Systematic Review and Evidence Integration for 16 Perfluoroalkyl and Polyfluoroalkyl Substances (PFAS)

Prepared by
Janet Petruska Hamilton, Ph.D., D.A.B.T.
Allison Jenkins, M.P.H.
Toxicology, Risk Assessment, and Research Division

Draft, 2025

TEXAS COMMISSION ON ENVIRONMENTAL QUALITY

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Acronyms and Abbreviations

Acronyms and Abbreviations	Definition
ACGIH	American Conference of Governmental Industrial Hygienists
ADME	absorption, distribution, metabolism and excretion
AI	artificial intelligence
ATSDR	Agency for Toxic Substances and Disease Registry
CASRN	Chemical Abstracts Service Registry Number
DSD	development support document
ESL	Effects Screening Level
GLP	Good Laboratory Practice
HQ	hazard quotient
HSDB	Hazardous Substance Data Base
IARC	International Agency for Research on Cancer
IPCS	International Programme on Chemical Safety
IRIS	USEPA Integrated Risk Information System
kg	kilogram
K_{ow}	n-octanol-water partition coefficient
LC_{50}	concentration causing lethality in 50% of test animals
LD_{50}	dose causing lethality in 50% of test animals
LOAEL	lowest-observed-adverse-effect-level
LOEL	lowest-observed-effect-level
MOA	mode of action
N	number
N/A, NA	not available
NOAEL	no-observed-adverse-effect-level
NR	not reported
NRC	National Research Council
NTP	National Toxicology Program

Acronyms and Abbreviations	Definition
OECD	Organization for Economic Co-operation and Development
OPPT	Office of Pollution Prevention and Toxics
OSHA	Occupational Safety and Health Administration
PBPK	physiologically-based pharmacokinetic
PECO	Population, Exposure, Comparator, Outcome
PFAS	perfluoroalkyl and polyfluoroalkyl substances
PFBA	perfluorobutanoic acid
PFBS	perfluorobutanesulfonic acid
PFDA	perfluorodecanoic acid
PFDoA	perfluorododecanoic acid
PFDS	perfluorodecane sulfonate
PFHpA	perfluoroheptanoic acid
PFHxA	perfluorohexanoic acid
PFHxS	perfluorohexanesulfonic acid
PFNA	perfluorononanoic acid
PFOA	perfluorooctanoic acid
PFOS	perfluorooctanesulfonic acid
PFOSA	perfluorooctanesulfonamide
PFPeA	perfluoropentanoic acid
PFTeDA	perfluorotetradecanoic acid
PFTrDA	perfluorotridecanoic acid
PFUnA	perfluoroundecanoic acid
PK	pharamacokinetics
POD	point of departure
QC	quality control
ReV	reference value
RfD	reference dose
SEM	systematic evidence map

Acronyms and Abbreviations	Definition
SFo	oral slope factor
SR	systematic review
TCEQ	Texas Commission on Environmental Quality
TD	Toxicology, Risk Assessment, and Research Division
USEPA	United States Environmental Protection Agency

Chapter 1 Introduction

1.1 Purpose of Systematic Review for 16 PFAS

In June 2011, the Texas Commission on Environmental Quality (TCEQ) derived reference doses (RfDs) for 16 perfluoroalkyl and polyfluoroalkyl substances (PFAS). Later, the TCEQ updated these values to reflect changes in the 2012 revision of the TCEQ Guidelines to Develop Toxicity Factors (RG-442), which were subsequently updated again in 2015 (TCEQ 2015). The most recent revision of some of these toxicity factors was completed in February 2023 (available at: https://www.tceq.texas.gov/downloads/toxicology/pfc/pfcs.pdf). The 16 PFAS are shown in Table 1.

A systematic review was not conducted previously for these 16 PFAS and new studies were published subsequent to the development of toxicity factors for the PFAS of interest. Therefore, the purpose of this systematic review was to re-evaluate the relevant literature available for these 16 PFAS with the goal of supporting the derivation of relevant oral toxicity factors, including RfDs, in development support documents (DSDs). A further goal was to support the derivation of inhalation toxicity factors (reference values, ReVs; effects screening levels, ESLs) if inhalation toxicity data were available for any of these 16 PFAS. This project began in 2021 and the first step was defining the body of toxicology and epidemiology literature by development of a systematic evidence map (SEM). This effort resulted in a full SEM to support the development of potential toxicity factors for the 16 PFAS compounds, where data were appropriate to do so. The protocol for the SEM is provided in Appendix A. The SEM was provided as an Excel file. Using the body of evidence identified by the SEM, a more focused systematic review (SR) was initiated to inform the development of toxicity factors in TCEQ DSDs. The protocol for the SR is provided in Appendix B. The specific methods used to identify, extract, and apply quality evaluation criteria to the relevant data for these 16 PFAS compounds are described in this report. The data obtained from this effort were compiled into an Excel file.

