

- For binary outcomes, such as the number of individuals that developed a disease or died, and with only one treatment evaluated, data can be represented in a 2 × 2 table with the odds ratio and its standard error. Note that when the value in any cell is zero, 0.5 is added to each cell to avoid problems with the computation of the standard error. For each comparison, the odds ratio can be calculated. Odds ratios are normally combined on a logarithmic scale.

Sometimes studies report mean outcomes without reporting variance, especially for animal studies in biomedical research (Vesterinen et al., 2014). In cases in which the evidence base is large, these studies may be excluded. When included, summary effect size estimates can often be presented using absolute difference in means or normalized difference in means. When sample size is not presented for individual groups, the mid value in a range will be used for effect size

calculations. In addition, for epidemiology studies, adjusted statistical estimates will be extracted rather than unadjusted or raw estimates when possible.

7.2. STANDARDIZING ADMINISTERED DOSE LEVELS/CONCENTRATIONS

Exposures will be standardized to common units when possible. For hazard characterization, exposure levels will typically be presented as reported in the study and standardized to common units (ppm or mg/m^3 for inhalation studies) as an initial phase in evidence synthesis and integration. For inhalation exposures to chloroform, concentration in air in ppm can be converted to concentration in air in mg/m^3 by multiplying ppm times (238.7 g/mol ÷ 24.45 L) at standard temperature (25°C) and pressure (1 atm). All assumptions used in performing dose conversions will be documented. Dosimetry adjustment factors will be applied as part of conducting the dose-response analysis (see Section 11).

8. PHYSIOLOGICALLY BASED PHARMACOKINETIC (PBPK) MODEL IDENTIFICATION, DESCRIPTIVE SUMMARY, AND EVALUATION

PBPK (or classical pharmacokinetic [PK]) models should be used in an assessment when an applicable one exists and no equal or better alternative for dosimetric extrapolation is available. Any models used should represent current scientific knowledge and accurately translate the science into computational code in a reproducible, transparent manner. For a specific target organ/tissue, it may be possible to employ or adapt an existing PBPK model, or develop a new PBPK model or an alternate quantitative approach. Data for PBPK models may come from studies with animals or humans, and may be in vitro or in vivo in design.

8.1. IDENTIFYING PHYSIOLOGICALLY BASED PHARMACOKINETIC (PBPK) MODELS

PBPK modeling is the preferred approach for calculating a human equivalent concentration (HEC) according to the hierarchy of approaches outlined in EPA guidance (U.S. EPA, 2011a). For chloroform, metabolism is a major component of target organ toxicity, and PBPK models are available to account for interspecies differences in metabolism between rats, mice, and humans (Sasso et al., 2013; Corley et al., 1990). Chloroform is metabolized to the reactive metabolites phosgene and dichloromethyl free radical in humans and animals by cytochrome P450-dependent pathways (Gemma et al., 2003; Constan et al., 1999).

Because of the role of metabolism in the production of target organ toxicity, and the reactive nature of the metabolites, local tissue bioactivation of chloroform will be modeled for the liver and kidney. A PBPK model is then used to convert the external chloroform concentration (in ppm) to an internal dose metric (average daily milligrams of chloroform metabolized per liter tissue) for toxicological data in animals. Because a PBPK model for exposure to chloroform and its bioactivation in the developing fetus is not available, alternative PBPK-derived internal dose metrics (i.e., area under the curve for chloroform in blood) may be used to evaluate developmental effects.

These PBPK-derived rodent internal doses simulate the intermittent exposure conditions in animal bioassays (i.e., 6 hours/day, 5 days/week). Benchmark dose modeling will be performed on the toxicological data based on internal dose. Once a benchmark dose lower confidence limit (BMDL) is derived for internal dose in the animal, the human PBPK model will then be used to predict the HEC of chloroform. This HEC represents the daily exposure, based on a continuous 24-

hour/day exposure period, that would result in a human internal dose equivalent to the corresponding animal internal dose BMDL. For more details on a candidate PBPK model for chloroform, and the derivation of tissue-specific metabolic rates for this chemical, see Sasso et al. (2013).

8.2. PHARMACOKINETIC (PK)/ PHYSIOLOGICALLY BASED PHARMACOKINETIC (PBPK) MODEL DESCRIPTIVE SUMMARY

Key information from identified PBPK models will be summarized in tabular format (see example in Table 7 below).

Table 7. Example descriptive summary for a physiologically based pharmacokinetic (PBPK) model

Author	Smith (2013)		
Contact Email	Smith@email.com		
Contact Phone			
Sponsor	N/A		
Model Summary			
Species	Human		
Strain			
Sex			
Lifestage	Adult		
Exposure Routes	Inhalation	Oral	
Tissue Dosimetry	Blood	Lung metabolism	
Model Evaluation			
Language	ACSL 11.8		
Code Available	YES	Effort to Recreate Model	COMPLETE
Code Received	YES	Effort to Migrate code	COMPLETE
Structure Evaluated	YES		
Math Evaluated	YES		
Code Evaluated	YES. Issue (minor): lung metabolism mislabeled as liver metabolism in code comments post-ACSLX migration.		
PK Data Available	NO		

8.3. PHARMACOKINETIC (PK)/ PHYSIOLOGICALLY BASED PHARMACOKINETIC (PBPK) MODEL EVALUATION

Once available PBPK models and related studies are summarized, EPA will undertake model evaluation. Judgments on the suitability of a model are separated into two categories: scientific and technical (see Table 8). The scientific criteria focus on whether the biology, chemistry, and other information available for a chemical's MOA are justified (i.e., preferably with citations to support

use) and represented by the model structure and equations. The scientific criteria are judged based on information presented in the publication or report that describes the model and do not require evaluation of the computer code. Preliminary technical criteria include availability of the computer code and completeness of parameter listing and documentation. Studies that meet the preliminary scientific and technical criteria are then subjected to an in-depth technical evaluation, which includes a thorough review and testing of the computational code. The in-depth technical and scientific analyses focus on the accurate implementation of the conceptual model in the computational code, use of scientifically supported and biologically consistent parameters in the model, and reproducibility of model results reported in journal publications and other documents. This approach stresses (1) clarity in the documentation of model purpose, structure, and biological characterization; (2) validation of mathematical descriptions, parameter values, and computer implementation; and (3) evaluation of each plausible dose metric. The in-depth analysis is used to evaluate the potential value and cost of developing a new model or substantially revising an existing one. PBPK models developed by EPA during the assessment will be peer reviewed, either as a component of the draft assessment or by publication in a journal article.

Table 8. Criteria for evaluation of physiologically based pharmacokinetic (PBPK) models

Criteria	Example information
Scientific	Biological basis for the model is accurate. • Consistent with mechanisms that significantly impact dosimetry. • Predicts dose-metrics expected to be relevant. • Applicable for relevant route(s) of exposure.
	Consideration of model fidelity to the biological system strengthens the scientific basis of the assessment relative to standard exposure-based extrapolation (default) approaches. • Can the model describe critical behavior, such as nonlinear kinetics in a relevant dose range, better than the default (i.e., $BW^{3/4}$ scaling)? • Is the available metric a better predictor of risk than the default? (Specifically, model-based metrics may correlate better than the applied doses with animal/human dose-response data.) The degree of certainty in model predictions vs. default is also a factor (e.g., while target tissue metrics are generally considered better than blood concentration metrics, lack of data to validate tissue predictions when blood data are available may lead to choosing the latter metric).
	Principle of parsimony • Model complexity or biological scale, including number and parameterization of (sub)compartments (e.g., tissue or subcellular levels) should be commensurate with data available to identify parameters.
	Model describes existing PK data reasonably well, both in "shape" (matches curvature, inflection points, peak concentration time, etc.) and quantitatively (e.g., within factor of 2–3).
	Model equations are consistent with biochemical understanding and biological plausibility.
Initial Technical	Well-documented model code is readily available to the EPA and the public.
	A set of published parameters is clearly identified, including origin/derivation.
	Parameters do not vary unpredictably with dose (e.g., any dose dependence in absorption constants is predictable across the dose ranges relevant for animal and human modeling).
	Sensitivity and uncertainty analysis has been conducted for relevant exposure levels (local sensitivity analysis is sufficient, but a global analysis provides more information). • If a sensitivity analysis was not conducted, EPA may decide to independently conduct this additional work before using the model in the assessment. • A sound explanation should be provided when sensitivity of the dose metric to model parameters differs from what is reasonably expected based on experience.

9. SYNTHESIS WITHIN LINES OF EVIDENCE

For each potential health effect (i.e., a health outcome; a grouping of related health outcomes; or a broad hazard category), EPA will separately synthesize the available human health effect evidence, animal health effect evidence, and relevant mechanistic data. Each synthesis will be written to emphasize considerations that may suggest causation, and that will ultimately support the evidence integration steps outlined in Section 10 (i.e., strengths and limitations of the individual studies or group of studies, consistency, exposure-response relationship, strength of the association, temporal relationship, biological plausibility, coherence, and "natural experiments" in humans (U.S. EPA, 2005a, 1994a)).

Specifically, the human and experimental animal evidence on potential health effects will first be analyzed and synthesized separately (see Section 9.1 and Figure 4). These syntheses (or the lack of data within these lines of evidence) help determine the approach to be taken in synthesizing the available mechanistic evidence. As discussed previously, in the current assessment of chloroform, , a synthesis of all identified mechanistic evidence is not anticipated to be critical for evaluation of carcinogenicity (see Section 9.2).

