



EPA/635/R-23/014
IRIS Assessment Protocol
www.epa.gov/iris

**Protocol for the Ethylbenzene IRIS Assessment
(Preliminary Assessment Materials)**

(CASRN 100-41-4)

February 2023

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

DISCLAIMER

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

CONTENTS

AUTHORS CONTRIBUTORS REVIEWERS.....	ix
1. INTRODUCTION	1-1
2. SCOPING AND INITIAL PROBLEM FORMULATION.....	2-1
2.1. BACKGROUND.....	2-1
2.1.1. Physical and Chemical Properties.....	2-1
2.1.2. Sources, Production, and Uses.....	2-2
2.1.3. Environmental Fate and Transportation	2-3
2.1.4. Potential for Human Exposure and Populations with Potentially Greater Exposure.....	2-3
2.2. SCOPING AND PROBLEM FORMULATION SUMMARY	2-3
2.3. KEY SCIENCE ISSUES	2-4
3. OVERALL OBJECTIVES AND SPECIFIC AIMS.....	3-1
3.1. OBJECTIVES	3-1
3.2. SPECIFIC AIMS	3-1
4. LITERATURE SEARCH, SCREENING, AND INVENTORY	4-1
4.1. POPULATIONS, EXPOSURES, COMPARATORS, AND OUTCOMES (PECO) CRITERIA FOR THE SYSTEMATIC EVIDENCE MAP	4-1
4.2. SUPPLEMENTAL CONTENT SCREENING CRITERIA.....	4-2
4.3. LITERATURE SEARCH STRATEGIES.....	4-7
4.3.1. Database Search Term Development.....	4-7
4.3.2. Database Searches	4-7
4.3.3. Searching Other Sources	4-8
4.3.4. Non-Peer-Reviewed Data	4-9
4.4. LITERATURE SCREENING	4-10
4.4.1. Title and Abstract Screening.....	4-10
4.4.2. Full-Text Screening	4-10
4.4.3. Multiple Citations with the Same Data	4-11
4.4.4. Literature Flow Diagrams	4-11
4.5. LITERATURE INVENTORY	4-13
4.5.1. Studies That Meet Problem Formulation PECO Criteria	4-13
4.5.2. Organizational Approach for Supplemental Material	4-13

This document is a draft for review purposes only and does not constitute Agency policy.

Protocol for the Ethylbenzene IRIS Assessment

4.6. SUMMARY-LEVEL LITERATURE INVENTORIES.....4-13

5. REFINE PROBLEM FORMULATION AND SPECIFY ASSESSMENT APPROACH..... 5-1

5.1. ASSESSMENT PECO CRITERIA..... 5-1

5.1.1. Other Exclusions Based on Full-Text Content 5-2

5.2. UNITS OF ANALYSES FOR DEVELOPING EVIDENCE SYNTHESIS AND INTEGRATION
JUDGMENTS FOR HEALTH EFFECT CATEGORIES..... 5-3

6. STUDY EVALUATION (RISK OF BIAS AND SENSITIVITY)..... 6-1

6.1. STUDY EVALUATION OVERVIEW FOR HEALTH EFFECT STUDIES..... 6-1

6.2. EPIDEMIOLOGY STUDY EVALUATION 6-5

6.2.1. Epidemiological Study Evaluation Considerations Specific to Exposure Domain for
Ethylbenzene 6-16

6.2.2. Exposure Assessment Approaches used in Epidemiology Studies of Ethylbenzene
and Potential Misclassification..... 6-16

6.2.3. ADME and Notes Relevant to Biomarkers..... 6-18

6.2.4. Time Frames Represented by Exposure Assessments 6-19

6.2.5. Correlation Between BTEX Compounds and Potential Confounding..... 6-19

6.2.6. Exposure Domain Evaluation Levels..... 6-19

6.3. CONTROLLED HUMAN EXPOSURE STUDY EVALUATION 6-22

6.4. EXPERIMENTAL ANIMAL STUDY EVALUATION 6-22

6.5. IN VITRO AND OTHER MECHANISTIC STUDY EVALUATION 6-31

6.6. PHYSIOLOGICALLY BASED PHARMACOKINETIC (PBPK) MODEL DESCRIPTIVE SUMMARY
AND EVALUATION 6-41

6.6.1. Pharmacokinetic (PK)/Physiologically Based Pharmacokinetic (PBPK) Model
Descriptive Summary..... 6-41

6.6.2. Pharmacokinetic (PK)/Physiologically Based Pharmacokinetic (PBPK) Model
Evaluation 6-43

6.6.3. Selection of the Appropriate Dose Metric 6-44

7. DATA EXTRACTION OF STUDY METHODS AND RESULTS..... 7-1

8. EVIDENCE SYNTHESIS AND INTEGRATION..... 8-1

8.1. EVIDENCE SYNTHESIS..... 8-5

8.2. EVIDENCE INTEGRATION..... 8-15

9. DOSE-RESPONSE ASSESSMENT: SELECTING STUDIES AND QUANTITATIVE ANALYSIS..... 9-1

9.1. OVERVIEW..... 9-1

9.2. SELECTING STUDIES FOR DOSE-RESPONSE ASSESSMENT 9-2

9.2.1. Hazard and MOA Considerations for Dose Response 9-2

This document is a draft for review purposes only and does not constitute Agency policy.

Protocol for the Ethylbenzene IRIS Assessment

9.3. CONDUCTING DOSE-RESPONSE ASSESSMENTS..... 9-8

 9.3.1. Dose-Response Analysis in the Range of Observation 9-8

 9.3.2. Extrapolation: Slope Factors and Unit Risk 9-11

 9.3.3. Extrapolation: Reference Values 9-11

10. PROTOCOL HISTORY 10-1

REFERENCES R-1

APPENDIX A. ELECTRONIC DATABASE SEARCH STRATEGIES A-1

APPENDIX B. PROCESS FOR SEARCHING AND COLLECTING EVIDENCE FROM SELECTED OTHER
 RESOURCES B-1

TABLES

Table 2-1. Predicted or experimental physicochemical properties of ethylbenzene.....	2-1
Table 2-2. EPA program and regional office interest in an updated ethylbenzene assessment.....	2-4
Table 4-1. Problem formulation populations, exposures, comparators, and outcomes (PECO) criteria for the ethylbenzene assessment	4-1
Table 4-2. Categories of potentially relevant supplemental material	4-3
Table 5-1. Assessment PECO criteria for the ethylbenzene assessment.....	5-1
Table 5-2. Human and animal endpoint grouping categories.	5-3
Table 6-1. Information relevant to evaluation domains for epidemiology studies	6-6
Table 6-2. Questions to guide the development of criteria for each domain in epidemiology studies.....	6-7
Table 6-3. Estimates representing total individual-level exposure based on personal or residential monitoring	6-19
Table 6-4. Exposure to ethylbenzene in ambient air	6-21
Table 6-5. Domains, questions, and general considerations to guide the evaluation of animal toxicology studies	6-23
Table 6-6. Domains, questions, and general considerations to guide the evaluation of in vitro studies.....	6-33
Table 6-7. Example descriptive summary for a physiologically based pharmacokinetic (PBPK) model study	6-42
Table 6-8. Criteria for evaluating physiologically based pharmacokinetic (PBPK) models.....	6-44
Table 8-1. Generalized evidence profile table to show the relationship between evidence synthesis and evidence integration to reach judgment of the evidence for hazard	8-3
Table 8-2. Generalized evidence profile table to show the key findings and supporting rationale from mechanistic analyses.....	8-4
Table 8-3. Considerations that inform judgments of the certainty of the evidence for hazard for each unit of analysis.....	8-7
Table 8-4. Framework for evidence synthesis judgments from studies in humans	8-11
Table 8-5. Framework for evidence synthesis judgments from studies in animals.....	8-13
Table 8-6. Considerations that inform evidence integration judgments.....	8-15
Table 8-7. Framework for summary evidence integration judgments in the evidence integration narrative.....	8-18
Table 9-1. Attributes used to evaluate studies for derivation of toxicity values (in addition to the health effect category-specific evidence integration judgment)	9-4
Table 9-2. Example table used in assessment to show endpoint consideration judgments for POD derivation.....	9-6
Table 9-3. Specific example of presenting endpoints considered for dose-response modeling and derivation of points of departure.	9-6
Table A-1. Database search strategy	A-1
Table B-1. Summary table for ethylbenzene other sources search results (12/2021)	B-4

FIGURES

Figure 1-1. IRIS systematic review problem formulation and method documents.....	1-2
Figure 4-1. Literature flow diagram for ethylbenzene.....	4-12
Figure 4-2. Inventory heatmap of PECO-relevant ethylbenzene human studies by study design and health system. An interactive version, which includes a list of citations with additional study details and summary of the results, is available here.	4-15
Figure 4-3. Inventory heatmap of PECO-relevant ethylbenzene animal studies by study design and health system. An interactive version, which includes a list of citations with additional study details and summary of the results, is available here.	4-16
Figure 4-4. Literature tag tree of the supplemental studies identified from the ethylbenzene literature searches. An interactive version, which includes a list of citations with additional study details and summary of the results, is available here.	4-17
Figure 4-5. High throughput screening bioactivity data from the CompTox Chemicals Dashboard. An interactive version, which includes a list of citations with additional study details and summary of the results, is available here.	4-18
Figure 6-1. Overview of IRIS study evaluation process. (a) An overview of the evaluation process. (b) The evaluation domains and definitions for ratings (i.e., domain and overall judgments, performed on an outcome-specific basis).	6-2

ABBREVIATIONS

AC50	activity concentration at 50%	CASRN	Chemical Abstracts Service registry number
ADME	absorption, distribution, metabolism, and excretion	CERCLA	Comprehensive Environmental Response, Compensation, and Liability Act
AIC	Akaike's information criterion	CHO	Chinese hamster ovary (cell line cells)
ALT	alanine aminotransferase	CI	confidence interval
AOP	adverse outcome pathway	CL	confidence limit
AST	aspartate aminotransferase	CMAQ	Community Multi-scale Air Quality model
atm	atmosphere	CNS	central nervous system
ATSDR	Agency for Toxic Substances and Disease Registry	COI	conflict of interest
BMC	benchmark concentration	CPAD	Chemical and Pollutant Assessment Division
BMCL	benchmark concentration lower confidence limit	CPHEA	Center for Public Health and Environmental Assessment
BMD	benchmark dose	CYP450	cytochrome P450
BMDL	benchmark dose lower confidence limit	DAF	dosimetric adjustment factor
BMDS	Benchmark Dose Software	DMSO	dimethylsulfoxide
BMR	benchmark response	DNA	deoxyribonucleic acid
BTEX	benzene, toluene, ethylbenzene, o-xylene, m-/p-xylene	EPA	Environmental Protection Agency
BUN	blood urea nitrogen	ER	extra risk
BW	body weight	FDA	Food and Drug Administration
BW ^{3/4}	body weight scaling to the 3/4 power	FEV ₁	forced expiratory volume of 1 second
CA	chromosomal aberration	GD	gestation day
CAA	Clean Air Act		
CAS	Chemical Abstracts Service		

This document is a draft for review purposes only and does not constitute Agency policy.

Protocol for the Ethylbenzene IRIS Assessment

GDH	glutamate dehydrogenase	QSAR	quantitative structure-activity relationship
GGT	γ -glutamyl transferase	RD	relative deviation
GLP	Good Laboratory Practice	RfC	inhalation reference concentration
GSH	glutathione	RfD	oral reference dose
GST	glutathione-S-transferase	RGDR	regional gas dose ratio
HAP	hazardous air pollutant	RNA	ribonucleic acid
HAWC	Health Assessment Workspace Collaborative	ROBINS I	Risk of Bias in Nonrandomized Studies of Interventions
Hb/g-A	animal blood:gas partition coefficient	SAR	structure-activity relationship
Hb/g-H	human blood:gas partition coefficient	SCE	sister chromatid exchange
HBCD	hexabromocyclododecane	SD	standard deviation
HEC	human equivalent concentration	SDH	sorbitol dehydrogenase
HED	human equivalent dose	SE	standard error
HERO	Health and Environmental Research Online	SGOT	serum glutamic oxaloacetic transaminase, also known as AST
i.p.	intraperitoneal	SGPT	serum glutamic pyruvic transaminase, also known as ALT
i.v.	intravenous	TIAB	title and abstract
IAP	IRIS Assessment Plan	TSCATS	Toxic Substances Control Act Test Submissions
IARC	International Agency for Research on Cancer	TWA	time-weighted average
IRIS	Integrated Risk Information System	UF	uncertainty factor
IUR	inhalation unit risk	UF _A	animal-to-human uncertainty factor
LC ₅₀	median lethal concentration	UF _D	database deficiencies uncertainty factor
LD ₅₀	median lethal dose	UF _H	human variation uncertainty factor
LOAEL	lowest-observed-adverse-effect level	UF _L	LOAEL-to-NOAEL uncertainty factor
LOEL	lowest-observed-effect level	UF _S	subchronic-to-chronic uncertainty factor
LUR	land use regression	WOS	Web of Science
MeSH	Medical Subject Headings		
MLE	maximum likelihood estimation		
MN	micronuclei		
MNPCE	micronucleated polychromatic erythrocyte		
MOA	mode of action		
MTD	maximum tolerated dose		
NCI	National Cancer Institute		
NMD	normalized mean difference		
NOAEL	no-observed-adverse-effect level		
NOEL	no-observed-effect level		
NTP	National Toxicology Program		
NZW	New Zealand White (rabbit breed)		
OAR	Office of Air and Radiation		
OECD	Organisation 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, exposures, comparators, and outcomes		
PK	pharmacokinetic		
PND	postnatal day		
POD	point of departure		
POD _[AD]	duration-adjusted POD		

This document is a draft for review purposes only and does not constitute Agency policy.

AUTHORS | CONTRIBUTORS | REVIEWERS

Assessment Managers

[Laura Dishaw](#), Ph.D. EPA/ORD/CPHEA
[Paul G. Reinhart](#), Ph.D.

Assessment Team

Timothy Anderson, Ph.D. EPA/ORD/CPHEA
Christine Cai, Ph.D.
[Ingrid Druwe](#), Ph.D.
[Yu-Sheng Lin](#), Ph.D.
Anuradha Mudipalli, Ph.D.
Rebecca Nachman, Ph.D.
[Rachel Shaffer](#), Ph.D.
John Stanek, Ph.D.
George Woodall, Ph.D.

[Brittany Schulz](#), B.S. Student Services Contractor, Oak Ridge
Associated Universities (ORAU)

Executive Direction

Wayne Cascio, M.D. (CPHEA Director) EPA/ORD/CPHEA
V. Kay Holt, M.S. (CPHEA Deputy Director)
Samantha Jones, Ph.D. (CPHEA Associate
Director)
Kristina Thayer, Ph.D. (CPAD Director)
Steve Dutton, Ph.D. (HEEAD Director)
Andrew Kraft, Ph.D. (CPAD Associate Director)
Ravi Subramaniam, Ph.D. (Acting CPAD Senior
Advisor)
Paul White, Ph.D. (CPAD Senior Science Advisor)
Andrew Hotchkiss, Ph.D. (Branch Chief)
Janice Lee, Ph.D. (Branch Chief)
Elizabeth Radke-Farabaugh, Ph.D. (Branch
Chief)
Viktor Morozov, Ph.D. (Branch Chief)
Garland Waleko, M.S. (Acting Branch Chief)

Contributors

Michelle Angrish, Ph.D. EPA/ORD/CPHEA
Andrew Shapiro

Production Team

Maureen Johnson (CPHEA Webmaster)
Ryan Jones (HERO Director)
Dahnish Shams (Project Management Team)
Vicki Soto (Project Management Team)
Jessica Soto-Hernandez (Project Management Team)
Samuel Thacker (HERO Team)

EPA/ORD/CPHEA

1. INTRODUCTION

1 The Integrated Risk Information System (IRIS) Program is undertaking a reassessment of
2 the health effects of ethylbenzene. IRIS assessments provide high quality, publicly available hazard
3 identification and dose-response analyses on chemicals to which the public might be exposed.
4 These assessments are not regulations but provide an important source of toxicity information
5 used by the Environmental Protection Agency (EPA), state and local health agencies, tribes, other
6 federal agencies, and international health organizations.

7 A draft IRIS assessment plan (IAP) for ethylbenzene was presented at a public science
8 meeting on September 27–28, 2017 (<https://sab.epa.gov/ords/sab/f?p=100:19:3574465722633>)
9 to seek input on the problem formulation components of the assessment plan. The 2017 IAP
10 specified the EPA need for an ethylbenzene assessment, described the objectives and specific aims
11 of the assessment, provided draft PECO (populations, exposures, comparators, and outcomes)
12 criteria, and described areas of scientific complexity. However, in April 2019 the ethylbenzene
13 assessment was suspended due to changes in how EPA identified priorities for the IRIS Program
14 ([April 2019 IRIS Program Outlook](#)). In June 2021, the assessment work was restarted after interest
15 was expressed by EPA’s Office of Land and Emergency Management (OLEM), Office of Chemical
16 Safety and Pollution Prevention (OCSPP), and Region 2. This assessment may also be used to
17 support actions in other EPA Program and Regional Offices and can inform efforts to address
18 ethylbenzene by tribes, states, and international health agencies (see Section 2.2).

19 This protocol document includes the IAP content, revised based on public input, and
20 updated EPA scoping needs and presents the methods for conducting the systematic review and
21 dose-response analysis for the assessment. While the IAP describes *what* the assessment will cover,
22 this protocol describes *how* the assessment will be conducted (see Figure 1-1). The methods
23 described in this protocol are based on the Office of Research and Development (ORD) Staff
24 Handbook for Developing Integrated Risk Information System (IRIS) Assessments (referred to as
25 the “IRIS Handbook”) ([U.S. EPA, 2022](#)).

Protocol for the Ethylbenzene IRIS Assessment

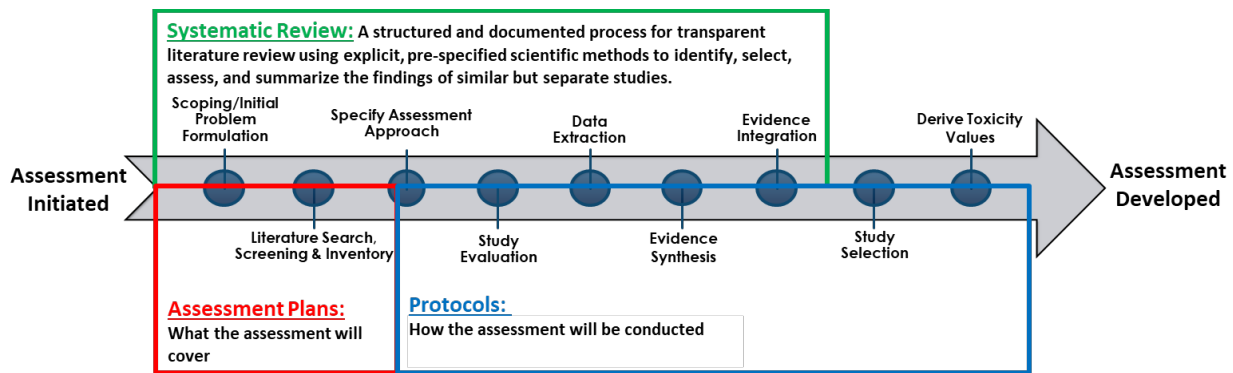


Figure 1-1. IRIS systematic review problem formulation and method documents.

2. SCOPING AND INITIAL PROBLEM FORMULATION

2.1. BACKGROUND

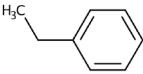
Section 2.1 provides a brief overview of aspects of the physicochemical properties, human exposure, and environmental fate characteristics of ethylbenzene that might provide useful context for this protocol. This overview is not intended to provide a comprehensive description of the available information on these topics and is not recommended for use in decision-making. The reader is encouraged to refer to the source materials cited below, more recent publications on these topics, and authoritative reviews or assessments focused on these topics.

A previous assessment of ethylbenzene is available on the IRIS website (https://cfpub.epa.gov/ncea/iris2/chemicalLanding.cfm?substance_nmbr=51) (U.S. EPA, 1991b). An oral RfD of 1×10^{-1} mg/kg-day was posted in 1987 based on hepatic and renal toxicity. An inhalation RfC of 1 mg/m³ was posted in 1991 based on developmental toxicity. In 1988 the cancer weight of evidence for ethylbenzene was categorized as “Group D,” that is, not classified concerning its potential to cause cancer in humans, due to a lack of animal and human data. Since then, several relevant studies on ethylbenzene toxicity have been completed and new data have become available.

2.1.1. Physical and Chemical Properties

Ethylbenzene is a colorless flammable liquid with a sweet, gasoline-like odor (ATSDR, 2010). Various physical and chemical properties are presented in Table 2-1 below.

Table 2-1. Predicted or experimental physicochemical properties of ethylbenzene

Characteristic or property (unit)	Value ^a	Reference
Chemical structure		U.S. EPA (2021)
CASRN	100-41-4	U.S. EPA (2021)
Synonyms	1-ethylbenzene, alpha-methyltoluene, ethylbenzol, phenylethane, EB	U.S. EPA (2021)
Color/form	colorless liquid	U.S. EPA (2021)

Characteristic or property (unit)	Value ^a	Reference
Molecular formula	C ₆ H ₅ CH ₂ CH ₃	U.S. EPA (2021)
Molecular weight (g/mol)	106.168	U.S. EPA (2021)
Density (g/cm ³)	0.879 ^b	U.S. EPA (2021)
Boiling point (°C)	136	U.S. EPA (2021)
Melting point (°C)	-95.0	U.S. EPA (2021)
Heat of formation (kJ/mol)	-12.55	ANL (2021)
Log K _{ow}	3.15	U.S. EPA (2021)
K _{oc} (L/kg)	170	U.S. EPA (2021)
Henry's law constant (atm·m ³ /mol)	7.88 × 10 ⁻³	U.S. EPA (2021)
Solubility in water (mol/L)	1.64 × 10 ⁻³	U.S. EPA (2021)
Vapor pressure (mmHg)	9.60	U.S. EPA (2021)

1 ppm = 4.34 mg/m³ at 25 °C ([ATSDR, 2010](#)).

^aWhen available, average experimental values are reported from [U.S. EPA \(2021\)](#) Chemicals Dashboard (Ethylbenzene DTXSID3020596): <https://comptox.epa.gov/dashboard/chemical/details/DTXSID3020596>.

^bPredicted values are provided when experimental values are not available but may be less reliable than experimental values.

2.1.2. Sources, Production, and Uses

1 Ethylbenzene can be found naturally in crude petroleum and in numerous man-made
 2 products for industrial and consumer use. Exposure to ethylbenzene can occur via releases to the
 3 air, water, and soil during the manufacturing process ([ATSDR, 2010](#)) and from burning fossil fuels
 4 (automobile exhaust and small gasoline engines).

5 Ethylbenzene is produced by the alkylation of benzene with ethylene in liquid-phase or by
 6 vapor-phase reaction of benzene with dilute ethylene ([Cannella, 2007](#); [Welch et al., 2005](#); [Ransley,
 7 1984](#); [Clayton and Clayton, 1981](#)). Newer methods employ synthetic zeolites for alkylation in the
 8 liquid phase or narrow pore synthetic zeolites in the vapor phase ([Welch et al., 2005](#)). Other
 9 methods include dehydrogenation of naphthenes, preparation from acetophenone, separation from
 10 mixed xylenes via fractionation, reaction of ethylmagnesium bromide and chlorobenzene,
 11 extraction from coal oil, and recovery from benzene-toluene-xylene (BTX) processing ([Clayton and
 12 Clayton, 1981](#)) ([Welch et al., 2005](#); [Ransley, 1984](#)).

13 Ethylbenzene can be found in a variety of products including gasoline, paints, inks,
 14 varnishes, pesticides, carpet glues, tobacco products, and automobile products. The majority of
 15 produced ethylbenzene is used in the production of styrene ([ATSDR, 2010](#)).

2.1.3. Environmental Fate and Transportation

1 While ethylbenzene is widespread in the environment and detected in air, water, and soil
2 but it is not considered to be highly persistent. In the air it is removed via photochemically
3 generated hydroxyl radicals with a half-life of approximately 1–2 days. Ethylbenzene undergoes
4 biodegradation under aerobic conditions and indirect photolysis in soil and water. Volatilization
5 from water and soil surfaces is expected to be an important environmental fate process for
6 ethylbenzene based on the vapor pressure and Henry’s law constant. On the basis of the soil
7 adsorption coefficient (K_{oc}), ethylbenzene is expected to possess moderate mobility ([ATSDR, 2010](#)).

2.1.4. Potential for Human Exposure and Populations with Potentially Greater Exposure

8 Exposure of the general population to ethylbenzene is from inhalation of contaminated air,
9 ingestion of contaminated drinking water and foods, and dermal contact from contaminated soil
10 and water. The predominate exposure to the general population is via inhalation of contaminated
11 air from automobile exhaust. Additionally, the general population can be exposed to ethylbenzene
12 from use of consumer products containing ethylbenzene [e.g., gasoline, paints, varnishes, inks,
13 solvents, pesticides, coatings, and tobacco smoke ([ATSDR, 2010](#))].

14 Populations with potentially greater exposure to ethylbenzene include people living near
15 facilities that manufacture, contain, or use ethylbenzene (e.g., petroleum refineries, hazardous
16 waste disposal sites, chemical plants) and people working or residing in high traffic areas. People
17 who obtain their drinking water from residential wells downstream from uncontrolled landfills,
18 leaking underground storage tanks, and hazardous waste sites, which are contaminated with
19 ethylbenzene, could potentially have a greater oral and dermal exposure. Populations that may
20 experience exposures greater than those of the general population may include individuals
21 employed in the petroleum refinery industry, paint, solvents, and inks industry, styrene producing
22 industries, as well as those involved in the manufacture of ethylbenzene and products that contain
23 ethylbenzene ([ATSDR, 2010](#)).

2.2. SCOPING AND PROBLEM FORMULATION SUMMARY

24 The IAP for ethylbenzene was released in September 2017 ([U.S. EPA, 2017b](#)). On September
25 27–28, 2017, the IAP was discussed at a Science Advisory Board Chemical Assessment Advisory
26 Committee (SAB CAAC) meeting (<https://sab.epa.gov/ords/sab/f?p=100:19:3574465722633>) in
27 which EPA sought input from the scientific community and interested parties.¹ This protocol
28 considers input received on the 2017 IAP. However, in 2019 the ethylbenzene assessment was

¹ Dissemination of scoping and problem formulation activities for public comment in IAPs began in 2017 as part of the IRIS Program’s implementation of systematic review. However, there were prior problem formulation efforts on ethylbenzene that informed the IAP. Earlier scoping and problem formulation materials were released in July 2014 ([U.S. EPA, 2014b](#)) and presented at a public science meeting on September 3, 2014 (<https://www.epa.gov/iris/iris-bimonthly-public-meeting-sep-2014>).

1 suspended due to changes in how EPA identified priorities for the IRIS Program ([April 2019 IRIS](#)
 2 [Program Outlook](#)). In 2021 the assessment work was restarted after it was nominated by EPA’s
 3 Office of Land and Emergency Management (OLEM) and Region 2 as a priority need (see Table 2-2).
 4 Interest was also expressed by the Office of Chemical Safety and Pollution Prevention (OCSP)
 5 because ethylbenzene is on the TSCA Work Plan list.

Table 2-2. EPA program and regional office interest in an updated ethylbenzene assessment

Program or regional office	Oral	Inhalation	Statutes/regulations	Anticipated uses/interest
OLEM	✓	✓	CERCLA	Ethylbenzene has been identified as a contaminant of concern at numerous contaminated waste sites. CERCLA authorizes EPA to conduct short- or long-term cleanups at Superfund sites and later recover cleanup costs from potentially responsible parties. Ethylbenzene toxicological information may be used to make risk determinations for response actions (e.g., short-term removals, long-term remedial response actions, RCRA Corrective Action).
Region 2	✓	✓	CERCLA	Region 2 contains 106 Superfund sites with ethylbenzene contamination. These include landfills, oil refineries, trucking facilities, former manufacturing facilities, and federal facilities.
OCSP	✓	✓	TSCA	Ethylbenzene was identified on the 2014 update of the TSCA Work Plan for Chemical Assessments.

CERCLA = Comprehensive Environmental Response, Compensation, and Liability Act; OCSP = Office of Chemical Safety and Pollution Prevention; OLEM = Office of Land and Emergency Management; RCRA = Resource Conservation and Recovery Act; TSCA = Toxic Substances Control Act.

2.3. KEY SCIENCE ISSUES

6 The 2017 IAP for ethylbenzene identified several key science issues that would require
 7 additional review and focus that were not covered in the previous assessment ([U.S. EPA, 1991b](#)).
 8 These key science issues continue to be of interest to EPA, as reflected in this protocol, in
 9 developing the ethylbenzene IRIS assessment:

- 10 • Interspecies difference in the pharmacokinetics of ethylbenzene. While there is evidence
 11 suggesting that ethylbenzene metabolism is critical to understanding its toxic effects,
 12 interspecies differences in the pharmacokinetics of ethylbenzene including metabolic
 13 biotransformation have been noted. Thus, one may need to apply toxicokinetic and

Protocol for the Ethylbenzene IRIS Assessment

- 1 dosimetry modeling (possibly including PBPK modeling) to account for interspecies
2 differences, as appropriate.
- 3 • The selection of appropriate dose metrics to inform the toxicity assessment and human
4 relevance for cancer and noncancer hazards observed in experimental systems (e.g., rat
5 renal toxicity and tumors, mouse lung toxicity and tumors).
- 6 • Mechanisms of neurotoxicity including ototoxicity.
- 7 ○ Reversibility, persistence, or potential for progression of the neurobehavioral or
8 ototoxic effects after humans are removed from ethylbenzene exposure.
- 9 ○ The relevance of ototoxicity to humans at lower exposure levels.

3. OVERALL OBJECTIVES AND SPECIFIC AIMS

3.1. OBJECTIVES

1 The overall objective of this assessment is to identify adverse health effects of ethylbenzene
2 exposure and characterize exposure-response relationships for these effects to support
3 development of toxicity values. This assessment will use systematic review methods to evaluate the
4 epidemiological and toxicological literature, including consideration of relevant mechanistic
5 evidence for ethylbenzene. The assessment methods described in this protocol utilize EPA
6 guidelines².

3.2. SPECIFIC AIMS

- 7 • Develop a systematic evidence map (SEM) to identify epidemiological (i.e., human),
8 toxicological (i.e., experimental animal), and supplemental literature pertinent to
9 characterizing the health effects of exposure to ethylbenzene. The PECO criteria used to
10 develop the SEM (referred to as “problem formulation PECO”) is intended to identify the
11 amount and type of evidence available to address a particular topic and is a useful scoping
12 tool for health effects assessments ([Thayer et al., 2022](#); [NASEM, 2021](#); [Wolffe et al., 2019](#)).
- 13 • Supplemental material content includes: mechanistic studies, including in vivo, in vitro, ex
14 vivo, or in silico models; nonmammalian model systems; pharmacokinetic and absorption,
15 distribution, metabolism, and excretion (ADME) studies; pharmacokinetic (PK) or
16 physiologically based pharmacokinetic (PBPK) models; exposure characteristics (no health
17 outcome); data pertinent to identify susceptible populations, mixture studies; non-PECO
18 routes of exposure; case studies; records with no original data; conference abstracts, and
19 errata.
- 20 • Use the results of the SEM to (1) develop PECO criteria for the assessment (referred to as
21 “assessment PECO”); (2) define the unit(s) of analysis at the level of endpoint or health
22 outcome for hazard characterization; and (3) identify priority analyses of supplemental
23 material to address the specific aims, uncertainties in hazard characterization,
24 susceptibility, and dose-response analysis.
- 25 • Conduct study evaluations (risk of bias and sensitivity) for individual epidemiological and
26 toxicological studies that meet assessment PECO criteria.
- 27 • Conduct a scientific and technical review for PBPK models considered for use in the
28 assessment. If a PBPK or PK model is selected for use, the most reliable dose metric will be

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

Protocol for the Ethylbenzene IRIS Assessment

- 1 applied based on analyses of the available dose metrics and the outcomes to which they are
2 being applied.
- 3 • Conduct data extraction (summarizing study methods and results) from epidemiological
4 and animal toxicological studies that meet the assessment PECO criteria.
 - 5 • For each evidence stream, and for each unit of analysis, use a structured framework to
6 develop and describe the certainty of evidence across studies and the supporting rationale
7 (“evidence synthesis”). Depending on the specific health endpoint or outcome, mechanistic
8 information and precursor events may be included in a unit of analysis.
 - 9 • For each health effect category, use a structured framework to develop and describe weight
10 of evidence judgments across evidence streams and the supporting rationale for those
11 judgments (“evidence integration”). The evidence integration analysis presents inferences
12 and conclusions on human relevance of findings in animals, cross-evidence stream
13 coherence, potentially susceptible populations and lifestages, biological plausibility, and
14 other critical inferences supported by mechanistic, ADME, or PK/PBPK analyses.
 - 15 • For each health effect category, summarize evidence synthesis (certainty of evidence) and
16 evidence integration (weight of evidence) conclusions in an evidence profile table.
 - 17 • As supported by the currently available evidence, derive chronic and subchronic inhalation
18 reference concentrations (RfCs) and reference doses (RfDs) and organ- or system-specific
19 RfCs and RfDs. Apply pharmacokinetic and dosimetry modeling (possibly including PBPK
20 modeling) to account for interspecies differences, as appropriate. Derive an inhalation unit
21 risk (IUR) and oral cancer slope factor (OSF) as appropriate. Characterize confidence in any
22 toxicity values that are derived.
 - 23 • Characterize uncertainties and identify key data gaps and research needs, such as
24 limitations of the evidence database, and consideration of dose relevance and
25 pharmacokinetic differences when extrapolating findings from higher dose animal studies
26 to lower levels of human exposure.