9.1. SYNTHESES OF HUMAN AND ANIMAL HEALTH EFFECTS EVIDENCE

To assess the weight of evidence regarding the potential for chemical exposure to cause a particular health effect, the syntheses of the human and animal health effects evidence will focus on describing aspects of the evidence that best inform causal interpretations. These syntheses will be based primarily on studies of *High* and *Medium* confidence. *Low* confidence studies will generally be used to evaluate consistency and coherence, but may only be used for hazard determination if no or few higher confidence studies are available. Any issues that stem from the evaluation of individual studies will be discussed (e.g., outstanding questions about bias or sensitivity, highlighting studies considered to be most informative for interpreting dose-response, results using exposure protocols, or assessments with the highest validity, etc.). If the evidence allows, consistency, dose-response, effect magnitude, precision, and coherence will each be addressed drawing from individual study results or groups of studies. If possible, results across studies will be compared using graphs and charts or other data visualization strategies; this will influence the selection of what analytic results to present. If possible, the analysis will include examination of results stratified by any or all of the following: study confidence classification (or specific issues within confidence evaluation domains), exposure level, sensitivity (e.g., low vs. high), and other factors that may have been identified in the preliminary analysis plan (e.g., sex, lifestage, or another

demographic). The number of studies and the differences encompassed by the studies will determine the extent to which specific types of factors can be examined to stratify study results.

Syntheses will articulate the strengths and the weaknesses of the available evidence in the context of the considerations described in Table 9, which are adapted from the paper by Austin Bradford Hill (Hill, 1965) (see Section 10). Overall confidence determinations for human and animal evidence streams are described using a framework (see Figure 4 for template) that includes similar considerations to those used by the Grading of Recommendations Assessment, Development, and Evaluation (GRADE) certainty in the evidence framework (Guyatt et al., 2011; Schünemann et al., 2011). Human and animal syntheses typically provide a foundation for the evidence integration decisions and both will be summarized in an evidence profile table (see Section 10 and Figure 4). In addition, to the extent the data allow, the syntheses will discuss factors relating to potential susceptible populations, based on demographics, genetic variability, lifestage, health status, behaviors, or practices, social determinants, and exposure to other pollutants.

For epidemiology evidence, the primary considerations used to inform causality and explore alternative explanations in the synthesis text are consistency (considering risk and direction of potential bias and sensitivity), biological gradient, strength (effect estimate magnitude and precision), coherence, natural experiments, and temporality. For experimental animal evidence, the primary considerations for the synthesis are consistency, dose-response gradient, strength (effect magnitude and precision), and coherence.

Consistency will often represent one of the most influential considerations, and the synthesis will specifically emphasize observations across populations (e.g., location) or exposure scenarios in human studies, and across laboratories, populations (e.g., species), or (more rarely) exposure scenarios (e.g., duration) in animal studies. When discussing the consistency of a set of study results, it is important to try to differentiate between conflicting evidence (unexplained positive and negative results in similarly exposed human populations or in similar animal models) and differing results [mixed results attributable to differences between human populations, animal models, or exposure conditions; (U.S. EPA, 2005a)]. Some study results that appear to be inconsistent may be explained by potential biases or other attributes that affect sensitivity, resulting in variations in the degree of confidence accorded to the study results. Additionally, the interpretation of the consistency of the evidence and the magnitude of the reported effects will emphasize biological significance as more relevant to the assessment than statistical significance. Statistical significance (as reported by *p*-values, etc.) provides no evidence about effect size or biological significance, and a lack of statistical significance will not be automatically interpreted as evidence of no effect. For both the human and animal evidence syntheses, if supported by the available data, additional analyses across studies (such as meta-analysis) may also be conducted.

Table 9. Primary considerations for human and animal health effect evidence syntheses[a]

Consideration	Description
Consistency	Repeated findings across different studies increase the evidence strength. When inconsistencies exist, the evaluator considers study confidence and whether results were "conflicting" or "differing" (U.S. EPA, 2005a). Conflicting results decrease evidence strength. *Stronger human evidence*: evidence in different populations and study designs. *Stronger animal evidence:* evidence in different species and strains, by different researchers.
Biological gradient (dose-response)[b]	Increases in risk, or in the frequency or severity of effects with increasing exposure (e.g., concentration; duration) increase the evidence strength. These associations can reflect simple or complex (i.e., nonlinear) relationships. Absence of a dose-response relationship does not necessarily decrease evidence strength, but it may after other studies and known biology are considered.
Strength (effect magnitude) and precision	Given what is known about the health outcome, larger effect sizes or higher relative risks, particularly for rare or severe effects, are more convincing of a causal relationship. Although small effect sizes are not grounds to dismiss an association, the evaluation of evidence strength may consider variability, historical data, or bias to assess the likelihood that effects are due to other explanations. Higher precision (reflected by narrow confidence bounds/smaller standard errors and statistical significance) also adds confidence in the observed associations. Analyzing results across studies can help to examine possible bias in individual studies or rule out chance (i.e., low precision) as an alternative explanation.
Mechanistic evidence related to biological plausibility	Supporting mechanistic evidence (e.g., associations with precursors or biomarkers related to effects; changes in established biological pathways or a theoretical mode of action) increases evidence strength. While a lack of mechanistic understanding does not decrease evidence strength, it may do so if findings demonstrate that effects are unlikely to occur in humans. *Human evidence:* studies in exposed humans or appropriately exposed human cells. *Animal evidence*: studies in exposed animals or appropriately exposed animal cells.
Coherence[c]	Findings across the database that fit into a consistent pattern as a whole and hold together (e.g., similarity in results for related effects within an organ system, or across systems; a temporal or dose-dependent progression of linked effects of increasing severity) increase evidence strength. Conversely, an observed lack of changes that would be expected to occur (e.g., in parallel, subsequently) with the effect of interest could decrease evidence strength. Coherence is informed by the known biological development of the health effect in question, as well as toxicokinetic/dynamic understanding of the chemical or related chemicals.[d]
Natural experiments	*Human evidence only*: Reductions in effect that occur after a clear reduction in exposure. Although rare, such reductions can provide compelling, highly persuasive evidence.
Temporality	*Human evidence only:* The exposure occurs before the effect (this issue is considered in the evaluation of exposure measures for each study).

[a]These ideas build upon the discussion for assessing causality of disease in Hill (1965), although there are some differences in the use or interpretations of the terms.

[b]While humans are "exposed" and not "dosed," and animals are not "dosed" via inhalation, "dose-response" is used for convention, although it is acknowledged that "exposure-response" may be more appropriate in many contexts.

[c]There is a clear overlap in the use of mechanistic evidence to interpret coherence (e.g., informing the relatedness or comparability of potentially coherent health findings) and biological plausibility. The available

mechanistic information is synthesized separately and considered during the subsequent step of evidence integration (see Section 10).

[d]Although it is not separately listed, Hill's consideration of "analogy" (information for a similar but different association that supports causation) is indirectly encompassed by the evaluation of coherence during the review of environmental health studies; however, this use of analogous chemicals or exposure scenarios is less common.

9.2. MECHANISTIC INFORMATION

Mechanistic information includes any experimental measurement related to a health outcome that informs the biological or chemical events associated with toxic effects, but is not itself an adverse outcome. This includes data from virtually all in vitro studies, and may also include data from human and animal studies. The synthesis of mechanistic information is used to inform the integration of health effects evidence for both hazard identification (i.e., biological plausibility or coherence within human or animal evidence streams; coherence or human relevance across streams of evidence) and dose-response evaluation.

In the current assessment of chloroform, a synthesis of all identified mechanistic evidence is not anticipated to be critical for evaluation of carcinogenicity. As outlined in Sections 2 and 3, the objective of this assessment is to determine whether the inhalation of chloroform results in adverse health effects and to derive an RfC for chloroform by using inhalation dose-response data from human or animal studies. Although both cancer and noncancer health outcomes are considered relevant to the PECO criteria, a detailed analysis of cancer-relevant mechanistic evidence is not included in the scope. Rather, the assessment will rely on the existing 2001 MOA analysis for cancer for chloroform posted on the IRIS website, which concluded that for cancer, chloroform exhibits a "threshold" by all routes of exposure, and thus a chloroform dose that does not elicit cytotoxicity exists and can be considered protective against cancer risk. Therefore, only new cancer MOA evidence will be screened to confirm those conclusions are still valid. In the absence of new information that may impact the 2001 conclusions, the current assessment will rely on other published authoritative sources like public health agency reports and expert review articles to summarize mechanistic information for chloroform. For specific health effects other than cancer, if there are remaining questions that could be informed by mechanistic studies for determining potential hazard, these studies will be synthesized, whereupon the process for determining the level of confidence in the results of individual studies will be developed, and the protocol will be updated.

Some examples of how the synthesis of the mechanistic evidence may be used to inform subsequent evidence integration decisions (see Section 10) are described in Table 10. Like Table 9 in Section 9.1, Table 10 provides examples of applying the synthesis of mechanistic information, including guiding the organization and focus of evidence integration, and informing potential implications for dose-response analysis (see Section 11), as well as hazard identification.

Table 10. Examples of the potential inferences and applications for mechanistic data that may be discussed in the mechanistic evidence synthesis

Mechanistic inferences considered	Potential applications of the mechanistic synthesis
Biological plausibility: Evidence that demonstrates plausible biological mechanisms, obtained from experimental studies or other sources including studies not directly investigating the health effect under evaluation, may strengthen (or weaken) the interpretation of the likelihood of an association between exposure and the health effect. Thus, in some instances, differing levels of biological plausibility (or certainty) might be drawn. It is important to note that the lack of mechanistic data explaining an association is not used to discount observations from human or animal studies. The interpretation of biological plausibility considers the existing knowledge for how the health effect develops and can involve analyses of information at different levels of biological organization (e.g., molecular or tissue).	***Evidence integration (within stream)*** • Observations of important mechanistic changes in exposed humans or animals that are plausibly associated with the health outcome in question can strengthen the confidence in the within-stream health effect findings, particularly when the changes are observed in the same exposed population presenting the health effect. • The absence of expected mechanistic changes in an exposed population might diminish the plausibility of an association. This considers the sensitivity of the mechanistic changes and the potential contribution of alternate or unidentified mechanisms of toxicity. • Inconsistent evidence (i.e., heterogeneous results) across different animal species or human populations might be explainable by evidence that different mechanisms are operant in the different populations (e.g., evidence demonstrating that certain populations cannot metabolize a reactive metabolite; evidence that variability in gene expression correlates with variability in response).
Human relevance of findings in animals: In the absence of sufficient MOA or ADME information, effects in animal models are assumed to be relevant to humans [e.g., U.S. EPA (2005a)]. For potential human health hazards supported by strong evidence from animal models, mechanistic evidence is considered in light of human relevance.	***Evidence integration (across stream)*** • Evidence establishing that the mechanisms underlying the animal response do not operate in humans, or that animal models do not suitably inform a specific human health outcome, can support the view that the animal response is not relevant to humans. In these cases, the animal response provides neither an argument for nor against an overall hazard judgment. • Observations of mechanistic changes in exposed humans that are similar or coherent with mechanistic or toxicological changes in experimental animals (and that are interpreted to be associated with the health outcome under evaluation) strengthen the human relevance of the animal findings.

1

Table 10. Examples of the potential inferences and applications for mechanistic data that may be discussed in the mechanistic evidence synthesis (continued)

Mechanistic inferences considered	Potential applications of the mechanistic synthesis
Potential vulnerabilities: Mechanistic understanding of how a health outcome develops, even without a full MOA, can clarify characteristics of important events (e.g., their presence or sensitivity across lifestages or across genetic variations) and helps identify susceptible populations.	***Susceptibility and dose-response analysis*** • Identification of lifestages or groups likely to be at greatest risk can clarify hazard descriptions, including whether the most susceptible populations have been adequately tested. • Knowledge of expected vulnerabilities can inform selection of studies for quantitative analysis (e.g., prioritizing studies including such populations). • When there is evidence of susceptibilities, but specific studies cannot be prioritized for quantitative analysis, susceptibility data may support refined human variability/uncertainty factors or probabilistic uncertainty analyses.
Biological understanding, including the identification of precursor events: Mechanistic data that reasonably describe how effects develop may clarify the exposure conditions expected to result in these effects. Further, well-studied MOAs can identify mechanistic precursor events linked qualitatively or quantitatively to apical health effect(s), increasing the strength of the hazard descriptor.	***Dose-response analysis*** MOA inferences can inform the use of: • Particular dose-response models (e.g., models integrating data across several related outcomes or incorporating toxicokinetic knowledge). • Proximal measures of exposure (e.g., external vs. internal dose metrics). • Improved characterization of responses (e.g., use of well-established precursors in lieu of direct observation of apical endpoints; combination of related outcomes [such as benign and malignant tumors] resulting from the same MOA).

10. INTEGRATION ACROSS LINES OF EVIDENCE

For the analysis of most health outcomes, IRIS assessments integrate the human, animal, and mechanistic evidence. Depending on the assessment scope and availability of human and animal evidence, conclusions for mechanistic evidence may be based on consideration of individual mechanistic studies or by relying on other sources. During evidence integration, three conclusions are drawn as follows (and depicted in Figure 3):

- First, a conclusion is made regarding the evidence for health effects associated with the chemical exposure from human ("human evidence stream") studies. The conclusion in this step incorporates mechanistic or MOA evidence informing the biological plausibility and coherence of the available human health effect studies.
- In parallel, a conclusion is made regarding the evidence for health effects associated with the chemical exposure from animal ("animal evidence stream") studies. The conclusion in this step also incorporates mechanistic or MOA evidence informing the biological plausibility and coherence of the available animal health effect studies.
- Finally, evidence integration combines the animal and human evidence stream conclusions, while taking into consideration the mechanistic or MOA information on the human relevance of the animal evidence, coherence across evidence streams, and susceptibility.

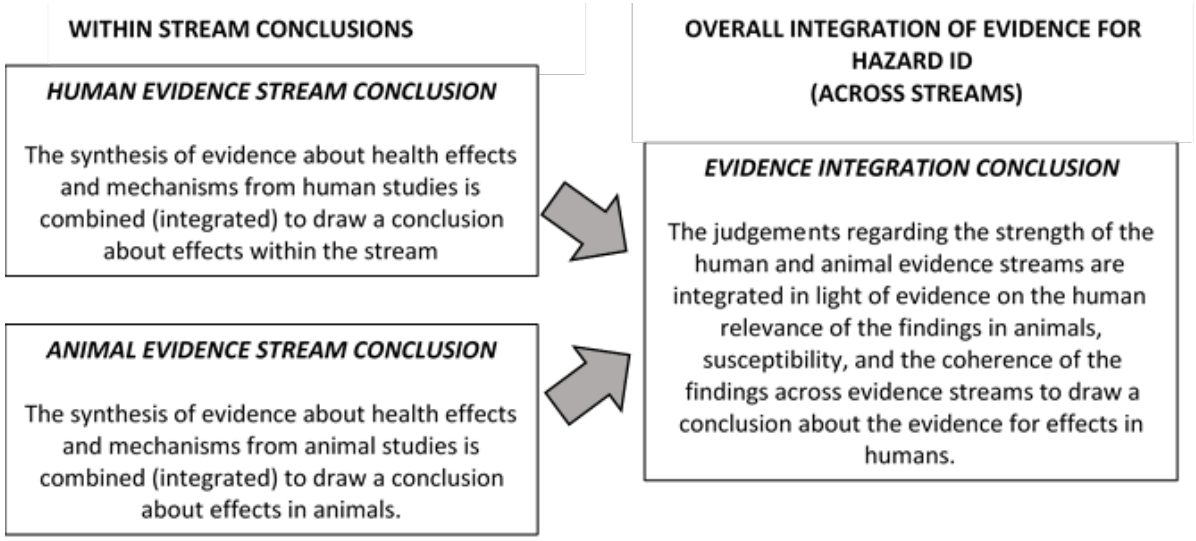

Figure 3. Process for evidence integration.

The within stream syntheses and conclusions and the overall integration of evidence for hazard identification will be summarized in an evidence profile table for each hazard (Figure 4).

Studies and interpretation	Factors that increase confidence	Factors that decrease confidence	Summary of findings	Within stream evidence judgements	Inference across evidence streams	Overall confidence conclusion
[Health Effect or Outcome Grouping]						
Evidence from Human Studies (Route)					*Human relevance of findings in animals* · *Cross-stream coherence (i.e. for both health effect-specific and mechanistic data)* · *Other inferences:* ○ *Information on susceptibility* ○ *MOA analysis inferences: precursors, cross-species inferences of toxicokinetics, or quantitative implications* ○ *Relevant information from other sources (e.g., read across; other, potentially related health hazards)*	*Describe conclusion(s) and primary basis for the integration of all available evidence (e.g., across human, animal, and mechanistic):* + + + Strongest conclusion + + ○ ↑ + ○ ○ Weakest conclusion ○ ○ ○ – ○ ○ Inadequate – – ○ – – – Conclusion of unlikely to be an effect Summarize the models and range of dose levels upon which the conclusions were primarily reliant
· *References* · *Study confidence (based on evaluation of risk of bias and sensitivity) and explanation* · *Study design description*	· *Consistency* · *Dose-response gradient* · *Coherence of observed effects (apical studies)* · *Effect size (magnitude, severity)* · *Biological plausibility* · *Low risk of bias/ high quality* · *Insensitivity of null/ negative studies* · *Natural experiments* · *Temporality*	· *Unexplained inconsistency* · *Imprecision* · *Indirectness/ applicability* · *Poor study quality/ high risk of bias* · *Other (e.g., Single/Few Studies; small sample size)* · *Evidence demonstrating implausibility*	· *Results information (general endpoints affected/ unaffected) across studies* · *Human evidence informing biological plausibility: discuss how mechanistic data influenced the within stream judgement (e.g., evidence of precursors in exposed humans).* *Could be multiple rows (e.g., grouped by study confidence or population) if this informs results heterogeneity*	*Describe confidence in evidence from human studies, and primary basis:* + + + Strongest evidence + + ○ ↑ + ○ ○ Weakest evidence ○ ○ ○ – ○ ○ Inadequate – – ○ – – – Convincing evidence of no effect		
Evidence for an Effect in Animals (Route)						
· *References* · *Study confidence (based on evaluation of risk of bias and sensitivity) and explanation* · *Study design description*	· *Consistency and Replication* · *Dose-response gradient* · *Coherence of observed effects (apical studies)* · *Effect size (magnitude, severity)* · *Biological plausibility* · *Low risk of bias/ high quality* · *Insensitivity of null/ negative studies*	· *Unexplained inconsistency* · *Imprecision* · *Indirectness/ applicability* · *Poor study quality/ high risk of bias* · *Other (e.g., Single/Few Studies; small sample size)* · *Evidence demonstrating implausibility*	· *Results information (general endpoints affected/ unaffected) across studies* · *Evidence informing biological plausibility for effects in animals: discuss how mechanistic data influenced the within stream judgement (e.g., evidence of coherent molecular changes in animal studies)* *Could be multiple rows (e.g., by study confidence, species, or exposure duration) if this informs results heterogeneity*	*Describe confidence in evidence for an effect in animals, and primary basis:* + + + Strongestevidence + + ○ ↑ + ○ ○ Weakest evidence ○ ○ ○ – ○ ○ Inadequate – – ○ – – – Convincing evidence of no effect		

Figure 4. Evidence profile table template.