4. LITERATURE SEARCH, SCREENING, AND INVENTORY

1 The literature search and screening processes described in this section were used to
 2 develop an SEM using the problem formulation PECO (see Section 4.1) and supplemental screening
 3 criteria (see Section 4.2) to guide the inclusion of studies. The resulting inventory of studies
 4 identified in the SEM was used to develop assessment PECO criteria and identify priority analyses
 5 of supplemental material (described in Section 5). The initial literature search as well as all
 6 subsequent literature search updates use the same literature search and screening process, and
 7 therefore the literature inventory is continually updated with new studies as the assessment
 8 progresses.

4.1. POPULATIONS, EXPOSURES, COMPARATORS, AND OUTCOMES (PECO) CRITERIA FOR THE SYSTEMATIC EVIDENCE MAP

9 PECO criteria are used to focus the assessment question(s), search terms, and inclusion
 10 criteria. To meet the PECO criteria a study must meet all PECO elements. The problem formulation
 11 PECO criteria used to develop the SEM were intentionally broad to identify all the available
 12 evidence in humans and animal models.

Table 4-1. Problem formulation populations, exposures, comparators, and outcomes (PECO) criteria for the ethylbenzene assessment

PECO element	Evidence
Populations	Human: All populations and life stages (e.g., children, general population, occupational, or high exposure from an environmental source). 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 more informative study designs are available. Case reports also can be used as supportive information to establish biological plausibility for some target organs and health outcomes.
	Animal: Nonhuman, mammalian, animal species (whole organism) of any life stage (including preconception, in utero, lactation, peripubertal, and adult stages).
Exposures	Human: Exposure to ethylbenzene (CASRN 100-41-4), including occupational exposures, alone or as a mixture by any route. Measures of metabolites used to estimate exposures to ethylbenzene.
	Animal: Exposure to ethylbenzene (CASRN 100-41-4) alone by the oral or inhalation route. Studies employing chronic exposures will be considered the most informative. Studies involving exposures to mixtures will be included only if they include a group with exposure to ethylbenzene alone.

PECO element	Evidence
<u>Comparators</u>	<p>Human: Any comparison or reference group exposed; lower levels of ethylbenzene, no exposure to ethylbenzene, or to ethylbenzene for shorter periods of time.</p> <p>Animal: Quantitative exposure vs. lower or no exposure with concurrent vehicle control group.</p>
<u>Outcomes</u>	<p>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.</p> <p><i>Notes: Studies meeting PECO criteria may also contain supplemental mechanistic content that describes biological or chemical events associated with phenotypic effects. When this occurs, these studies are also tagged as having supplemental mechanistic information. This typically happens during full-text review. Full-text retrieval is performed for studies of transgenic model systems that meet E and C criteria because they may present phenotypic information in wildtype animals that meet P and O criteria but is not reported in the abstract.</i></p>

CASRN = Chemical Abstract Service registry number.

4.2. SUPPLEMENTAL CONTENT SCREENING CRITERIA

1 During the literature screening process, studies containing information that may be
2 potentially relevant to the specific aims of the assessment are tagged as supplemental material by
3 category. Some studies could emerge as being critically important to the assessment and may need
4 to be evaluated and summarized at the individual study level (e.g., certain cancer MOA or ADME
5 studies), or might be helpful to provide context (e.g., provide hazard evidence from routes or
6 durations of exposure not meeting the assessment PECO), or might not be cited at all in the
7 assessment (e.g., individual studies that contribute to a well-established scientific conclusion).
8 Because it is often difficult to assess the impact of individual studies tagged as supplemental
9 material on assessment conclusions at the screening stage, the tagging structure, described in
10 Table 4-2, allows for easy retrieval later in the assessment process.

Table 4-2. Categories of potentially relevant supplemental material

Category (tag)	Description	Typical assessment use
Pharmacokinetics data potentially informative to assessment analyses		
<p>Classical pharmacokinetic (PK) or physiologically based pharmacokinetic (PBPK) model studies</p>	<p>Classical Pharmacokinetic or Dosimetry Model Studies: Classical PK or dosimetry modeling usually divides the body into just one or two compartments, which are not specified by physiology, where movement of a chemical into, between, and out of the compartments is quantified empirically by fitting model parameters to ADME (absorption, distribution, metabolism, and excretion) data. This category is for papers that provide detailed descriptions of PK models but are not PBPK models.</p> <ul style="list-style-type: none"> The data are typically the concentration time course in blood or plasma after inhalation exposure, but other exposure routes (i.e., oral and or intravenous administration) can be described. <p>Physiologically Based Pharmacokinetic or Mechanistic Dosimetry Model Studies: PBPK models represent the body as various compartments (e.g., liver, lung, slowly perfused tissue, richly perfused tissue) to quantify the movement of chemicals or particles into and out of the body (compartments) by defined routes of exposure, metabolism, and elimination, and thereby estimate concentrations in blood or target tissues.</p> <ul style="list-style-type: none"> A defining characteristic is that key parameters are determined from a substance’s physicochemical parameters (e.g., particle size and distribution, octanol-water partition coefficient) and physiological parameters (e.g., ventilation rate, tissue volumes). 	<p>PBPK and PK model studies are included in the assessment and evaluated for possible use in conducting quantitative extrapolations. PBPK/PK models are categorized as supplemental material with the expectation that each one will be evaluated for applicability to address assessment extrapolation needs and technical conduct. Specialized expertise is required for their evaluation.</p> <p>Standard operating procedures for PBPK/PK model evaluation and the identification, organization, and evaluation of ADME studies are outlined in <i>An Umbrella Quality Assurance Project Plan (QAPP) for PBPK models</i> (U.S. EPA, 2018b).</p>
<p>Pharmacokinetic (ADME)</p>	<p>Pharmacokinetic (ADME) studies are primarily controlled experiments, where defined exposures usually occur by intravenous, oral, inhalation, or dermal routes, and the concentration of particles, a chemical, or its metabolites in blood or serum, other body tissues, or excreta are then measured.</p> <ul style="list-style-type: none"> These data are used to estimate the amount absorbed (A), distributed (D), metabolized (M), and/or excreted (E). ADME data can also be collected from human subjects who have had environmental or workplace exposures that are not quantified or fully defined. ADME data, especially metabolism and tissue partition coefficient information, can be generated using in vitro model systems. Although in vitro data may not be as definitive as in vivo data, these studies should also be tracked as 	<p>ADME studies are inventoried and prioritized for possible inclusion in an ADME synthesis section on the chemical’s PK properties and for conducting quantitative adjustments or extrapolations (e.g., animal-to-human). Specialized expertise in PK is necessary for inventory and prioritization.</p> <p>Standard operating procedures for PBPK/PK model evaluation and the identification, organization, and</p>

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

Protocol for the Ethylbenzene IRIS Assessment

Category (tag)	Description	Typical assessment use
Non-PECO animal model	<p>Studies reporting outcomes in animal models that meet the outcome criteria but do not meet the population criteria in the PECO.</p> <p>Depending on the endpoints measured in these studies, they can also provide mechanistic information (in these cases studies should also be tagged “mechanistic endpoints”).</p> <p>*This categorization generally does not apply to studies that use species with limited human health relevance (e.g., ecotoxicity-focused studies are typically excluded).</p>	<p>Studies of non-PECO animals, exposures, or durations can be summarized to inform evaluations of consistency (e.g., across species or routes or durations), coherence, or adversity; subsets of the most informative studies may be included in the unit of analysis. These studies may also be used to inform evidence integration judgments of biological plausibility and/or MOA analyses and thus may be summarized as part of the mechanistic evidence synthesis.</p>
Non-PECO route of exposure	<p>Epidemiological or animal studies that use a non-PECO route of exposure, e.g., injection studies or dermal studies if the dermal route is not part of the exposure criteria.</p> <p>*This categorization generally does not apply to epidemiological studies where the exposure route is unclear; such studies are considered to meet PECO criteria if the relevant route(s) of exposure are plausible, with exposure being more thoroughly evaluated at later steps.</p>	
Susceptible population	<p>Studies that help to identify potentially susceptible subgroups, including studies on the influence of intrinsic factors such as sex, lifestage, or genotype to toxicity, as well as some other factors (e.g., health status). These are often co-tagged with other supplemental material categories, such as mechanistic or ADME. Studies meeting PECO criteria that also address susceptibility should be co-tagged as supplemental.</p> <p><i>*Susceptibility based on most extrinsic factors, such as increased risk for exposure due to residential proximity to exposure sources, is not considered an indicator of susceptible populations for the purposes of IRIS assessments.</i></p>	<p>Provides information on factors that might predispose sensitive populations or lifestages to a higher risk of adverse health effects following exposure to the chemical. This information is summarized during evidence integration for each health effect and is considered during dose-response, where it can directly impact modeling decisions.</p>
<p>Background information potentially useful to problem formulation and protocol development (These studies fall outside the scope of IRIS assessment analyses)</p>		
Human exposure and biomonitoring (no health outcome)	<p>Information regarding exposure monitoring methods and reporting that are unrelated to health outcomes, but which provide information on the following: methods for measuring human exposure, biomonitoring (e.g., detection of chemical in blood, urine, hair), defining exposure sources, or modeled estimates of exposure (e.g., in occupational settings). Studies that compare exposure levels to a reference value, risk threshold or assessment points of departure are also included in this</p>	<p>This information may be useful for developing exposure criteria for study evaluation or refining problem formulation decisions.</p>

This document is a draft for review purposes only and does not constitute Agency policy.

Protocol for the Ethylbenzene IRIS Assessment

Category (tag)	Description	Typical assessment use
	category. Studies related to environmental fate and transport are typically tagged as background materials unless otherwise described in the assessment-specific protocol. *Assessment teams may want to subtag studies that describe or predict exposure levels versus those that present exposure assessment methods.	Notably, providing an assessment of typical human exposures (e.g., sources, levels) falls outside the scope of an IRIS assessment.
Mixture study	Mixture studies use methods that do not allow investigation of the health effects of exposure to the chemical of interest by itself (e.g., animal studies that lack exposure to chemical of interest alone or epidemiology studies that do not evaluate associations of the chemical of interest with relevant health outcome(s)). *Methods used to assess investigation of the exposure by itself may not be clear from the abstract, in particular for epidemiology studies. When unclear, the study is advanced to full-text review to determine eligibility.	Mixture studies are tracked to help inform cumulative risk analyses, which may provide useful context for risk assessment but fall outside the scope of an IRIS assessment.
Case reports or case series	Human studies that present an investigation of a single exposed individual or group of ≤3 subjects who describe health outcomes after exposure but lack a comparison group (i.e., do not meet the “C” in the PECO) and typically do not include reliable exposure estimates.	Tracking case studies can facilitate awareness of potential human health issues missed by other types of studies during problem formulation.
Reference materials		
Records with no original data	Records that do not contain original data, such as other agency assessments, informative scientific literature reviews, editorials, or commentaries.	Studies that are tracked for potential use in identifying missing studies, background information, or current scientific opinions (e.g., hypothesized MOAs).
Posters or conference abstracts	Records that do not contain sufficient documentation to support study evaluation and data extraction.	

4.3. LITERATURE SEARCH STRATEGIES

4.3.1. Database Search Term Development

1 Literature search strategies are developed using key terms and words related to the PECO
2 criteria. Development of the search strategy for each topic area is conducted by identifying relevant
3 search terms through the following approaches: (1) reviewing PubMed’s Medical Subject Headings
4 (MeSH) for relevant and appropriate terms, (2) extracting key terminology from relevant reviews
5 and a set of previously identified primary data studies known to be relevant to the topic (“test set”),
6 and (3) reviewing search strategies presented in other reviews. Relevant subject headings and text-
7 words are crafted into a search strategy designed to maximize the sensitivity and specificity of the
8 search results. The search strategy is run, and the results assessed to ensure that all previously
9 identified relevant primary studies are retrieved in the search. The database search terms focused
10 only on the chemical name (and synonyms or trade names) with no additional limits. Because each
11 database has its own search architecture, the resulting search strategy is tailored to account for
12 each database’s unique search functionality.

4.3.2. Database Searches

13 Searches are not restricted by publication date and no language restrictions are applied.
14 The detailed search strategies are presented in Appendix A. Literature searches are conducted
15 using EPA’s Health and Environmental Research Online (HERO) database.³

16 The following databases are searched as described in the IRIS Handbook ([U.S. EPA, 2022](#)):

- 17 • [PubMed](#) (National Library of Medicine)
- 18 • [Web of Science](#) (Thomson Reuters)
- 19 • [Toxline \(National Library of Medicine\)](#) – Searched through December 2019, after which
20 Toxline content was moved to PubMed (National Library of Medicine) products.
- 21 • Toxic Substances Control Act Test Submissions (TSCATS) database

22 The literature searches are updated throughout the assessment’s development and review
23 process to identify newly published literature. During this period, studies are screened according to
24 both the problem formulation and assessment PECO criteria. Thus, the literature inventory is
25 updated during the process of developing the draft assessment. The last full literature search
26 update is conducted several months prior to the planned release of the draft document for public
27 comment. Studies identified after peer review begins are only considered for inclusion if they are

³Health and Environmental Research Online: <https://hero.epa.gov/hero/>.

1 directly relevant to the assessment PECO criteria and are expected to fundamentally alter the draft
2 assessment conclusions.

4.3.3. Searching Other Sources

3 The literature search strategy described above was designed to be broad, but like any
4 search strategy, studies can be missed [e.g., cases where the specific chemical is not mentioned in
5 title, abstract, or keyword content; ability to capture “gray” literature (studies not reported in the
6 peer-reviewed literature) that is not indexed in the databases listed above]. Thus, in addition to the
7 database searches, the sources below are used to identify studies that could have been missed
8 based on the database search. Searching of these resources occurs during preparation of the initial
9 literature inventory when assembling the SEM. After preparation of the initial literature inventory,
10 references can be identified during public comment periods, by technical consultants, and during
11 peer review. Records that appear to meet the problem formulation PECO criteria are uploaded into
12 a screening software, annotated with respect to source of the record, and screened using the
13 methods described in Section 4.4. Appendix B describes the specific methods and results for
14 searching the sources below. Searching of these sources is summarized to include the source type
15 or name, the search string (when applicable), number of results present within the resource, and
16 the URL (uniform resource locator, when available and applicable). The list of other sources
17 consulted includes:

- 18 • Manual review (at the title level) of reference list in studies screened as meeting problem
19 formulation PECO after full-text review.
- 20 • Manual review (at the title level) of the reference list from other publicly available final or
21 draft assessments from other non-EPA Agencies (e.g., ATSDR [Agency for Toxic Substances
22 and Disease Registry] Toxicological Profile) or published journal review specifically focused
23 on human health. Reviews can be identified from the database search or from the resources
24 listed in Appendix B.
- 25 • European Chemicals Agency (ECHA) registration dossiers to identify data submitted by
26 registrants [http://echa.europa.eu/information-on-chemicals/information-from-existing-](http://echa.europa.eu/information-on-chemicals/information-from-existing-substances-regulation)
27 [substances-regulation](http://echa.europa.eu/information-on-chemicals/information-from-existing-substances-regulation).
- 28 • EPA ChemView database ([U.S. EPA, 2019a](#)) to identify unpublished studies, information
29 submitted to EPA under Toxic Substances Control Act (TSCA) Section 4 (chemical testing
30 results), Section 8(d) (health and safety studies), Section 8(e) (substantial risk of injury to
31 health or the environment notices), and FYI (For Your Information, voluntary documents).
32 Other databases accessible via ChemView include EPA’s High Production Volume (HPV)
33 Challenge database and the Toxic Release Inventory database.
- 34 • The National Toxicology Program (NTP) database of study results and research projects
35 (<https://ntp.niehs.nih.gov/results/index.html>).

- 1 • The Organization for Economic Cooperation and Development (OECD) Screening
2 Information DataSet (SIDS) High Production Volume Chemicals
3 <https://www.echemportal.org/echemportal/substance-search>.
- 4 • The EPA CompTox (Computational Toxicology Program) Chemical Dashboard ([U.S. EPA,](#)
5 [2019b](#)) to retrieve a summary of any ToxCast or Tox21 high throughput screening
6 information. This data can be used to generate mechanistic insight, predict outcome using
7 appropriate models, and potentially inform dose-response modeling. Their importance for
8 outcome prediction and dose-response modeling depends on the context, size and quality of
9 retrieved results and the lack of availability of other data typically used for these purposes.
- 10 • Review of the list of references in the [ECOTOX database](#) for the chemical(s) of interest.
- 11 • References identified during public comment periods, by technical consultants, and during
12 peer review.

4.3.4. Non-Peer-Reviewed Data

13 IRIS assessments rely mainly on publicly accessible, peer-reviewed studies. However, it is
14 possible that unpublished data directly relevant to the PECO may be identified during assessment
15 development. In these instances, the EPA will try to get permission to make the data publicly
16 available (e.g., in HERO); data that cannot be made publicly available are not used in IRIS
17 assessments. In addition, on rare occasions where unpublished data would be used to support key
18 assessment decisions (e.g., deriving a toxicity value), EPA may obtain external peer review if the
19 owners of the data are willing to have the study details and results made publicly accessible, or if an
20 unpublished report is publicly accessible (or submitted to EPA in a nonconfidential manner) ([U.S.](#)
21 [EPA, 2015](#)). This independent, contractor driven, peer review would include an evaluation of the
22 study similar to that for peer review of a journal publication. The contractor would identify and
23 typically select three scientists knowledgeable in scientific disciplines relevant to the topic as
24 potential peer reviewers. Persons invited to serve as peer reviewers would be screened for conflict
25 of interest. In most instances, the peer review would be conducted by letter review. The study and
26 its related information, if used in the IRIS assessment, would become publicly available. In the
27 assessment, EPA would acknowledge that the document underwent external peer review managed
28 by the EPA, and the names of the peer reviewers would be identified. In certain cases, IRIS will
29 assess the utility of a data analysis of accessible raw data (with descriptive methods) that has
30 undergone rigorous quality assurance/quality control review (e.g., ToxCast/Tox21 data, results of
31 NTP studies not yet published) but that have not yet undergone external peer review.

32 Unpublished data from personal author communication can supplement a peer-reviewed
33 study as long as the information is made publicly available. If such ancillary information is acquired,
34 it is documented in the Health Assessment Workspace Collaborative (HAWC) or HERO project page
35 (depending on the nature of the information received).

4.4. LITERATURE SCREENING

1 Records identified from the literature searches are housed in HERO. After deduplication in
2 HERO, records are imported into SWIFT Review software ([Howard et al., 2016](#)) to identify those
3 references most likely to be applicable to a human health assessment. Briefly, SWIFT Review has
4 preset literature search strategies (“filters”) developed and applied by information specialists to
5 identify studies more likely to be useful for identifying human health content from those that likely
6 are not (e.g., analytical methods). The filters function like a typical search strategy in which studies
7 are tagged as belonging to a certain filter if the terms in the filter literature search strategy appear
8 in title, abstract, keyword or medical subject headings (MeSH) fields content. The applied SWIFT
9 Review filters focused on lines of evidence: human, animal models for human health, and in vitro
10 studies. The details of the search strategies that underlie the filters are available online
11 (<https://www.sciome.com/swift-review/searchstrategies/>). Studies not retrieved using these
12 filters are not considered further. Studies that included one or more of the search terms in the title,
13 abstract, keyword, or MeSH fields are exported as a RIS (Research Information System) file for title
14 and abstract (TIAB) and full-text screening in DistillerSR (Evidence Partners;
15 <https://distillercer.com/products/distillersr-systematic-review-software/>), as described below.
16 The impact of application of the SWIFT evidence stream filters on the number of studies for TIAB
17 screening is presented in Figure 4-1.

4.4.1. Title and Abstract Screening

18 The studies prioritized by SWIFT Review are imported into DistillerSR software for TIAB
19 screening by two independent reviewers. Reviewers complete a structured form asking whether a
20 study meets PECO criteria or contains potentially relevant supplemental material. Studies
21 considered relevant or “unclear” based on meeting all PECO criteria at the TIAB level are
22 considered for inclusion and advanced to full-text screening.

23 Any screening conflicts are resolved by discussion between the primary screeners with
24 consultation by a third reviewer, if needed. For citations with no abstract, articles are initially
25 screened based on the following: title relevance (title should indicate clear relevance), and page
26 length (articles two pages in length or less are assumed to be conference reports, editorials, or
27 letters). Eligibility status of non-English studies is assessed using the same approach with online
28 translation tools or engagement with a native speaker.

4.4.2. Full-Text Screening

29 Full-text references are sought through EPA’s HERO database for studies screened as
30 meeting problem formulation PECO criteria, potentially relevant supplemental material, or
31 “unclear” based on TIAB screening. Full-text screening occurs in Distiller SR. Full-text copies of
32 these citations are retrieved, stored in the HERO database, and independently assessed by two
33 screeners using a structured form in DistillerSR to confirm eligibility. Screening conflicts are

1 resolved by discussion among the primary screeners with consultation by a third reviewer or
2 technical advisor (as needed to resolve any remaining disagreements). Rationales for excluding
3 citations are documented, e.g., study did not meet problem formulation PECO, full-text not
4 available. Approaches for language translation include online translation tools or engagement of a
5 native speaker. Fee-based translation services for non-English studies are typically reserved for
6 studies that are anticipated as being useful for toxicity value derivation. Conflicts between
7 screeners in applying the supplemental material tags are resolved similarly, erring on the side of
8 over tagging. Note that more granular sub-tagging of supplemental material occurs during
9 preparation of the literature inventory as described in Section 4.5.2.

4.4.3. Multiple Citations with the Same Data

10 When there are multiple citations using the same or overlapping data, all citations are
11 included, with one selected for use as the primary citation; the others are considered as secondary
12 publications with annotation in HAWC and HERO indicating their relationship to the primary
13 citation during data extraction. For epidemiology studies, the primary citation is generally the one
14 with the longest follow-up, the largest number of cases, or the most recent publication date. For
15 animal studies, the primary citation is typically the one with the longest duration of exposure, the
16 largest sample size, or with the outcome(s) most informative to the problem formulation PECO. For
17 both epidemiology and animal studies, the assessments include relevant data from all citations of
18 the study, although if the same data are reported in more than one citation, the data
19 are only extracted once (see Section 7). For corrections, retractions, and other companion
20 documents to the included citations, a similar approach to annotation is taken and the most
21 recently published data are incorporated into the assessments.

4.4.4. Literature Flow Diagrams

22 The results of the screening process are posted on the project page for the assessment in
23 the HERO database (https://heronet.epa.gov/heronet/index.cfm/project/page/project_id/59).
24 Results for SEM screening against the problem formulation PECO are also summarized in a
25 literature flow diagram (see Figure 4-1) and interactive HAWC literature tag trees (see Figure 4-4).

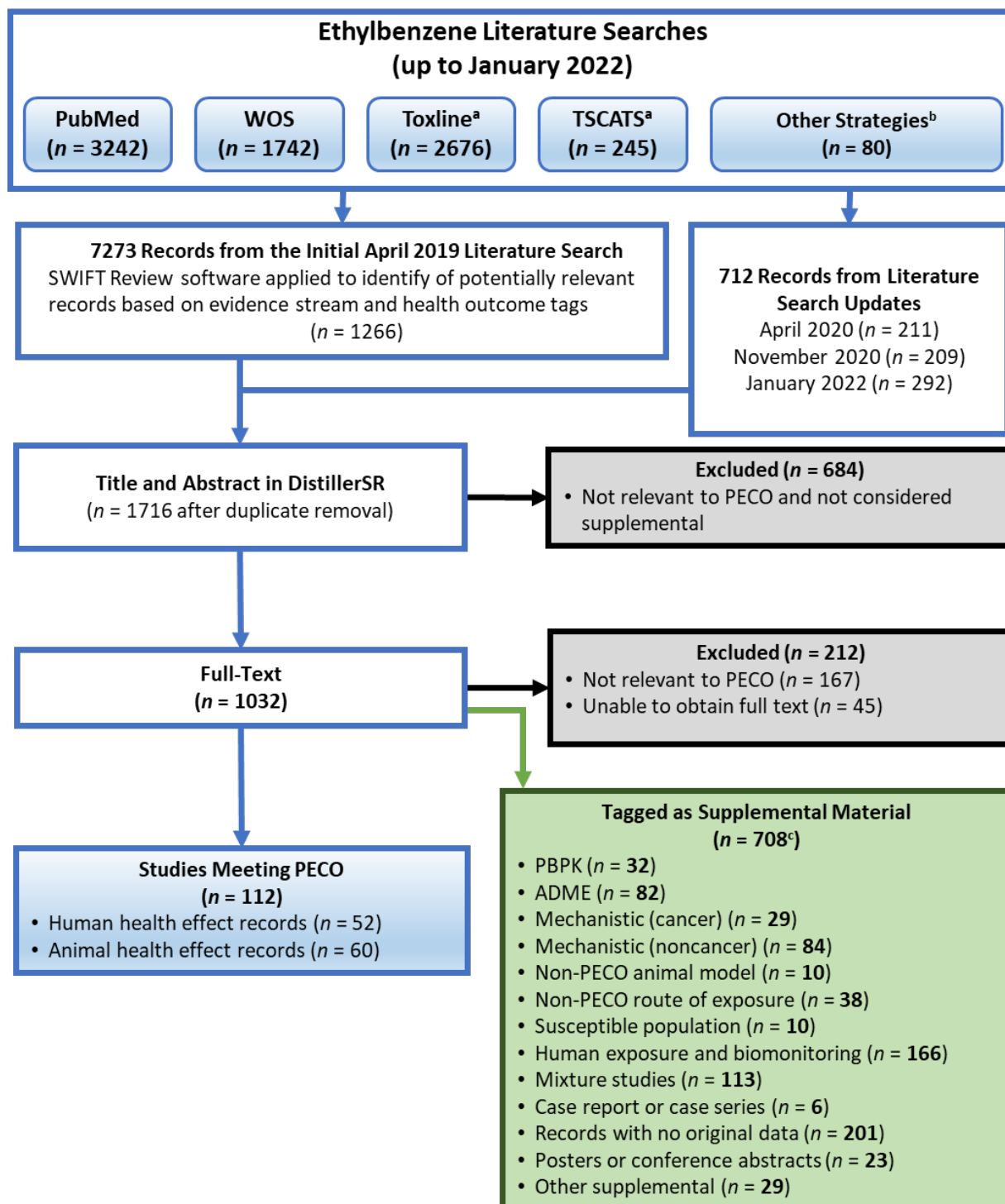


Figure 4-1. Literature flow diagram for ethylbenzene.

^aToxline and TSCATS only included in Apr 2019 search.

^bOther strategies include the following sources of gray literature: ToxVal, CEBS, ECHA, ChemView, and OECD SIDS); Jan 2022 = 3; Nov 2020 = 77.

^cIndicates the total number of unique citations that were identified; because some citations are given multiple tags, the sum of the individual supplemental material tags is greater than the total number of unique citations.

4.5. LITERATURE INVENTORY

1 During full-text-level screening, citations that meet problem formulation PECO criteria are
2 categorized by evidence type (human or animal) or category of supplemental information (e.g.,
3 mechanistic, PBPK, ADME). Next, study design details for citations that meet problem formulation
4 PECO criteria are summarized as described in Section 4.5.1. A more granular tagging of
5 supplemental material may be conducted as described in Section 4.5.2. The results of this
6 categorization and tagging are referred to as the literature inventory and is the key analysis output
7 of the SEM.

4.5.1. Studies That Meet Problem Formulation PECO Criteria

8 Human and animal studies that met problem formulation PECO criteria after full-text
9 review are briefly summarized using DistillerSR Hierarchical Data Extraction (HDE) forms to create
10 literature inventories which were used to display the extent and nature of the available evidence.
11 Data extraction details for the literature inventory are presented in Section 7. These study
12 summaries are exported from DistillerSR in Excel format and imported into Tableau software
13 (<https://www.tableau.com/>) to create interactive literature inventory visualizations. The literature
14 inventories are used to inform the assessment PECO criteria and evaluation plan. More detail on the
15 process of summarizing studies is presented in Section 7 (Data Extraction of Study Methods and
16 Results).

4.5.2. Organizational Approach for Supplemental Material

17 The results of the supplemental material tagging conducted in DistillerSR are imported into
18 the literature review module in HAWC, where more granular sub-tagging within a type of
19 supplemental material content category may be conducted if determined to be useful to support
20 assessment conclusions. A single study can have multiple tags. The degree of sub-tagging depends
21 on the extent of content for a given type of supplemental material and needs of the assessment with
22 respect to developing human health hazard conclusions and derivation of toxicity values. Typically,
23 more granular tagging is most useful for supplemental content classified as mechanistic, ADME,
24 PK/PBPK models, routes of administration not meeting the PECO, and nonmammalian model
25 studies. Tagging judgments in HAWC are made by one assessment member and confirmed during
26 preparation of draft assessment by another member of the assessment team. The overall approach
27 for supplemental material content was previously described in Section 4.2.

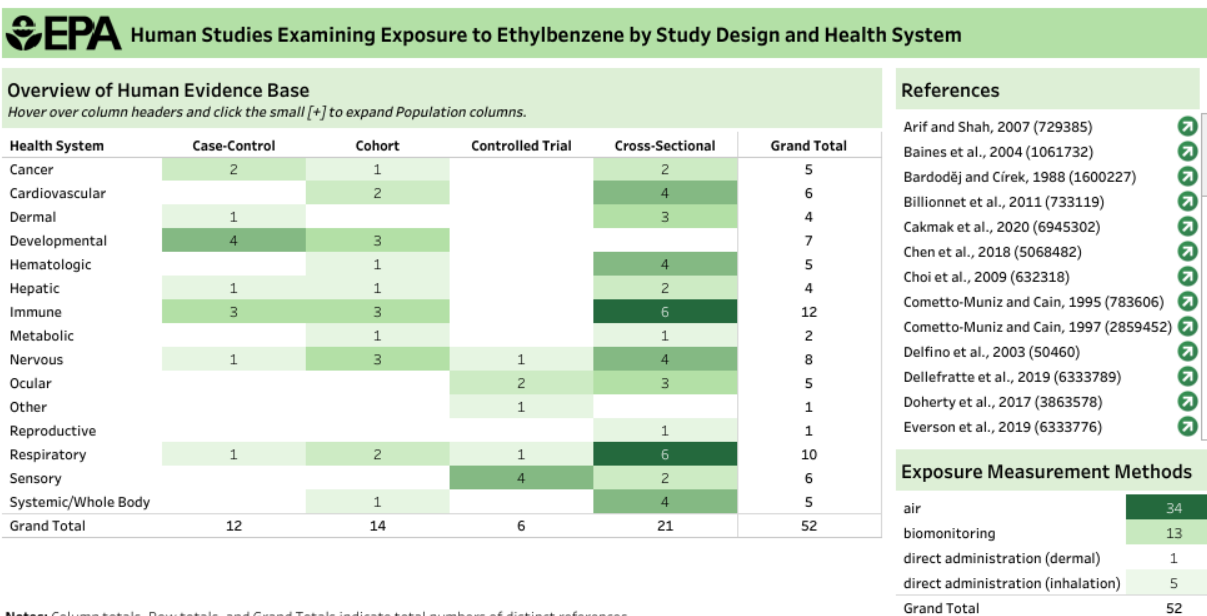
4.6. SUMMARY-LEVEL LITERATURE INVENTORIES

28 During TIAB or full-text-level screening, citations tagged based on problem formulation
29 PECO eligibility were further categorized based on features such as evidence type (i.e., human,
30 animal), health outcome(s), and/or endpoint measure(s) included in the citation. Literature
31 inventories for PECO-relevant citations were created to develop summary-level, sortable lists that

Protocol for the Ethylbenzene IRIS Assessment

1 include some basic study design information (e.g., study population, exposure information such as
2 doses administered or biomarkers analyzed, age/life stage⁴ of exposure, endpoints examined).
3 These literature inventories facilitate subsequent review of individual studies or sets of studies by
4 topic-specific experts. The summary results are presented in Figures 4-2 and 4-3 for human and
5 animal studies, respectively. An interactive version of these figures, including additional study
6 design details and a high-level summary of the results is available [here](#).

⁴Age/life stage of chemical exposure are considered according to EPA's [Guidance on Selecting Age Groups for Monitoring and Assessing Childhood Exposures to Environmental Contaminants](#) and EPA's [A Framework for Assessing Health Risk of Environmental Exposures to Children](#).



Study Details

■ statistically significant association
 ■ potential beneficial association
 ■ no effect(s) reported

Health System	Population	Sex	Exposure Measurement	Endpoints	Reference		
Cancer	Children	both	air	germ cell tumors (GCTs), yolk s..	Hall et al., 2019	■	
				retinoblastoma	Heck et al., 2015	■	
	General population (adults)	both	air	lifetime cancer risk	Tunsaringkarn et al., 2015	■	
				Lung cancer	Khorrani et al., 2021	■	
Cardiovascular	General population (adults)	both	biomonitoring	cancer mortality	Wenzhen et al., 2022	■	
				Infants	both	air	germ cell tumors (GCTs), yolk s..
	Occupational	both	air	cardiovascular disease	Xu et al., 2009	■	
				heart disease mortality	Wenzhen et al., 2022	■	
	Pregnant women	female	air	systolic blood pressure (SBP), ..	Everson et al., 2019	■	
				heart symptoms associated wi..	Sakellaris et al., 2020	■	
	Dermal	General population (adults)	both	air	increased heart rate	Moradi et al., 2019	■
					any cardiovascular event - con..	Männistö et al., 2015	■
				itching, dry, flushed, or erupte..	Saijo et al., 2004	■	
				itchina/rash. dry and cracked s..	Tunsaringkarn et al., 2015	■	

Figure 4-2. Inventory heatmap of PECO-relevant ethylbenzene human studies by study design and health system. An interactive version, which includes a list of citations with additional study details and summary of the results, is available [here](#).