10.1. INTEGRATION WITHIN HUMAN AND ANIMAL EVIDENCE STREAMS

Prior to drawing the hazard conclusion about whether a chemical is likely to cause a particular health effect(s) in humans, given relevant exposure circumstances, interim judgments are drawn regarding the evidence for humans and animals with regard to each hazard assessed. Tables 10 and 11 describe the evidence bases for human and animal studies, respectively, for each of the standardized conclusions. Briefly, the terms *Robust* and *Moderate* are shorthand characterizations of the standardized conclusions reached for an evidence base that supports the judgment that a hazard is associated with human or animal (depending on the evidence type) exposure to the agent. These terms are differentiated by the quantity and quality of information available to rule out alternative explanations for the results. *Slight* evidence includes situations in which there is some evidence to support a hazard, but with substantial uncertainties in the data and for which a conclusion of *Moderate* does not apply. *Indeterminate* describes a situation in which no studies are available for that evidence stream or in which the evidence is inconsistent and of low confidence, and cannot provide a basis for making a conclusion in either direction. *Compelling evidence of no effect* represents a situation in which extensive evidence across a range of populations and exposures has identified no association. This scenario is rare.

Table 11. Framework for evidence conclusions from studies in humans

Extent of support for hazard	Within-stream strength of evidence conclusion	Description
Supports hazard	*Robust … human evidence of an effect*	A set of *high* or *medium* confidence independent studies reporting an association between the exposure and the health outcome, with reasonable confidence that alternative explanations, including chance, bias, and confounding, can be ruled out across studies. The set of studies is primarily consistent, with reasonable explanations when results differ; an exposure-response gradient is demonstrated; and the set of studies includes varied populations. Additional supporting evidence, such as associations with biologically related endpoints in human studies (coherence) or large estimates of risk, may increase confidence, but are not required. Selective reporting and publication bias are not a reasonable explanation for results. In exceptional circumstances, a finding in one study may be *robust*, even when other studies are not available (e.g., analogous to the finding of angiosarcoma, an exceedingly rare liver cancer, in the vinyl chloride industry). Mechanistic evidence from exposed humans or human cells, if available, may add support informing considerations such as exposure-response, temporality, coherence, and MOA, thus raising the level of certainty to *robust* for a set of studies that otherwise would be described as *moderate*.
	Moderate … human evidence of an effect	A smaller number of studies (at least one *high* or *medium* confidence study with supporting evidence), or with some heterogeneous results, that do not reach the degree of confidence required for *robust.* There is some consistent evidence of an association, but alternative explanations, including chance, bias, and confounding, have not been ruled out. Associations with related endpoints, including mechanistic evidence from exposed humans or human cells, if available, may add support based on considerations such as exposure-response, temporality, coherence, and MOA, thus raising the level of certainty to *moderate* for a set of studies that otherwise would be described as borderline *moderate*/*slight*.
Could support hazard or no hazard	*Slight … human evidence of an effect*	One or more studies reporting an association between exposure and the health outcome, where alternative explanations exist. In general, only *low* confidence studies may be available, or considerable heterogeneity across studies may exist, and a MOA is not understood. Strong biological support from mechanistic studies in exposed humans or human cells may also be independently interpreted as slight. More rarely, a single *high* confidence study that is the initial evaluation of the study question (e.g., a hypothesis-generating vs. hypothesis-testing analysis) would also be described as *slight*. This category serves primarily to encourage additional study where evidence does exist that might provide some support for an association, but for which the evidence does not reach the degree of confidence required for *moderate*.
	Indeterminate … human evidence of an effect	No studies available in humans or only a set of weak studies that are not consistent or are not informative to the hazard question under evaluation.

1

Table 11. Framework for evidence conclusions from studies in humans (continued)

Extent of support for hazard	Within-stream strength of evidence conclusion	Description
Supports no hazard	*Compelling evidence of no effect … in human studies*	Several *high* confidence studies, showing consistently null results (for example, an odds ratio of 1.0) ruling out alternative explanations such as chance, bias, and confounding with reasonable confidence. Each of the studies should have used an optimal outcome and exposure assessment and adequate sample size (specifically for higher exposure groups and for susceptible populations). The set should include the full range of levels of exposures that human beings are known to encounter, an evaluation of an exposure-response gradient, and an examination of at-risk populations and lifestages. The studies should be mutually consistent in not showing any indication of effect at any level of exposure.

Table 12. Framework for evidence conclusions from studies in animals

Extent of support for hazard	Within-stream strength of evidence conclusion	Description
Supports hazard	*Robust … evidence of an effect in animals*	Consistent evidence for effects in animals has been observed in *high* or *medium* confidence studies[a] of varied design; any inconsistent evidence (evidence that cannot be reasonably explained by the respective study design or differences in animal model) is from a set of weaker studies. The set of studies supporting a hazard includes consistent findings of adverse or toxicologically significant effects across multiple laboratories or species, and the design of the studies can reasonably rule out the potential for nonspecific effects (e.g., toxicity) to have resulted in the findings. Multiple lines of additional evidence in the set of studies support a causal association, such as coherent effects across multiple related endpoints (may include mechanistic endpoints); an unusual magnitude of effect, rarity, age at onset, or severity; a strong dose-response relationship; and/or consistent observations across exposure scenarios (e.g., route; timing; duration), sexes, or animal strains. Mechanistic data in animals or animal cells that address the above considerations or that provide experimental support for a MOA that defines a causal relationship with reasonable confidence may raise the level of certainty to *robust* for evidence that otherwise would be described as *moderate* or, exceptionally, *slight, or indeterminate*.
	Moderate … evidence of an effect in animals	A set of evidence that does not reach the degree of certainty required for *robust*, but which includes at least one *high* or *medium* confidence study and supporting information. Although the results are largely consistent, notable uncertainties remain regarding the causal nature of the observed association. However, while inconsistent evidence and/or evidence indicating nonspecific effects may exist, it is not sufficient to reduce or discount the level of concern regarding the positive findings from the supportive studies. Additionally, the set of supportive studies provide evidence supporting a causal association, such as consistent effects across laboratories or species; coherent effects across multiple related endpoints (may include mechanistic endpoints); an unusual magnitude of effect, rarity, age at onset, or severity; a strong dose-response relationship; and/or consistent observations across exposure scenarios (e.g., route, timing, duration), sexes, or animal strains. Mechanistic data in animals or animal cells that address the above considerations or that provide information supporting an association between exposure and effect with reasonable confidence may raise the level of certainty to *moderate* for evidence that otherwise would be described as *slight*.

Table 12. Framework for evidence conclusions from studies in animals (continued)

Extent of support for hazard	Within-stream strength of evidence conclusion	Description
Could support hazard or no hazard	*Slight … evidence in animals*	A lack of relevant studies or a set of experiments for which none of the other conclusions apply. This includes situations in which only *low* confidence experiments are available and a MOA is not understood. Strong biological support from mechanistic studies in exposed animals or animal cells may also be independently interpreted as *slight*. Notably, to encourage additional research, it is important to describe situations for which evidence does exist that might provide some support for an association, but is insufficient for a conclusion of *moderate*.
	Indeterminate …evidence in animals	No animal studies were available, or a set of *low* confidence animal studies exist that are not reasonably consistent or are not informative to the hazard question under evaluation.
Supports no hazard	*Compelling evidence of no effect … in animals*	A set of *high* confidence experiments examining the full spectrum of related endpoints within a type of toxicity, with multiple species, and testing a reasonable range of exposure levels and adequate sample size in both sexes, with none showing any indication of effects. The data are compelling in that the experiments have examined the range of scenarios across which health effects in animals could be observed, and an alternative explanation (e.g., inadequately controlled features of the studies' experimental designs) for the observed lack of effects is not available. The experiments were designed to specifically test for effects of interest, including suitable exposure timing and duration, post-exposure latency, and endpoint evaluation procedures, and to address potentially susceptible populations and lifestages.

[a]"Experiment" is used here to refer to measurements in a single cohort of exposed animals (e.g., a study that included separate evaluations of rats and of mice, or separate cohorts exposed at different lifestages, would be considered as multiple experiments). Conversely, two papers or studies that report on the same cohort of exposed animals (e.g., examining different endpoints) would not be separate experiments. This language is used to reduce confusion regarding the use of the term "study."

10.2. OVERALL INTEGRATION OF EVIDENCE FOR HAZARD IDENTIFICATION

In an IRIS assessment, EPA integrates evidence through a structured process that involves using scientific judgment, applying a standardized approach for evaluating the weight of the evidence, and using clear and consistent summary language (NRC, 2011). As the IRIS Program evaluates multiple health outcomes of many chemical agents, the terms used in these conclusions should be consistent across health outcomes and across assessments. The goal is to communicate

the hazard conclusions clearly and consistently, maintaining the rigor and transparency that systematic review brings to the early stages of an assessment.