EPA Toxicological Studies Examining Exposure to Ethylbenzene by Study Design and Health System



Notes: Column totals, Row totals, and Grand Totals indicate total numbers of distinct references.

Study Details										effect(s) observed	no effect(s) reported			
Health System	Study Design	Route	Dosing Duration	Species	Strain	Sex	All Dose Levels	Dose Units	Reference					
Cancer	Chronic	inhalation	103 wk (6 h/d x 5 d/wk)	Mouse	B6C3F1	female	0, 75, 250, 750	ppm	Chan et al., 1998	■				
									NTP, 1999	■				
						male	0, 75, 250, 750	ppm	Chan et al., 1998	■				
						NTP, 1999	■							
						104 wk (6 h/d x 5 d/wk)	Rat	Fischer 344/N	female	0, 75, 250, 750	ppm	Chan et al., 1998	■	
									male	0, 75, 250, 750	ppm	Chan et al., 1998	■	
									NTP, 1999	■				
			104 wk (4 d/wk)	Rat	Sprague-D.	both	0, 500, 800	mg/kg-d	Maltoni et al., 1997	■				

Figure 4-3. Inventory heatmap of PECO-relevant ethylbenzene animal studies by study design and health system. An interactive version, which includes a list of citations with additional study details and summary of the results, is available [here](#).

1 HAWC literature trees are created for citations that are tagged as “potentially relevant
 2 supplemental material” during screening, including mechanistic studies (e.g., in vitro or in silico
 3 models), ADME studies, and studies on endpoints or routes of exposure that do not meet the
 4 specific PECO criteria but that may still be relevant to the research question(s). Here, the objective
 5 is to create an inventory of citations that can be tracked and further summarized as needed—for
 6 example, by model system, key characteristic [e.g., of carcinogens; [Smith et al. \(2016\)](#)], mechanistic
 7 endpoint, or key event—to support analyses of critical mechanistic questions that arise at various
 8 stages of the systematic review (see Section 9.2 for a description of the process for determining the
 9 specific questions and pertinent mechanistic studies to be analyzed). ADME data and related
 10 information can be critical to the next steps of prioritizing or evaluating individual PECO-specific

- 1 studies and are reviewed by subject-matter experts early in the assessment process. A literature
- 2 tree of the supplemental material identified from the literature searches (as of 1/2022) is
- 3 presented in Figure 4-4.

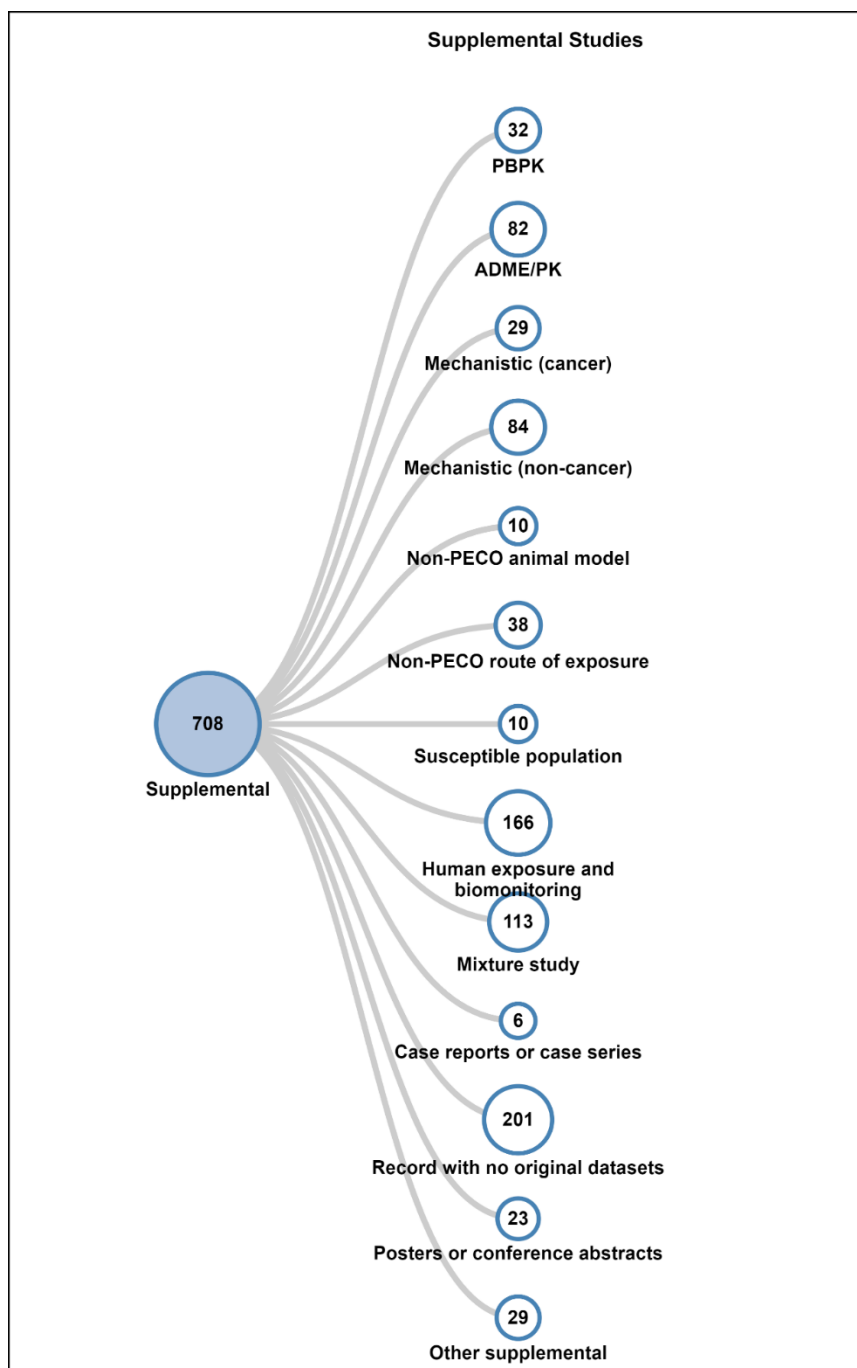


Figure 4-4. Literature tag tree of the supplemental studies identified from the ethylbenzene literature searches. An interactive version, which includes a list of citations with additional study details and summary of the results, is available [here](#).

- 1 A single active high throughput screening assay was reported for ethylbenzene on the
- 2 CompTox Chemicals Dashboard ([U.S. EPA, 2019b](https://comptox.epa.gov/dashboard/)). The TOXCAST summary plot is shown in Figure
- 3 4-5 and an interactive version can be found online
- 4 (<https://comptox.epa.gov/dashboard/chemical/invitrodb/DTXSID3020596>).

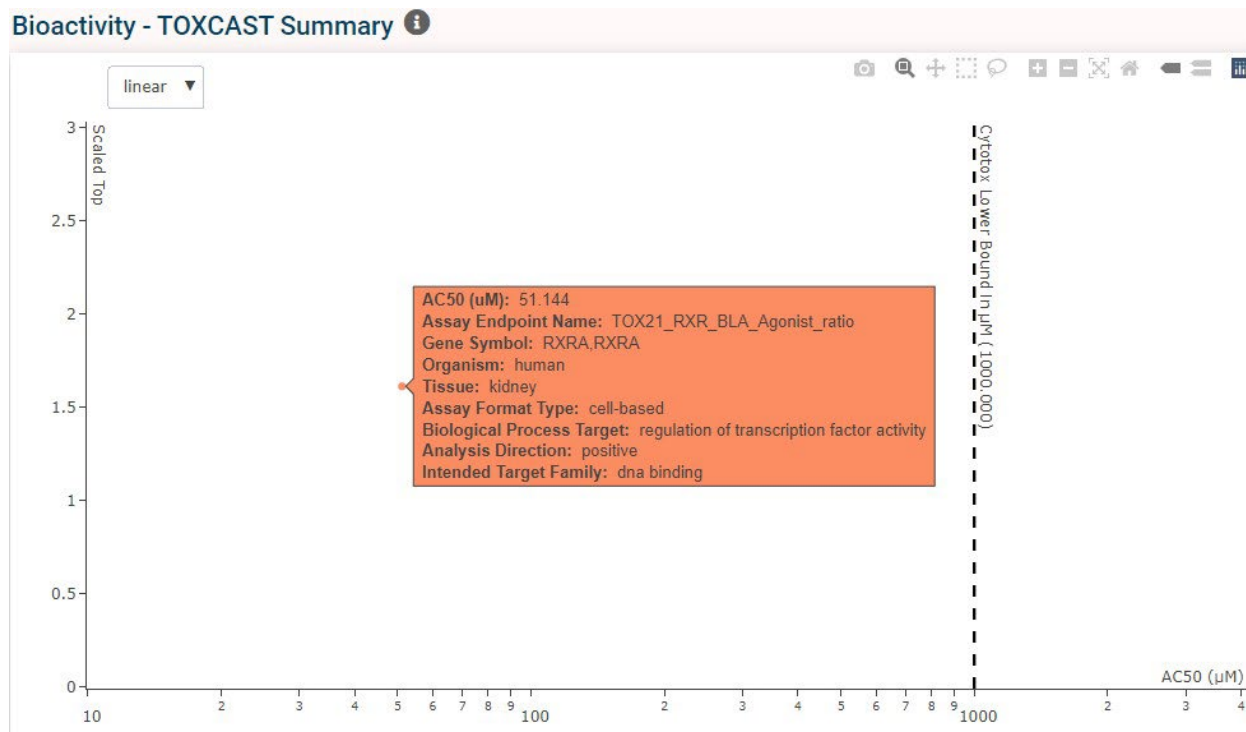


Figure 4-5. High throughput screening bioactivity data from the CompTox Chemicals Dashboard. An interactive version, which includes a list of citations with additional study details and summary of the results, is available [here](#).

5. REFINE PROBLEM FORMULATION AND SPECIFY ASSESSMENT APPROACH

5.1. ASSESSMENT PECO CRITERIA

1 The primary purpose of this step is to provide further specification to the assessment
 2 methods based on characterization of the extent and nature of the evidence identified from the
 3 literature inventory. This includes refinements to PECO criteria and defining the unit(s) of analysis
 4 for health endpoints/outcomes during evidence synthesis, and presenting analysis approaches for
 5 mechanistic, ADME or other types of supplemental material content. A unit of analysis is an
 6 outcome or group of related outcomes within a health effect category that are considered together
 7 during evidence synthesis (see Section 8). In some assessments, the units of analysis may include
 8 predefined categories of mechanistic evidence (e.g., biomarkers or precursors relating to other
 9 outcomes within the unit of analysis, evidence that provides support for grouping together
 10 biologically linked endpoints into a unit of analysis).

Table 5-1. Assessment PECO criteria for the ethylbenzene assessment

PECO element	Evidence
Populations	Human: Any population and lifestage (occupational or general population, including children and other sensitive populations).
	Animal: Nonhuman mammalian animal species (whole organism) of any lifestage (including preconception, in utero, lactation, peripubertal, and adult stages). Studies of transgenic animals are tracked as mechanistic studies under “potentially relevant supplemental material.”
Exposures	Human: Exposure to ethylbenzene (CASRN 100-41-4), including occupational exposures, alone or as a mixture by any route. Measures of metabolites used to estimate exposures to ethylbenzene.
	Animal: Exposure to ethylbenzene (CASRN 100-41-4) alone by the oral or inhalation route. Studies employing chronic exposures will be considered the most informative. Studies involving exposures to mixtures will be included only if they include a group with exposure to ethylbenzene alone.
Comparators	Human: Any comparison or reference group exposed; lower levels of ethylbenzene, no exposure to ethylbenzene, or to ethylbenzene for shorter periods of time.
	Animal: Quantitative exposure vs. lower or no exposure with concurrent vehicle control group.
Outcomes	Health Outcomes: <u>Cancer, cardiovascular, developmental, general toxicity (systemic / whole body), hematologic, hepatic, immune/lymphatic, metabolic, nervous system/auditory, renal/urinary, reproductive, respiratory system, thyroid (endocrine).</u> 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.

Underlined text show modifications in the assessment PECO criteria compared with the problem formulation PECO criteria.

CASRN = Chemical Abstract Service registry number.

5.1.1. Other Exclusions Based on Full-Text Content

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

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

13
14 Epidemiology studies

- 15 • Sample size
- 16 • Exposure characterization and/or measurement method
- 17 • Outcome ascertainment method
- 18 • Study design

19 Animal studies

- 20 • Species
- 21 • Test article name
- 22 • Levels and duration of exposure
- 23 • Route of exposure
- 24 • Quantitative or qualitative (e.g., photomicrographs; author-reported lack of an effect on the
25 outcome) results for at least one endpoint of interest

26 In vitro studies prioritized for focused analysis

- 27 • Cell/tissue type(s) or test system
- 28 • Test article name

- 1 • Concentration and duration of treatment
- 2 • Quantitative or qualitative results for at least one endpoint of interest

5.2. UNITS OF ANALYSES FOR DEVELOPING EVIDENCE SYNTHESIS AND INTEGRATION JUDGMENTS FOR HEALTH EFFECT CATEGORIES

3 The planned units of analysis based on outcomes identified in the assessment PECO are
 4 summarized in Table 5-2. General considerations for defining the units of analysis are presented in
 5 the IRIS Handbook ([U.S. EPA, 2022](#)). Each unit of analysis is initially synthesized and judged
 6 separately within an evidence stream (see Section 8.1). Depending on the specific health endpoint
 7 or outcome, PK data, mechanistic information, and other supporting evidence (e.g., from studies of
 8 non-PECO routes of exposure) may be included in a unit of analysis.

9 Evidence integration judgments focus on the stronger within evidence stream synthesis
 10 conclusions when multiple units of analysis are synthesized. The evidence synthesis judgments are
 11 used alongside other key considerations (i.e., human relevance of findings in animal evidence,
 12 coherence across evidence streams, information on susceptible populations or lifestages, and other
 13 critical inferences that draw on mechanistic evidence) to draw an overall evidence integration
 14 judgment for each health effect category or more granular health outcome grouping (see
 15 Section 8.2).

Table 5-2. Human and animal endpoint grouping categories.

Relevant human health effect category ^a	Units of analysis for evidence synthesis that inform evidence integration for the ethylbenzene assessment (each bullet represents a unit of analysis)	
	Human evidence	Animal evidence
Cancer	<ul style="list-style-type: none"> • Lifetime cancer risk • Tumors and precancerous lesions 	<ul style="list-style-type: none"> • Tumors and precancerous lesions
Cardiovascular	<ul style="list-style-type: none"> • Heart disease • Blood pressure, vascular dilation, and pulse 	<ul style="list-style-type: none"> • Heart weight • Histopathology
Developmental	<ul style="list-style-type: none"> • Birth defects • Birth weight • Preeclampsia • Age at use of academic support services 	<ul style="list-style-type: none"> • Offspring mortality/ survival • Body weight, body weight change • Developmental milestones (e.g., eye opening, incisor eruption, pinna detachment) • Skeletal and visceral malformations/ variations
Hematologic	<ul style="list-style-type: none"> • Red blood cells, hematocrit or hemoglobin, cell volume 	<ul style="list-style-type: none"> • Red blood cells, hematocrit or hemoglobin, cell volume • Blood platelets, reticulocytes

Protocol for the Ethylbenzene IRIS Assessment

Relevant human health effect category ^a	Units of analysis for evidence synthesis that inform evidence integration for the ethylbenzene assessment (each bullet represents a unit of analysis)	
	Human evidence	Animal evidence
	<ul style="list-style-type: none"> Blood platelets, reticulocytes 	<ul style="list-style-type: none"> Blood biochemical measures (e.g., sodium, calcium)
Hepatic	<ul style="list-style-type: none"> Serum or liver enzymes (e.g., ALT, AST) 	<ul style="list-style-type: none"> Liver weight and histopathology Serum or liver enzymes (e.g., ALT, AST) Liver tissue biochemical markers/biochemistry
Immune/lymphatic	<ul style="list-style-type: none"> Asthma incidence/ severity Respiratory infection Immune cell counts Inflammation (c-reactive protein) 	<ul style="list-style-type: none"> Immune organ weight/histopathology Immune cell counts
Metabolic	<ul style="list-style-type: none"> Serum glucose, insulin; A1C 	<ul style="list-style-type: none"> Serum glucose
Nervous system/auditory	<ul style="list-style-type: none"> Neurodevelopmental disorders (e.g., autism, learning disabilities) Neurobehavioral function (e.g., reaction time, emotional changes) Headache, fatigue, sensory irritation Hearing loss, tinnitus 	<ul style="list-style-type: none"> Brain weight/histopathology Functional observational battery, including motor activity and reflex responses Learning and memory Seizures/tremors Neurotransmitters Histopathology (hair cell loss) Auditory function (e.g., MER, auditory threshold)
Renal/urinary	<ul style="list-style-type: none"> No studies 	<ul style="list-style-type: none"> Organ weight/ histopathology Blood and urine biomarkers (e.g., BUN, CREA, CK) Urinalysis measures (e.g., specific gravity, protein)
Reproductive <i>Note: Evidence synthesis and integration conclusions in the assessment are developed separately for male and female reproductive effects</i>	<ul style="list-style-type: none"> Menstrual disorders 	<ul style="list-style-type: none"> Reproductive organ weight/ histopathology Reproductive hormones Puberty onset Fertility and pregnancy outcomes (e.g., sperm measures, estrous cyclicity, litter size, gestation length, mating/fertility index) Dam body weight/body weight gain

Protocol for the Ethylbenzene IRIS Assessment

Relevant human health effect category ^a	Units of analysis for evidence synthesis that inform evidence integration for the ethylbenzene assessment (each bullet represents a unit of analysis)	
	Human evidence	Animal evidence
Respiratory	<ul style="list-style-type: none"> Measures of respiratory function (e.g., FEV, FVC, MEF) Acute respiratory symptoms (e.g., wheezing, irritation, shortness of breath) 	<ul style="list-style-type: none"> Respiratory organ weight/histopathology Respiratory irritation Respiratory rate
Thyroid (endocrine)	<ul style="list-style-type: none"> No studies 	<ul style="list-style-type: none"> Hormone levels Histopathology Thyroid weight
General toxicity (systemic/whole body)	<ul style="list-style-type: none"> Sick building syndrome Worker health status Adverse health symptoms (e.g., fatigue, nausea) 	<ul style="list-style-type: none"> Mortality and clinical observations (e.g., lethargy, weakness, labored breathing)^b Growth and body weight^b Food consumption

ALT = alanine aminotransferase; AST = aspartate aminotransferase; A1C = glycated hemoglobin; BUN = blood urea nitrogen; CREA = creatinine; CK = creatine kinase; FEV = forced expiratory volume; FVC = forced vital capacity; MEF = maximal expiratory flow; MER = middle ear reflex.

^aBased on the currently available evidence base, other health outcomes will not be formally evaluated in this assessment. However, short summaries of the evidence might be included for context. These decisions may be reevaluated if literature search updates identify additional data that may warrant further evaluation.

^bEffects in dams/pups or animals exposed only during development will be discussed in the developmental and reproductive sections.

6. STUDY EVALUATION (RISK OF BIAS AND SENSITIVITY)

1 The general approach for evaluating primary health effect studies that meet PECO is
2 described in Section 5.1. Instructional and informational materials for study evaluations are
3 available at <https://hawcprd.epa.gov/assessment/100000039/>. The approach is conceptually the
4 same for epidemiology, controlled human exposure, animal toxicology, and in vitro studies but the
5 application specifics differ; thus, they are described separately in Sections 6.2, 6.3 and 6.4,
6 respectively. Any physiologically based PBPK models used in the assessment are evaluated using
7 methods described in the Quality Assurance Project Plan for PBPK models ([U.S. EPA, 2018b](#)), which
8 is summarized below (see Section 6.6).

6.1. STUDY EVALUATION OVERVIEW FOR HEALTH EFFECT STUDIES





9 The IRIS Program uses a domain-based approach to evaluate studies. Key concerns for the
10 review of epidemiology and animal toxicology studies are potential bias (factors that affect the
11 magnitude or direction of an effect in either direction) and insensitivity (factors that limit the
12 ability of a study to detect a true effect; low sensitivity is a bias toward the null when an effect
13 exists). The study evaluations are aimed at discerning the expected magnitude of any identified
14 limitations (focusing on limitations that could substantively change a result), considering the
15 expected direction of the bias. The study evaluation approach is designed to address a range of
16 study designs, health effects, and chemicals. The general approach for reaching an overall judgment
17 regarding confidence in the reliability of the results is illustrated in Figure 6-1.

(a) Individual evaluation domains

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

(b) Domain level judgments and overall study rating

Domain judgments

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

Overall study rating for an outcome

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

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

- To calibrate the assessment-specific considerations, the study evaluation process includes a
- pilot phase to assess and refine the evaluation process. Following this pilot, at least two reviewers
- independently evaluate studies to identify characteristics that bear on the informativeness
- (i.e., validity and sensitivity) of the results. The independent reviewers use structured web-forms

This document is a draft for review purposes only and does not constitute Agency policy.

1 for study evaluation housed within the EPA's version of HAWC
2 (<https://hawcprd.epa.gov/assessment/100000039/>) to record separate judgments for each
3 domain and the overall study for each outcome and unit of analysis, to reach consensus between
4 reviewers, and when necessary, resolve differences by discussion between the reviewers or
5 consultation with additional independent reviewers. As reviewers examine a group of studies,
6 additional chemical-specific knowledge or methodological concerns could emerge, and a second
7 pass of all pertinent studies might become necessary.

8 In general, considerations for reviewing a study with regard to its conduct for specific
9 health outcomes are based on considerations presented in the IRIS Handbook ([U.S. EPA, 2022](#)) and
10 use of existing guideline documents when available, including EPA guidelines for carcinogenicity,
11 neurotoxicity, reproductive toxicity, and developmental toxicity ([U.S. EPA, 2005a](#), [1998](#), [1996](#),
12 [1991a](#)).

13 Authors might be queried to obtain critical information, particularly that involving missing
14 key study design or results information that or additional analyses that could address potential
15 study limitations. During study evaluation, the decision on whether to seek missing information
16 focuses on information that could result in a reevaluation of the overall study confidence for an
17 outcome. Outreach to study authors is documented in HAWC and considered unsuccessful if
18 researchers do not respond to an email or phone request within one month of the attempt to
19 contact. Only information or data that can be made publicly available (e.g., within HAWC or HERO)
20 will be considered.

21 When evaluating studies that examine more than one outcome, the evaluation process is
22 explicitly conducted at the individual outcome level within the study. Thus, the same study may
23 have different outcome domain judgments for different outcomes. These measures could still be
24 grouped for evidence synthesis.

25 During review, for each evaluation domain, reviewers reach a consensus judgment of *good*,
26 *adequate*, *deficient*, *not reported*, or *critically deficient*. If a consensus is not reached, a third
27 reviewer performs conflict resolution. It is important to emphasize that evaluations are performed
28 in the context of the study's utility for identifying individual hazards. Limitations specific to the
29 usability of the study for dose-response analysis are useful to note and applicable to selecting
30 studies for that purpose (see Section 9), but they do not contribute to the study confidence
31 classifications. These four categories are applied to each evaluation domain for each outcome
32 considered within a study, as follows:

- 33 • *Good* represents a judgment that the study was conducted appropriately in relation to the
34 evaluation domain, and any minor deficiencies noted are not expected to influence the
35 study results or interpretation of the study findings.
- 36 • *Adequate* indicates a judgment that methodological limitations related to the evaluation
37 domain are (or are likely to be) present, but those limitations are unlikely to be severe or to
38 notably impact the study results or interpretation of the study findings

- 1 • *Deficient* denotes identified biases or deficiencies interpreted as likely to have had a notable
2 impact on the results, or that limit interpretation of the study findings.
- 3 • *Not reported* indicates the information necessary to evaluate the domain question was not
4 available in the study. Depending on the expected impact, the domain may be interpreted as
5 *adequate* or *deficient* for the purposes of the study confidence rating.
- 6 • *Critically deficient* reflects a judgment that the study conduct relating to the evaluation
7 domain introduced a serious flaw that is interpreted to be the primary driver of any
8 observed effect(s) or makes the study uninterpretable. Studies with *critically deficient*
9 judgments in any evaluation domain are almost always classified as overall *uninformative*
10 for the relevant outcome(s).

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

- 19 • *High* confidence: No notable deficiencies or concerns were identified; the potential for bias
20 is unlikely or minimal, and the study used sensitive methodology. High confidence studies
21 generally reflect judgments of good across all or most evaluation domains.
- 22 • *Medium* confidence: Possible deficiencies or concerns were identified, but the limitations
23 are unlikely to have a significant impact on the study results or their interpretation.
24 Generally, medium confidence studies include adequate or good judgments across most
25 domains, with the impact of any identified limitation not being judged as severe.
- 26 • *Low* confidence: Deficiencies or concerns are identified, and the potential for bias or
27 inadequate sensitivity is expected to have a significant impact on the study results or their
28 interpretation. Typically, low confidence studies have a deficient evaluation for one or more
29 domains, although some medium confidence studies might have a deficient rating in
30 domain(s) considered to have less influence on the magnitude or direction of effect
31 estimates. Low confidence results are given less weight compared with high or medium
32 confidence results during evidence synthesis and integration (see Sections 7 and 8) and are
33 generally not used as the primary sources of information for hazard identification or
34 derivation of toxicity values unless they are the only studies available (in which case, this
35 significant uncertainty would be emphasized during dose-response analysis). Studies rated
36 low confidence only because of sensitivity concerns are asterisked or otherwise noted
37 because they often require additional consideration during evidence synthesis. Effects
38 observed in studies that are biased toward the null may increase confidence in the results,
39 assuming the study is otherwise well conducted (see Section 8).
- 40 • *Uninformative*: Serious flaw(s) are judged to make the study results uninterpretable for use
41 in the assessment. Studies with critically deficient judgments in any evaluation domain are

1 almost always rated uninformative. Studies with multiple deficient judgments across
2 domains may also be considered uninformative. Given that the findings of interest are
3 considered uninterpretable based on the identified flaws (see above definition of critically
4 deficient) and do not provide information of use to assessment interpretations, these
5 studies have no impact on evidence synthesis or integration judgments and are not usable
6 for dose-response analyses but may be used to highlight research gaps.

7 As previously noted, study evaluation determinations reached by each reviewer and the
8 consensus judgment between reviewers are recorded in HAWC. Final study evaluations housed in
9 HAWC are made available when the draft is publicly released. The study confidence classifications
10 and their rationales are carried forward and considered as part of evidence synthesis (see
11 Section 11) to help interpret the results across studies. *Critically deficient* and *Uninformative* ratings
12 are uncommon; these ratings are reserved for critical flaws where the study findings are truly
13 uninterpretable due to identified biases. The most frequent situation where they are used for
14 epidemiology studies is when potential confounding has not been considered using any method
15 (e.g., adjustment, stratification, restriction), including unadjusted correlation coefficients or means
16 in cases/controls in a heterogeneous population where confounding is likely.

6.2. EPIDEMIOLOGY STUDY EVALUATION

17 Evaluation of epidemiology studies of health effects to assess risk of bias and study
18 sensitivity are conducted for the following domains: exposure measurement, outcome
19 ascertainment, participant selection, potential confounding, analysis, study sensitivity, and selective
20 reporting. Bias can result in false positives and negatives (i.e., Types I and II errors), whereas study
21 sensitivity is typically concerned with identifying the latter.

22 The principles and framework used for evaluating epidemiology studies are adapted from
23 the principles in the Cochrane Risk of Bias in Nonrandomized Studies of Interventions [ROBINS-I;
24 [Sterne et al. \(2016\)](#)] but modified to address environmental and occupational exposures. The types
25 of information that may be the focus of those criteria are listed in Table 6-1. Core and prompting
26 questions, presented in Table 6-2, are used to collect information to guide evaluation of each
27 domain. Core questions represent key concepts while the prompting questions help the reviewer
28 focus on relevant details under each key domain. Exposure- and outcome-specific criteria to use
29 during study evaluation are developed using the core and prompting questions and refined during a
30 pilot phase with engagement from topic-specific experts. The protocol may also be adjusted in the
31 early phases of the study evaluation process if corrections are identified based on initial literature
32 reviews. Exposure and confounding domain considerations specific to ethylbenzene are presented
33 in Sections 6.2.1 to 6.2.6.

Table 6-1. Information relevant to evaluation domains for epidemiology studies

Domain	Types of information that may need to be collected or are important for evaluating the domain
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 repeated-measures studies, 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, 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 or life stages (see discussion in Section 9)?
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.
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 the level of exposure contrast between groups (i.e., the extent to which the “unexposed group” is truly unexposed, and the prevalence of exposure in the group designated as “exposed”).
Selective reporting	Are results presented with adequate detail for all the endpoints and exposure measures reported in the methods section, and are they relevant to 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?

1

Table 6-2. Questions to guide the development of criteria for each domain in epidemiology studies

Domain and core question	Prompting questions	Follow-up questions	Considerations that apply to most exposures and outcomes
<p>Exposure measurement</p> <p>Does the exposure measure reliably distinguish between levels of exposure in a time window considered most relevant for a causal effect with respect to the development of the outcome?</p>	<p>For all:</p> <ul style="list-style-type: none"> Does the exposure measure capture the variability in exposure among the participants, considering intensity, frequency, and duration of exposure? Does the exposure measure reflect a relevant time window? If not, can the relationship between measures in this time and the relevant time window be estimated reliably? Is the exposure measurement likely to be affected by a knowledge of the outcome? Is the exposure measurement likely to be affected by the presence of the outcome (i.e., reverse causality)? <p>For case-control studies of occupational exposures:</p> <ul style="list-style-type: none"> Is exposure based on a comprehensive job history describing tasks, setting, time period, and use of specific materials? <p>For biomarkers of exposure, general population:</p> <ul style="list-style-type: none"> Is a standard assay used? What are the intra- and inter-assay 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? 	<p>Is the degree of exposure misclassification likely to vary by exposure level?</p> <p>If the correlation between exposure measurements is moderate, is there an adequate statistical approach to ameliorate variability in measurements?</p> <p>If there is a concern about the potential for bias, what is the predicted direction or distortion of the bias on the effect estimate (if there is enough information)?</p>	<p>These considerations require customization to the exposure and outcome (relevant timing of exposure).</p> <p>Good</p> <ul style="list-style-type: none"> Valid exposure assessment methods used, which represent the etiologically relevant time period of interest. Exposure misclassification is expected to be minimal. <p>Adequate</p> <ul style="list-style-type: none"> Valid exposure assessment methods used, which represent the etiologically relevant time period of interest. Exposure misclassification may exist but is not expected to greatly change the effect estimate. <p>Deficient</p> <ul style="list-style-type: none"> Valid exposure assessment methods used, which represent the etiologically relevant time period of interest. Specific knowledge about the exposure and outcome raises concerns about reverse causality, but there is uncertainty whether it is influencing the effect estimate. Exposed groups are expected to contain a notable proportion of unexposed or minimally exposed individuals, the method did not capture important temporal or spatial variation, or there is other evidence of exposure misclassification that would be expected to notably change the effect estimate. <p>Critically deficient</p> <ul style="list-style-type: none"> Exposure measurement does not characterize the etiologically relevant time period of exposure or is not valid. There is evidence that reverse causality is very likely to account for the observed association. Exposure measurement was not independent of outcome status.

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

Domain and core question	Prompting questions	Follow-up questions	Considerations that apply to most exposures and outcomes
analysis sample) is jointly related to exposure and to outcome?	<p>For occupational cohort:</p> <ul style="list-style-type: none"> • Did entry into the cohort begin with the start of the exposure? • Was follow-up or outcome assessment incomplete, and if so, was follow-up related to both exposure and outcome status? • Could exposure produce symptoms that would result in a change in work assignment/work status (“healthy worker survivor effect”)? <p>For case-control study:</p> <ul style="list-style-type: none"> • Were controls representative of population and time periods from which cases were drawn? • Are hospital controls selected from a group whose reason for admission is independent of exposure? • Could recruitment strategies, eligibility criteria, or participation rates result in differential participation relating to both disease and exposure? 	<p>If there is a concern about the potential for bias, what is the predicted direction or distortion of the bias on the effect estimate (if there is enough information)?</p> <p>Are appropriate analyses performed to address changing exposures over time in relation to symptoms?</p> <p>Is there a comparison of participants and nonparticipants to address whether differential selection is likely?</p>	<ul style="list-style-type: none"> • Minimal concern for selection bias based on description of recruitment process (e.g., selection of comparison population, population based random sample selection, recruitment from sampling frame including current and previous employees). • Exclusion and inclusion criteria are specified and do not induce bias. • Participation rate is reported at all steps of study (e.g., initial enrollment, follow-up, selection into analysis sample). If rate is not high, there is appropriate rationale for why it is unlikely to be related to exposure (e.g., comparison between participants and nonparticipants or other available information indicates differential selection is not likely). <p>Adequate</p> <ul style="list-style-type: none"> • Enough of a description of the recruitment process to be comfortable that there is no serious risk of bias. • Inclusion and exclusion criteria are specified and do not induce bias. • Participation rate is incompletely reported but available information indicates participation is unlikely to be related to exposure. <p>Deficient</p> <ul style="list-style-type: none"> • Little information on recruitment process, selection strategy, sampling framework and/or participation or aspects of these processes raise the potential for bias (e.g., healthy worker effect, survivor bias).
	<p>For population-based survey:</p> <ul style="list-style-type: none"> • Was recruitment based on advertisement to people with knowledge of exposure, outcome, and hypothesis? 		<p>Critically deficient</p> <ul style="list-style-type: none"> • Aspects of the processes for recruitment, selection strategy, sampling framework, or participation result in concern that selection bias resulted in a large impact on effect estimates (e.g., convenience sample with no information about recruitment and selection, cases

Protocol for the Ethylbenzene IRIS Assessment

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

Protocol for the Ethylbenzene IRIS Assessment

Domain and core question	Prompting questions	Follow-up questions	Considerations that apply to most exposures and outcomes
			<p>Adequate</p> <ul style="list-style-type: none"> Similar to <i>good</i> but may not have included all key confounders, or less detail may be available on the evaluation of confounders (e.g., subbullets in <i>good</i>). It is possible that residual confounding could explain part of the observed effect, but concern is minimal.
			<p>Deficient</p> <ul style="list-style-type: none"> Does not include variables in the models that are likely to be influential colliders or intermediates on the causal pathway. <p>And any of the following:</p> <ul style="list-style-type: none"> The potential for bias to explain some of the results is high based on an inability to rule out residual confounding, such as a lack of demonstration that key confounders of the exposure outcome relationships are considered; Descriptive information on key confounders (e.g., their relationship relative to the outcomes and exposure levels) are not presented; or Strategy of evaluating confounding is unclear or is not recommended (e.g., only based on statistical significance criteria or stepwise regression [forward or backward elimination]). <p>Critically deficient</p> <ul style="list-style-type: none"> Includes variables in the models that are colliders and/or intermediates in the causal pathway, indicating that substantial bias is likely from this adjustment or Confounding is likely present and not accounted for, indicating that all of the results are most likely due to bias. <ul style="list-style-type: none"> Presenting a progression of model results with adjustments for different potential confounders, if warranted.