This second stage of evidence integration involves combining the animal and human evidence conclusions while also considering mechanistic or MOA information on the human relevance of the animal evidence, coherence across evidence streams, and susceptibility. Coherent results across multiple species, even in the absence of mechanistic understanding, also increases confidence that the animal results are relevant to humans.

Based on the totality of the evidence, this stage culminates in a summary narrative conclusion for each potential health outcome (i.e., each noncancer health effect and specific type of cancer, or broader grouping of related outcomes). This narrative describes the judgements and rationale regarding a chemical exposure's potential to cause each outcome, and the level of confidence in each conclusion. Thus, the evidence integration narrative will include:

- Conclusions about the potential for health effects in exposed humans;
- A summary of key evidence supporting these conclusions, highlighting the line(s) of evidence that were the primary drivers of these conclusions as well as any notable issues with data quality or coherence of the results;
- Information on the conditions of expression of these health effects (e.g., exposure routes);
- A summary of potential MOAs and how they reinforce the conclusions;
- Indications of potentially susceptible populations or lifestages;
- A summary of key assumptions used in the analysis, which are often based on EPA guidelines;
- A narrative expression of confidence in the hazard characterization; and
- Strengths and limitations of the conclusions, including key uncertainties and data gaps.

The current assessment will rely on the conclusions of the 2001 assessment which classified chloroform as likely to be carcinogenic to humans by all routes of exposure under high-exposure conditions that lead to cytotoxicity and regenerative hyperplasia in susceptible tissues (https://cfpub.epa.gov/ncea/iris2/chemicalLanding.cfm?substance_nmbr=25). The 2001 assessment also concluded that chloroform is not likely to be carcinogenic to humans by any route of exposure under exposure conditions that do not cause cytotoxicity and cell regeneration. Currently, EPA does not have guidance on use of standardized descriptors for noncancer hazards, so none will be applied although conclusions will indicate confidence in the body of evidence with exposure context provided.

EPA's standardized hazard descriptors for cancer

Carcinogenic to humans: convincing epidemiologic evidence of a causal association between human exposure and cancer; or strong evidence between human exposure and either cancer or the key precursor events of the agent's MOA, extensive animal evidence of carcinogenicity, identification of mode of action and its key precursors in animals, and strong evidence that the key precursor events that precede the cancer response in animals are anticipated to occur in humans and progress to tumors, based on biological information.

Likely to be carcinogenic to humans: evidence that demonstrates carcinogenic potential to humans but that does not reach the WOE for the prior descriptor. Examples include demonstration of a plausible association between human exposure and cancer with supporting experimental evidence; positive results in animal experiments in more than one species, sex, strain, site or exposure route; a positive tumor study that raises additional biological concerns beyond statistical significance (e.g., a high degree of malignancy, or an early age at onset); a rare animal tumor response that is assumed to be relevant to humans; or a positive tumor study strengthened by other lines of evidence (e.g., plausible association between human exposure and cancer, or evidence that the agent or important metabolite causes events generally known to be associated with tumor formation likely related to tumor response in this case).

Suggestive evidence of carcinogenic potential: evidence that raises a concern for humans but that is judged not sufficient for a stronger WOE conclusion. Examples include a single positive result that may not be statistically significant but is not contradicted by other studies of equal quality in the same population group or experimental system; a small increase in a tumor with a high background rate in that sex and strain, when there is insufficient evidence that the observed tumors may be due to intrinsic factors that cause background tumors and not due to the agent being assessed; evidence of a positive response in a study whose power, design, or conduct limits the ability to draw a confident conclusion, but there the carcinogenic potential is strengthened by other lines of evidence (e.g., structure-activity relationships); or a statistically significant increase at only one dose, but not significant response at the other doses and no overall trend.

Inadequate information to assess carcinogenic potential: no other descriptors apply. Examples include little or no pertinent information, conflicting evidence, or negative results not sufficiently robust for the "*Not Likely" descriptor*.

Not likely to be carcinogenic to humans: robust data for deciding that there is no basis for human hazard concern. Examples include no effects in well-conducted and well-designed studies in both sexes of at least two appropriate animal species (without data in other animals or humans suggesting a potential for carcinogenicity), convincing and extensive evidence showing that the only carcinogenic effects observed in animals are not relevant to humans, or convincing evidence that carcinogenic effects are not likely by a particular exposure route or below a defined dose range.

10.3. SUMMARY OF SUSCEPTIBLE POPULATIONS AND LIFESTAGES

This section will draw from Sections 9 and 10 to describe the evidence (i.e., human, animal, mechanistic) regarding populations and lifestages susceptible to the hazards identified and factors that increase risk of the hazards. Background information about biological mechanisms or ADME, as well as biochemical and physiological differences among lifestages may be used to guide the selection of populations and lifestages to consider. At a minimum, particular consideration will be given to infants and children, pregnant women, and women of childbearing age. Evidence on factors that contribute to some population groups having increased responses to chemical exposure and/or factors that contribute to increases in exposure or dose will be summarized and evaluated

with respect to patterns across studies pertinent to consistency, coherence, and the magnitude and direction of effect measures. Relevant factors may include intrinsic factors (e.g., age, sex, genetics), extrinsic factors (e.g., SES, access to health care), and differential exposure levels or frequency (e.g., occupation-related exposure, residential proximity to locations with greater exposure intensity).

11. DOSE-RESPONSE ASSESSMENT: STUDY SELECTION AND QUANTITATIVE ANALYSIS

The previous sections of this protocol describe how systematic review principles are applied to support transparent identification of health outcomes (or hazards) associated with exposure to the chemical of interest in conjunction with evaluation of the quality of the studies considered during hazard identification. Selection of specific data for dose-response assessment and performance of the dose-response assessment is conducted after hazard identification is complete, and builds off this step in developing the complete IRIS assessment. The dataset selection process involves database- and chemical-specific biological judgments that are beyond the scope of this protocol, but are discussed in existing EPA guidance and support documents. This section of the protocol provides an overview of points to consider when conducting the dose-response assessment, particularly statistical considerations specific to dose response analysis that support quantitative risk assessment. Importantly, the considerations outlined in this protocol do not supersede existing EPA guidance. Several EPA guidance and support documents provide more detailed considerations for the development of EPA's traditional dose-response values, especially EPA's *Review of the Reference Dose and Reference Concentration Processes* (U.S. EPA, 2002), EPA's *Benchmark Dose Technical Guidance* (U.S. EPA, 2012b), *Guidelines for Carcinogen Risk Assessment* (U.S. EPA, 2005a), and *Supplemental Guidance for Assessing Susceptibility from Early-Life Exposure to Carcinogens* (U.S. EPA, 2005b).

For IRIS toxicological reviews, dose-response assessments are typically performed for both noncancer and cancer hazards, and for both oral and inhalation routes of exposure following chronic exposure[9] to the chemical of interest if supported by existing data. As outlined in Section 2 and Section 3, the objective of this assessment is to derive an RfC for chloroform by using inhalation dose-response data from human or animal studies. An RfC is an estimate of an exposure to the human population (including susceptible subgroups) that is likely to be without an appreciable risk of deleterious health effects over a lifetime (U.S. EPA, 2002). For chloroform, an RfC approach can be protective of cancer hazards because the 2001 MOA analysis concluded that for cancer, chloroform exhibits a "threshold" by all routes of exposure, and thus a chloroform dose exists that does not elicit cytotoxicity and presents no cancer risk. Reference values are not predictive risk values; that is, they provide no information about risks at higher or lower exposure levels. The MOA analysis for cancer for chloroform posted on the IRIS website in 2001 will be used to

[9]Dose-response assessments may also be conducted for shorter durations, particularly where the evidence base for an agent indicates the importance of considering such durations to risks posed by an agent (U.S. EPA, 2002).

determine whether this newly derived RfC is protective with respect to cancer. The results of this evaluation is anticipated to result in a new RfC that would replace the existing IUR from 1987.

11.1. SELECTING STUDIES FOR DOSE-RESPONSE ASSESSMENT

The dose-response assessment begins with a re-examination of the studies highlighted in the hazard identification, because some studies that are used qualitatively for hazard identification may or may not be useful quantitatively for dose-response assessment due to such factors as the lack of quantitative measures of exposure or relevant details on responses (e.g., lack of variability measures for continuous data).

Attributes of the studies identified for each hazard are reviewed considering such factors as (1) human relevance of the test species; (2) human relevance of exposure route, duration and magnitude; (3) subject selection methods; (4) controls for possible confounding; (5) methods employed to measure exposure and health outcomes; (6) study size and design; and (7) studies representative of the most susceptible populations. Other aspects of study utility that may be important include investigation of early effects that precede overt toxicity, and adequate reporting of related effects that help characterize overall toxicity (e.g., combining effects that comprise a syndrome, or explicitly describing benign and malignant tumors in a specific tissue). Statistically, confidence in a study considered for dose-response assessment increases with the number of exposure levels tested, especially in the low-dose region of the exposure-response curve, and with increasing sample sizes (U.S. EPA, 2012b). Studies of low sensitivity may be less useful if they fail to detect a true effect or yield toxicity values with wide confidence limits. These attributes support a more complete characterization of the shape of the exposure-response curve, and decrease the uncertainty in the associated exposure-response metric (e.g., RfC) by reducing variability and minimizing the need for low-dose extrapolation. In addition to the more general considerations described above, specific issues that may impact the feasibility of dose-response modeling for individual data sets are described in more detail in the *Benchmark Dose Technical Guidance* (U.S. EPA, 2012b).