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

Domain and core question	Prompting questions	Follow-up questions	Considerations that apply to most exposures and outcomes
			<p>Deficient</p> <ul style="list-style-type: none"> Does not conduct analysis using optimal characterization of the outcome variable. Descriptive information about exposure levels is not provided (where applicable). Effect estimate and <i>p</i>-value are presented, without standard error or confidence interval. Results are presented as statistically “significant”/“not significant.” <p>Critically deficient</p> <ul style="list-style-type: none"> Results of analyses of effect modification are examined without clear a priori rationale and without providing main/principal effects (e.g., presentation only of statistically significant interactions that were not hypothesis driven). Analysis methods are not appropriate for design or data of the study.
<p><u>Selective reporting</u> Is there reason to be concerned about selective reporting?</p>	<ul style="list-style-type: none"> Are results provided for all the primary analyses described in the methods section? Is there appropriate justification for restricting the amount and type of results that are shown? Are only statistically significant results presented? 	<p>If there is a concern about the potential for bias, what is the predicted direction or distortion of the bias on the effect estimate (if there is enough information)?</p>	<p>These considerations generally do not require customization and may have fewer than four levels.</p> <p>Good</p> <ul style="list-style-type: none"> The results reported by study authors are consistent with the primary and secondary analyses described in a registered protocol or methods paper. <p>Adequate</p> <ul style="list-style-type: none"> The authors described their primary (and secondary) analyses in the methods section and results are reported for all primary analyses. <p>Deficient</p> <ul style="list-style-type: none"> Concerns are raised based on previous publications, a methods paper, or a registered protocol indicating that analyses are planned or conducted that are not reported, or that hypotheses

Protocol for the Ethylbenzene IRIS Assessment

Domain and core question	Prompting questions	Follow-up questions	Considerations that apply to most exposures and outcomes
			<p>originally considered to be secondary are represented as primary in the reviewed paper.</p> <ul style="list-style-type: none"> • Only subgroup analyses are reported suggesting that results for the entire group are omitted. • Only statistically significant results are reported.
<p>Sensitivity Is there a concern that sensitivity of the study is not adequate to detect an effect?</p>	<ul style="list-style-type: none"> • Is the exposure range adequate to detect associations and exposure-response relationships? • Was the appropriate population included? • Was the length of follow-up adequate? Is the time/age of outcome ascertainment optimal given the interval of exposure and the health outcome? • Are there other aspects related to risk of bias or otherwise that raise concerns about sensitivity? 		<p>These considerations may require customization to the exposure and outcome. Depending on the needs of the assessment, there may be fewer than four rating levels. Some study features that affect study sensitivity may have already been included in the other evaluation domains; these should be noted in this domain again, along with any features that have not been addressed elsewhere so that the rating provides an overall summary of factors that may impact sensitivity. When determining the overall study confidence rating, the evaluator should be conscious that a limitation could contribute to multiple domains and not double-penalize the study. Some considerations include:</p> <p>Good</p> <ul style="list-style-type: none"> • The range of exposure levels provides sufficient variability in exposure distribution and/or sufficient range or contrasts (e.g., across groups or exposure categories) to detect associations or exposure-response relationships that may be present. • The population was exposed to levels expected to have an impact on response. • The study population was at risk of developing the outcomes of interest (e.g., ages, life stage, sex). • The timing of outcome ascertainment was appropriate given expected latency for outcome development (i.e., adequate follow-up interval). • There was evidence of sufficient statistical power (which may include formal power calculations) to observe an effect if it exists.

Protocol for the Ethylbenzene IRIS Assessment

Domain and core question	Prompting questions	Follow-up questions	Considerations that apply to most exposures and outcomes
			<ul style="list-style-type: none"> • No other concerns raised regarding study sensitivity (e.g., no evidence that results would be attenuated enough to preclude detection of an adverse health effect). <p>Adequate</p> <ul style="list-style-type: none"> • Same considerations as <i>good</i>, except: <ul style="list-style-type: none"> ○ Issues are identified that could reduce sensitivity, but they are unlikely to impact the overall findings of the study. <p>Deficient</p> <ul style="list-style-type: none"> • Concerns were raised about the issues described for <i>good</i> that are expected to notably decrease the sensitivity of the study to detect associations for the outcome (i.e., reasonably high likelihood of a false null result). • Note: <i>Deficient</i> sensitivity indicates that null findings should be interpreted with caution and may not represent a lack of association. <p>Critically deficient</p> <ul style="list-style-type: none"> • Severe concerns were raised about the sensitivity of the study such that any observed association is uninterpretable (e.g., exposure gradients/contrasts that precluded an ability to distinguish exposure levels between study participants).

6.2.1. Epidemiological Study Evaluation Considerations Specific to Exposure Domain for Ethylbenzene

1 Ethylbenzene is present in solvents, inks, paint, pesticides and other household products,
2 and concentrations indoors are typically higher than levels measured outdoors. Traffic emissions,
3 escape of vapors at gas stations or car repair garages, car and truck idling in parking lots and
4 border crossings, and emissions from the petrochemical industry are primary contributors to
5 ethylbenzene concentrations in ambient air. While ethylbenzene from ambient air contributes to
6 indoor levels, variability of ethylbenzene levels in residences primarily is due to indoor sources,
7 such as the presence of a smoker in the home ([Wallace et al., 1987](#)) or product use ([Adgate et al.,](#)
8 [2004](#)), as well as housing characteristics, such as attached garages and ventilation ([Sexton et al.,](#)
9 [2007](#)). Because there are unique sources both indoors and outdoors, individual-level exposure
10 assessments for health effects studies ideally would capture contributions from time at home,
11 school, or work, and in transit.

6.2.2. Exposure Assessment Approaches used in Epidemiology Studies of Ethylbenzene and Potential Misclassification

12 A few of the epidemiology studies in the ethylbenzene inventory characterized individual
13 exposures using personal monitoring over a few days. Most of these studies measured average
14 concentrations in the home over a period of days to a few weeks during one or more seasons.
15 Because indoor levels typically are higher than outdoor concentrations and people typically spend
16 the majority of their time indoors, measurements of exposure levels in the home are likely to
17 adequately characterize personal exposure. A comparison of exposure estimates in children or
18 nonsmoking adults in Minnesota using personal sampling and a time-weighted model, based on
19 indoor measurements in their homes and schools (or work) and outdoors at school (or
20 community), found that the model with only the home measurements was comparable to the model
21 containing all microenvironments in explaining the variation in the personal exposure
22 measurements ([Adgate et al., 2004](#); [Sexton et al., 2004a](#)). The degree to which indoor residential
23 measurements explain personal exposure likely depends on the local and meteorological
24 characteristics in different locations. Within communities, variation in indoor aromatic VOC
25 concentrations is primarily due to variability between residences and between seasons, with much
26 lower variability due to variation between cities or measurement error ([Jia et al., 2012](#)). Within a
27 residence, concentrations measured in different rooms (e.g., living room, bedroom) are highly
28 correlated ([Wallace et al., 1991](#)). Therefore, to characterize average indoor exposure to
29 ethylbenzene over longer timeframes (e.g., the previous year), sampling from at least one room
30 would be adequate, but multiple sampling periods in different seasons would provide an estimate
31 with less exposure misclassification compared with estimates based on measurements during one
32 season.

33 Exposure estimates based on outdoor concentrations or air quality models capture a small
34 portion of an individual's average exposure. Studies have demonstrated that estimates based on

1 ambient exposures are an underestimate of an individual’s personal exposure ([Sexton et al.,](#)
2 [2004b](#)), and, similarly, increasing evidence suggests the importance of indoor sources ([Konkle et al.,](#)
3 [2020](#)). However, health effect studies may be able to identify associations with ambient
4 ethylbenzene exposure using methods to characterize the spatial or temporal variation in
5 communities, primarily due to traffic and industrial point sources. Annual exposure estimates
6 based on land use regression that capture finer scale concentration gradients across a community
7 are expected to result in less exposure misclassification compared with methods based on
8 measurements from central site monitors accounting for the relative distance to subject’s homes
9 ([Mukerjee et al., 2009](#); [Aguilera et al., 2008](#)). However, the use of exposure estimates from land use
10 regression (LUR) models in epidemiology studies of air pollution can introduce measurement error
11 with attenuated effect estimates and inflated variance, if the spatial variation within a community
12 has not been adequately characterized ([Basagaña et al., 2013](#)). Publications reporting studies of
13 ambient ethylbenzene exposure should describe the approach to model development and present
14 information about the sources of ethylbenzene emissions and their impact on spatial variation. Still,
15 due to concerns about misclassification from ambient exposure assessment approaches, these
16 studies will be unable to reach the “good” ranking in the exposure domain.

17 Some of the studies of ambient exposure in the ethylbenzene inventory used annual average
18 exposure estimates for each census tract generated by regional air quality models based on the
19 National Emissions Inventory (e.g., National Air Toxics Assessment data)
20 (<https://www.epa.gov/national-air-toxics-assessment/2014-nata-assessment-results>). NATA
21 estimates are based on the National Emissions Inventory (NEI) for a specific year, which uses
22 empirical and engineering factors, not measurements, but the models account for spatial variation
23 incorporating secondary formation and decay, pollutant dispersion, meteorology, population
24 activity data, and several sources of exposure. NATA is a screening level tool to look at annual
25 population exposures. NATA has been found to underpredict concentrations of many VOCs due to
26 missing and underestimated emission sources and other reasons ([U.S. EPA, 2010a](#)). For
27 ethylbenzene, a comparison of annual average concentrations at 242 specific sites estimated using
28 the model outputs from 2005 and monitoring data found a median model-to-monitor ratio of 0.471
29 with 85% of the modeled estimates underestimating those based on monitoring data. Twenty
30 percent of the modeled values were within 30% of the values based on monitors and 41% of the
31 modeled values were within a factor of 2 of those based on monitoring data
32 (<https://www3.epa.gov/ttn/chief/conference/ei19/session1/oommen.pdf>). These analyses
33 indicate that exposure misclassification may be a concern for individual exposure estimates based
34 on the 2005 (and previous) NATA models. Other sources of exposure misclassification in
35 epidemiology studies that use NATA estimates include the use of exposure assignments at the
36 census tract level (not individual level) and the use of annual average concentration estimates for
37 only one year (i.e., 1996 or 2005).

1 Exposure estimates in a few of the epidemiology studies of ethylbenzene exposure were
2 derived using the Community Multi-scale Air Quality (CMAQ) model, which also uses emissions
3 data, as well as meteorological and atmospheric chemistry inputs. CMAQ models concentrations
4 over large regions using a 36-km horizontal resolution domain but has also been used to model
5 concentrations at a finer resolution (i.e., 1 km). Exposure estimates based on a grid size of 36 km
6 would have limited spatial resolution, and therefore exposure misclassification would be of greater
7 concern.

6.2.3. ADME and Notes Relevant to Biomarkers

8 Similar to many VOCs, ethylbenzene is rapidly distributed in the body and can undergo
9 metabolism prior to elimination unchanged as the parent compound in exhaled breath or its
10 metabolic derivatives in urine. Thus, ethylbenzene is generally not persistent in the body: the half-
11 life in blood is less than a half-hour ([ATSDR, 2010](#)). A complex multiexponential elimination curve
12 for ethylbenzene was measured in the blood of four individuals after a six-hour exposure to a
13 mixture of VOCs, including ethylbenzene. While declines after exposure ended were rapid during
14 the first hour, subsequent decline slowed and a three-compartment model appeared to be the best
15 fit to the data ([Ashley and Prah, 1997](#)). Although bioaccumulation may occur, the concentration in
16 blood primarily signifies recent exposure levels and is not considered a relevant exposure measure
17 for chronic disease (e.g., prevalent cardiovascular disease). Analyses of matched blood values and
18 personal air measurements of BTEX compounds (benzene, toluene, ethylbenzene, o-xylene, m-/p-
19 xylene) have found relatively low correlations, possibly due to mistiming of the air sampling or
20 other unknown factors ([Su et al., 2011](#); [Sexton et al., 2005](#)). Because of rapid clearance, blood
21 concentrations would reflect exposures occurring just prior to a blood draw.

22 In contrast to blood biomarkers, urinary biomarkers of VOCs have delayed clearance and
23 therefore may be representative of exposures in the period of hours to days ([Heinrich-Ramm et al.,
24 2000](#)). Therefore, urinary biomarkers are preferable to blood biomarkers to assess daily exposures
25 to VOCs potentially relevant to chronic health outcomes, though it is important to adjust for kidney
26 function when using urinary measures ([Heinrich-Ramm et al., 2000](#)). However, it should be noted
27 that the primary measurable metabolites for ethylbenzene (mandelic acid and phenylglyoxylic acid)
28 are not specific to ethylbenzene and are also derived from styrene, which is commonly detected in
29 conjunction with ethylbenzene ([Capella et al., 2019](#)). As such, the use of urinary biomarkers should
30 be restricted to cases where substantial co-exposure to styrene can be ruled out. Overall, in
31 comparison to outdoor or indoor air measurement alone, the use of biomarkers can account for
32 exposures from multiple routes and sources and may have smaller variance ratios than air
33 measurements ([Lin et al., 2005](#)). They may also better capture the growing importance of exposure
34 from to VOCs from volatile chemical products ([Mcdonald et al., 2018](#)), which may not be accounted
35 for in traditional ambient exposure models.

6.2.4. Time Frames Represented by Exposure Assessments

1 The time frame represented by the exposure estimates should correspond to the period in
 2 which the health outcomes were expected to have developed. Indoor exposure assessments
 3 representing a period of week(s) in more than one season could reasonably characterize average
 4 exposure over the previous year and would be relevant to immune-related or other symptoms (e.g.,
 5 asthma, wheezing illness, allergy symptoms, sensory irritation) occurring over the previous several
 6 weeks to a year. Daily sampling is best, but periodic sampling on a less than daily basis may be
 7 sufficient depending on the variability in air concentrations. Developmental outcomes should be
 8 evaluated in relation to the relevant critical exposure periods during pregnancy if they are known.
 9 Exposure measurements with shorter time frames are less informative for studying the prevalence
 10 or incidence of chronic disease, such as physician-diagnosed asthma, cardiovascular disease,
 11 cancer.

6.2.5. Correlation Between BTEX Compounds and Potential Confounding

12 BTEX compounds, all traffic pollutants, are correlated in ambient air ($r = 0.43 - 0.59$)
 13 ([Sexton et al., 2004a](#)). Ethylbenzene and o-xylene concentrations in blood were correlated in the
 14 NHANES III and continuous NHANES cohorts ($r = 0.81$ and 0.89 , respectively) (see Appendix C in [Su
 15 et al. \(2011\)](#)). Confounding of observed associations with health outcomes by other BTEX
 16 compounds is best considered when interpreting results across studies if they analyzed exposures
 17 from different locations or settings (e.g., traffic-related, indoor product use).

6.2.6. Exposure Domain Evaluation Levels

18 The following exposure domain rating levels will be applied. The exposure assessment
 19 methods will be evaluated for how well they characterize either (1) total personal/residential or
 20 (2) outdoor (ambient) ethylbenzene exposure to the individuals in the study.

Table 6-3. Estimates representing total individual-level exposure based on personal or residential monitoring

Rating	Criteria
Good	Integrated personal measurements using passive monitors, over multiple 24-hr periods (since there could be relevant daily variations), or time-weighted summary concentrations incorporating concentrations in residence and school/workplace. Sampling details provided including type of samplers, placement of samplers, sampling periods, status of activities in structures, chemical analysis methods (or citation provided). Time frame of measurements appropriate to development of health outcome. OR Area measurements in home using passive or active monitors, average of measurements in one or more rooms; average over longer periods is better (weeks) and multiple seasons if estimating annual average. Sampling details provided including type of samplers, placement of samplers, sampling periods, status of activities in structures, chemical

Protocol for the Ethylbenzene IRIS Assessment

Rating	Criteria
	<p>analysis methods (or citation provided). Time frame of measurements appropriate to development of health outcome.</p> <p>OR</p> <p>In cases where co-exposure to styrene can be ruled out, urinary biomarkers collected via standardized procedures (e.g., gas chromatography–mass spectrometry, GC/MS) and appropriate QC.</p>
Adequate	<p>Area measurements in home using passive or active monitors, average of measurements in one or more rooms; average of shorter duration (less than 1 wk) with information about monitoring protocol, and multiple seasons if estimating annual average. Sampling details provided including type of samplers, placement of samplers, sampling periods, status of activities in structures, chemical analysis methods (or citation provided). Time frame of measurements appropriate to development of health outcome.</p>
Deficient	<p>Area measurements in home obtained on one occasion if estimating annual average. (A single measure does not capture daily variations in the relative proportion of time in different microenvironments nor variations in concentrations of VOCs (Kim et al., 2002). Sampling details provided including type of samplers, placement of samplers, sampling periods, status of activities in structures, chemical analysis methods (or citation provided). Time frame of measurements appropriate to development of health outcome.</p> <p>OR</p> <p>Use of questionnaires or observations of VOC products in the home by trained study personnel</p> <p>OR</p> <p>Blood biomarkers collected via standardized procedures (e.g., GC/MS) and appropriate QC</p> <p>OR</p> <p>Urinary biomarkers (not specific to ethyl benzene and where there is concern for co-exposure to styrene) collected via standardized procedures (e.g., GC/MS) and appropriate QC</p> <p>OR</p> <p>Air sampling with gas chromatography-flame ionization detection (preferred method would utilize mass spectrometry detection) (e.g., gas chromatography–mass spectrometry).</p>
Critically deficient	<p>Time frame for exposure estimation was not appropriate to development of health outcome.</p>

Table 6-4. Exposure to ethylbenzene in ambient air

Rating	Criteria
Good	No studies using ambient exposure assessment approaches can reach classification of “good” due to concerns regarding misclassification of personal/individual-level exposure.
Adequate	Average estimates based on land use regression models developed for location where study was conducted including description of model development and sufficient information about how the model adequately characterizes spatial variation in the community due to what was known about sources. Time frame of measurements appropriate to development of health outcome. Potentially other methods besides LUR might fall into this category if detailed validation information was provided to ensure model adequately characterizes spatial variation.
Deficient	<p>Average estimates based on land use regression models developed for location where study was conducted, but some uncertainties remain regarding how the model was developed or how the model adequately characterizes spatial variation in the community due to what was known about sources.</p> <p>OR</p> <p>Annual average estimates or other time-period-specific averages appropriate to development of health outcome based on NATA data linked to residential census tract.</p> <p>OR</p> <p>Annual average estimates or other time-period-specific averages appropriate to development of health outcome based on chemical transport models (CMAQ) using spatially resolved grid size (i.e., 1 km).</p> <p>OR</p> <p>Annual average estimates based on proximity to central monitor for homes, with multiple sampling locations in a community, with some description of how well the monitoring network characterizes variation due to sources. Time frame of measurements averages appropriate to development of health outcome.</p>
Critically deficient	<p>Annual average estimates or other time-period-specific averages appropriate to development of health outcome based on CMAQ using large grid (resolution) size (i.e., 36 km).</p> <p>OR</p> <p>Time frame for exposure estimation was not appropriate to development of health outcome</p> <p>OR</p> <p>Air sampling with gas chromatography-flame ionization detection (preferred method would utilize mass spectrometry detection) (e.g., gas chromatography-mass spectrometry).</p>

6.3. CONTROLLED HUMAN EXPOSURE STUDY EVALUATION

1 This study design involves human volunteers to test specific hypotheses about short-term
2 exposures and biological responses that inform potential mechanisms and understanding of
3 exposure-response patterns. The exposures are generated in the laboratory to achieve
4 predetermined concentrations for periods of minutes to hours. For study evaluation, a process
5 incorporating aspects of the approaches used for epidemiology studies and experimental animal
6 studies, as well as the ROBINS-I tool discussed in Section 6.2 ([Sterne et al., 2016](#)), are used to
7 evaluate controlled exposure studies in humans. Controlled human exposure studies are evaluated
8 for important attributes of experimental studies, including randomization of exposure assignments,
9 blinding of subjects and investigators, exposure generation, inclusion of a clean air control
10 exposure (if applicable), study sensitivity, and other aspects of the exposure protocol. Sample size is
11 considered, as is the process of recruitment and selection of study subjects and differences in
12 characteristics between groups reflecting potential differences in sensitivity.

6.4. EXPERIMENTAL ANIMAL STUDY EVALUATION

13 Using the principles described in Section 6.1, the animal studies of health effects to assess
14 risk of bias and sensitivity are evaluated for the following domains: allocation, observational
15 bias/blinding, confounding, selective reporting, attrition, chemical administration and
16 characterization, endpoint measurement and validity, results presentation and comparisons, and
17 sensitivity (see Table 6-5).

18 The rationale for judgments is documented at the outcome level. The evaluation
19 documentation in HAWC includes the identified limitations and their expected impact on the overall
20 confidence level. To the extent possible, the rationale will reflect an interpretation of the potential
21 influence on the outcome-specific results, including the direction or magnitude of influence
22 (or both).

Table 6-5. Domains, questions, and general considerations to guide the evaluation of animal toxicology studies

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

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

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

Domain and core question	Prompting questions	General considerations
<p>endpoint(s)/outcome(s) of interest? <i>Notes:</i> <i>Considerations related to the sensitivity of the animal model and timing of endpoint measurement are evaluated under Sensitivity</i> <i>Considerations related to adjustments/corrections to endpoint measurements (e.g., organ weight corrected for body weight) are addressed under results presentation.</i></p>	<p>Are there concerns regarding the methodology selected for endpoint evaluation? Are there concerns about the specificity of the experimental design? Are there serious concerns regarding the sample size or how endpoints were sampled? Are appropriate control groups for the study/assay type included?</p>	<p>Good:</p> <ul style="list-style-type: none"> • Adequate description of methods and animal models. • Use of generally accepted and reliable endpoint methods. • Sample sizes are generally considered adequate for the assay or protocol of interest and there are no notable concerns about sampling in the context of the endpoint protocol (e.g., sampling procedures for histological analysis). • Includes appropriate control groups and any use of nonconcurrent or historical control data (e.g., for evaluation of rare tumors) is justified (e.g., authors or evaluators considered the similarity between current experimental animals and laboratory conditions to historical controls). <p>Ratings of Adequate, Deficient, and Critically Deficient are generally defined as follows:</p> <p>Adequate: Issues are identified that may affect endpoint measurement but are considered unlikely to substantially impact the overall findings or the ability to reliably interpret those findings.</p> <p>Deficient: Concerns are raised that are expected to notably affect endpoint measurement and reduce the reliability of the study findings</p> <p>Critically deficient: Severe concerns are raised about endpoint measurement and any findings are likely to be largely explained by these limitations</p> <p>The following specific examples of relevant concerns are typically associated with a Deficient rating, but Adequate or Critically Deficient might be applied depending on the expected impact of limitations on the reliability and interpretation of the results:</p> <ul style="list-style-type: none"> • Study report lacks important details that are necessary to evaluate the appropriateness of the study design (e.g., description of the assays or protocols; information on the strain, sex, or lifestage of the animals) • Selection of protocols that are nonpreferred or lack specificity for investigating the endpoint of interest. This includes omission of additional experimental criteria (e.g., inclusion of a positive control or dosing up to levels causing minimal toxicity) when required by specific testing guidelines/protocols.* • Overt toxicity (e.g., mortality, extreme weight loss) is observed or expected based on findings from similarly designed studies and may mask interpretation of outcome(s) of interest.

Domain and core question	Prompting questions	General considerations
		<ul style="list-style-type: none"> • Sample sizes are smaller than is generally considered adequate for the assay or protocol of interest. Inadequate sampling can also be raised within the context of the endpoint protocol (e.g., in a pathology study, bias that is introduced by only sampling a single tissue depth or an inadequate number of slides per animal)** • Control groups are not included, considered inappropriate, or comparisons to nonconcurrent or historical controls are not adequately justified <p>*These limitations typically also raise a concern for insensitivity</p> <p>** Sample size alone is not a reason to conclude an individual study is critically deficient.</p>
<p>Results presentation Are the results presented and compared in a way that is appropriate and transparent?</p>	<p>For each endpoint/outcome or grouping of endpoints/outcomes in a study: Does the level of detail allow for an informed interpretation of the results? Are the data compared, or presented, in a way that is inappropriate or misleading?</p>	<p>Considerations for this domain are highly variable depending on the outcomes of interest and typically must be refined by assessment teams. A judgment and rationale for this domain should be given for each endpoint/outcome or group of endpoints/outcomes investigated in the study. Some considerations include the following: Good:</p> <ul style="list-style-type: none"> • No concerns with how the data are presented. • Results are quantified or otherwise presented in a manner that allows for an independent consideration of the data (assessments do not rely on author interpretations). • No concerns with completeness of the results reporting.* <p>Ratings of Adequate, Deficient, and Critically Deficient are generally defined as follows:</p> <p>Adequate: Concerns are identified that may affect results presentation but are considered unlikely to substantially impact the overall findings or the ability to reliably interpret those findings.</p> <p>Deficient: Concerns with results presentation are identified and expected to substantially impact results interpretation and reduce the reliability of the study findings.</p> <p>Critically deficient: Severe concerns about results presentation were identified and study findings are likely to be largely explained by these limitations.</p>

Domain and core question	Prompting questions	General considerations
		<p>The following specific examples of relevant concerns are typically associated with a Deficient rating but Adequate or Critically Deficient might be applied depending on expected impact of limitations on the reliability and interpretation of the results:</p> <ul style="list-style-type: none"> • Nonpreferred presentation of data (e.g., developmental toxicity data averaged across pups in a treatment group, when litter responses are more appropriate; presentation of only absolute organ weight data when relative weights are more appropriate). • Pooling data when responses are known or expected to differ substantially (e.g., across sexes or ages). • Incomplete presentation of the data* (e.g., presentation of mean without variance data; concurrent control data are not presented; dichotomizing or truncating continuous data). <p>*Failure to describe <u>any</u> findings for assessed outcomes (i.e., report lacks any qualitative or quantitative description of the results in tables, figures, or text) is addressed under Selective Reporting.</p>
<p>Selective reporting Did the study report results for all prespecified outcomes? <i>Note:</i> <i>This domain does not consider the appropriateness of the analysis/results presentation. This aspect of study quality is evaluated in another domain.</i></p>	<p>For each study: Are results presented for all endpoints/outcomes described in the methods (see note)? If unexplained results omissions are identified, what is the expected impact on the interpretation of the results?</p>	<p>These considerations typically do not need to be refined by assessment teams. A judgment and rationale for this domain should be given for each cohort or experiment in the study.</p> <p>Good: Quantitative or qualitative results were reported for all prespecified outcomes (explicitly stated or inferred), exposure groups and evaluation time points. Data not reported in the primary article is available from supplemental material. If results omissions are identified, the authors provide an explanation, and these are not expected to impact the interpretation of the results.</p> <p>Adequate: Quantitative or qualitative results are reported for most prespecified outcomes (explicitly stated or inferred) and evaluation time points. Omissions and are not explained but are not expected to significantly impact the interpretation of the results.</p> <p>Deficient: Quantitative or qualitative results are missing for many prespecified outcomes (explicitly stated or inferred), omissions are not explained and may significantly impact the interpretation of the results.</p> <p>Critically deficient: Extensive results omission is identified and prevents comparisons of results across treatment groups.</p>

Domain and core question	Prompting questions	General considerations
<p>Sensitivity Are there concerns that sensitivity in the study is not adequate to detect an effect? <i>Note:</i> <i>Consideration of exposure level selection (e.g., were levels sufficiently high to elicit effects) is addressed during evidence synthesis and is not a study sensitivity consideration.</i></p>	<p>Was the exposure period, timing (e.g., lifestage), frequency, and duration sensitive for the outcome(s) of interest? Given knowledge of the health hazard of concern, did the selection of species, strain, and/or sex of the animal model reduce study sensitivity? Are there concerns regarding the timing (e.g., lifestage) of the outcome evaluation? Are there aspects related to risk of bias domains that raise concerns about insensitivity (e.g., selection of protocols that are known to be insensitive or nonspecific for the outcome(s) of interest)</p>	<p>These considerations may require customization to the specific exposure and outcomes. Some study design features that affect study sensitivity may have already been included in the other evaluation domains; these should be noted in this domain, along with any features that have not been addressed elsewhere. Some considerations include:</p> <p>Good</p> <ul style="list-style-type: none"> • The experimental design (considering exposure period, timing, frequency, and duration) is appropriate and sensitive for evaluating the outcome(s) of interest. • The selected animal model (considering species, strain, sex, and/or lifestage) is known or assumed to be appropriate and sensitive for evaluating the outcome(s) of interest. • No significant concerns with the ability of the experimental design to detect the specific outcome(s) of interest. (e.g., outcomes evaluated at the appropriate lifestage; study designed to address known endpoint variability that is unrelated to treatment, such as estrous cyclicity or time of day). • Timing of endpoint measurement in relation to the chemical exposure is appropriate and sensitive (e.g., behavioral testing is not performed during a transient period of test chemical-induced depressant or irritant effects; endpoint testing does not occur only after a prolonged period, such as weeks or months, of nonexposure). • Potential sources of bias toward the null are not a substantial concern. <p>Adequate Same considerations as <i>Good</i>, except:</p> <ul style="list-style-type: none"> • The duration and frequency of the exposure was appropriate, and the exposure covered most of the critical window (if known) for the outcome(s) of interest. • Potential issues are identified that could reduce sensitivity, but they are unlikely to impact the overall findings of the study. <p>Deficient</p> <ul style="list-style-type: none"> • Concerns were raised about the considerations described for <i>Good</i> or <i>Adequate</i> that are expected to notably decrease the sensitivity of the study to detect a response in the exposed group(s).

Domain and core question	Prompting questions	General considerations
		<p>Critically deficient</p> <ul style="list-style-type: none"> Severe concerns were raised about the sensitivity of the study and experimental design such that any observed associations are likely to be explained by bias. The rationale should indicate the specific concern(s).
<p>Overall confidence Considering the identified strengths and limitations, what is the overall confidence rating for the endpoint(s)/outcome(s) of interest?</p>	<p>For each endpoint/outcome or grouping of endpoints/outcomes in a study: Were concerns (i.e., limitations or uncertainties) related to the risk of bias or sensitivity identified? If yes, what is their expected impact on the overall interpretation of the reliability and validity of the study results, including (when possible) interpretations of impacts on the magnitude or direction of the reported effects?</p>	<p>The overall confidence rating considers the likely impact of the noted concerns (i.e., limitations or uncertainties) in reporting, bias and sensitivity on the results. Reviewers should mark studies that are rated lower than high confidence only due to low sensitivity (i.e., bias toward the null) for additional consideration during evidence synthesis. If the study is otherwise well conducted and an effect is observed, it may increase the certainty of evidence judgment. A confidence rating and rationale should be given for each endpoint/outcome or group of endpoints/outcomes investigated in the study. Confidence ratings are described above (see Section 6.1.1).</p>

6.5. IN VITRO AND OTHER MECHANISTIC STUDY EVALUATION

1 As described in Section 4.4, the initial literature screening identifies sets of other potentially
2 informative studies, including mechanistic studies, as “potentially relevant supplemental
3 information.” Mechanistic information includes any experimental measurement related to a health
4 outcome that informs the biological or chemical events associated with phenotypic effects. These
5 measurements can improve understanding of the mechanisms involved in the biological effects
6 following exposure to a chemical but are not generally considered by themselves adverse outcomes.
7 Mechanistic data are reported in a diverse array of observational and experimental studies across
8 species, model systems, and exposure paradigms, including in vitro, in vivo (by various routes of
9 exposure), ex vivo, and in silico studies.

10 Individual study-level evaluations of mechanistic endpoints are not typically pursued. To
11 undergo a full reporting quality, risk of bias, and sensitivity evaluation of every identified study that
12 may report mechanistic information before the relevant toxicity pathways have been identified or
13 the needs of the assessment are better understood would not be an effective use of time. However,
14 for some chemical assessments, it may be necessary to identify assay-specific considerations for
15 study endpoint evaluations, on a case-by-case basis, to provide a more detailed summary and
16 evaluation for the most relevant individual studies. This may be done, for example, when the
17 scientific understanding of a critical mechanistic event or MOA is less established or lacks scientific
18 consensus, when the reported findings on a mechanistic endpoint are conflicting, when the
19 available mechanistic evidence addresses a complex and influential aspect of the assessment, or
20 when in vitro or in silico data make up the bulk of the evidence base and there is little or no
21 evidence from epidemiological studies or animal bioassays.