11.2. CONDUCTING DOSE-RESPONSE ASSESSMENTS

EPA uses a two-step approach for dose-response assessment that distinguishes analysis of the dose-response data in the range of observation from any inferences about responses at lower environmentally-relevant exposure levels (U.S. EPA, 2012b, 2005a):

1. Within the observed dose range, the preferred approach is to use dose-response modeling to incorporate as much of the data set as possible into the analysis. This modeling yields a point of departure (POD), an exposure level near the lower end of the range of observation, without significant extrapolation to lower exposure levels

2. To derive reference values or a cancer risk estimate, extrapolation to exposures lower than the POD may be necessary.

When both sufficient and appropriate human data and laboratory animal data are available, human data are generally preferred for the dose-response assessment as their use eliminates the need to perform interspecies extrapolations.

For reference values, IRIS assessments typically derive a candidate value from each suitable data set whether it be human or animal. Evaluation of these candidate values grouped within a particular organ/system yields a single organ/system-specific value for each organ/system under consideration. Next, evaluation of these organ/system-specific values results in the selection of a single overall reference value to cover all health outcomes across all organs/systems. While this overall reference value is the focus of the assessment, the organ/system-specific values can be useful for subsequent cumulative risk assessments that consider the combined effect of multiple agents acting at a common anatomical site.

For cancer, if there are multiple tumor sites, final cancer risk estimates will typically address overall cancer risk.

11.2.1. Dose-Response Analysis in the Range of Observation

For conducting a dose-response assessment, toxicodynamic ("biologically based") modeling can be used when there are sufficient data to ascertain the mode of action and quantitatively support model parameters that represent rates and other quantities associated with the key precursor events of the mode of action. Toxicodynamic modeling is potentially the most comprehensive way to account for the biological processes involved in a response. Such models seek to reflect the sequence of key precursor events that lead to a response. Toxicodynamic models can contribute to dose-response assessment by revealing and describing nonlinear relationships between internal dose and response. Such models may provide a useful approach for analysis in the range of observation, provided the purpose of the assessment justifies the effort involved.

When a toxicodynamic model is not available for dose-response assessment or when the purpose of the assessment does not warrant developing such a model, empirical modeling should be used to fit the data (on the apical outcome or a key precursor event) in the range of observation. For this purpose, EPA has developed a standard set of models for modeling animal data (https://www.epa.gov/bmds) that can be applied to typical data sets. Modeling epidemiologic studies is highly specific to the features of a study, so modeling of epidemiologic data will be described in the draft assessment if this type of data is considered for dose-response analysis. In situations where there are alternative models with significant biological support, the decision-maker can be informed by the presentation of these alternatives along with the models' strengths and uncertainties. EPA has developed guidance on modeling dose-response data, assessing model fit, selecting suitable models, and reporting modeling results (see the EPA's Benchmark Dose Technical Guidance) (U.S. EPA, 2012b). Additional judgment or alternative analyses are used if 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.

For each modeled response, a POD from the observed data should be estimated to mark the beginning of extrapolation to lower doses. The POD is an estimated dose (expressed in human-equivalent terms) near the lower end of the observed range without significant extrapolation to lower doses. The POD is used as the starting point for subsequent extrapolations and analyses. For linear extrapolation of cancer risk, the POD is used to calculate an OSF or IUR, and for nonlinear extrapolation the POD is used in the calculation of an RfD or RfC.

The response level at which the POD is calculated is guided by the severity of the endpoint. If 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.

EPA has developed standard approaches for determining the relevant dose to be used in the dose-response modeling in the absence of appropriate toxicokinetic modeling. These standard approaches also facilitate comparison across exposure patterns and species. These standard approaches include:

- 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, 2005a, 1991).
- Doses are standardized to equivalent human terms to facilitate comparison of results from different species. Oral doses are scaled allometrically using $mg/kg^{3/4}$-day as the equivalent dose metric across species. Allometric scaling pertains to equivalence across species, not across lifestages, and is not used to scale doses from adult humans or mature animals to infants or children (U.S. EPA, 2011a, 2005a). Inhalation exposures are scaled using dosimetry models that apply species-specific physiologic and anatomic factors and consider whether the effect occurs at the site of first contact or after systemic circulation (U.S. EPA, 2012a, 1994a).
- 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).

11.2.2. Extrapolation: Slope Factors and Unit Risks

An OSF or IUR facilitates estimation of human cancer risks when low-dose linear extrapolation for cancer effects is supported. This is appropriate for agents with direct mutagenic activity and other agents for which the data indicate a linear component below the POD, and is also used as a default when the data are insufficient to establish the mode of action (U.S. EPA, 2005a). If data are sufficient to ascertain the mode of action and to conclude that it is not linear at low doses, extrapolation may use the reference-value approach (U.S. EPA, 2005a); see Section 11.2.3 below.

Differences in susceptibility may warrant derivation of multiple slope factors or unit risks, with separate estimates for susceptible populations and lifestages (U.S. EPA, 2005a, b). If appropriate chemical-specific data on susceptibility from early life exposures are available, then these data are used to develop cancer slope factors or unit risks that specifically address any potential for differential potency in early lifestages (U.S. EPA, 2005a, b). If such data are not available, the WOE analysis supports a mutagenic MOA for carcinogenicity, and the extrapolation approach is linear, the dose-response assessment should indicate that in the development of risk estimates, the default age-dependent adjustment factors should be used with the cancer slope factor or unit risk and age-specific estimates of exposure (U.S. EPA, 2005a, b).

11.2.3. Extrapolation: Reference Values

Reference value derivation is EPA's most frequently used type of nonlinear extrapolation method, and is most commonly used for noncancer effects. This approach is also used for cancer effects if there are sufficient data to ascertain the MOA and conclude that it is not linear at low doses. For these cases, reference values for each relevant route of exposure are developed following EPA's established practices (U.S. EPA, 2005a, 2002, 1994a); in general, the reference value is based not on tumor incidence, but on a key precursor event in the MOA that is necessary for tumor formation.

For each data set selected for reference value derivation, reference values are estimated by applying relevant adjustments to the PODs to account for the conditions of the reference value definition—for human variation, extrapolation from animals to humans, extrapolation to chronic exposure duration, and extrapolation to a minimal level of risk (if not observed in the data set). Increasingly, data-based adjustments are feasible (U.S. EPA, 2014) and distinguished from the uncertainty factor (UF) considerations outlined below. The assessment will discuss the scientific bases for estimating these data-based adjustments and UFs.

- *Animal-to-human extrapolation*: If animal results are used to make inferences about humans, the reference value derivation incorporates the potential for cross-species differences, which may arise from differences in toxicokinetics or toxicodynamics. If

available, a biologically-based model that adjusts fully for toxicokinetic and toxicodynamic differences across species may be used. Otherwise, the POD is standardized to equivalent human terms or is based on toxicokinetic or dosimetry modeling, that may range from detailed chemical-specific to default approaches (U.S. EPA, 2014, 2011a), and a factor of $10^{1/2}$ (rounded to 3) is applied to account for the remaining uncertainty involving toxicokinetic and toxicodynamic differences.

- *Human variation*: The assessment accounts for variation in susceptibility across the human population and the possibility that the available data may not be representative of individuals who are most susceptible to the effect using a data-based adjustment or UF or a combination of the two. Where appropriate data or models for the effect or for characterizing the internal dose are available, the potential for data-based adjustments for toxicodynamics or toxicokinetics is considered (U.S. EPA, 2014, 2002).[10, 11] Where the use of such data or modeling is not supported, an UF, with a default value of 10 is considered. When sufficient data are available, an intraspecies UF either less than or greater than 10x may be justified (U.S. EPA, 2002). However, this factor is generally reduced only if the POD is derived or adjusted specifically for susceptible individuals (not for a general population that includes both susceptible and non-susceptible individuals) (U.S. EPA, 2002, 1998, 1996, 1994a, 1991). Otherwise, a factor of 10 is generally used to account for this UF.
- *LOAEL to NOAEL*: If a POD is based on a LOAEL, the assessment includes an adjustment to an exposure level where such effects are not expected. . A factor of 3 or 10 is generally applied to extrapolate to a lower exposure expected to be without appreciable effects. A factor other than 10 may be used depending on the magnitude and nature of the response and the shape of the dose-response curve (U.S. EPA, 2002, 1998, 1996, 1994a, 1991).
- *Subchronic-to-chronic exposure*: If a chronic reference value is being developed and a POD is based on subchronic evidence, the assessment considers whether lifetime exposure could have effects at lower levels of exposure. A factor of up to 10 is applied when using subchronic studies to make inferences about chronic/lifetime exposure. A factor other than 10 may be used, depending on the duration of the studies and the nature of the response (U.S. EPA, 2002, 1998, 1994a).
- *Database deficiencies*: In addition to the adjustments above, if database deficiencies raise concern that further studies might identify a more sensitive effect, organ system, or lifestage, the assessment may apply a database UF (U.S. EPA, 2002, 1998, 1996, 1994a,

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

[11]Note that when a PBPK model is available for relating human internal dose to environmental exposure, relevant portions of this UF may be more usefully applied prior to dose-response modeling, depending on the correspondence of any nonlinearities (e.g., saturation levels) between species.