22 If a subset of individual mechanistic studies is identified for evaluation, the study evaluation
23 considerations will differ depending on the type of endpoints, study designs, and model systems or
24 populations evaluated. Note that because the evaluation process is outcome specific, overall
25 confidence classifications for human or animal studies that have already been determined will not
26 automatically apply to mechanistic endpoints if reported in the same study; instead, a separate
27 evaluation of the mechanistic endpoints should be performed because the utility of a study may
28 vary for the different outcomes reported. Developing specific considerations requires a familiarity
29 with the studies to be evaluated and cannot be conducted in the absence of knowledge of the
30 relevant study designs, measurements, and analytic issues. Knowledge of issues related to the
31 hazards and the outcomes identified in the revised evaluation plan is also important for developing
32 specific evaluation considerations. One challenge is that novel methodologies for studying
33 mechanistic evidence are continually being developed and implemented and often no “standard
34 practices” exist.

35 The evaluation of mechanistic studies applies similar principles as those described above
36 for the evaluation of experimental animal studies. Table 6-6 provides the standard domains and
37 core questions for evaluating studies conducted in in vitro test systems, along with some basic

Protocol for the Ethylbenzene IRIS Assessment

1 considerations for guiding the evaluation. The evaluation process focuses on assessing aspects of
2 the study design and conduct through three broad types of evaluations: reporting quality, risk of
3 bias, and study sensitivity. Some domain considerations are tailored to the chemical, as well as the
4 assay(s) and/or endpoint(s) being evaluated. Assessment teams work with subject-matter experts
5 to develop specific considerations. These specific considerations are determined before performing
6 the study evaluation, although they may be refined as the study evaluation proceeds (e.g., during
7 pilot testing). Assessment-specific and/or assay-specific considerations are documented and made
8 publicly available in the assessment.

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

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

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

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

Domain and core question	Prompting questions	General considerations
<p>Endpoint measurement Are the selected protocols, procedures, and test systems adequately described and appropriate for evaluating the endpoint(s) of interest? <i>Notes:</i> <i>Considerations related to adjustments or corrections to endpoint measurements are addressed under results presentation.</i> <i>Considerations related to the sensitivity of the animal model and timing of endpoint measurement are evaluated under sensitivity.</i></p>	<p>For each endpoint or grouping of endpoints in a study: Are the evaluation methods and test systems adequately described and appropriate? Are there concerns regarding the methodology selected (e.g., accepted guidelines, established criteria) for endpoint evaluation? Are there concerns about the specificity of the experimental design? Did the study address features inherent to the test system or experiment that have the potential to lead to bias away from the null? Are there serious concerns about the number of replicates or sample size in the study? Are appropriate control groups for the study/assay type included? Was there a need for the assay to include specific controls to reduce potential sources of underlying bias? Did the test compound induce cytotoxicity (known, or expected based on other studies of similar design) to a degree that is expected to affect interpretation of results?</p>	<p>Considerations for this domain are highly variable depending on the assay or endpoint(s) of interest and must be refined by assessment teams. A judgment and rationale for this domain should be given for each assay or endpoint or group of endpoints investigated in the study. Some considerations include the following: Good:</p> <ul style="list-style-type: none"> • Adequate description of methods and test system. • Use of generally accepted and reliable endpoint methods that are consistent with accepted guidelines or established criteria for the assay(s)/endpoint(s) of interest. • Sample sizes are generally considered adequate for the assay or protocol of interest and there are no notable concerns about sampling in the context of the endpoint protocol. • Includes appropriate control groups (e.g., use of loading controls) and any use of nonconcurrent or historical control data (e.g., for comparison to background levels in negative controls) is justified (e.g., authors or evaluators considered the similarity between current cell cultures and laboratory conditions to historical controls). <p>Ratings of Adequate, Deficient, and Critically Deficient are generally defined as follows: Adequate: Issues are identified that may affect endpoint measurement but are considered unlikely to substantially impact the overall findings or the ability to reliably interpret those findings. Deficient: Concerns are raised that are expected to notably affect endpoint measurement and reduce the reliability of the study findings Critically deficient: Severe concerns are raised about endpoint measurement and any findings are likely to be largely explained by these limitations The following specific examples of relevant concerns are typically associated with a Deficient rating, but Adequate or Critically Deficient might be applied depending on the expected impact of limitations on the reliability and interpretation of the results:</p>

Domain and core question	Prompting questions	General considerations
		<ul style="list-style-type: none"> • Study report lacks important details that are necessary to evaluate the appropriateness of the study design (e.g., description of the assays or protocols; information on the cell line, passage number). • Selection of protocols that are nonpreferred or lack specificity for investigating the endpoint of interest. This includes omission of additional experimental criteria (e.g., inclusion of a positive control or dosing up to levels causing minimal toxicity) when required by specific testing guidelines/protocols.* • Cytotoxicity is observed or expected based on findings from similarly designed studies and may mask interpretation of outcome(s) of interest. • Sample sizes are smaller than is generally considered adequate for the assay or protocol of interest. Inadequate sampling can also be raised within the context of the endpoint protocol (e.g., in a pathology study, bias that is introduced by only sampling a single tissue depth or an inadequate number of slides per animal)** • Controls are not included or considered inappropriate. <p>*These limitations typically also raise a concern for insensitivity **Sample size alone is not a reason to conclude an individual study is critically deficient.</p>
<p>Results presentation Are the results presented and compared in a way that is appropriate and transparent and makes the data usable?</p>	<p>For each assay/endpoint or grouping of endpoints in a study: Does the level of detail allow for an informed interpretation of the results? If applicable, was the assay signal normalized to account for nonbiological differences across replicates and exposure groups? Are the data compared or presented in a way that is inappropriate or misleading (e.g., presenting western blot images without including numerical values for densitometry analysis, or vice versa)? Flag potentially</p>	<p>Considerations for this domain are highly variable depending on the endpoints of interest and must be refined by assessment teams. A judgment and rationale for this domain should be given for each assay or endpoint or group of endpoints investigated in the study. Some considerations include the following: Good:</p> <ul style="list-style-type: none"> • No concerns with how the data are presented. • Results are quantified or otherwise presented in a manner that allows for an independent consideration of the data (assessments do not rely on author interpretations). • No concerns with completeness of the results reporting.* <p>Ratings of Adequate, Deficient, and Critically Deficient are generally defined as follows:</p>

Domain and core question	Prompting questions	General considerations
	<p>inappropriate statistical comparisons for further review.</p>	<p>Adequate: Concerns are identified that may affect results presentation but are considered unlikely to substantially impact the overall findings or the ability to reliably interpret those findings.</p> <p>Deficient: Concerns with results presentation are identified and expected to substantially impact results interpretation and reduce the reliability of the study findings.</p> <p>Critically deficient: Severe concerns about results presentation were identified and study findings are likely to be largely explained by these limitations.</p> <p>The following specific examples of relevant concerns are typically associated with a Deficient rating but Adequate or Critically Deficient might be applied depending on expected impact of limitations on the reliability and interpretation of the results:</p> <ul style="list-style-type: none"> • Nonpreferred presentation of data (e.g., averaging technical replicates rather than independent replicates). • Failure to present quantitative results • Pooling data when responses are known or expected to differ substantially (e.g., across cell types or passage number). • Incomplete presentation of the data* (e.g., presentation of mean without variance data; concurrent control data are not presented; failure to report or address overt cytotoxicity). <p>*Failure to describe <i>any</i> findings for assessed outcomes (i.e., report lacks any qualitative or quantitative description of the results in tables, figures, or text) will result in a critically deficient rating for the outcome(s) of interest for Results Presentation; overall completeness of reporting at the study level is addressed under Selective Reporting.</p>
<p>Sensitivity Are there concerns that sensitivity in the study is not adequate to detect an effect?</p>	<p>Was the exposure period, timing (i.e., cell passage number, insufficient culture maturity for the adequate expression of mature cell markers; insufficient treatment and/or measurement duration for the production of protein above the level of detection),</p>	<p>Are there concerns regarding the need for positive controls (e.g., concerns that the effects of interest may be inhibited or otherwise poorly manifest in the test system, for example due to differences from in vivo biology)? If used, was the selected positive test substance (and dose) reasonable and appropriate and was the intended positive response induced?</p>

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

Domain and core question	Prompting questions	General considerations
		<ul style="list-style-type: none"> Severe concerns were raised about the sensitivity of the study and experimental design such that any observed associations are likely to be explained by bias. The rationale should indicate the specific concern(s).
<p>Overall confidence Considering the identified strengths and limitations, what is the overall confidence rating for the assay(s) or endpoint(s) of interest? <i>Note:</i> <i>Reviewers should mark studies for additional consideration during evidence synthesis if, due to low sensitivity only (i.e., bias toward the null), these studies are rated as lower than high confidence. If the study is otherwise well conducted and an effect is observed, the confidence may be increased.</i></p>	<p>For each assay or endpoint or grouping of endpoints in a study:</p> <ul style="list-style-type: none"> Were concerns (i.e., limitations or uncertainties) related to the risk of bias or sensitivity identified? If yes, what is their expected impact on the overall interpretation of the reliability and validity of the study results, including (when possible) interpretations of impacts on the magnitude or direction of the reported effects? 	<p>The overall confidence rating considers the likely impact of the noted concerns (i.e., limitations or uncertainties) in reporting, bias and sensitivity on the results. A confidence rating and rationale should be given for each assay or endpoint or group of endpoints investigated in the study. Confidence rating definitions are described above (see Section 4.1).</p>

6.6. PHYSIOLOGICALLY BASED PHARMACOKINETIC (PBPK) MODEL DESCRIPTIVE SUMMARY AND EVALUATION

1 PBPK (or classical pharmacokinetic [PK]) models should be used in an assessment when a
2 validated and applicable one exists and no equal or better alternative for dosimetric extrapolation
3 is available. Any models used should represent current scientific knowledge and accurately
4 translate the science into computational code in a reproducible, transparent manner. For a specific
5 target organ/tissue, it may be possible to employ or adapt an existing PBPK model or develop a new
6 PBPK model or an alternate quantitative approach. Data for PBPK models may come from studies
7 across various species and may be in vitro or in vivo in design. Specific details for this evaluation
8 are provided below and in the Umbrella Quality Assurance Project Plan (QAPP) for dosimetry and
9 mechanism-based models ([U.S. EPA, 2020b](#)) and Umbrella QAPP for PBPK models ([U.S. EPA,
10 2018b](#)).

11 As interspecies difference in ethylbenzene pharmacokinetics have been noted, a major
12 strength of a PBPK model is its capacity to account for physiological, biochemical, and metabolic
13 determinants when extrapolating findings from higher dose animal studies to lower levels of
14 human exposure. Note that a nonlinear ethylbenzene metabolism has been observed, suggesting
15 high-dose saturation of metabolic processes ([Sweeney et al., 2015](#); [Nong et al., 2007](#)). Hence the
16 internal dose responsible for observed toxicity is a nonlinear function of the exposure levels.
17 Therefore, the PBPK model(s) selected for assessing ethylbenzene toxicity should account for this
18 dose saturation as well as reflect the current state of knowledge of toxicological mechanisms or
19 MOA for specific toxicological endpoints when estimating relevant dose metrics ([U.S. EPA, 2018b](#)).

20 Over a dozen scientific publications or reports describing the development or application of
21 PBPK models since 2000 have been identified and will be evaluated for quality and potential use in
22 the assessment. This evaluation will be conducted according to EPA's Umbrella QAPP for Dosimetry
23 and Mechanism-Based Models ([U.S. EPA, 2020b](#)) and Umbrella QAPP for PBPK models ([U.S. EPA,
24 2018b](#)). It may be that none of the existing PBPK models adequately fulfills all of the assessment
25 applications. In this case, a hybrid model could be created which merges elements from the existing
26 models to achieve this objective if needed and feasible under the time constraints for the
27 assessment.

6.6.1. Pharmacokinetic (PK)/Physiologically Based Pharmacokinetic (PBPK) Model Descriptive Summary

28 PBPK modeling is the preferred approach for calculating a human equivalent concentration
29 (HEC) according to the hierarchy of approaches outlined in EPA guidelines ([U.S. EPA, 2020a, 2002](#)).

30 Following literature searches, a stepwise approach is taken that includes conducting an
31 initial scoping of the supplemental material studies categorized as PK/PBPK models. Then, an in-
32 depth full model evaluation is implemented to identify PBPK models that are potentially suitable
33 for deriving toxicity values for the ethylbenzene assessment.

1 The initial scoping process is distinct from the full model evaluation. The scoping process
 2 provides a rapid assessment and communication of the availability, structure, and potential uses of
 3 PBPK/PK models, but is not a full evaluation. Full model evaluation—the complete and thorough
 4 assessment of the quality and utility of a particular model—is conducted if the initial scoping
 5 identifies one or more models that are available and considered appropriate for one or more
 6 applications in the assessment. The model evaluation is then conducted for the selected
 7 application(s). As shown below in Table 6-7, for example, key information from identified PBPK
 8 models during the scoping process is summarized in tabular format for further in-depth model
 9 evaluation following the evaluation approaches summarized in Section 6.6.2.

Table 6-7. Example descriptive summary for a physiologically based pharmacokinetic (PBPK) model study

Study detail	Description/notes				
Author	Smith et al. (2003)				
Contact email	xxxxx@email.com				
Contact phone	xxx-xxx-xxxx				
Sponsor	N/A				
Model summary					
Species	Rat				
Strain	F433				
Sex	Male and female				
Life stage	Adult				
Exposure routes	Inhalation	Oral	I.V.	Skin	
Tissue dosimetry	Blood	Liver	Kidney	Urine	Lung
Model evaluation					
Language	ACSL 11.8				
Code available	YES	Effort to recreate model		COMPLETE	
Code received	YES	Effort to migrate to open software		SIGNIFICANT	
Structure evaluated	YES				
Math evaluated	YES				
Code evaluated	YES. Issue (minor): Incorrect units listed in comments for liver metabolism (line 233). Issue (major): Mass balance error in stomach compartment.				
Available PK data	Urine (cumulative amount excreted) and blood (concentration) time course data for oral (gavage) and inhalation (6 hr/d for 4 d) exposure. In vitro skin permeation.				

6.6.2. Pharmacokinetic (PK)/Physiologically Based Pharmacokinetic (PBPK) Model Evaluation

1 Once available PBPK models are summarized, the assessment team undertakes model
2 evaluation in accordance with criteria outlined by [U.S. EPA \(2018b\)](#). Judgments on the suitability of
3 a model are separated into two categories: scientific and technical (see Table 6-8). The scientific
4 criteria focus on whether the biology, chemistry, and other information available for chemical
5 MOA(s) are justified (i.e., preferably with citations to support use) and represented by the model
6 structure and equations. The scientific criteria are judged based on information presented in the
7 publication or report that describes the model and do not require evaluation of the computer code.
8 Preliminary technical criteria include availability of the computer code and completeness of
9 parameter listing and documentation. Studies that meet the preliminary scientific and technical
10 criteria are then subjected to an in-depth technical evaluation, which includes a thorough review
11 and testing of the computational code. The in-depth technical and scientific analyses focus on the
12 accurate implementation of the conceptual model in the computational code, use of scientifically
13 supported and biologically consistent parameters in the model, and reproducibility of model results
14 reported in journal publications and other documents. This approach stresses (1) clarity in the
15 documentation of model purpose, structure, and biological characterization; (2) validation of
16 mathematical descriptions, parameter values, and computer implementation; and (3) evaluation of
17 each plausible dose metric. The in--depth analysis is used to evaluate the potential value and cost of
18 developing a new model or substantially revising an existing one. PBPK models developed by EPA
19 during the course of the assessment are peer reviewed, either as a component of the draft
20 assessment or by publication in a journal article.

Table 6-8. Criteria for evaluating physiologically based pharmacokinetic (PBPK) models

Category	Specific criteria
Scientific	Biological basis for the model is accurate. <ul style="list-style-type: none"> • Consistent with mechanisms that substantially impact dosimetry. • Predicts dose metric(s) 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. <ul style="list-style-type: none"> • Ability of model to describe critical behavior, such as nonlinear kinetics in a relevant dose range, better than the default (i.e., BW^{3/4} scaling). • Model parameterization for critical life stages or windows of susceptibility. Evaluation of these criteria should also consider the model’s fidelity vs. default approaches and possible use of an intraspecies uncertainty factor (UF_H) in conjunction with the model to account for variations in sensitivity between life stages. • Predictive power of model-based dose metric vs. default approach, based on exposure. <ul style="list-style-type: none"> ○ 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. For example, 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.
	Principle of parsimony <ul style="list-style-type: none"> • 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 EPA and public.
	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 global analysis provides more information). <ul style="list-style-type: none"> • 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.

6.6.3. Selection of the Appropriate Dose Metric

- 1 The level of confidence in using a pharmacokinetic (PK) or PBPK model depends on its
- 2 ability to provide a reliable estimation of dose metrics based on biological plausibility and MOA

Protocol for the Ethylbenzene IRIS Assessment

1 considerations. Thus, one needs to take into consideration mechanism(s) relevant to the
2 endpoint(s) of interest, data availability and uncertainties in estimating that dose metric.
3 Compared to liver and kidney toxicity, it remains less understood what the appropriate dose metric
4 for other toxicities should be, including lung and ototoxicity endpoints. Therefore, various dose
5 metrics (e.g., the area under the curve (AUC) for arterial blood concentration of ethylbenzene or its
6 metabolites) will be explored to inform dose-response extrapolation of animal data to humans.

7. DATA EXTRACTION OF STUDY METHODS AND RESULTS

1 The process of summarizing study methods and results is referred to as data extraction.
2 Studies that met problem formulation PECO criteria after full-text review are briefly summarized in
3 DistillerSR HDE forms. These study summaries are exported from DistillerSR in Excel format and
4 imported into Tableau software (<https://www.tableau.com/>) to create interactive literature
5 inventory visualizations used to display the extent and nature of the available evidence. (see below
6 for studies decisions related to studies meeting the assessment PECO).

7 For experimental animal studies, which are typically studies in rodents, the following
8 information is captured: chemical form, study type (acute [<24 hours], short term [<7 days], short
9 term [7–27 days], subchronic [28–90 days], chronic [>90 days⁵] and developmental, which includes
10 multigeneration studies), duration of treatment, route, species, strain, sex, dose or concentration
11 levels tested, dose units, health system and specific endpoints assessed. Animal studies that meet
12 the assessment PECO undergo a subsequent phase of full data extraction in HAWC that includes
13 detailed presentation of results (described below). For studies that meet problem formulation
14 PECO criteria (but not the assessment PECO) the SEM (initial) literature inventory summary
15 includes the no-observed-effect level/low-observed-effect level (NOEL/LOEL) based on author-
16 reported statistical significance. Expert judgment may be used to identify NOEL/LOELs in cases
17 where only qualitative results are reported (e.g., “no effects on liver weight were observed at any
18 dose level”) or when the findings indicate an apparent clear and strong effect of exposure (e.g.,
19 large magnitude of change) but the authors did not present a statistical comparison. When findings
20 are not analyzed by the authors and are not readily interpretable, then NOEL/LOELs are not
21 identified, and the extraction field entry indicates “not reported.”

22 For human studies, the following information is summarized in DistillerSR HDE forms:
23 chemical form, population type (e.g., general population-adult, occupational, pregnant women,
24 infants and children), study type (e.g., cross-sectional, cohort, case-control), sex, major route of
25 exposure (if known), description of how exposure was assessed, health system studied, specific
26 endpoints assessed and a quantitative summary of findings at the endpoint level (or narrative only
27 if the finding was qualitatively presented). In contrast to the animal studies, epidemiological studies

⁵EPA considers chronic exposure to be more than approximately 10% of the life span in humans. For typical laboratory rodent species, this can lead to consideration of exposure durations of approximately 90 days to 2 years. However, studies in duration of 1–2 years are typical of what is considered representative of chronic exposure rather than durations just over 90 days.

1 that met assessment PECO did not undergo additional more detailed data extraction in HAWC
2 because that module in HAWC was under development at the time of preparation of this protocol.

3 For animal studies that met the assessment PECO criteria, HAWC is used for full extraction
4 of study methods and results. For animal studies, compared with the literature inventory forms
5 used to described studies that meet problem formulation PECO criteria, full data extraction in
6 HAWC includes summarizing more details of study design (e.g., diet, chemical purity) and gathering
7 effect size information. Instructions on how to conduct data extraction in HAWC are available at
8 <https://hawcproject.org/resources/>. Over 100 distinct extraction fields are collected for each
9 animal study and endpoint (for list of data extraction fields, see Downloads > Animal Bioassay Data
10 > Complete Export at the HAWC Ethylbenzene Project
11 <https://hawc.epa.gov/assessment/100000059/>), An additional resource used to implement use of
12 a consistent vocabulary to summarize endpoints assessed in animal studies is available in the
13 HAWC project “[IRIS PPRTV SEM Template Figures and Resources](#)” (see “Attachments,” then select
14 the “Environmental Health Vocabulary (EHV) – a recommended terminology for
15 outcomes/endpoints” file).

16 In some cases, EPA may conduct their own statistical analysis of human and animal
17 toxicology data (assuming the data are amenable to doing so and the study is otherwise well
18 conducted) during evidence synthesis.

19 Data extraction for *in vivo* and *in vitro* studies prioritized to assess key mechanistic analyses
20 is conducted in Microsoft Word and presented in tabular format.

21 All findings are considered for extraction, regardless of statistical significance. The level of
22 extraction for specific outcomes within a study could differ (i.e., narrative only if the finding was
23 qualitative). For quality control, studies were summarized by one member of the evaluation team
24 and independently verified by at least one other member. Discrepancies were resolved by
25 discussion or consultation within the evaluation team. Data extraction results are presented via
26 figures, tables, or interactive web-based graphics in the assessment. The information is also made
27 available for download in Excel format when the draft is publicly released. The literature
28 inventories are presented in the HAWC Visualization module, with options to link to the native
29 Tableau application where the underlying information is available for download. Download of full
30 data extraction for animal studies is done directly in HAWC.

31 For non-English studies online translation tools (e.g., Google translator) or engagement with
32 a native speaker can be used to summarize studies at the level of the literature inventory. Fee-based
33 translation services for non-English studies are typically reserved for studies considered potentially
34 informative for dose response, a consideration that occurs after preparation of the initial literature
35 inventory during draft assessment development. Digital rulers, such as WebPlotDigitizer
36 (<https://automeris.io/WebPlotDigitizer/>), are used to extract numerical information from figures,
37 and their use is be documented during extraction. For studies that evaluate endpoints at multiple
38 time points (e.g., 7 days, 3 weeks, 3 months) data are generally summarized for the longest duration

1 in the study report, but other durations may be summarized if they provide important contextual
2 information for hazard characterization (e.g., an effect was present at an interim time point but did
3 not appear to persist or the magnitude of the effect diminished). A free text field is available in
4 HAWC to describe cases when the approach for summarizing results requires explanation.

5 Author queries may be conducted for studies considered for dose-response to facilitate
6 quantitative analysis (e.g., information on variability or availability of individual animal data).
7 Outreach to study authors or designated contact persons is documented and considered
8 unsuccessful if researchers do not respond to email or phone requests within 1 month of initial
9 attempt(s) to contact. Only information or data that can be made publicly available (e.g., within
10 HAWC or HERO) will be considered.

11 Exposures are standardized to common units when possible. For hazard characterization,
12 exposure levels are typically presented as reported in the study and standardized to common units
13 (e.g., ppm or mg/m³ for inhalation studies) as an initial phase in evidence synthesis and integration.
14 For inhalation exposures to ethylbenzene, concentration in air in ppm can be converted to
15 concentration in air in mg/m³ by multiplying ppm times 4.344 (106.2 g/mol ÷ 24.45 L) at standard
16 temperature (25°C) and pressure (1 atm).

8. EVIDENCE SYNTHESIS AND INTEGRATION

1 Evidence synthesis⁶ is a within-stream analysis, conducted separately for human, animal,
 2 and mechanistic evidence. Findings from human and animal evidence for each unit of analysis are
 3 separately judged to reach an expression of certainty in the evidence for a hazard (*robust, moderate,*
 4 *slight, indeterminate, or compelling evidence of no effect*). Within-stream evidence synthesis
 5 conclusions directly inform the integration across the evidence streams to draw overall conclusions
 6 for each of the assessed health effect categories (*evidence demonstrates, evidence indicates, evidence*
 7 *suggests, evidence inadequate, or strong evidence supports no effect*). A structured framework
 8 approach is used to guide both evidence synthesis and integration. While there are circumstances
 9 where specific mechanistic evidence (typically biological precursors) is included in the unit of
 10 analysis for human or animal evidence synthesis, in most cases mechanistic findings are presented
 11 separately from the human and animal evidence and used to inform conclusions on (1) the
 12 coherence, directness of outcome measures, and biological significance of findings within the
 13 animal or human evidence streams during evidence synthesis and, (2) evidence integration
 14 judgments on the human relevance of findings in animals, coherence across evidence streams
 15 (“cross-stream coherence”), information on susceptible populations or lifestyles, understanding of
 16 biological plausibility and MOA, and possibly other critical inferences (e.g., read-across analyses).
 17 The structured framework also accommodates consideration of supplemental information (e.g.,
 18 ADME, non-PECO route of exposure) that can inform evidence synthesis and integration judgments.

- 19 • Evidence synthesis: A summary of findings and judgment(s) regarding the certainty in the
 20 evidence for hazard for each unit of analysis from the human and animal studies are made
 21 in parallel, but separately. A unit of analysis is an outcome or group of related outcomes
 22 within a health effect category that are considered together during evidence synthesis.
 23 These judgments can incorporate mechanistic and other supplemental evidence when the
 24 unit of analysis is defined as such (see Section 3). The units of analysis can also include or be
 25 framed to focus on precursor events (e.g., biomarkers). In addition, this can include an
 26 evaluation of coherence across units of analysis within an evidence stream. At this stage, the
 27 animal evidence judgment(s) does not yet consider the human relevance of that evidence.
- 28 • Evidence integration: The animal and human evidence judgments are combined to draw an
 29 overall evidence integration judgment(s) that incorporates inferences drawn based on
 30 information on the human relevance of the animal evidence, coherence across evidence

⁶The phrases “evidence synthesis” and “evidence integration” used here are analogous to the phrases “strength of evidence” and “weight of evidence,” respectively, used in some other assessment processes ([EFSA, 2017](#); [U.S. EPA, 2017a](#); [NRC, 2014](#); [U.S. EPA, 2005a](#)).

1 streams, potential susceptibility, understanding of biological plausibility and MOA and other
2 critical inferences informed by mechanistic, ADME, or other supplemental data.

3 Evidence synthesis and integration judgments are expressed both narratively in the
4 assessment and summarized in tabular format in evidence profile tables (see Table 8-1). Key
5 findings and analyses of mechanistic and other supplemental content are also summarized in
6 narrative and tabular format to inform evidence synthesis and integration judgments (see
7 Table 8-2). In brief, after synthesis a certainty in the evidence judgment is drawn for each unit of
8 analysis summarized as *robust, moderate, slight, indeterminate, or compelling evidence of no effect*
9 (see Section 8.1). Next, these judgments are used to inform evidence integration judgments
10 summarized as ***evidence demonstrates, evidence indicates, evidence suggests, evidence***
11 ***inadequate, or strong evidence supports no effect*** (see Section 8.2). These summary judgments
12 are included as part of the evidence synthesis and integration narratives. When multiple units of
13 analysis are synthesized, the main evidence integration judgments typically focus on the unit of
14 analysis with the strongest evidence synthesis judgments, although exceptions may occur.⁷ Health
15 outcomes or endpoints where the unit of analysis is considered to present *slight, indeterminate* or
16 *compelling evidence of no effect* can inform the evidence integration hazard judgment but would
17 typically not be used as the basis for deriving a toxicity value. Structured evidence profile tables are
18 used to summarize these analyses and foster consistency within and across assessments.
19 Instructions for using HAWC to create these tables are available at the HAWC project "[IRIS PPRTV](#)
20 [SEM Template Figures and Resources](#)" (see "Attachments," then select the "Creating Evidence
21 Profile Tables in HAWC").

⁷In some cases, it may be appropriate to draw multiple evidence integration judgments within a given health effect category. This is generally dependent on data availability (i.e., more narrowly defined categories may be possible with more evidence) and the ability to integrate the different evidence streams at the level of these more granular categories. More granular categories will generally be organized by pre-defined manifestations of potential toxicity. For example, within the health effect category of immune effects, separate and different evidence integration judgments might be appropriate for immunosuppression, immunostimulation, and sensitization and allergic response (i.e., the three types of immunotoxicity described in the [IPCS \(2012\)](#)). Likewise, within the category of developmental effects, it may be appropriate to draw separate judgments for potential effects on fetal death, structural abnormality, altered growth, and functional deficits (i.e., the four manifestations of developmental toxicity described in EPA guidelines ([U.S. EPA, 1991a](#))). These separate judgments are particularly important when the evidence supports that the different manifestations might be based on different toxicological mechanisms. As described for the evidence synthesis judgments, the strongest evidence integration judgment will typically be used to reflect certainty in the broader health effect category.

Table 8-1. Generalized evidence profile table to show the relationship between evidence synthesis and evidence integration to reach judgment of the evidence for hazard

Evidence synthesis (certainty of evidence) judgments (note that many factors and judgments require elaboration or evidence-based justification; see IRIS Handbook for details)					Evidence integration (weight of evidence) judgment(s)
Studies	Summary of key findings	Factors that increase certainty (applied to each unit of analysis)	Factors that decrease certainty (applied to each unit of analysis)	Evidence synthesis judgment(s)	Describe overall evidence integration judgment(s): ⊕⊕⊕ Evidence demonstrates ⊕⊕⊖ Evidence indicates (likely) ⊕⊖⊖ Evidence suggests ⊖⊖⊖ Evidence inadequate --- Strong evidence supports no effect Highlight the primary supporting evidence for each integration judgment* Present inferences and conclusions on:
Evidence from human studies					
Unit of analysis #1 Studies considered and study confidence	Description of the primary results	<ul style="list-style-type: none"> All/Mostly <i>medium</i> or <i>high</i> confidence studies Consistency Dose-response gradient Large or concerning magnitude of effect Coherence* 	<ul style="list-style-type: none"> All/Mostly <i>low</i> confidence studies Unexplained inconsistency Imprecision Concerns about biological significance* Indirect outcome measures* Lack of expected coherence* 	Judgment reached for each unit of analysis* ⊕⊕⊕ <i>Robust</i> ⊕⊕⊖ <i>Moderate</i> ⊕⊖⊖ <i>Slight</i> ⊖⊖⊖ <i>Indeterminate</i> --- <i>Compelling evidence of no effect</i>	<ul style="list-style-type: none"> Human relevance of findings in animals* Cross-stream coherence* Potential susceptibility* Biological plausibility* Other critical inferences (e.g., from ADME or other supplemental information)*
Evidence from animal studies					
Unit of analysis #1 Studies considered and study confidence	Description of the primary results	<ul style="list-style-type: none"> All/Mostly <i>medium</i> or <i>high</i> confidence studies Consistency Dose-response gradient Large or concerning magnitude of effect Coherence* 	<ul style="list-style-type: none"> All/Mostly <i>low</i> confidence studies Unexplained inconsistency Imprecision Concerns about biological significance* Indirect outcome measures* Lack of expected coherence* 	Judgment reached for each unit of analysis ⊕⊕⊕ <i>Robust</i> ⊕⊕⊖ <i>Moderate</i> ⊕⊖⊖ <i>Slight</i> ⊖⊖⊖ <i>Indeterminate</i> --- <i>Compelling evidence of no effect</i>	(This section is merged with the 'Evidence from human studies' row in the original table)

*Can be informed by key findings from the mechanistic analyses (see Table 8-2).

Table 8-2. Generalized evidence profile table to show the key findings and supporting rationale from mechanistic analyses.

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

1

8.1. EVIDENCE SYNTHESIS

1 IRIS assessments synthesize the evidence separately for each unit of analysis by focusing on
 2 factors that increase or decrease certainty in the reported findings (see Table 8-1). These factors
 3 are adapted from considerations for causality introduced by Austin Bradford Hill ([Hill, 1965](#)) with
 4 some expansion and adaptation of how they are applied to facilitate transparent application to
 5 chemical assessments that consider multiple streams of evidence. Specifically, the factors
 6 considered are confidence in study findings (risk of bias and sensitivity), consistency across studies
 7 or experiments, dose-/exposure-response gradient, strength (effect magnitude) of the association,
 8 directness of outcome or endpoint measures, and coherence [Table 8-3; see additional discussion in
 9 U.S. EPA ([2005a](#)), U.S. EPA ([1994](#)), and U.S. EPA ([2020a](#))]. These factors are similar to the domains
 10 considered in the GRADE Quality of Evidence framework ([Schünemann et al., 2013](#)). Each of the
 11 considered factors and the certainty of evidence judgments require elaboration or evidence-based
 12 justification in the synthesis narrative. Analysis of evidence synthesis considerations is qualitative
 13 (i.e., numerical scores are not developed, summed, or subtracted).

14 Biological understanding (e.g., knowledge of how an effect manifests or progresses) or
 15 mechanistic inference (e.g., dependency on a conserved key event across outcomes) can be used to
 16 define which related outcomes are considered as a unit of analysis. The units of analysis may also
 17 include predefined categories of mechanistic evidence (typically precursor events). When
 18 mechanistic evidence is included in the units of analysis, it is evaluated against all evidence
 19 synthesis factors. Mechanistic and other supplemental evidence not included in the units of analysis
 20 can be analyzed to inform select evidence synthesis factors (i.e., coherence, directness of outcome
 21 measures, or biological significance) within the animal and human evidence synthesis. Additional
 22 mechanistic evaluations (e.g., biological plausibility) as considered as part of across stream
 23 evidence integration (see Section 8.2).

24 Five levels of certainty in the evidence for a hazard are used to summarize evidence
 25 synthesis judgments: robust ($\oplus\oplus\oplus$, very little uncertainty exists), moderate ($\oplus\oplus\ominus$, some
 26 uncertainty exists), slight ($\oplus\ominus\ominus$, large uncertainty exists), indeterminate ($\ominus\ominus\ominus$), or compelling
 27 evidence of no effect (- - -, little to no uncertainty exists for lack of hazard) (see Tables 8.4 and 8.5
 28 for descriptions). Conceptually, before the evidence synthesis framework is applied, certainty in the
 29 evidence is neutral (i.e., functionally equivalent to indeterminate). Next, the level of certainty
 30 regarding the evidence for (or against) hazard is increased or decreased depending on
 31 interpretations using the factors described in Table 8-3. Level of certainty analyses are conducted
 32 for each unit of analysis within an evidence stream. Observations that increase certainty are having
 33 an evidence base exhibiting a signal of an effect on the health outcome based on evaluation of
 34 consistency across studies or experiments, the presence of a dose or exposure-response gradient,
 35 observing a large or concerning magnitude of effect, and coherent findings for closely related
 36 endpoints (can include mechanistic endpoints). These patterns are more compelling when
 37 observed among high or medium confidence studies. Observations that decrease certainty are

Protocol for the Ethylbenzene IRIS Assessment

1 having an evidence base of mostly low confidence studies, unexplained inconsistency, imprecision,
2 concerns about biological significance, indirect measures of outcomes, and lack of expected
3 coherence. Study sensitivity considerations can be expressed as a factor that can either increase or
4 decrease certainty in the evidence, depending on whether an association is observed. An evidence
5 base of mostly null findings where insensitivity is a serious concern decreases certainty that the
6 evidence is sufficient to support a lack of health effect or association. Conversely, there may be an
7 increase in the evidence certainty in cases where an association is observed although the expected
8 impact of study sensitivity is toward the null.

Table 8-3. Considerations that inform judgments of the certainty of the evidence for hazard for each unit of analysis

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

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

Consideration	Increased evidence certainty (of the human or animal evidence for hazard ^a)	Decreased evidence certainty (of the human or animal evidence for hazard ^a)
Directness of outcome/endpoint measures	<ul style="list-style-type: none"> • Not applicable 	<ul style="list-style-type: none"> • If the evidence base primarily includes outcomes or endpoints that are indirect measures (e.g., biomarkers) of the unit of analysis, certainty (for that unit of analysis) is typically decreased. Judgments to decrease certainty based on indirectness should focus on findings that have an unclear linkage to an apical or clinical (adverse^b) outcome. Scenarios where the magnitude of the response is not considered to reflect a biologically meaningful level of change (i.e., biological significance; see ‘effect magnitude and imprecision’ row above) are not considered under indirectness. • Related to indirectness, certainty in the evidence may be decreased when the findings are determined to be nonspecific to the hazard under evaluation. This consideration is generally only applicable to animal evidence and the most common example is effects only with exposures (level, duration) shown to cause excessive toxicity in that species and lifestage (including consideration of maternal toxicity in developmental evaluations). This does not apply when an effect is viewed as secondary to other changes (e.g., effects on pulmonary function because of disrupted immune responses).
Coherence	<ul style="list-style-type: none"> • Biologically related findings within or across studies, within an organ system or across populations (e.g., sex), increase certainty (generally only when observed in <i>medium</i> or <i>high</i> confidence studies). Certainty is further increased when a temporal or dose-dependent progression of related effects is observed within or across studies, or when related findings of increasing severity are observed with increasing exposure. • Coherence across findings within a unit of analysis (e.g., consistent changes in disease markers and biological precursors in exposed humans) can increase certainty in the evidence for an effect. • Coherence within or across biologically related units of analysis can also increase certainty for a given (or multiple) unit(s) of analysis. This considers certainty in the biological relationships between the endpoints 	<ul style="list-style-type: none"> • An observed lack of expected coherent changes (e.g., in well-established biological relationships) within or across biologically related units of analysis typically decrease evidence certainty. This includes mechanistic changes when included in the unit of analysis. However, as described for decisions to increase certainty in the biological relationships between the endpoints being compared, and the sensitivity and specificity of the measures used, need to be carefully examined. The decision to decrease depends on the availability of evidence across multiple related endpoints for which changes would be anticipated, and it considers factors (e.g., dose and duration of exposure, strength of expected relationship) across the studies of related changes.

Consideration	Increased evidence certainty (of the human or animal evidence for hazard ^a)	Decreased evidence certainty (of the human or animal evidence for hazard ^a)
	being compared, and the sensitivity and specificity of the measures used. <ul style="list-style-type: none"> • Mechanistic support for, or biological understanding of, the relatedness between different endpoints within (or across different) units of analysis, can inform an understanding of coherence. 	
Other factors	<ul style="list-style-type: none"> • Unusual scenarios that cannot be addressed by the considerations above, e.g., read-across inferences supporting the adversity of observed changes. 	<ul style="list-style-type: none"> • Unusual scenarios that cannot be addressed by the considerations above, e.g., strong evidence of publication bias.^c

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

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

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

1 A structured framework approach is used to draw evidence synthesis judgments for human
 2 and animal evidence. Tables 8-4 and 8-5 (for human and animal evidence, respectively) provide the
 3 example-based criteria that guide how to draw the certainty of evidence judgments for each unit of
 4 analysis within a health effect category and the terms used to summarize those judgments. These
 5 terms are applied to human and animal evidence separately. The terms *robust* and *moderate* are
 6 characterizations for judgments that the evidence (across studies) supports that the effect(s)
 7 results from the exposure being assessed. These two terms are differentiated by the quality and
 8 amount of information available to rule out alternative explanations for the results. For example,
 9 repeated observations of effects by independent studies or experiments examining various aspects
 10 of exposure or response (e.g., different exposure settings, dose levels or patterns, populations or
 11 species, biologically related endpoints) result in a stronger certainty of evidence judgment. The
 12 term *slight* indicates situations in which there is some evidence supporting an association within
 13 the evidence stream, but substantial uncertainties in the data exist to prevent judgments that the
 14 effect(s) can be reliably attributed to the exposure being assessed. *Indeterminate* reflects judgments
 15 for a wide variety of evidence scenarios, including when no studies are available or when the
 16 evidence from studies of similar confidence has a high degree of unexplained inconsistency.
 17 *Compelling evidence of no effect* represents a rare situation in which extensive evidence across a
 18 range of populations and exposures has demonstrated that no effects are likely to be attributable to
 19 the exposure being assessed. This category is applied at the health effect level (e.g., hepatic effects)
 20 rather than more granular units of analysis level to avoid giving the impression of confidence in
 21 lack of a health effect when aspects of potential toxicity have not been adequately examined.
 22 Reaching this judgment is infrequent because it requires both a high degree of confidence in the
 23 conduct of individual studies, including consideration of study sensitivity, as well as comprehensive
 24 assessments of outcomes and lifestages of exposure that adequately address concern for the hazard
 25 under evaluation.

Table 8-4. Framework for evidence synthesis judgments from studies in humans

Evidence synthesis judgment	Description
<p><i>Robust</i> (⊕⊕⊕) ...evidence in human studies <i>(strong signal of effect with very little uncertainty)</i></p>	<p>A set of <i>high</i> or <i>medium</i> confidence independent studies (e.g., in different populations) reporting an association between the exposure and the health outcome(s), with reasonable confidence that alternative explanations, including chance, bias, and confounding, can be ruled out across studies. The set of studies is primarily consistent, with reasonable explanations when results differ; the findings are considered adverse (i.e., biologically significant and without notable concern for indirectness); and an exposure-response gradient is demonstrated. Additional supporting evidence, such as associations with biologically related endpoints in human studies (coherence) or large estimates of risk or severity of the response, can increase confidence but are not required. Supplemental evidence included in the unit of analysis (e.g., mechanistic studies in exposed humans or human cells) may raise</p>

Protocol for the Ethylbenzene IRIS Assessment

Evidence synthesis judgment	Description
	the certainty of evidence to robust for a set of studies that otherwise would be described as moderate. Such evidence not included in the unit of analysis can also inform evaluations of the coherence of the human evidence, the directness of the outcome measures, and the biological significance of the findings. Causality is inferred for a human evidence base of robust.
<p><i>Moderate</i> (⊕⊕⊖) ...evidence in human studies (<i>signal of effect with some uncertainty</i>)</p>	<p>A set of evidence that does not reach the degree of certainty required for <i>Robust</i>, but which includes at least one <i>high</i> or <i>medium</i> confidence study reporting an association and additional information increasing the certainty of evidence. For multiple studies, there is primarily consistent evidence of an association with reasonable support for adversity, but there may be some uncertainty due to potential chance, bias, or confounding or because of the indirectness of some measures.</p> <p>For a single study, there is a large magnitude or severity of the effect, or a dose-response gradient, or other supporting evidence, and there are no serious residual methodological uncertainties. Supporting evidence could include associations with related endpoints, including mechanistic evidence from exposed humans when included within the unit of analysis.</p> <p>When available and included in the unit of analysis, mechanistic data in humans that address the above considerations may raise the certainty of evidence to <i>Moderate</i> for a set of studies that otherwise would be described as <i>Slight</i>. In exceptional cases, biological support from mechanistic evidence in exposed humans may support raising the certainty of evidence to <i>Moderate</i> for evidence that would otherwise be described as <i>Indeterminate</i>.</p>
<p><i>Slight</i> (⊕⊖⊖) ...evidence in human studies (<i>signal of effect with large amount of uncertainty</i>)</p>	<p>One or more studies reporting an association between exposure and the health outcome, but considerable uncertainty exists and supporting coherent evidence is sparse. In general, the evidence is limited to a set of consistent <i>low</i> confidence studies, or higher confidence studies with significant unexplained heterogeneity or other serious residual uncertainties. It also applies when one <i>medium</i> or <i>high</i> confidence study is available without additional information strengthening the likelihood of a causal association (e.g., coherent findings within the same study or from other studies). This category serves primarily to encourage additional study where evidence does exist that might provide some support for an association, but for which the evidence does not reach the degree of confidence required for moderate.</p>
<p><i>Indeterminate</i> (⊖⊖⊖) ...evidence in human studies (<i>signal cannot be determined for or against an effect</i>)</p>	<p>No studies available in humans or situations when the evidence is inconsistent and primarily of <i>low</i> confidence. In addition, this may include situations where higher confidence studies exist, but there are major concerns with the evidence base such as unexplained inconsistency, a lack of expected coherence from a stronger set of studies, very small effect magnitude (i.e., major concerns about biological significance), or uncertainties or methodological limitations that result in an inability to discern effects from exposure. It also applies for a single <i>low</i> confidence study in the absence of factors that increase certainty. A set of largely null studies could be concluded to be <i>Indeterminate</i> if the evidence does not reach the level required for <i>Compelling evidence of no effect</i>.</p>

Evidence synthesis judgment	Description
<p><i>Compelling evidence of no effect</i> (- - -) ...in human studies</p> <p><i>(strong signal for lack of an effect with little uncertainty)</i></p>	<p>A set of <i>high</i> confidence studies examining a reasonable spectrum of endpoints showing null results (for example, an odds ratio of 1.0), ruling out alternative explanations including chance, bias, and confounding) with reasonable confidence. Each of the studies should have used an optimal outcome and exposure assessment and adequate sample size (specifically for higher exposure groups and for susceptible populations). The set as a whole should include diverse sampling (across sexes [if applicable] and different populations) and include the full range of levels of exposures that human beings are known to encounter, an evaluation of an exposure-response gradient, and an examination of at-risk populations and lifestages.</p> <p>Mechanistic data in humans that address the above considerations or that provide information supporting the lack of an association between exposure and effect with reasonable confidence may provide additional support for this judgment.</p>

Table 8-5. Framework for evidence synthesis judgments from studies in animals

Evidence synthesis judgment	Description
<p><i>Robust</i> (⊕⊕⊕) ...evidence in animal studies</p> <p><i>(strong signal of effect with very little uncertainty)</i></p>	<p>The set of <i>high</i> or <i>medium</i> confidence, independent experiments (i.e., across laboratories, exposure routes, experimental designs [for example, a subchronic study and a multigenerational study], or species) reporting effects of exposure on the health outcome(s). The set of studies is primarily consistent, with reasonable explanations when results differ (i.e., due to differences in study design, exposure level, animal model, or study confidence), and the findings are considered adverse (i.e., biologically significant and without notable concern for indirectness).</p> <p>At least two of the following additional factors in the set of experiments increase the certainty of evidence: coherent effects across multiple related endpoints (within or across biologically related units of analysis and may include mechanistic endpoints); an unusual magnitude of effect, rarity, age at onset, or severity; a strong dose-response relationship; or consistent observations across animal lifestages, sexes, or strains. Mechanistic evidence from animals included in the unit of analysis or used to assess coherence of findings in the animal evidence may raise the certainty of evidence to <i>robust</i> for a set of studies that otherwise would be described as <i>moderate</i>.</p>

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

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

8.2. EVIDENCE INTEGRATION

1 The phase of evidence integration combines animal and human evidence synthesis
2 judgments while also considering information on the human relevance of findings in animal
3 evidence, coherence across evidence streams (“cross-stream coherence”), information on
4 susceptible populations or lifestyles, understanding of biological plausibility and MOA, and
5 possibly other critical inferences (e.g., read-across analyses) that generally draw on mechanistic
6 and other supplemental evidence (see Table 8-6). This analysis culminates in an evidence
7 integration judgment and narrative for each potential health effect (i.e., each noncancer health
8 effect and specific type of cancer, or broader grouping of related outcomes as defined in the
9 evaluation plan). To the extent it can be characterized prior to conducting dose-response analyses,
10 exposure context is provided.

Table 8-6. Considerations that inform evidence integration judgments

Judgment	Description
<p>Human relevance of findings</p>	<ul style="list-style-type: none"> • Used to describe and justify the interpretation of the relevance of the animal data to humans. This can include consideration of mechanistic or other supplemental information. When human evidence is lacking or has results that differ from animals, analyses of the mechanisms underlying the animal response in relation to those presumed to operate in humans, and the chemical’s pharmacokinetics, can inform the extent to which the animal response is likely to be relevant to humans and potentially strengthen overall confidence in the evidence integration conclusion. Conversely, evidence for a mechanistic pathway that is expected to only occur in animals and not in humans can provide support for a conclusion that the animal evidence for an effect is not relevant to humans. • In the absence of chemical-specific evidence informing human relevance, the evidence integration narrative will briefly describe the interpreted comparability of experimental animal organs/systems to humans based on underlying biological similarity (e.g., thyroid signaling processes are well conserved across rodents and humans). Generally, a high-

Protocol for the Ethylbenzene IRIS Assessment

Judgment	Description
	<p>level systems summary should be possible for most encountered effects. In some cases, however, it may be appropriate to use a statement such as, ‘without evidence to the contrary, [health effect described in the table] responses in animals are presumed to be relevant to humans.’ As noted in EPA guidelines (U.S. EPA, 2005a), there needs to be evidence or a biological explanation to support an interpreted lack of human relevance for findings in animals, and site concordance is neither expected nor required.</p>
Cross-stream coherence	<ul style="list-style-type: none"> • Addresses the concordance of findings known to be biologically related across human, animal, and mechanistic studies, considering factors such as exposure timing and levels. Notably, for many health effects (e.g., some nervous system and reproductive effects; cancer), it is not necessary (or expected) that effects manifest in humans are identical to those observed in animals, although this typically provides stronger evidence. For example, tumors in one animal species can be predictive of carcinogenic potential in humans or other species, but not necessarily at the same site. EPA guidelines and other resources (e.g., OECD guidelines) are consulted when drawing these inferences. • Mechanistic support for, or biological understanding of, the relatedness between different outcomes (and the manner in which they are manifest) in different species can inform an understanding of coherence across evidence streams. Evidence supporting a biologically plausible mechanistic pathway across species adds coherence (see below).
Potential susceptibility Susceptible populations and lifestages	<ul style="list-style-type: none"> • Used to summarize analyses relating to individual and social factors that may increase susceptibility to exposure-related health effects in certain populations or lifestages, or to highlight the lack of such information. These analyses are based on knowledge about the health outcome or organ system affected and focus primarily on the influence of intrinsic biological factors such as race/ethnicity, genetic variability, sex, lifestage, and pre-existing health conditions (which can also have an extrinsic basis). Information on extrinsic factors potentially influencing susceptibility (e.g., proximity to exposure; certain lifestyle factors including subsistence living) are not considered in evidence integration judgments on potential susceptibility; these exposure-focused factors are considered by risk managers after the human health assessment is complete. Evaluation of potential susceptibility can also include consideration of mechanistic and ADME evidence.
Biological plausibility or MOA understanding	<ul style="list-style-type: none"> • Support for the biological plausibility of an association between exposure and the health effect increases evidence certainty, particularly when observed across species. This may be provided by data from experimental studies of mechanistic pathways, particularly when support is provided for key events or is conserved across multiple components of the pathway. Mechanisms or biological changes with broad scientific acceptance for their relevance to chemical toxicity or the health effect (e.g., key characteristics, hallmarks of cancer) may be used to organize the chemical-specific evidence and identify key events leading from exposure to the health effect. For each key event and key event relationship, the evidence is considered regarding the consistency of experimental data and the generalizability, or likelihood of similarities (e.g., in presence or function) across species, as well as the strength of the support for the biological mechanism. • Mechanistic evidence from well conducted studies that demonstrates that the health effect is unlikely to occur (i.e., species-specific effects, irrelevant exposure conditions) can support a judgment that the effects from animal or human studies are not biologically relevant, which weakens the summary evidence integration judgment. Such a decision depends on an evaluation of the certainty of the information supporting vs. opposing biological plausibility, as well as the certainty of the health effect specific findings (e.g., stronger health effect data require more certainty in mechanistic evidence opposing

Judgment	Description
	<p>plausibility). Importantly, because understanding biological plausibility is dependent on expert knowledge and canonical scientific knowledge, the lack of such understanding does not provide a rationale to decrease the certainty of the evidence for an effect (NTP, 2015; NRC, 2014).</p> <ul style="list-style-type: none"> • These analyses are typically conducted separately to establish MOA understanding and referenced in the evidence integration judgment. If sufficiently supported, MOA understanding can serve to increase (e.g., strong support for mutagenicity) or increase (e.g., critical dependence on a key event not likely to be operant in humans) certainty in the evidence integration judgments.
Other critical inferences (optional)	<ul style="list-style-type: none"> • Consideration of other evidence or nonchemical-specific information that informs evidence integration judgments (e.g., read-across analyses, ADME understanding used to inform other considerations; judgments on other health effects expected to be linked to the health effect under evaluation; read-across analyses or inferences) may be separately described as “other critical inferences.”

1 Using a structured framework approach, one of five phrases is used to summarize the
2 evidence integration judgment based on the within evidence stream integration of the human and
3 animal evidence, and supplemental (mechanistic) evidence: evidence demonstrates, evidence
4 indicates, evidence suggests, evidence is inadequate, or strong evidence supports no effect (see
5 Table 8-7). The five integration judgment levels reflect the differences in the amount and quality of
6 the data that inform the evaluation of whether exposure may cause the health effect(s). As it is
7 assumed that any identified health hazards will only be manifest given exposures of a certain type
8 and amount (e.g., a specific route; a minimal duration, periodicity, and level), the evidence
9 integration narrative and summary judgment levels include the generic phrase, “given sufficient
10 exposure conditions.” This highlights that, for those assessment-specific health effects identified as
11 potential hazards, the exposure conditions associated with those health effects will be defined (as
12 will the uncertainties in the ability to define those conditions) during dose-response analysis. More
13 than one descriptor can be used when the evidence base is able to support that a chemical’s effects
14 differ by exposure level or route ([U.S. EPA, 2005a](#)). The analyses and judgments are summarized in
15 the evidence profile table (see Table 8-1).

Table 8-7. Framework for summary evidence integration judgments in the evidence integration narrative

Summary evidence integration judgment ^a in narrative	Evidence integration judgment level	Explanation and example scenarios ^b
The currently available evidence demonstrates that [chemical] causes [health effect] in humans ^c given sufficient exposure conditions. This conclusion is based on studies of [humans or animals] that assessed [exposure or dose] levels of [range of concentrations or specific cutoff level concentration ^d].	Evidence demonstrates	A strong evidence base demonstrating that [chemical] exposure causes [health effect] in humans. <ul style="list-style-type: none"> This conclusion level <u>is</u> used if there is <i>robust</i> human evidence supporting an effect. This conclusion level <u>could also be</u> used with <i>moderate</i> human evidence and <i>robust</i> animal evidence if there is strong mechanistic evidence that MOAs and key precursors identified in animals are anticipated to occur and progress in humans.
The currently available evidence indicates that [chemical] likely causes [health effect] in humans given sufficient exposure conditions. This conclusion is based on studies of [humans or animals] that assessed [exposure or dose] levels of [range of concentrations or specific cutoff level concentration].	Evidence indicates (likely^e)	An evidence base that indicates that [chemical] exposure likely causes [health effect] in humans, although there may be outstanding questions or limitations that remain, and the evidence is insufficient for the higher conclusion level. <ul style="list-style-type: none"> This conclusion level <u>is</u> used if there is <i>robust</i> animal evidence supporting an effect and <i>slight-to-indeterminate</i> human evidence, or with <i>moderate</i> human evidence when strong mechanistic evidence is lacking. This conclusion level <u>could also be</u> used with <i>moderate</i> human evidence supporting an effect and <i>moderate-to-indeterminate</i> animal evidence, or with <i>moderate</i> animal evidence supporting an effect and <i>moderate-to-indeterminate</i> human evidence. In these scenarios, any uncertainties in the <i>moderate</i> evidence are not sufficient to substantially reduce confidence in the reliability of the evidence, or mechanistic evidence in the <i>slight</i> or <i>indeterminate</i> evidence base (e.g., precursors) exists to increase confidence in the reliability of the <i>moderate</i> evidence.
The currently available evidence suggests that [chemical] may cause [health effect] in humans This conclusion is based on studies of [humans or animals] that assessed [exposure or dose] levels of [range of concentrations or specific cutoff level concentration].	Evidence suggests	An evidence base that suggests that [chemical] exposure may cause [health effect] in humans, but there are very few studies that contributed to the evaluation, the evidence is very weak or conflicting, and/or the methodological conduct of the studies is poor. <ul style="list-style-type: none"> This conclusion level <u>is</u> used if there is <i>slight</i> human evidence and <i>indeterminate-to-slight</i> animal evidence. This conclusion level <u>is</u> also used with <i>slight</i> animal evidence and <i>indeterminate-to-slight</i> human evidence. This conclusion level <u>could also be</u> used with <i>moderate</i> human evidence and <i>slight</i> or <i>indeterminate</i> animal evidence, or with <i>moderate</i> animal evidence and <i>slight</i>

Summary evidence integration judgment ^a in narrative	Evidence integration judgment level	Explanation and example scenarios ^b
		<p>or <i>indeterminate</i> human evidence. In these scenarios, there are outstanding issues or uncertainties regarding the <i>moderate</i> evidence (i.e., the synthesis judgment was borderline with <i>slight</i>), or mechanistic evidence in the <i>slight</i> or <i>indeterminate</i> evidence base (e.g., null results in well-conducted evaluations of precursors) exists to decrease confidence in the reliability of the <i>moderate</i> evidence.</p> <ul style="list-style-type: none"> • Exceptionally, when there is general scientific understanding of mechanistic events that result in a health effect, this conclusion level <u>could also be</u> used if there is strong mechanistic evidence that is sufficient to highlight potential human toxicity^f—in the absence of informative conventional studies in humans or in animals (i.e., <i>indeterminate</i> evidence in both).
<p>The currently available evidence is inadequate to assess whether [chemical] may cause [health effect] in humans.</p>	<p>Evidence inadequate</p>	<p>This conveys either a lack of information or an inability to interpret the available evidence for [health effect]. On an assessment-specific basis, a single use of this “inadequate” conclusion level might be used to characterize the evidence for multiple health effect categories (i.e., all health effects that were examined and did not support other conclusion levels).^g</p> <ul style="list-style-type: none"> • This conclusion level <u>is</u> used if there is <i>indeterminate</i> human and animal evidence. • This conclusion level <u>is</u> also used with <i>slight</i> animal evidence and <i>compelling evidence of no effect</i> human evidence. • This conclusion level <u>could also be</u> used with <i>slight-to-robust</i> animal evidence and <i>indeterminate</i> human evidence if strong mechanistic information indicated that the animal evidence is unlikely to be relevant to humans. A conclusion of inadequate is not a determination that the agent does not cause the indicated health effect(s). It simply indicates that the available evidence is insufficient to reach conclusions.

Summary evidence integration judgment ^a in narrative	Evidence integration judgment level	Explanation and example scenarios ^b
<p>Strong evidence supports no effect in humans. This conclusion is based on studies of [humans or animals] that assessed [exposure or dose] levels of [range of concentrations].</p>	<p>Strong evidence supports no effect</p>	<p>This represents a situation in which extensive evidence across a range of populations and exposure levels has identified no effects/associations. This scenario requires a <i>high</i> degree of confidence in the conduct of individual studies, including consideration of study sensitivity, and comprehensive assessments of the endpoints and lifestages of exposure relevant to the health effect of interest.</p> <ul style="list-style-type: none"> • This conclusion level <u>is</u> used if there is compelling evidence of no effect in human studies and compelling evidence of no effect to indeterminate in animals. • This conclusion level <u>is</u> also used if there is <i>indeterminate</i> human evidence and <i>compelling evidence of no effect</i> animal evidence in models concluded to be relevant to humans. • This conclusion level <u>could also be</u> used with <i>compelling evidence of no effect</i> in human studies and <i>moderate to robust</i> animal evidence if strong mechanistic information indicated that the animal evidence is unlikely to be relevant to humans.

^aEvidence integration judgments are typically developed at the level of the health effect when there are sufficient studies on the topic to evaluate the evidence at that level; this should always be the case for “evidence demonstrates” and “strong evidence supports no effect,” and typically for “evidence indicates (likely).” However, some databases only allow for evaluations at the category of health effects examined; this will more frequently be the case for conclusion levels of “evidence suggests” and “evidence inadequate.” A judgment of “strong evidence supports no effect” is drawn at the health effect level.

^bTerminology of “is” refers to the default option; terminology of “could also be” refers to situational options dependent on mechanistic understanding.

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

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

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

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

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

1 For evaluations of carcinogenicity, consistent with EPA’s cancer guidelines ([U.S. EPA,](#)
2 [2005a](#)), one of EPA’s standardized cancer descriptors is used to describe the overall potential for
3 carcinogenicity within the evidence integration narrative for carcinogenicity. These descriptors are:
4 (1) ***carcinogenic to humans***, (2) ***likely to be carcinogenic to humans***, (3) ***suggestive evidence of***
5 ***carcinogenic potential***, (4) ***inadequate information to assess carcinogenic potential***, or (5) ***not***
6 ***likely to be carcinogenic to humans***. The standardized cancer descriptors will often align with the
7 evidence integration judgments (i.e., “evidence demonstrates” aligns with “carcinogenic to
8 humans”) but not in all cases. For example, the evidence integration judgments are generally used
9 for individual tumor or cancer types and the standardized EPA descriptors are used to characterize
10 overall cancer hazard.

11 For each type of cancer evaluated (e.g., lung cancer; renal cancer) or sets of related cancer
12 types, an evidence integration narrative and summary judgment level are provided as described
13 above for noncancer health effects. When considering evidence on carcinogenicity across human
14 and animal evidence, site concordance is not required ([U.S. EPA, 2005a](#)). If a systematic review of
15 more than one cancer type was conducted, then the strongest evidence integration judgment(s) is
16 used as the basis for selecting the standardized cancer descriptor in accordance with the EPA
17 cancer guidelines ([U.S. EPA, 2005a](#)).

9. DOSE-RESPONSE ASSESSMENT: SELECTING STUDIES AND QUANTITATIVE ANALYSIS

9.1. OVERVIEW

1 Selection of specific data sets for dose-response assessment and performance of the
2 dose-response assessment is conducted after hazard identification is complete and involves
3 database- and chemical-specific biological judgments. A number of EPA guidelines and support
4 documents detail data requirements and other considerations for dose-response modeling,
5 especially EPA's *Benchmark Dose Technical Guidance* ([U.S. EPA, 2012b](#)), EPA's *Review of the*
6 *Reference Dose and Reference Concentration Processes* [([U.S. EPA, 2005a, 2002](#)), *Guidelines for*
7 *Carcinogen Risk Assessment* ([U.S. EPA, 2005a](#)), and *Supplemental Guidance for Assessing*
8 *Susceptibility from Early-Life Exposure to Carcinogens* ([U.S. EPA, 2005b](#)). This section of the protocol
9 provides an overview of considerations for conducting the dose-response assessment, particularly
10 statistical considerations specific to dose-response analysis that support quantitative risk
11 assessment. Importantly, these considerations do not supersede existing EPA guidelines.

12 For IRIS assessments, dose response assessments are typically performed for both
13 noncancer and cancer hazards, and for both oral and inhalation routes of exposure following
14 chronic exposure⁸ to the chemical of interest, if supported by existing data. For noncancer hazards,
15 an inhalation reference concentration (RfC) and an oral reference dose (RfD) will be derived. In
16 addition to an RfC and RfD, this assessment will attempt to derive organ- or system-specific toxicity
17 values when the data are sufficiently strong (i.e., noncancer conclusions of evidence demonstrate or
18 evidence indicates [likely]). A reference value may also be derived for cancer effects in cases where
19 a nonlinear MOA is concluded that indicates a key precursor event necessary for carcinogenicity
20 does not occur below a specific exposure level (([U.S. EPA, 2005a](#)), Section 3.3.4). In addition, when
21 feasible and if the available data are appropriate for doing so, the assessment will derive a less-
22 than-lifetime toxicity value (a "subchronic" reference value) for noncancer hazards. Both less-than-
23 lifetime and hazard-specific values may be useful to EPA risk assessors within specific decision
24 contexts.

25 When low-dose linear extrapolation for cancer effects is supported, particularly for
26 chemicals with direct mutagenic activity or those for which the data indicate a linear component
27 below the point of departure (POD), an inhalation unit risk (IUR) facilitates estimation of human
28 cancer risks. Low-dose linear extrapolation is also used as a default when the data are insufficient

⁸Dose-response assessments may also be conducted for shorter durations, particularly if the evidence base for a chemical indicates risks associated with shorter exposures to the chemical ([U.S. EPA, 2002](#)).

1 to establish the mode of action ([U.S. EPA, 2005a](#)). An IUR is a plausible upper-bound lifetime cancer
2 risk from chronic inhalation of a chemical per unit of air concentration (expressed as ppm or
3 $\mu\text{g}/\text{m}^3$). In contrast with RfCs, an IUR can be used in conjunction with exposure information to
4 estimate cancer risk at a given dose.

5 The derivation of toxicity values also depends on the nature of the hazard conclusion.
6 Specifically, EPA generally conducts dose-response assessments and derives cancer values for
7 chemicals that are classified as *carcinogenic* or *likely to be carcinogenic* to humans. When there is
8 *suggestive evidence* of carcinogenic potential to humans, EPA generally would not conduct a
9 dose-response assessment and derive a cancer value. Similarly, for noncancer outcomes dose-
10 response is conducted based on having stronger evidence of a hazard (generally, “*evidence*
11 *demonstrates*” and “*evidence indicates [likely]*”. When the noncancer outcome is considered,
12 *evidence suggests* of potential hazard to humans, EPA generally would not conduct a dose-response
13 assessment and derive a RfC or RfD. Cases where suggestive evidence might be used to develop
14 cancer risk estimates or noncancer toxicity value include when the evidence base includes a
15 well-conducted study (overall *medium* or *high* confidence for the outcome), quantitative analyses
16 may be useful for some purposes, (e.g., providing a sense of the magnitude and uncertainty of
17 potential risks, ranking potential hazards, or setting research priorities) ([U.S. EPA, 2005a](#)).

9.2. SELECTING STUDIES FOR DOSE-RESPONSE ASSESSMENT

9.2.1. Hazard and MOA Considerations for Dose Response

18 The assessment presents a summary of hazard identification conclusions to transition to
19 dose response considerations, highlighting (1) information used to inform the selection of
20 outcomes or broader health effect categories for which toxicity values will be derived, (2) whether
21 toxicity values can be derived to protect specific populations or life stages, (3) how dose response
22 modeling will be informed by pharmacokinetic information, and (4) the identification of
23 biologically based BMR levels. The pool of outcomes and study-specific endpoints is discussed to
24 identify which categories of effects and study designs are considered the strongest and most
25 appropriate for quantitative assessment of a given health effect, particularly among the studies that
26 exemplify the study attributes summarized in Table 9-1.

27 Also considered is whether there are opportunities for quantitative evidence integration.
28 Examples of quantitative integration, from simplest to more complex, include (1) combining results
29 for an outcome across sex (within a study); (2) characterizing overall toxicity, as in combining
30 effects that comprise a syndrome, or occur on a continuum (e.g., precursors and eventual overt
31 toxicity, benign tumors that progress to malignant tumors); and (3) conducting a meta-analysis or
32 meta-regression of all studies addressing a category of important health effects.

33 Some studies that are used qualitatively for hazard identification may or may not be useful
34 quantitatively for dose-response assessment due to such factors as the lack of quantitative
35 measures of exposure or lack of variability measures for response data. If the needed information

1 cannot be located, semiquantitative analysis may be feasible (e.g., via NOAEL/LOAEL). In the draft
2 and final assessments, specific endpoints considered for dose-response are summarized in a tabular
3 format that includes rationales for decisions to proceed (or not) for POD derivation (see Table 9-2
4 for example format) selection.