1991). The size of the factor depends on the nature of the database deficiency. For example, the EPA typically follows the suggestion that a factor of 10 be applied if both a prenatal toxicity study and a two-generation reproduction study are missing and a factor of $10^{1/2}$ (i.e., 3) if either one or the other is missing (U.S. EPA, 2002).

The derivation of an RfC for chloroform, and any subsequent cancer analyses conducted as part of the current assessment will be performed consistent with EPA guidance summarized above.

12. PROTOCOL HISTORY

Release date: (January 2018 [*chloroform protocol version 1*])

REFERENCES

Constan, AA; Sprankle, CS; Peters, JM; Kedderis, GL; Everitt, JI; Wong, BA; Gonzalez, FL; Butterworth, BE. (1999). Metabolism of chloroform by cytochrome P450 2E1 is required for induction of toxicity in the liver, kidney, and nose of male mice. Toxicol Appl Pharmacol 160: 120-126. http://dx.doi.org/10.1006/taap.1999.8756.

Corley, RA; Mendrala, AL; Smith, FA; Staats, DA; Gargas, ML; Conolly, RB; Andersen, ME; Reitz, RH. (1990). Development of a physiologically based pharmacokinetic model for chloroform. Toxicol Appl Pharmacol 103: 512-527.

Gemma, S; Vittozzi, L; Testai, E. (2003). Metabolism of chloroform in the human liver and identification of the competent P450s. Drug Metab Dispos 31: 266-274. http://dx.doi.org/10.1124/dmd.31.3.266.

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

Hill, AB. (1965). The environment and disease: Association or causation? Proc R Soc Med 58: 295-300.

Hines, RN; Sargent, D; Autrup, H; Birnbaum, LS; Brent, RL; Doerrer, NG; Hubal, EAC; Juberg, DR; Laurent, C; Luebke, R; Olejniczak, K; Portier, CJ; Slikker, W. (2010). Approaches for Assessing Risks to Sensitive Populations: Lessons Learned from Evaluating Risks in the Pediatric Population. Toxicol Sci 113: 4-26. http://dx.doi.org/10.1093/toxsci/kfp217.

NRC (National Research Council). (2011). Review of the Environmental Protection Agency's draft IRIS assessment of formaldehyde (pp. 1-194). Washington, DC: The National Academies Press. http://dx.doi.org/10.17226/13142.

Sasso, AF; Schlosser, PM; Kedderis, GL; Genter, MB; Snawder, JE; Li, Z; Rieth, S; Lipscomb, JC. (2013). Application of an updated physiologically based pharmacokinetic model for chloroform to evaluate CYP2E1-mediated renal toxicity in rats and mice. Toxicol Sci 131: 360-374. http://dx.doi.org/10.1093/toxsci/kfs320.

Schünemann, H; Hill, S; Guyatt, G; Akl, EA; Ahmed, F. (2011). The GRADE approach and Bradford Hill's criteria for causation. J Epidemiol Community Health 65: 392-395. http://dx.doi.org/10.1136/jech.2010.119933.

Sterne, J; Higgins, J; Reeves, B. (2016). ROBINS-I: a tool for assessing risk of bias in non-randomized studies of interventions, Version 7 March 2016 [Computer Program]. Ottawa, Canada: Cochrane Methods Bias. Retrieved from http://www.riskofbias.info

U.S. EPA (U.S. Environmental Protection Agency). (1988). Recommendations for and documentation of biological values for use in risk assessment (pp. 1-395). (EPA/600/6-87/008). Cincinnati, OH: U.S. Environmental Protection Agency, Office of Research and Development, Office of Health and Environmental Assessment. http://cfpub.epa.gov/ncea/cfm/recordisplay.cfm?deid=34855.

U.S. EPA (U.S. Environmental Protection Agency). (1991). Guidelines for developmental toxicity risk assessment (pp. 1-71). (EPA/600/FR-91/001). Washington, DC: U.S. Environmental Protection Agency, Risk Assessment Forum. http://cfpub.epa.gov/ncea/cfm/recordisplay.cfm?deid=23162.

U.S. EPA (U.S. Environmental Protection Agency). (1994a). Methods for derivation of inhalation reference concentrations and application of inhalation dosimetry [EPA Report] (pp. 1-409). (EPA/600/8-90/066F). Research Triangle Park, NC: U.S. Environmental Protection Agency, Office of Research and Development, Office of Health and Environmental Assessment, Environmental Criteria and Assessment Office. https://cfpub.epa.gov/ncea/risk/recordisplay.cfm?deid=71993&CFID=51174829&CFTOKEN=25006317.

U.S. EPA (U.S. Environmental Protection Agency). (1994b). National primary drinking water regulations: Disinfectants and disinfection byproducts: Proposed rule. Fed Reg 59: 38668.

U.S. EPA (U.S. Environmental Protection Agency). (1996). Guidelines for reproductive toxicity risk assessment (pp. 1-143). (EPA/630/R-96/009). Washington, DC: U.S. Environmental Protection Agency, Risk Assessment Forum.

U.S. EPA (U.S. Environmental Protection Agency). (1998). Guidelines for neurotoxicity risk assessment [EPA Report] (pp. 1-89). (EPA/630/R-95/001F). Washington, DC: U.S. Environmental Protection Agency, Risk Assessment Forum. http://www.epa.gov/risk/guidelines-neurotoxicity-risk-assessment.

U.S. EPA (U.S. Environmental Protection Agency). (2002). A review of the reference dose and reference concentration processes (pp. 1-192). (EPA/630/P-02/002F). Washington, DC: U.S. Environmental Protection Agency, Risk Assessment Forum. http://www.epa.gov/osa/review-reference-dose-and-reference-concentration-processes.

U.S. EPA (U.S. Environmental Protection Agency). (2004). Toxicological review of boron and compounds. In support of summary information on the Integrated Risk Information System (IRIS) [EPA Report]. (EPA/635/04/052). Washington, DC: U.S. Environmental Protection Agency, IRIS. http://nepis.epa.gov/exe/ZyPURL.cgi?Dockey=P1006CK9.txt.

U.S. EPA (U.S. Environmental Protection Agency). (2005a). Guidelines for carcinogen risk assessment [EPA Report] (pp. 1-166). (EPA/630/P-03/001F). Washington, DC: U.S. Environmental Protection Agency, Risk Assessment Forum. http://www2.epa.gov/osa/guidelines-carcinogen-risk-assessment.

U.S. EPA (U.S. Environmental Protection Agency). (2005b). Supplemental guidance for assessing susceptibility from early-life exposure to carcinogens [EPA Report]. (EPA/630/R-03/003F). Washington, DC: U.S. Environmental Protection Agency, Risk Assessment Forum. https://www3.epa.gov/airtoxics/childrens_supplement_final.pdf.

U.S. EPA (U.S. Environmental Protection Agency). (2011a). Recommended use of body weight 3/4 as the default method in derivation of the oral reference dose (pp. 1-50). (EPA/100/R11/0001). Washington, DC: U.S. Environmental Protection Agency, Risk Assessment Forum, Office of the Science Advisor. https://www.epa.gov/risk/recommended-use-body-weight-34-default-method-derivation-oral-reference-dose.

U.S. EPA (U.S. Environmental Protection Agency). (2011b). Toxicological review of trichloroethylene (CASRN 79-01-6) in support of summary information on the Integrated Risk Information System (IRIS) [EPA Report]. (EPA/635/R-09/011F). Washington, DC. http://www.epa.gov/iris/supdocs/0199index.html.

U.S. EPA (U.S. Environmental Protection Agency). (2012a). Advances in inhalation gas dosimetry for derivation of a reference concentration (RfC) and use in risk assessment (pp. 1-140). (EPA/600/R-12/044). Washington, DC. https://cfpub.epa.gov/ncea/risk/recordisplay.cfm?deid=244650&CFID=50524762&CFTOKEN=17139189.

U.S. EPA (U.S. Environmental Protection Agency). (2012b). Benchmark dose technical guidance. (EPA/100/R-12/001). Washington, DC: U.S. Environmental Protection Agency, Risk Assessment Forum. https://www.epa.gov/risk/benchmark-dose-technical-guidance.

U.S. EPA (U.S. Environmental Protection Agency). (2014). Guidance for applying quantitative data to develop data-derived extrapolation factors for interspecies and intraspecies extrapolation [EPA Report]. (EPA/100/R-14/002F). Washington, DC: Risk Assessment Forum, Office of the Science Advisor. https://www.epa.gov/risk/guidance-applying-quantitative-data-develop-data-derived-extrapolation-factors-interspecies-and.

Vesterinen, HM; Sena, ES; Egan, KJ; Hirst, TC; Churolov, L; Currie, GL; Antonic, A; Howells, DW; Macleod, MR. (2014). Meta-analysis of data from animal studies: a practical guide. J Neurosci Methods 221: 92-102. http://dx.doi.org/10.1016/j.jneumeth.2013.09.010.