5 In addition, mechanistic evidence that influences the dose-response analyses is highlighted,
6 for example, evidence related to susceptibility or potential shape of the dose-response curve (i.e.,
7 linear, nonlinear, or threshold model). Mode(s) of action is summarized including any interactions
8 between them relevant to understanding overall risk. For cancer dose-response, biological
9 considerations relevant to dose-response for cancer are:

- 10 • Is there evidence for direct mutagenicity?
- 11 • Does tumor latency decrease with increasing exposure?
- 12 • If there are multiple tumor types, which cancers have a longer latency period?
- 13 • Is incidence data available (incidence data are preferred to mortality data)?
- 14 • Were there different background incidences in different (geographic) populations?
- 15 • While benign and malignant tumors of the same cell of origin are generally evaluated
16 together, was there an increase only in malignant tumors?

Table 9-1. Attributes used to evaluate studies for derivation of toxicity values (in addition to the health effect category-specific evidence integration judgment)

Study attributes		Considerations	
		Human studies	Animal studies
Study confidence		High or medium confidence studies are highly preferred over low confidence studies. The available high and medium confidence studies are further differentiated based on the study attributes below as well as a reconsideration of the specific limitations identified and their potential impact on dose-response analyses.	
Rationale for choice of species		Human data are preferred over animal data to eliminate interspecies extrapolation uncertainties (e.g., in pharmacodynamics, relevance of specific health outcomes to humans).	Animal studies provide supporting evidence when adequate human studies are available and are considered principal studies when adequate human studies are not available. For some hazards, studies of particular animal species known to respond similarly to humans would be preferred over studies of other species.
Relevance of exposure paradigm	Exposure route	Studies involving human environmental exposures (oral, inhalation).	Studies by a route of administration relevant to human environmental exposure are preferred. A validated pharmacokinetic or PBPK model can also be used to extrapolate across exposure routes.
	Exposure durations	When developing a chronic toxicity value, chronic or subchronic studies are preferred over studies of acute exposure durations. Exceptions exist, such as when a susceptible population or life stage is more sensitive in a particular time window (e.g., developmental exposure).	
	Exposure levels	Exposures near the range of typical environmental human exposures are preferred. Studies with a broad exposure range and multiple exposure levels are preferred to the extent that they can provide information about the shape of the exposure-response relationship (see the EPA <i>Benchmark Dose Technical Guidance</i> , (U.S. EPA, 2012b), Section 2.1.1) and facilitate extrapolation to more relevant (generally lower) exposures.	
Subject selection		Studies that provide risk estimates in the most susceptible groups are preferred. Attempts are made to highlight where it might be possible to develop separate risk estimates for a specific population or life stage or determine whether evidence is available to select a data-derived uncertainty factor (UF).	
Controls for possible confounding ^a		Studies with a design (e.g., matching procedures, blocking) or analysis (e.g., covariates or other procedures for statistical adjustment) that adequately address the relevant sources of potential critical confounding for a given outcome are preferred.	

Study attributes	Considerations	
	Human studies	Animal studies
Measurement of exposure	<p>Studies that can reliably distinguish between levels of exposure in a time window considered most relevant for development of a causal effect are preferred.</p> <p>Exposure assessment methods that provide measurements at the level of the individual and that reduce measurement error are preferred.</p> <p>Measurements of exposure should not be influenced by knowledge of health outcome status.</p>	<p>Studies providing actual measurements of exposure (e.g., analytical inhalation concentrations vs. target concentrations) are preferred.</p> <p>Relevant internal dose measures may facilitate extrapolation to humans, as would availability of a suitable animal PBPK model in conjunction with an animal study reported in terms of administered exposure.</p>
Measurement of health outcome(s)	<p>Studies that can reliably distinguish the presence or absence (or degree of severity) of the outcome are preferred. Outcome ascertainment methods using generally accepted or standardized approaches are preferred.</p>	
	<p>Studies with individual data are preferred in general. Examples include: to characterize experimental variability more realistically, to characterize overall incidence of individuals affected by related outcomes (e.g., phthalate syndrome).</p>	
	<p>Among several relevant health outcomes, preference is generally given to those with greater biological significance. When there are multiple endpoints for an organ/system, characterizing the overall impact on this organ/system is considered. For example, if there are multiple histopathological alterations relevant to liver function changes, liver necrosis may be selected as the most representative endpoint to consider for dose-response analysis. For cancer types, consideration is given to the overall risk of multiple types of tumors. Multiple tumor types (if applicable) are discussed, and a rationale given for any grouping.</p>	
Study size and design	<p>Preference is given to studies using designs reasonably expected to have power to detect responses of suitable magnitude.^b This does not mean that studies with substantial responses but low power would be ignored, but that they should be interpreted in light of a confidence interval or variance for the response. Studies that address changes in the number at risk (through decreased survival, loss to follow-up) are preferred.</p>	

^aAn exposure or other variable that is associated with both exposure and outcome but is not an intermediary between the two.

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

Table 9-2. Example table used in assessment to show endpoint consideration judgments for POD derivation.

Endpoint	Study reference/ confidence	Exposure route duration	Human population or strain/species	Sexes studied	POD derivation	Rationale
Endpoint 1	Study citation and confidence (endpoint-specific level)	e.g., Gestational (route)	e.g., Wistar rats	males, females, or both	✓	e.g., Exposure-related increase
Endpoint 2	Study citation and confidence (endpoint-specific level)	e.g., Gestational (route)	e.g., Sprague-Dawley rats	males, females, or both	✗	e.g., No exposure-related effect; response not considered biologically significant (<5%)
Endpoint 3	Study citation and confidence (endpoint-specific level)	e.g., ongoing, measured during gestation	e.g., Children aged 7 yr	Both males and females	✓	e.g., Consistent associations across studies, minimal concerns for exposure measurement

Table 9-3. Specific example of presenting endpoints considered for dose-response modeling and derivation of points of departure.

Endpoint	Study reference/ confidence	Exposure route and duration	Human population or test species and strain	Lifestage and sex	POD derivation	Rationale
Endocrine Effects (hazard judgment of evidence indicates [likely])						
Decreased serum free and total T4	NTP (2018) ; high confidence	Gavage, 28 d	S-D rat	Adult female	Yes ü	Dose-dependent effects in free and total T4 in females and free T4 in males; large magnitude of effect in both sexes (91% reduction in free T4 in males at low dose where body weight unaffected, and 36%–53% reduction in free and total T4 in females at ≥3.12 mg/kg-d); effects in males were not prioritized due to elevated weight loss at higher doses.
	NTP (2018) ; high confidence	Gavage, 28 d	S-D rat	Adult male	No, ✗	

Endpoint	Study reference/ confidence	Exposure route and duration	Human population or test species and strain	Lifestage and sex	POD derivation	Rationale
Endocrine Effects (hazard judgment of evidence indicates [likely])						
Add a second endpoint, maybe not modeled due to large insensitivity vs. T4				Adult males and females	No, X	

1

9.3. CONDUCTING DOSE-RESPONSE ASSESSMENTS

1 EPA uses a two-step approach for dose-response assessment that distinguishes analysis of
2 the dose-response data in the range of observation from any inferences about responses at lower,
3 generally more environmentally relevant, exposure levels (([U.S. EPA, 2012b](#)); ([U.S. EPA, 2005a](#)),
4 Section 3):

- 5 1) Within the observed dose range, the preferred approach is to use dose-response modeling
6 to incorporate as much of the data set as possible into the analysis for the purpose of
7 deriving a POD, see Section 9.3.1 for more details.
- 8 2) Derivation of cancer risk estimates or reference values nearly always involves extrapolation
9 to exposures lower than the POD and is described in more detail in Sections 9.3.2 and 9.3.3,
10 respectively.

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

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

22 For cancer analyses, if there are multiple tumor types in a study population (human or
23 animal), final cancer risk estimates will typically address overall cancer risk.

9.3.1. Dose-Response Analysis in the Range of Observation

24 For conducting a dose response assessment, pharmacodynamic (“biologically based”)
25 modeling can be used when there are sufficient data to ascertain the mode of action and
26 quantitatively support model parameters that represent rates and other quantities associated with
27 the key precursor events of the modes of action. When pharmacodynamic modeling is not available
28 to assess health effects associated with exposure to ethylbenzene, empirical dose-response
29 modeling is used to fit the data (on the apical outcomes or a key precursor events) in the ranges of
30 observation. For this purpose of empirical dose-response modeling, EPA has developed a standard
31 set of models (<https://www.epa.gov/bmds>) that can be applied to typical dichotomous and
32 continuous data sets, including those that are nonlinear. In situations where there are alternative
33 models with significant biological support, the users of the assessment can be informed by the
34 presentation of these alternatives along with the models’ strengths and uncertainties. The EPA has

1 developed guidelines on modeling dose-response data, assessing model fit, selecting suitable
2 models, and reporting modeling results [see the *EPA Benchmark Dose Technical Guidance* ([U.S. EPA,
3 2012b](#))].

4 U.S. EPA Benchmark Dose Software (BMDS) is designed to model dose-response datasets in
5 accordance with EPA Benchmark Dose Technical Guidance ([U.S. EPA, 2012b](#)). For noncancer (and
6 nonlinear cancer), a BMDL is computed from a model selected from the BMDS suite of models using
7 statistical and graphical criteria. Linear analysis of cancer datasets is generally based on the
8 Multistage model, with degree selected following a U.S. EPA Statistical Workgroup technical memo
9 available on the BMDS website (<https://cfpub.epa.gov/ncea/bmds/recordisplay.cfm?deid=308382>
10). Modeling of cancer data may in some cases involve additional, specialized methods, particularly
11 for multiple tumors or early removal from observation (due to death or morbidity). Additional
12 judgments or alternative analyses may be used if initial modeling procedures fail to yield results in
13 reasonable agreement with the data. For example, modeling may be restricted to the lower doses,
14 especially if there is competing toxicity at higher doses.

15 For noncancer (and nonlinear cancer) datasets, EPA recommends (1) application of a
16 preferred set of models that use maximum likelihood estimation (MLE) methods (default models in
17 BMDS) and (2) selection of a POD from a single model based on criteria designed to limit model
18 selection subjectivity (auto implemented in BMDS version 3 and higher). For the linear analysis of
19 cancer datasets, EPA recommends (1) application of the Multistage MLE model; (2) selection of a
20 single Multistage degree; and (3) in cases where tumors are observed in multiple organ systems,
21 use of a multi-tumor model (i.e., MS-Combo) that appropriately estimates combined tumor risk
22 (both (2) and (3) are available in BMDS).⁹

23 Version 3.2 and higher of BMDS also provides an alternative modeling approach that uses
24 Bayesian model averaging for dichotomous modeling average (DMA). EPA makes DMA available as
25 alternative approaches but has not yet finalized guidelines for their use.

26 For each modeled dataset for an outcome, a POD from the observed data should be
27 estimated to mark the beginning of extrapolation to lower doses. The POD is an estimated dose
28 (expressed in human equivalent terms) near the lower end of the observed range without
29 significant extrapolation to lower doses. For linear extrapolation of cancer risk, the POD is used to
30 calculate an OSF or IUR, and for nonlinear extrapolation, the POD is used in calculating an RfD
31 or RfC.

32 The selection of the response level at which the POD is calculated is guided by the severity
33 of the endpoint. If linear extrapolation is used, selection of a response level corresponding to the
34 POD is not highly influential, so standard values near the low end of the observable range are
35 generally used (for example, 10% extra risk for cancer bioassay data, 1% for epidemiologic data,

⁹The Multistage degree selection process outlined in the memo is auto-implemented in the BMDS multitumor model, which can be run on one or more tumor data sets, but only the noncancer model selection process is auto-implemented for individual Multistage model runs in the current version, BMDS 3.2).

1 lower for rare cancers). Nonlinear approaches consider both statistical and biologic considerations.
2 For dichotomous data, a response level of 10% extra risk is generally used for minimally adverse
3 effects, 5% or lower for more severe effects. For continuous data, a response level is ideally based
4 on an established definition of biologic significance. In the absence of such definition, one control
5 standard deviation from the control mean is often used for minimally adverse effects, 1/2 standard
6 deviation for more severe effects. The POD is the 95% lower bound on the dose associated with the
7 selected response level.

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

- 11 • Intermittent study exposures are standardized to a daily average over the duration of
12 exposure. For chronic effects, daily exposures are averaged over the lifespan. Exposures
13 during a critical period, however, are not averaged over a longer duration (([U.S. EPA,](#)
14 [2005a](#)), Section 3.1.1; ([U.S. EPA, 1991a](#)), Section 3.2). Note that this will typically be done
15 after modeling because the conversion is linear.
- 16 • Doses are standardized to equivalent human terms to facilitate comparison of results from
17 different species. Oral doses are scaled allometrically using $\text{mg}/\text{kg}^{3/4}$ day as the equivalent
18 dose metric across species. Allometric scaling pertains to equivalence across species, not
19 across life stages, and is not used to scale doses from adult humans or mature animals to
20 infants or children (([U.S. EPA, 2011](#)) ([U.S. EPA, 2005a](#)), section 3.1.3). Inhalation exposures
21 are scaled using dosimetry models that apply species-specific physiologic and anatomic
22 factors and consider whether the effect occurs at the site of first contact or after systemic
23 circulation ([U.S. EPA, 2012a, 1994](#)), Section 3).
- 24 • It can be informative to convert doses across exposure routes. If this is done, the assessment
25 describes the underlying data, algorithms, and assumptions (([U.S. EPA, 2005a](#)), Section
26 3.1.4).
- 27 • In the absence of study specific data on, for example, intake rates or body weight, the EPA
28 has developed recommended values for use in dose response analysis ([U.S. EPA, 1988](#)).
- 29 • The preferred approach for dosimetry extrapolation from animals to humans is through
30 PBPK modeling.
- 31 • Briefly, PBPK model simulations can be used to estimate internal dose metrics (e.g.,
32 ethylbenzene on blood or its oxidative metabolite produced in the liver) corresponding to
33 the applied doses for each experimental animal bioassay. By simulating the exposure
34 scenario for each toxicity study (e.g., 6 hours/day, 5 days/week inhalation exposure), the
35 resulting internal metric effectively accounts for the difference between the pattern and a
36 nominal 24 hours/day, 7 days/week exposure. The set of internal dose metrics for each
37 toxicity study and endpoint can then be used in dose-response analysis to identify a BMDL
38 or other POD for individual animal toxicity studies. In this assessment, the internal dose
39 metric is either the tissue-specific rate of oxidative metabolism or a daily average blood
40 concentration of ethylbenzene. The human version of the PBPK model can then be used to
41 estimate the exposure concentration in air which, given continuous (24 hours/day,

1 7 days/week) inhalation exposure, would result in internal dose PODs aforementioned. Any
2 remaining uncertainty factors, including the factor of 10 for human inter-individual
3 variability (UFH) will then be applied for derivation of the HECs.

- 4 • If needed, a similar approach can be applied for oral-to-inhalation route extrapolation for
5 endpoints where toxicity data are available from oral dosimetry studies but not from
6 inhalation.

9.3.2. Extrapolation: Slope Factors and Unit Risk

7 An OSF or *IUR* facilitates estimation of human cancer risks when low-dose linear
8 extrapolation for cancer effects is supported, particularly for chemicals with direct mutagenic
9 activity or those for which the data indicate a linear component below the POD. Low-dose linear
10 extrapolation is also used as a default when the data are insufficient to establish the mode of action
11 ([U.S. EPA, 2005a](#)). If data are sufficient to ascertain one or more modes of action consistent with
12 low-dose nonlinearity, or to support their biological plausibility, low-dose extrapolation may use
13 the reference value approach when suitable data are available ([U.S. EPA, 2005a](#)).

9.3.3. Extrapolation: Reference Values

14 Reference value derivation is EPA's most frequently used type of nonlinear extrapolation
15 method. Although it is most commonly used for noncancer effects, this approach is also used for
16 cancer effects if there are sufficient data to ascertain the MOA and conclude that it is not linear at
17 low doses. For these cases, reference values for each relevant route of exposure are developed
18 following EPA's established practices (([U.S. EPA, 2005a](#)), Section 3.3.4). In general, it has been the
19 IRIS Program's preference to base cancer reference values on key precursor events in the MOA that
20 are necessary for tumor formation rather than on the incidence of tumors themselves. For example,
21 see the ethylene glycol monobutyl ether (EGBE) assessment where the cancer RfD was based on
22 hemosiderin deposition in the liver vs. liver tumor incidence ([2010b](#)).

23 For each data set selected for reference value derivation, reference values are estimated by
24 applying relevant adjustments to the PODs to account for the conditions of the reference value
25 definition—for human variation, extrapolation from animals to humans, extrapolation to chronic
26 exposure duration, and extrapolation to a minimal level of risk (if not observed in the data set).
27 Increasingly, data-based adjustments ([U.S. EPA, 2014a](#)) and Bayesian methods for characterizing
28 population variability ([NRC, 2014](#)) are feasible and may be distinguished from the UF
29 considerations outlined below. The assessment will discuss the scientific bases for estimating these
30 data-based adjustments and UFs:

- 31 • *Animal-to-human* extrapolation: If animal results are used to make inferences about
32 humans, the reference value derivation incorporates the potential for cross-species
33 differences, which may arise from differences in pharmacokinetics or pharmacodynamics. If
34 available, a biologically based model that adjusts fully for pharmacokinetic and
35 pharmacodynamic differences across species may be used. Otherwise, the POD is
36 standardized to equivalent human terms or is based on pharmacokinetic or dosimetry

1 modeling, which may range from detailed chemical-specific to default approaches ([U.S. EPA, 2014a, 2011](#)), and a factor of 10^{1/2} (rounded to 3) is applied to account for the remaining
2 uncertainty involving pharmacokinetic and pharmacodynamic differences.
3

- 4 • Human variation: The assessment accounts for variation in susceptibility across the human
5 population and the possibility that the available data may not represent individuals who are
6 most susceptible to the effect, by using a data-based adjustment or UF or a combination of
7 the two. Where appropriate data or models for the effect or for characterizing the internal
8 dose are available, the potential for data-based adjustments for pharmacodynamics or
9 pharmacokinetics is considered 9, 10 ([U.S. EPA, 2014a, 2002](#)). When sufficient data are
10 available, an intraspecies UF either less than or greater than 10-fold may be justified ([U.S.
11 EPA, 2002](#)). This factor may be reduced if the POD is derived from or adjusted specifically
12 for susceptible individuals [not for a general population that includes both susceptible and
13 nonsusceptible individuals; (see ([U.S. EPA, 2002](#)), Section 4.4.5; ([U.S. EPA, 1998](#)), Section
14 4.2; ([U.S. EPA, 1996](#)), Section 4; ([U.S. EPA, 1994](#)), Section 4.3.9.1; ([U.S. EPA, 1991a](#)), Section
15 3.4). When the use of such data or modeling is not supported, an UF with a default value of
16 10 is considered.
- 17 • LOAEL to NOAEL: If a POD is based on a LOAEL, the assessment includes an adjustment to
18 an exposure level where such effects are not expected. This can be a matter of great
19 uncertainty if there is no evidence available at lower exposures. A factor of 3 or 10 is
20 generally applied to extrapolate to a lower exposure expected to be without appreciable
21 effects. A factor other than 10 may be used depending on the magnitude and nature of the
22 response and the shape of the dose-response curve ([U.S. EPA, 2002, 1998, 1996, 1994,
23 1991a](#)).
- 24 • Subchronic-to-chronic exposure: When using subchronic studies to make inferences about
25 chronic/lifetime exposure, the assessment considers whether lifetime exposure could have
26 effects at lower levels of exposure. A factor of up to 10 may be applied to the POD,
27 depending on the duration of the studies and the nature of the response ([U.S. EPA, 2002,
28 1998, 1994](#)).
- 29 • Database deficiencies: In addition to the adjustments above, if database deficiencies raise
30 concern that further studies might identify a more sensitive effect, organ system, or life
31 stage, the assessment may apply a database UF ([U.S. EPA, 2002, 1998, 1996, 1994, 1991a](#)).
32 The size of the factor depends on the nature of the database deficiency. For example, the
33 EPA typically follows the recommendation that a factor of 10 be applied if both a prenatal
34 toxicity study and a two-generation reproduction study are missing and a factor of 10^{1/2}
35 (i.e., 3) if either one or the other is missing (([U.S. EPA, 2002](#)), Section 4.4.5).

36 The POD for a reference value is divided by the product of these factors (([U.S. EPA, 2002](#)),
37 Section 4.4.5), recommends that any composite factor that exceeds 3,000 represents excessive
38 uncertainty and recommends against relying on the associated reference value.

10. PROTOCOL HISTORY

REFERENCES

- 1 [Adgate, JL; Church, TR; Ryan, AD; Ramachandran, G; Fredrickson, AL; Stock, TH; Morandi, MT;](#)
 2 [Sexton, K.](#) (2004). Outdoor, indoor, and personal exposure to VOCs in children. *Environ*
 3 *Health Perspect* 112: 1386-1392. <http://dx.doi.org/10.1289/ehp.7107>.
- 4 [Aguilera, I; Sunyer, J; Fernández-Patier, R; Hoek, G; Aguirre-Alfaro, A; Meliefste, K; Bomboi-](#)
 5 [Mingarro, MT; Nieuwenhuijsen, MJ; Herce-Garraleta, D; Brunekreef, B.](#) (2008). Estimation of
 6 outdoor NO_x, NO₂, and BTEX exposure in a cohort of pregnant women using land use
 7 regression modeling. *Environ Sci Technol* 42: 815-821.
 8 <http://dx.doi.org/10.1021/es0715492>.
- 9 [ANL](#) (Argonne National Laboratory). (2021). Active thermochemical tables (ATcT), version 1.122d:
 10 Ethylbenzene enthalpy of formation. Available online at
 11 [https://atct.anl.gov/Thermochemical%20Data/version%201.122d/species/?species_num](https://atct.anl.gov/Thermochemical%20Data/version%201.122d/species/?species_number=1092)
 12 [ber=1092](https://atct.anl.gov/Thermochemical%20Data/version%201.122d/species/?species_number=1092) (accessed December 16, 2021).
- 13 [Ashley, DL; Prah, JD.](#) (1997). Time dependence of blood concentrations during and after exposure to
 14 a mixture of volatile organic compounds. *Arch Environ Health* 52: 26-33.
 15 <http://dx.doi.org/10.1080/00039899709603796>.
- 16 [ATSDR](#) (Agency for Toxic Substances and Disease Registry). (2010). Toxicological profile for
 17 ethylbenzene [ATSDR Tox Profile]. (PB2010100004). Atlanta, GA: U.S. Department of Health
 18 and Human Services, Public Health Service.
 19 <http://www.atsdr.cdc.gov/ToxProfiles/tp.asp?id=383&tid=66>.
- 20 [Basagaña, X; Aguilera, I; Rivera, M; Agis, D; Foraster, M; Marrugat, J; Elosua, R; Künzli, N.](#) (2013).
 21 Measurement error in epidemiologic studies of air pollution based on land-use regression
 22 models. *Am J Epidemiol* 178: 1342-1346. <http://dx.doi.org/10.1093/aje/kwt127>.
- 23 [Cannella, W.](#) (2007). Xylenes and ethylbenzene [Encyclopedia]. In Kirk-Othmer Encyclopedia of
 24 Chemical Technology. Michigan: John Wiley & Sons.
- 25 [Capella, KM; Roland, K; Geldner, N; Rey deCastro, B; De Jesús, VR; van Bommel, D; Blount, BC.](#)
 26 (2019). Ethylbenzene and styrene exposure in the United States based on urinary mandelic
 27 acid and phenylglyoxylic acid: NHANES 2005-2006 and 2011-2012. *Environ Res* 171: 101-
 28 110. <http://dx.doi.org/10.1016/j.envres.2019.01.018>.
- 29 [Clayton, GD; Clayton, FE.](#) (1981). "Alcohols". In Patty's industrial hygiene and toxicology:
 30 Toxicology. New York, NY: John Wiley & Sons.
- 31 [Dickersin, K.](#) (1990). The existence of publication bias and risk factors for its occurrence. *JAMA* 263:
 32 1385-1389.
- 33 [EFSA](#) (European Food Safety Authority). (2017). Guidance on the use of the weight of evidence
 34 approach in scientific assessments. *EFSA J* 15: 1-69.
 35 <http://dx.doi.org/10.2903/j.efsa.2017.4971>.
- 36 [Heinrich-Ramm, R; Jakubowski, M; Heinzow, B; Christensen, JM; Olsen, E; Hertel, O.](#) (2000).
 37 Biological monitoring for exposure to volatile organic compounds (VOCs). *Pure Appl Chem*
 38 72: 385-436. <http://dx.doi.org/10.1351/pac200072030385>.
- 39 [Higgins, JPT; Thomas, J; Chandler, J; Cumpston, M; Li, T; Page, MJ; Welch, VA.](#) (2022). Cochrane
 40 handbook for systematic reviews of interventions version 6.3. Higgins, JPT; Thomas, J;
 41 Chandler, J; Cumpston, M; Li, T; Page, MJ; Welch, VA.
 42 <http://www.training.cochrane.org/handbook>.

- 1 [Hill, AB.](#) (1965). The environment and disease: Association or causation? Proc R Soc Med 58: 295-
2 300.
- 3 [Hoenig, JM; Heisey, DM.](#) (2001). The abuse of power: The pervasive fallacy of power calculations for
4 data analysis. Am Stat 55: 19-24.
- 5 [Howard, BE; Phillips, J; Miller, K; Tandon, A; Mav, D; Shah, MR; Holmgren, S; Pelch, KE; Walker, V;
6 Rooney, AA; Macleod, M; Shah, RR; Thayer, K.](#) (2016). SWIFT-Review: A text-mining
7 workbench for systematic review. Syst Rev 5: 87. <http://dx.doi.org/10.1186/s13643-016-0263-z>.
- 8
- 9 [IPCS](#) (International Programme on Chemical Safety). (2012). Harmonization project document no.
10 10: Guidance for immunotoxicity risk assessment for chemicals. (Harmonization Project
11 Document No. 10). Geneva, Switzerland: World Health Organization.
12 <http://www.inchem.org/documents/harmproj/harmproj/harmproj10.pdf>.
- 13 [Jia, C; Batterman, SA; Relyea, GE.](#) (2012). Variability of indoor and outdoor VOC measurements: an
14 analysis using variance components. Environ Pollut 169: 152-159.
15 <http://dx.doi.org/10.1016/j.envpol.2011.09.024>.
- 16 [Kim, YM; Harrad, S; Harrison, RM.](#) (2002). Levels and sources of personal inhalation exposure to
17 volatile organic compounds. Environ Sci Technol 36: 5405-5410.
18 <http://dx.doi.org/10.1021/es010148y>.
- 19 [Konkle, SL; Zierold, KM; Taylor, KC; Riggs, DW; Bhatnagar, A.](#) (2020). National secular trends in
20 ambient air volatile organic compound levels and biomarkers of exposure in the United
21 States. Environ Res 182: 108991. <http://dx.doi.org/10.1016/j.envres.2019.108991>.
- 22 [Lin, YS; Kupper, LL; Rappaport, SM.](#) (2005). Air samples versus biomarkers for epidemiology. Occup
23 Environ Med 62: 750-760. <http://dx.doi.org/10.1136/oem.2004.013102>.
- 24 [Mcdonald, BC; de Gouw, JA; Gilman, JB; Jathar, SH; Akherati, A; Cappa, CD; Jimenez, JL; Lee-Taylor, J;
25 Hayes, PL; Mckeen, SA; Cui, YY; Kim, SW; Gentner, DR; Isaacman-Vanwertz, G; Goldstein, AH;
26 Harley, RA; Frost, GJ; Roberts, JM; Ryerson, TB; Trainer, M.](#) (2018). Volatile chemical
27 products emerging as largest petrochemical source of urban organic emissions. Science 359:
28 760-764. <http://dx.doi.org/10.1126/science.aag0524>.
- 29 [Mukerjee, S; Smith, LA; Johnson, MM; Neas, LM; Stallings, CA.](#) (2009). Spatial analysis and land use
30 regression of VOCs and NO(2) from school-based urban air monitoring in Detroit/Dearborn,
31 USA. Sci Total Environ 407: 4642-4651. <http://dx.doi.org/10.1016/j.scitotenv.2009.04.030>.
- 32 [NASEM](#) (National Academies of Sciences, Engineering, and Medicine). (2021). Review of U.S. EPA's
33 ORD staff handbook for developing IRIS assessments: 2020 version. Washington, DC:
34 National Academies Press. <http://dx.doi.org/10.17226/26289>.
- 35 [Nong, A; Charest-Tardif, G; Tardif, R; Lewis, DF; Sweeney, LM; Gargas, ML; Krishnan, K.](#) (2007).
36 Physiologically based modeling of the inhalation pharmacokinetics of ethylbenzene in
37 B6C3F1 mice. J Toxicol Environ Health A 70: 1838-1848.
38 <http://dx.doi.org/10.1080/15287390701459239>.
- 39 [NRC](#) (National Research Council). (2014). Review of EPA's Integrated Risk Information System
40 (IRIS) process. Washington, DC: The National Academies Press.
41 <http://dx.doi.org/10.17226/18764>.
- 42 [NTP](#) (National Toxicology Program). (2015). Handbook for conducting a literature-based health
43 assessment using OHAT approach for systematic review and evidence integration. Research
44 Triangle Park, NC: U.S. Department of Health and Human Services, National Toxicology
45 Program, Office of Health Assessment and Translation.
46 https://ntp.niehs.nih.gov/ntp/ohat/pubs/handbookjan2015_508.pdf.
- 47 [NTP](#) (National Toxicology Program). (2018). 28-day evaluation of the toxicity (C04049) of
48 perfluorononanoic acid (PFNA) (375-95-1) on Harlan Sprague-Dawley rats exposed via
49 gavage [NTP]. <http://dx.doi.org/10.22427/NTP-DATA-002-02655-0003-0000-3>.

- 1 [NTP](#) (National Toxicology Program). (2019). Handbook for conducting a literature-based health
2 assessment using OHAT approach for systematic review and evidence integration. Research
3 Triangle, NC: National Institute of Environmental Health Sciences.
4 https://ntp.niehs.nih.gov/ntp/ohat/pubs/handbookmarch2019_508.pdf.
- 5 [Ransley, DL](#). (1984). Xylenes and ethylbenzene. In HF Mark; DF Othmer; CG Overberger; GT
6 Seaborg; M Grayson (Eds.), Kirk-Othmer encyclopedia of chemical technology (Vol 24) (3rd
7 ed., pp. 709-744). New York, NY: John Wiley & Sons.
- 8 [Schünemann, H; Brožek, J; Guyatt, G; Oxman, A](#). (2013). GRADE handbook. Available online at
9 <https://gdt.gradepro.org/app/handbook/handbook.html> (accessed April 22, 2022).
- 10 [Sexton, K; Adgate, JL; Church, TR; Ashley, DL; Needham, LL; Ramachandran, G; Fredrickson, AL;](#)
11 [Ryan, AD](#). (2005). Children's exposure to volatile organic compounds as determined by
12 longitudinal measurements in blood. *Environ Health Perspect* 113: 342-349.
13 <http://dx.doi.org/10.1289/ehp.7412>.
- 14 [Sexton, K; Adgate, JL; Mongin, SJ; Pratt, GC; Ramachandran, G; Stock, TH; Morandi, MT](#). (2004a).
15 Evaluating differences between measured personal exposures to volatile organic
16 compounds and concentrations in outdoor and indoor air. *Environ Sci Technol* 38: 2593-
17 2602. <http://dx.doi.org/10.1021/es030607q>.
- 18 [Sexton, K; Adgate, JL; Ramachandran, G; Pratt, GC; Mongin, SJ; Stock, TH; Morandi, MT](#). (2004b).
19 Comparison of personal, indoor, and outdoor exposures to hazardous air pollutants in three
20 urban communities. *Environ Sci Technol* 38: 423-430.
21 <http://dx.doi.org/10.1021/es030319u>.
- 22 [Sexton, K; Mongin, SJ; Adgate, JL; Pratt, GC; Ramachandran, G; Stock, TH; Morandi, MT](#). (2007).
23 Estimating volatile organic compound concentrations in selected microenvironments using
24 time-activity and personal exposure data. *J Toxicol Environ Health A* 70: 465-476.
25 <http://dx.doi.org/10.1080/15287390600870858>.
- 26 [Smith, MT; Guyton, KZ; Gibbons, CF; Fritz, JM; Portier, CJ; Rusyn, I; DeMarini, DM; Caldwell, JC;](#)
27 [Kavlock, RJ; Lambert, PF; Hecht, SS; Bucher, JR; Stewart, BW; Baan, RA; Cogliano, VJ; Straif,](#)
28 [K](#). (2016). Key characteristics of carcinogens as a basis for organizing data on mechanisms
29 of carcinogenesis [Review]. *Environ Health Perspect* 124: 713-721.
30 <http://dx.doi.org/10.1289/ehp.1509912>.
- 31 [Sterne, JAC; Hernán, MA; Reeves, BC; Savović, J; Berkman, ND; Viswanathan, M; Henry, D; Altman,](#)
32 [DG; Ansari, MT; Boutron, I; Carpenter, JR; Chan, AW; Churchill, R; Deeks, JJ; Hróbjartsson, A;](#)
33 [Kirkham, J; Jüni, P; Loke, YK; Pigott, TD; Ramsay, CR; Regidor, D; Rothstein, HR; Sandhu, L;](#)
34 [Santaguida, PL; Schünemann, HJ; Shea, B; Shrier, I; Tugwell, P; Turner, L; Valentine, JC;](#)
35 [Waddington, H; Waters, E; Wells, GA; Whiting, PF; Higgins, JPT](#). (2016). ROBINS-I: A tool for
36 assessing risk of bias in non-randomised studies of interventions. *BMJ* 355: i4919.
37 <http://dx.doi.org/10.1136/bmj.i4919>.
- 38 [Su, FC; Mukherjee, B; Batterman, S](#). (2011). Trends of VOC exposures among a nationally
39 representative sample: Analysis of the NHANES 1988 through 2004 data sets. *Atmos*
40 *Environ* 45: 4858-4867. <http://dx.doi.org/10.1016/j.atmosenv.2011.06.016>.
- 41 [Sweeney, LM; Kester, JE; Kirman, CR; Gentry, RP; Banton, MI; Bus, JS; Gargas, ML](#). (2015). Risk
42 assessments for chronic exposure of children and prospective parents to ethylbenzene (CAS
43 No. 100-41-4) [Review]. *Crit Rev Toxicol* 45: 662-726.
44 <http://dx.doi.org/10.3109/10408444.2015.1046157>.
- 45 [Thayer, KA; Shaffer, RM; Angrish, M; Arzuaga, X; Carlson, LM; Davis, A; Dishaw, L; Druwe, I; Gibbons,](#)
46 [C; Glenn, B; Jones, R; Kaiser, JP; Keshava, C; Keshava, N; Kraft, A; Lizarraga, L; Markey, K;](#)
47 [Persad, A; Radke, EG; ... Yost, E](#). (2022). Use of systematic evidence maps within the US
48 environmental protection agency (EPA) integrated risk information system (IRIS) program:
49 Advancements to date and looking ahead [Comment]. *Environ Int* 169: 107363.
50 <http://dx.doi.org/10.1016/j.envint.2022.107363>.