APPENDICES

APPENDIX A. ELECTRONIC DATABASE SEARCH STRATEGIES

Table A-1. Database search strategy

Search	Search Strategy	Date and Results
PUBMED	((((("chloroform"[MeSH Terms] OR "1,1,1-trichloromethane"[All Fields]) OR "chloroforme"[All Fields]) OR "trichloromethane"[All Fields]) OR "67-66-3"[EC/RN Number]) AND ("2009"[PDAT] : "3000"[PDAT])	10/26/2017: 1,133
WEB OF SCIENCE	(TS="chloroform" OR TS="1,1,1-trichloromethane" OR TS="chloroforme" OR TS="trichloromethane") AND PY=(2009-2017) NOT (SU="PHYSICS" OR SU="PLANT SCIENCES" OR SU="ENERGY FUELS" OR SU="INSTRUMENTS INSTRUMENTATION" OR SU="COMPUTER SCIENCE" OR SU="LEGAL MEDICINE" OR SU="METALLURGY METALLURGICAL ENGINEERING" OR SU="MECHANICS" OR SU="EDUCATION EDUCATIONAL RESEARCH" OR SU="ACOUSTICS" OR SU="GEOCHEMISTRY GEOPHYSICS" OR SU="MATHEMATICS" OR SU="FORESTRY" OR SU="AUTOMATION CONTROL SYSTEMS" OR SU="MINING MINERAL PROCESSING" OR SU="CONSTRUCTION BUILDING TECHNOLOGY" OR SU="ASTRONOMY ASTROPHYSICS" OR SU="ARCHAEOLOGY" OR SU="OPERATIONS RESEARCH MANAGEMENT SCIENCE" OR SU="ANTHROPOLOGY" OR SU="SPORT SCIENCES" OR SU="ART" OR SU="PALEONTOLOGY" OR SU="TELECOMMUNICATIONS" OR SU="CHEMISTRY" OR SU="POLYMER SCIENCE" OR SU="ENGINEERING" OR SU="ENVIRONMENTAL SCIENCES ECOLOGY" OR SU="FOOD SCIENCE TECHNOLOGY" OR SU="SCIENCE TECHNOLOGY OTHER TOPICS" OR SU="BIOTECHNOLOGY APPLIED MICROBIOLOGY" OR SU="AGRICULTURE" OR SU="SPECTROSCOPY" OR SU="CRYSTALLOGRAPHY" OR SU="INTEGRATIVE COMPLEMENTARY MEDICINE" OR SU="WATER RESOURCES" OR SU="NUTRITION DIETETICS" OR SU="LIFE SCIENCES BIOMEDICINE OTHER TOPICS" OR SU="PARASITOLOGY" OR SU="THERMODYNAMICS" OR SU="OPTICS" OR SU="BIOPHYSICS" OR SU="TROPICAL MEDICINE" OR SU="VETERINARY SCIENCES" OR SU="RESEARCH EXPERIMENTAL MEDICINE" OR SU="MARINE FRESHWATER BIOLOGY" OR SU="METEOROLOGY ATMOSPHERIC SCIENCES" OR SU="GEOLOGY" OR SU="ELECTROCHEMISTRY" OR SU="GENERAL INTERNAL MEDICINE" OR SU="DENTISTRY ORAL SURGERY MEDICINE" OR SU="ENTOMOLOGY" OR SU="NUCLEAR SCIENCE TECHNOLOGY" OR SU="INFECTIOUS DISEASES" OR SU="FISHERIES" OR SU="OCEANOGRAPHY" OR SU="ANESTHESIOLOGY" OR SU="ZOOLOGY" OR SU="VIROLOGY" OR SU="RADIOLOGY NUCLEAR MEDICINE MEDICAL IMAGING" OR SU="MEDICAL LABORATORY TECHNOLOGY" OR SU="MYCOLOGY" OR SU="SURGERY" OR SU="BIODIVERSITY CONSERVATION" OR SU="OBSTETRICS GYNECOLOGY" OR SU="EVOLUTIONARY BIOLOGY" OR SU="PSYCHIATRY" OR SU="REMOTE SENSING" OR SU="PEDIATRICS" OR SU="MINERALOGY" OR SU="TRANSPLANTATION" OR SU="MICROSCOPY" OR SU="RHEUMATOLOGY" OR SU="GERIATRICS GERONTOLOGY" OR SU="ORTHOPEDICS" OR SU="MATERIALS SCIENCE")	10/26/2017: 1,283

Table A-1. Database search strategy (continued)

Search	Search Strategy	Date and Results
TOXLINE March 2017	@AND+@OR+(chloroform+"1,1,1+trichloromethane"+chloroforme+trichloromethane+@TERM+@rn+"67+66+3")+@RANGE+yr+2009+2017+@NOT+@org+"nih+reporter"	3/2017: 1,283
TOXLINE October 26, 2017 update	@AND+@OR+(chloroform+chloroforme+trichloromethane+@TERM+@rn+67+66+3)+@RANGE+yr+2009+2017+@NOT+@org+pubmed+pubdart+"nih+reporter"	10/26/2017: 1,283

APPENDIX B. TYPICAL DATA EXTRACTION FIELDS

Table B-1. Typical data extraction fields

Field label	Data extraction elements
HUMAN	
Funding	Funding source(s)
	Reporting of conflict of interest (COI) by authors
Subjects	Study population name/description
	Dates of study and sampling time frame
	Geography (country, region, state, etc.)
	Demographics (sex, race/ethnicity, age, or lifestage at exposure and at outcome assessment)
	Number of subjects (target, enrolled, *n* per group in analysis, and participation/follow-up rates)
	Inclusion/exclusion criteria/recruitment strategy
	Description of reference group
Methods	Study design (e.g., prospective or retrospective cohort, nested case-control study, cross-sectional, population-based case-control study, intervention, case report, etc.)
	Length of follow-up
	Health outcome category (e.g., cardiovascular)
	Health outcome (e.g., blood pressure)
	Diagnostic or methods used to measure health outcome
	Confounders or modifying factors and how considered in analysis (e.g., included in final model, considered for inclusion but determined not needed)
	Chemical name and CAS number
	Exposure assessment (e.g., blood, urine, hair, air, drinking water, job classification, residence, administered treatment in controlled study, etc.)
	Methodological details for exposure assessment (e.g., HPLC-MS/MS, limit of detection)
	Statistical methods

Table B-1. Typical data extraction fields (continued)

Field label	Data extraction elements
Results	Exposure levels (e.g., mean, median, measures of variance as presented in paper, such as SD, SEM, 75^{th}/90^{th}/95^{th} percentile, minimum/maximum); range of exposure levels, number of exposed cases
	Statistical findings (e.g., adjusted β, standardized mean difference, adjusted odds ratio, standardized mortality ratio, relative risk, etc.) or description of qualitative results. When possible, convert measures of effect to a common metric with associated 95% confidence intervals (CI). Most often, measures of effect for continuous data are expressed as mean difference, standardized mean difference, and percentage control response. Categorical data are typically expressed as odds ratio, relative risk (RR, also called risk ratio), or β values, depending on what metric is most commonly reported in the included studies and ability to obtain information for effect conversions from the study or through author query.
	Observations on dose-response (e.g., trend analysis, description of whether dose-response shape appears to be monotonic, nonmonotonic)
Other	Documentation of author queries, use of digital rulers to estimate data values from figures, exposure unit, and statistical result conversions, etc.
ANIMAL	
Funding	Funding source(s)
	Reporting of COI by authors
Animal Model	Sex
	Species
	Strain
	Source of animals
	Age or lifestage at start of dosing and at health outcome assessment
	Diet and husbandry information (e.g., diet name/source)
Treatment	Chemical name and CAS number
	Source of chemical
	Purity of chemical
	Dose levels or concentration (as presented and converted to mg/kg bw/day when possible)
	Other dose-related details, such as whether administered dose level was verified by measurement, information on internal dosimetry
	Vehicle used for exposed animals
	Route of administration (e.g., oral, inhalation, dermal, injection)
	Duration and frequency of dosing (e.g., hours, days, weeks when administration was ended, days per week)

Table B-1. Typical data extraction fields (continued)

Field label	Data extraction elements
Methods	Study design (e.g., single treatment, acute, subchronic [e.g., 90 days in a rodent], chronic, multigenerational, developmental, other)
	Guideline compliance (i.e., use of EPA, OECD, NTP, or another guideline for study design, conducted under GLP guideline conditions, non-GLP but consistent with guideline study, non-guideline peer-reviewed publication)
	Number of animals per group (and dams per group in developmental studies)
	Randomization procedure, allocation concealment, blinding during outcome assessment
	Method to control for litter effects in developmental studies
	Use of negative controls and whether controls were untreated, vehicle-treated, or both
	Report on data from positive controls—was expected response observed?
	Endpoint health category (e.g., reproductive)
	Endpoint (e.g., infertility)
	Diagnostic or method to measure endpoint
	Statistical methods
Results	Measures of effect at each dose or concentration level (e.g., mean, median, frequency, and measures of precision or variance) or description of qualitative results. When possible, convert measures of effect to a common metric with associated 95% confidence intervals (CI). Most often, measures of effect for continuous data will be expressed as mean difference, standardized mean difference, and percentage control response. Categorical data will be expressed as relative risk (also called risk ratio).
	No observed effect level (NOEL), lowest observed effect level (LOEL), benchmark dose (BMD) analysis, statistical significance of other dose levels, or other estimates of effect presented in paper. *Note*: The NOEL and LOEL are highly influenced by study design, do not give any quantitative information about the relationship between dose and response, and can be subject to author's interpretation (e.g., a statistically significant effect may not be considered biologically important). Also, a NOEL does not necessarily mean zero response. Ideally, the response rate at specific dose levels is used as the primary measure to characterize the response.
	Observations on dose-response (e.g., trend analysis, description of whether dose-response shape appears to be monotonic, nonmonotonic)
	Data on internal concentration, toxicokinetics, or toxicodynamics (when reported)
Other	Documentation of author queries, use of digital rulers to estimate data values from figures, exposure unit, and statistical result conversions, etc.