Protocol for the Ethylbenzene IRIS Assessment

- 1 [U.S. EPA](#) (U.S. Environmental Protection Agency). (1988). Recommendations for and documentation
2 of biological values for use in risk assessment [EPA Report]. (EPA600687008). Cincinnati,
3 OH. <http://cfpub.epa.gov/ncea/cfm/recordisplay.cfm?deid=34855>.
- 4 [U.S. EPA](#) (U.S. Environmental Protection Agency). (1991a). Guidelines for developmental toxicity
5 risk assessment. Fed Reg 56: 63798-63826.
- 6 [U.S. EPA](#) (U.S. Environmental Protection Agency). (1991b). Integrated Risk Information System
7 (IRIS): Chemical assessment summary for ethylbenzene (CASRN 100-41-4) [EPA Report].
8 Washington DC. <http://www.epa.gov/iris/subst/0051.htm>.
- 9 [U.S. EPA](#) (U.S. Environmental Protection Agency). (1994). Methods for derivation of inhalation
10 reference concentrations and application of inhalation dosimetry [EPA Report].
11 (EPA600890066F). Research Triangle Park, NC.
12 <https://cfpub.epa.gov/ncea/risk/recordisplay.cfm?deid=71993&CFID=51174829&CFTOKEN=25006317>.
- 13
14 [U.S. EPA](#) (U.S. Environmental Protection Agency). (1996). Guidelines for reproductive toxicity risk
15 assessment (pp. 1-143). (EPA/630/R-96/009). Washington, DC: U.S. Environmental
16 Protection Agency, Risk Assessment Forum.
17 [https://www.epa.gov/sites/production/files/2014-
18 11/documents/guidelines_repro_toxicity.pdf](https://www.epa.gov/sites/production/files/2014-11/documents/guidelines_repro_toxicity.pdf).
- 19 [U.S. EPA](#) (U.S. Environmental Protection Agency). (1998). Guidelines for neurotoxicity risk
20 assessment [EPA Report] (pp. 1-89). (ISSN 0097-6326/EISSN 2167-2520
21 EPA/630/R-95/001F). Washington, DC: U.S. Environmental Protection Agency, Risk Assessment
22 Forum. <http://www.epa.gov/risk/guidelines-neurotoxicity-risk-assessment>.
- 23 [U.S. EPA](#) (U.S. Environmental Protection Agency). (2002). A review of the reference dose and
24 reference concentration processes. (EPA630P02002F). Washington, DC.
25 <https://www.epa.gov/sites/production/files/2014-12/documents/rfd-final.pdf>.
- 26 [U.S. EPA](#) (U.S. Environmental Protection Agency). (2005a). Guidelines for carcinogen risk
27 assessment [EPA Report]. (EPA630P03001F). Washington, DC.
28 [https://www.epa.gov/sites/production/files/2013-
29 09/documents/cancer_guidelines_final_3-25-05.pdf](https://www.epa.gov/sites/production/files/2013-09/documents/cancer_guidelines_final_3-25-05.pdf).
- 30 [U.S. EPA](#) (U.S. Environmental Protection Agency). (2005b). Supplemental guidance for assessing
31 susceptibility from early-life exposure to carcinogens [EPA Report]. (EPA/630/R-03/003F).
32 Washington, DC: U.S. Environmental Protection Agency, Risk Assessment Forum.
33 [https://www.epa.gov/risk/supplemental-guidance-assessing-susceptibility-early-life-
34 exposure-carcinogens](https://www.epa.gov/risk/supplemental-guidance-assessing-susceptibility-early-life-exposure-carcinogens).
- 35 [U.S. EPA](#) (U.S. Environmental Protection Agency). (2010a). Comparison of 1999 model-predicted
36 concentrations to monitored data.
37 <https://archive.epa.gov/airtoxics/nata1999/web/html/99compare.html>.
- 38 [U.S. EPA](#) (U.S. Environmental Protection Agency). (2010b). Toxicological review of ethylene glycol
39 monobutyl ether (EGBE) (CAS no. 111-76-2) in support of summary information on the
40 integrated risk information system (IRIS), march 2010. (EPA/635/R-08/006F).
41 <https://iris.epa.gov/static/pdfs/0500tr.pdf>.
- 42 [U.S. EPA](#) (U.S. Environmental Protection Agency). (2011). Recommended use of body weight 3/4 as
43 the default method in derivation of the oral reference dose. (EPA100R110001). Washington,
44 DC. [https://www.epa.gov/sites/production/files/2013-09/documents/recommended-use-
45 of-bw34.pdf](https://www.epa.gov/sites/production/files/2013-09/documents/recommended-use-of-bw34.pdf).
- 46 [U.S. EPA](#) (U.S. Environmental Protection Agency). (2012a). Advances in inhalation gas dosimetry for
47 derivation of a reference concentration (RfC) and use in risk assessment (pp. 1-140).
48 (EPA/600/R-12/044). Washington, DC.

- 1 <https://cfpub.epa.gov/ncea/risk/recordisplay.cfm?deid=244650&CFID=50524762&CFTOKEN=17139189>.
- 2
- 3 [U.S. EPA](#) (U.S. Environmental Protection Agency). (2012b). Benchmark dose technical guidance
4 [EPA Report]. (EPA100R12001). Washington, DC: U.S. Environmental Protection Agency,
5 Risk Assessment Forum. <https://www.epa.gov/risk/benchmark-dose-technical-guidance>.
- 6 [U.S. EPA](#) (U.S. Environmental Protection Agency). (2014a). Guidance for applying quantitative data
7 to develop data-derived extrapolation factors for interspecies and intraspecies
8 extrapolation [EPA Report]. (EPA/100/R-14/002F). Washington, DC: Risk Assessment
9 Forum, Office of the Science Advisor. <https://www.epa.gov/sites/production/files/2015-01/documents/ddef-final.pdf>.
- 10
- 11 [U.S. EPA](#) (U.S. Environmental Protection Agency). (2014b). Scoping and problem formulation for the
12 identification of potential health hazards for the Integrated Risk Information System (IRIS)
13 toxicological review of ethylbenzene [CASRN 100-41-4] [EPA Report]. (EPA/635/R-
14 14/198). Washington, DC. <http://nepis.epa.gov/Exe/ZyPURL.cgi?Dockey=P100L2B0.txt>.
- 15 [U.S. EPA](#) (U.S. Environmental Protection Agency). (2015). Peer review handbook [EPA Report] (4th
16 ed.). (EPA/100/B-15/001). Washington, DC: U.S. Environmental Protection Agency, Science
17 Policy Council. <https://www.epa.gov/osa/peer-review-handbook-4th-edition-2015>.
- 18 [U.S. EPA](#) (U.S. Environmental Protection Agency). (2017a). Guidance to assist interested persons in
19 developing and submitting draft risk evaluations under the Toxic Substances Control Act.
20 (EPA/740/R17/001). Washington, DC: U.S. Environmental Protection Agency, Office of
21 Chemical Safety and Pollution Prevention.
22 https://www.epa.gov/sites/production/files/2017-06/documents/tsca_ra_guidance_final.pdf.
- 23
- 24 [U.S. EPA](#) (U.S. Environmental Protection Agency). (2017b). IRIS assessment plan for ethylbenzene
25 [CASRN 100-41-4].
- 26 [U.S. EPA](#) (U.S. Environmental Protection Agency). (2018a). Chemistry Dashboard. Washington, DC.
27 Retrieved from <https://comptox.epa.gov/dashboard>
- 28 [U.S. EPA](#) (U.S. Environmental Protection Agency). (2018b). An umbrella Quality Assurance Project
29 Plan (QAPP) for PBPK models [EPA Report]. (ORD QAPP ID No: B-0030740-QP-1-1).
30 Research Triangle Park, NC.
- 31 [U.S. EPA](#) (U.S. Environmental Protection Agency). (2019a). ChemView [Database]. Retrieved from
32 <https://chemview.epa.gov/chemview>
- 33 [U.S. EPA](#) (U.S. Environmental Protection Agency). (2019b). CompTox Chemicals Dashboard
34 [Database]. Research Triangle Park, NC. Retrieved from
35 <https://comptox.epa.gov/dashboard>
- 36 [U.S. EPA](#) (U.S. Environmental Protection Agency). (2020a). ORD staff handbook for developing IRIS
37 assessments (public comment draft) [EPA Report]. (EPA/600/R-20/137). Washington, DC:
38 U.S. Environmental Protection Agency, Office of Research and Development, Center for
39 Public Health and Environmental Assessment.
40 https://cfpub.epa.gov/ncea/iris_drafts/recordisplay.cfm?deid=350086.
- 41 [U.S. EPA](#) (U.S. Environmental Protection Agency). (2020b). Umbrella quality assurance project plan
42 (QAPP) for dosimetry and mechanism-based models. (EPA QAPP ID Number: L-CPAD-
43 0032188-QP-1-2). Research Triangle Park, NC.
- 44 [U.S. EPA](#) (U.S. Environmental Protection Agency). (2021). CompTox chemicals dashboard.
45 Washington, DC. Retrieved from <https://comptox.epa.gov/dashboard>
- 46 [U.S. EPA](#) (U.S. Environmental Protection Agency). (2022). ORD staff handbook for developing IRIS
47 assessments [EPA Report]. (EPA 600/R-22/268). Washington, DC: U.S. Environmental
48 Protection Agency, Office of Research and Development, Center for Public Health and
49 Environmental Assessment.
50 https://cfpub.epa.gov/ncea/iris_drafts/recordisplay.cfm?deid=356370.

Protocol for the Ethylbenzene IRIS Assessment

- 1 [Wallace, L; Nelson, W; Ziegenfus, R; Pellizzari, E; Michael, L; Whitmore, R; Zelon, H; Hartwell, T;](#)
2 [Perritt, R; Westerdahl, D.](#) (1991). The Los Angeles TEAM Study: personal exposures, indoor-
3 outdoor air concentrations, and breath concentrations of 25 volatile organic compounds. J
4 Expo Anal Environ Epidemiol 1: 157-192.
- 5 [Wallace, LA; Pellizzari, ED; Hartwell, TD; Sparacino, C; Whitmore, R; Sheldon, L; Zelon, H; Perritt, R.](#)
6 (1987). The TEAM study: Personal exposures to toxic substances in air, drinking water, and
7 breath of 400 residents of New Jersey, North Carolina, and North Dakota. Environ Res 43:
8 290-307. [http://dx.doi.org/10.1016/S0013-9351\(87\)80030-0](http://dx.doi.org/10.1016/S0013-9351(87)80030-0).
- 9 [Welch, VA; Fallon, KJ; Gelbke, HP.](#) (2005). Ethylbenzene. In Ullmann's Encyclopedia of Industrial
10 Chemistry. http://dx.doi.org/10.1002/14356007.a10_035.pub2.
- 11 [Wolffe, TAM; Whaley, P; Halsall, C; Rooney, AA; Walker, VR.](#) (2019). Systematic evidence maps as a
12 novel tool to support evidence-based decision-making in chemicals policy and risk
13 management. Environ Int 130: 104871. <http://dx.doi.org/10.1016/j.envint.2019.05.065>.

APPENDIX A. ELECTRONIC DATABASE SEARCH STRATEGIES

Table A-1. Database search strategy

Search	Search strategy	Date and results
PubMed		
Chemical terms	(100-41-4[rn] OR "ethylbenzene"[tw] OR "Ethylbenzol"[tw] OR "4-Ethylphenetole"[tw] OR "Ethyl(benzene-d5)"[tw] OR "Ethyl-1,1-d2 benzene-d5"[tw] OR "Ethyl, 2-phenyl-"[tw] OR "α-Methyltoluene"[tw] OR "Phenylurethane"[tw] OR "Ethyl-d5-benzene"[tw] OR "Ethylbenzene-d10"[tw] OR "NCI-C56393"[tw] OR "NSC 406903"[tw] OR "Phenylethane"[tw] OR "UNII-L5I45M5G00"[tw] OR "Ethylbenzol"[tw] OR "Etilbenzene"[tw] OR "Etylobenzen"[tw] OR "HSDB 84"[tw] OR "EC 202-849-4"[tw] OR "EINECS 202-849-4"[tw] OR "Ethylbenzene"[tw] OR "Ethylbenzeen"[tw] OR "Aethylbenzol"[tw] OR "A13-09057"[tw] OR "CCRIS 916"[tw] OR "DA0700000"[tw] OR "Phenylethane"[tw] OR "C004912"[tw] OR "ethylbenzene"[tw]) NOT medline	Date: 4/22/2019 Results: 2,765 Batch: 31018
	(100-41-4[rn] OR "ethylbenzene"[tw] OR "Ethylbenzol"[tw] OR "4-Ethylphenetole"[tw] OR "Ethyl(benzene-d5)"[tw] OR "Ethyl-1,1-d2 benzene-d5"[tw] OR "Ethyl, 2-phenyl-"[tw] OR "α-Methyltoluene"[tw] OR "Phenylurethane"[tw] OR "Ethyl-d5-benzene"[tw] OR "Ethylbenzene-d10"[tw] OR "NCI-C56393"[tw] OR "NSC 406903"[tw] OR "Phenylethane"[tw] OR "UNII-L5I45M5G00"[tw] OR "Ethylbenzol"[tw] OR "Etilbenzene"[tw] OR "Etylobenzen"[tw] OR "HSDB 84"[tw] OR "EC 202-849-4"[tw] OR "EINECS 202-849-4"[tw] OR "Ethylbenzene"[tw] OR "Ethylbenzeen"[tw] OR "Aethylbenzol"[tw] OR "A13-09057"[tw] OR "CCRIS 916"[tw] OR "DA0700000"[tw] OR "Phenylethane"[tw] OR "C004912"[tw] OR "ethylbenzene"[tw]) AND ("2019/04/01"[PDAT] : "3000"[PDAT]) NOT medline	Date: 4/13/2020 Results: 180 Batch: 37652

Protocol for the Ethylbenzene IRIS Assessment

Search	Search strategy	Date and results
	(100-41-4[rn] OR "ethylbenzene"[tw] OR "Ethylbenzol"[tw] OR "4-Ethylphenetole"[tw] OR "Ethyl(benzene-d5)"[tw] OR "Ethyl-1,1-d2 benzene-d5"[tw] OR "Ethyl, 2-phenyl-"[tw] OR "α-Methyltoluene"[tw] OR "Phenylurethane"[tw] OR "Ethyl-d5-benzene"[tw] OR "Ethylbenzene-d10"[tw] OR "NCI-C56393"[tw] OR "NSC 406903"[tw] OR "Phenylethane"[tw] OR "UNII-L5I45M5G00"[tw] OR "Ethylbenzol"[tw] OR "Etilbenzene"[tw] OR "Etylobenzen"[tw] OR "HSDB 84"[tw] OR "EC 202-849-4"[tw] OR "EINECS 202-849-4"[tw] OR "Ethyl benzene"[tw] OR "Ethylbenzeen"[tw] OR "Aethylbenzol"[tw] OR "A13-09057"[tw] OR "CCRIS 916"[tw] OR "DA0700000"[tw] OR "Phenylethane"[tw] OR "C004912"[tw] OR "ethylbenzene"[tw])	Date: 11/13/2020 Results: 164
	(100-41-4[rn] OR "ethylbenzene"[tw] OR "Ethylbenzol"[tw] OR "4-Ethylphenetole"[tw] OR "Ethyl(benzene-d5)"[tw] OR "Ethyl-1,1-d2 benzene-d5"[tw] OR "Ethyl, 2-phenyl-"[tw] OR "α-Methyltoluene"[tw] OR "Phenylurethane"[tw] OR "Ethyl-d5-benzene"[tw] OR "Ethylbenzene-d10"[tw] OR "NCI-C56393"[tw] OR "NSC 406903"[tw] OR "Phenylethane"[tw] OR "UNII-L5I45M5G00"[tw] OR "Ethylbenzol"[tw] OR "Etilbenzene"[tw] OR "Etylobenzen"[tw] OR "HSDB 84"[tw] OR "EC 202-849-4"[tw] OR "EINECS 202-849-4"[tw] OR "Ethyl benzene"[tw] OR "Ethylbenzeen"[tw] OR "Aethylbenzol"[tw] OR "A13-09057"[tw] OR "CCRIS 916"[tw] OR "DA0700000"[tw] OR "Phenylethane"[tw] OR "C004912"[tw] OR "ethylbenzene"[tw]) AND (2020/11/01:3000[dp])	Date: 1/21/2022 Results: 232 Batch: 46084
Web of Science		
Chemical terms^a	TS=("100-41-4" OR "Benzene, ethyl-" OR "4-Ethylphenetole" OR "Ethyl(benzene-d5)" OR "Ethyl-1,1-d2 benzene-d5" OR "Ethyl, 2-phenyl-" OR "α-Methyltoluene" OR "Phenylurethane" OR "Ethyl-d5-benzene" OR "Ethylbenzene-d10" OR "NCI-C56393" OR "NSC 406903" OR "Phenylethane" OR "UNII-L5I45M5G00" OR "Ethylbenzol" OR "Etilbenzene" OR "Etylobenzen" OR "HSDB 84" OR "EC 202-849-4" OR "EINECS 202-849-4" OR "Ethyl benzene" OR "Ethylbenzeen" OR "Aethylbenzol" OR "A13-09057" OR "CCRIS 916" OR "DA0700000" OR "Phenylethane" OR "C004912" OR "ethylbenzene")	Date: 4/22/2019 Results: 1,585 Batch: 31051

Protocol for the Ethylbenzene IRIS Assessment

Search	Search strategy	Date and results
	TS=("100-41-4" OR "Benzene, ethyl-" OR "4-Ethylphenetole" OR "Ethyl(benzene-d5)" OR "Ethyl-1,1-d2 benzene-d5" OR "Ethyl, 2-phenyl-" OR "α-Methyltoluene" OR "Phenylurethane" OR "Ethyl-d5-benzene" OR "Ethylbenzene-d10" OR "NCI-C56393" OR "NSC 406903" OR "Phenylethane" OR "UNII-L5I45M5G00" OR "Ethylbenzol" OR "Etilbenzene" OR "Etylobenzen" OR "HSDB 84" OR "EC 202-849-4" OR "EINECS 202-849-4" OR "Ethyl benzene" OR "Ethylbenzeen" OR "Aethylbenzol" OR "A13-09057" OR "CCRIS 916" OR "DA0700000" OR "Phenylethane" OR "C004912" OR "ethyl-benzene") AND PY=(2019-2020)	Date: 4/13/2020 Results: 73 Batch: 37653
	TS=("100-41-4" OR "Benzene, ethyl-" OR "4-Ethylphenetole" OR "Ethyl(benzene-d5)" OR "Ethyl-1,1-d2 benzene-d5" OR "Ethyl, 2-phenyl-" OR "α-Methyltoluene" OR "Phenylurethane" OR "Ethyl-d5-benzene" OR "Ethylbenzene-d10" OR "NCI-C56393" OR "NSC 406903" OR "Phenylethane" OR "UNII-L5I45M5G00" OR "Ethylbenzol" OR "Etilbenzene" OR "Etylobenzen" OR "HSDB 84" OR "EC 202-849-4" OR "EINECS 202-849-4" OR "Ethyl benzene" OR "Ethylbenzeen" OR "Aethylbenzol" OR "A13-09057" OR "CCRIS 916" OR "DA0700000" OR "Phenylethane" OR "C004912" OR "ethyl-benzene")	Date: 4/13/2020 Results: 50

Protocol for the Ethylbenzene IRIS Assessment

Search	Search strategy	Date and results
	<p>(TI=("100-41-4" OR "Benzene, ethyl-" OR "4-Ethylphenetole" OR "Ethyl(benzene-d5)" OR "Ethyl-1,1-d2 benzene-d5" OR "Ethyl, 2-phenyl-" OR "α-Methyltoluene" OR "Phenylurethane" OR "Ethyl-d5-benzene" OR "Ethylbenzene-d10" OR "NCI-C56393" OR "NSC 406903" OR "Phenylethane" OR "UNII-L5I45M5G00" OR "Ethylbenzol" OR "Etilbenzene" OR "Etylobenzen" OR "HSDB 84" OR "EC 202-849-4" OR "EINECS 202-849-4" OR "Ethyl benzene" OR "Ethylbenzeen" OR "Aethylbenzol" OR "AI3-09057" OR "CCRIS 916" OR "DA0700000" OR "Phenylethane" OR "C004912" OR "ethyl-benzene") OR AB=("100-41-4" OR "Benzene, ethyl-" OR "4-Ethylphenetole" OR "Ethyl(benzene-d5)" OR "Ethyl-1,1-d2 benzene-d5" OR "Ethyl, 2-phenyl-" OR "α-Methyltoluene" OR "Phenylurethane" OR "Ethyl-d5-benzene" OR "Ethylbenzene-d10" OR "NCI-C56393" OR "NSC 406903" OR "Phenylethane" OR "UNII-L5I45M5G00" OR "Ethylbenzol" OR "Etilbenzene" OR "Etylobenzen" OR "HSDB 84" OR "EC 202-849-4" OR "EINECS 202-849-4" OR "Ethyl benzene" OR "Ethylbenzeen" OR "Aethylbenzol" OR "AI3-09057" OR "CCRIS 916" OR "DA0700000" OR "Phenylethane" OR "C004912" OR "ethyl-benzene") OR AK=("100-41-4" OR "Benzene, ethyl-" OR "4-Ethylphenetole" OR "Ethyl(benzene-d5)" OR "Ethyl-1,1-d2 benzene-d5" OR "Ethyl, 2-phenyl-" OR "α-Methyltoluene" OR "Phenylurethane" OR "Ethyl-d5-benzene" OR "Ethylbenzene-d10" OR "NCI-C56393" OR "NSC 406903" OR "Phenylethane" OR "UNII-L5I45M5G00" OR "Ethylbenzol" OR "Etilbenzene" OR "Etylobenzen" OR "HSDB 84" OR "EC 202-849-4" OR "EINECS 202-849-4" OR "Ethyl benzene" OR "Ethylbenzeen" OR "Aethylbenzol" OR "AI3-09057" OR "CCRIS 916" OR "DA0700000" OR "Phenylethane" OR "C004912" OR "ethyl-benzene")) AND DOP=2020-11-01/2022-01-30</p>	<p>1/21/2022 56 results Batch: 46083</p>

Protocol for the Ethylbenzene IRIS Assessment

Search	Search strategy	Date and results
Toxline		
Chemical terms	@AND+@OR+("Benzene, ethyl-"+"4-Ethylphenetole"+"Ethyl(benzene-d5) "+"Ethyl-1,1-d2 benzene-d5"+"Ethyl, 2-phenyl-"+"α-Methyltoluene"+"Phenylurethane"+"Ethyl-d5-benzene"+"Ethylbenzene-d10"+"NCI-C56393"+"NSC 406903"+"Phenylethane"+"UNII-L5I45M5G00"+"Ethylbenzol"+"Etilbenzene"+"Etylobenzen"+"HSDB 84"+"EC 202-849-4"+"EINECS 202-849-4"+"Ethylbenzene"+"Ethylbenzeen"+"Aethylbenzol"+"A13-09057"+"CCRIS 916"+"DA0700000"+"Phenylethane"+"C004912"+"ethyl-benzene"+@TERM+@rn+100-41-4)	Date: 4/22/2019 Results: 2,780
TSCATS		
Chemical terms	@AND+@OR+@rn+"100-41-4"+@AND+@org+TSCATS+@NOT+@org+pubmed	Date: 4/22/2019 Results: 245

^aThe search conducted on 1/21/2022 utilized an updated Web of Science search process. Previous searches used only the topic (TS) field tag, which searches title, abstract, author-keywords, and keywords Plus. The updated process searches title (TI), abstract (AB), and author-keywords (AK) tags filtering out references that only matched in the keywords plus that are WOS-generated keywords and typically are not relevant to assessments.

APPENDIX B. PROCESS FOR SEARCHING AND COLLECTING EVIDENCE FROM SELECTED OTHER RESOURCES

1 Review of the citation reference lists is typically done manually because they are not
2 available in a file format (e.g., RIS) that permits uploading into screening software applications.
3 Manual review entails scanning the title, study summary, or study details as presented in the
4 resource for those that appear to meet the PECO criteria. Any records identified that are not
5 identified from the other sources are formatted in an RIS file format, imported into DistillerSR,
6 annotated with respect to source, and screened as outlined in Section 3.2. For tracking assessments
7 or reviews, the name of the source citation and the number of records imported into DistillerSR are
8 noted. The reference list of any study included in the literature inventory is reviewed manually to
9 identify titles that appear relevant to the PECO criteria. These citations are tracked in a
10 spreadsheet, compared against the literature base to determine whether they are unique to the
11 project, and then added to DistillerSR to be screened at the title and abstract stage for PECO
12 relevance.

B.1 EPA COMPTOX CHEMICALS DASHBOARD (TOXVAL)

13 ToxVal is searched in the EPA CompTox Chemicals Dashboard ([U.S. EPA, 2018a](#)), and data
14 available from the “Hazard” tab is exported from the CompTox File Transfer Protocol site. Using
15 both the human health POD summary file and the Record Source file, citations are identified that
16 apply to human health PODs. A citation for each referenced study is generated in HERO and verified
17 that it is not already identified from the database search (or searches of “other sources consulted”)
18 prior to moving forward to screening in DistillerSR. Full texts are retrieved where possible; if full
19 texts are not available, data from the ToxVal dashboard are entered and the citation is annotated
20 accordingly for Tableau and HAWC visualizations by adding “(ToxVal)” to the citation.

B.2 EUROPEAN CHEMICALS AGENCY (ECHA)

21 A search of the ECHA registered substances database is conducted using the CASRN. The
22 registration dossier associated with the CASRN is retrieved by navigating to and clicking the eye-
23 shaped view icon displayed in the chemical summary panel. The general information page and all
24 subpages included under the Toxicological Information tab are downloaded in Portable Document
25 Format (PDF), including all nested reports having unique URLs. In addition, the data are extracted
26 from each dossier page and used to populate an Excel tracking sheet. Extracted fields include data

1 from the general information page regarding the registration type and publication dates, and on a
2 typical study summary page the primary fields reported in the administrative data, data source, and
3 effect levels sections. Each study summary results in more than one row in the tracking sheet if
4 more than one data source or effect level is reported.

5 At this stage, each study summary is reviewed for inclusion based on PECO criteria. Study
6 summaries identified as without administrative data information are excluded from review, and
7 study summaries labeled “read-across” (if any) are screened and considered supplemental material.
8 When a study summary considered relevant reports data from a study or lab report, a citation for
9 the full study is generated in HERO and verified that it was not already identified from the database
10 search (or searches of “other sources consulted”) prior to moving forward to screening. When
11 citation information is not available and a full text could not be retrieved, the generated PDF is used
12 as the full text for screening and extraction and the citation is annotated accordingly for Tableau
13 and HAWC visualizations by adding “(ECHA Summary)” to the citation.

B.3 EPA CHEMVIEW

14 The EPA ChemView database ([U.S. EPA, 2019a](#)) using the chemical CASRN is searched. The
15 prepopulated CASRN match and the “Information Submitted to EPA” output option filter are
16 selected before generating results. If results are available, the square-shaped icon under the “Data
17 Submitted to EPA” column is selected, and the following records are included:

- 18 • High Production Volume Challenge Database (HPVIS)
- 19 • Human Health studies (Substantial Risk Reports)
- 20 • Monitoring (includes environmental, occupational, and general entries)
- 21 • TSCA Section 4 (chemical testing results)
- 22 • TSCA Section 8(d) (health and safety studies)
- 23 • TSCA Section 8(e) (substantial risk)
- 24 • FYI (voluntary documents)

25 All records for ecotoxicology and physical and chemical property entries are excluded.
26 When results are available, extractors navigate into each record until a substantial risk report link
27 is identified and saved as a PDF file. If the report cannot be saved, due to file corruption or broken
28 links, the record is excluded during full-text review as “unable to obtain record.” Most substantial
29 risk reports contain multiple document IDs, so citations are derived by concatenating the unique
30 report numbers (OTS; 8EHD Num; DCN; TSCATS RefID; and CIS) associated with each document,
31 along with the typical author organization, year, and title. Once a citation is generated, the study

1 moves forward to DistillerSR where it is screened according to PECO and supplemental material
2 criteria.

B.4 NTP CHEMICAL EFFECTS IN BIOLOGICAL SYSTEMS

3 This database is searched using the chemical CASRN
4 (<https://manticore.niehs.nih.gov/cebssearch>). All non-NTP data are excluded using the “NTP Data
5 Only” filter. Data tables for reports undergoing peer review are also searched for studies that have
6 not been finalized (<https://ntp.niehs.nih.gov/data/tables/index.html>) based on a manual review of
7 chemical names.

B.5 OECD ECHEMPORTAL

8 The OECD eChemPortal (<https://hvpchemicals.oecd.org/UI/Search.aspx>) is searched using
9 the chemical CASRN. Only database entries from the following sources are included and entries
10 from all other databases are excluded in the search. Final assessment reports and other relevant
11 SIDS reports embedded in the links are captured and saved as PDF files.

- 12 • OECD HPV
- 13 • OECD SIDS IUCLID
- 14 • SIDS United Nations Environment Programme (UNEP)

B.6 ECOTOX DATABASE

15 EPA’s ECOTOX Knowledgebase (<https://cfpub.epa.gov/ecotox/search.cfm>) is searched
16 using the CASRN. Results are refined to terrestrial mammalian studies by selecting the terrestrial
17 tab at the top of the search page and sorting the results by species group. A citation for each
18 referenced study is generated in HERO and verified that it is not already identified from the
19 database search (or searches of “other sources consulted”) search prior to moving forward to
20 screening in DistillerSR.

B.7 EPA COMPTOX CHEMICAL DASHBOARD VERSION TO RETRIEVE A SUMMARY OF ANY TOXCAST OR TOX21 HIGH-THROUGHPUT SCREENING INFORMATION

21 Version 3.0.9 of the CompTox Chemicals Dashboard ([U.S. EPA, 2019b](#)) is accessed for
22 high-throughput screening (HTS) data by searching the Dashboard by CASRN. Next, the
23 “Bioactivity” section is selected and the availability of ToxCast/Tox21 HTS data for active and
24 inactive assays is examined in the “TOXCAST: Summary” tab. If active assays are reported, the
25 figure is copied for presentation in the systematic evidence map. This figure presents (1) a
26 scatterplot of scaled assay responses versus AC50 values for each active assay endpoint and (2) a

1 cytotoxicity limit as a vertical line. More detailed information on the results of ToxCast and Tox21
 2 assays are available in the CompTox Chemicals Dashboard section “ToxCast/Tox21,” which includes
 3 chemical analysis data, dose-response data and model fits, and “flags” assigned by an automated
 4 analysis, which might suggest false positivity/negativity or indicate other anomalies in the data.
 5 This information is not summarized further for the purposes of the systematic evidence map, which
 6 is focused on identifying the extent of available evidence.

B.8 ETHYLBENZENE GRAY LITERATURE SEARCH SUMMARY

7 **Dates Run:** All gray literature searches were conducted in 2020 (between 11/1/2020-
 8 12/1/2020) and on 1/21/2022.

9 **Search Limits:** No date limits were applied to the gray literature search.

10 **Search Terms:**

- 11 • CASRN: 100-41-4
- 12 • "EC 202-849-4"
- 13 • "ethylbenzene"
- 14 • "1-ethylbenzene"

15 **Sources Searched:** The following sources were searched:

- 16 • **ECHA Registration Dossiers**
- 17 • **ChemView**
- 18 • **OECD eChem Portal**
- 19 • **NTP Chemical Effects In Biological Systems (CEBS)**
- 20 • **EPA ToxVal** – Searched using internal data files provided by CCTE
- 21 • **EPA ECOTOX**

Table B-1. Summary table for ethylbenzene other sources search results (12/2021)

Source	Search method	Total results retrieved (2020)	Total results retrieved (2022)	Unique results
ECHA	Automated Webscraping	359	3	60
ChemView	Manual Searching	23	0	5
OECD eChem Portal	Manual Searching	2	0	0
CEBS	Manual Searching	1	0	1

Protocol for the Ethylbenzene IRIS Assessment

Source	Search method	Total results retrieved (2020)	Total results retrieved (2022)	Unique results
ToxVal	Manual Searching	83	–	14
ECOTOX	Manual Searching	–	3	0
Total	N/A	468	6	80

CEBS = Chemical Effects in Biological Systems; ECHA = European Chemicals Agency; NA = not applicable;
OECD = Organisation for Economic Co-operation and Development.