Table 1. List of 16 PFAS compounds evaluated by TCEQ

PFAS Compound Name	CASRN	PFAS Acronym used by TCEQ
Perfluorobutanoic acid	375-22-4	PFBA
Perfluorobutanesulfonic acid	375-73-5	PFBS
Perfluoropentanoic acid	2706-90-3	PFPeA
Perfluorohexanoic acid	307-24-4	PFHxA
Perfluorohexanesulfonic acid	355-46-4	PFHxS
Perfluoroheptanoic acid	375-85-9	PFHpA
Perfluorooctanoic acid	335-67-1	PFOA
Perfluorooctanesulfonic acid	1763-23-1	PFOS
Perfluorooctanesulfonamide	754-91-6	PFOSA

PFAS Compound Name	CASRN	PFAS Acronym used by TCEQ
Perfluorononanoic acid	375-95-1	PFNA
Perfluorodecanoic acid	335-76-2	PFDA
Perfluorodecane sulfonate	67906-42-7	PFDS
Perfluoroundecanoic acid	2058-94-8	PFUnA
Perfluorododecanoic acid	307-55-1	PFDoA
Perfluorotridecanoic acid	72629-94-8	PFTrDA
Perfluorotetradecanoic acid	376-06-7	PFTeDA

Abbreviations: CASRN, Chemical Abstracts Service Registry Number; TCEQ, Texas Commission on Environmental Quality

In support of developing toxicity factors, it is standard practice for the TCEQ to review all available relevant data for a particular chemical. Based on the database, a toxicologist then identifies the critical effect that occurs at the lowest human equivalent dose or concentration (TCEQ 2015). The critical effect is the basis for the development of a given toxicity factor.

1.2 Problem Formulation

During the problem formulation phase of the SR, the project team was defined, and the SR workflow was developed based on the final goal of supporting the development of toxicity factors for 16 PFAS chemicals in TCEQ DSDs. These included developing a literature identification strategy, inclusion and exclusion criteria, data extraction process, study evaluation considerations, and piloting and reviewer calibration. These aspects were defined *a priori* and documented in the SEM protocol (Appendix A) and SR protocol (Appendix B). This process was iterative in nature and involved a series of discussions as well as piloting and calibration exercises for full text screening, data extraction, and study quality evaluation. The purpose of the piloting and calibration exercises was to ensure consistency among the project team and to reduce response conflicts. Based on feedback from the exercises, preliminary screening, data extraction, and study quality templates were revised for clarity or to improve functionality.

During problem formulation, the project team selected DistillerSR, a web-based literature review software platform, as the primary software for execution of the workflow. DistillerSR allows each stage of the workflow to be completely customized via project-specific forms. This flexibility was of high importance due to the volume and type of information collected during the SR process.

Refinements to the inclusion and exclusion criteria were made based on results of piloting, to ensure that the SR process used was fit-for-purpose for TCEQ's dose-response assessment and toxicity factor development goals.

Chapter 2 Systematic Evidence Map

2.1 Literature Search

The initial objective was to develop a systematic map at the title/abstract level for use as a tool for characterizing the types of human, animal, and mechanistic studies available for each of the relevant PFAS compounds. A search syntax was developed to query the PubMed citation database. Initial drafting of syntax used PFAS compound names, acronyms, and CAS numbers as shown in Table 1. Additional synonyms and acronyms for PFBA, PFHxA, PFHxS, PFNA, and PFDA were added to the syntax based on USEPA's systematic review protocol for the IRIS assessments for each of these PFAS (anionic and acid forms, refer to IRIS documents cited in Chapter 5 References). For the remaining compounds, PubChem was consulted, and any available synonyms and acronyms were added to the search syntax. Initially, in order to perform a broad and comprehensive search, no additional terms restricting the query were added (e.g., outcome, species, route). This initial search syntax is included in the Systematic Map protocol (Appendix A: refer to Appendix A.A Identification of Additional PFAS Synonyms and Acronyms).

A pilot query using the concatenated search string was conducted in PubMed. This search generated > 3,000,000 results, because PubMed failed to recognize several PFAS synonyms that were included as quoted phrases. Subsequently, the search syntax was adjusted by removing several synonyms (refer to Appendix A.B.1 List of PFAS synonyms not recognized by PubMed and removed from search syntax). The finalized search syntax of May 19, 2021 is in the Systematic Map protocol (refer to Appendix A.B.2 Final syntax for literature search conducted on May 19, 2021, in PubMed). Search results were validated by comparison to approximately 50 primary publications from USEPA assessments of PFAS (USEPA 2016a, USEPA 2021a). All publications used in validation that are also indexed in PubMed were identified by the final syntax. References cited by USEPA that are unpublished or not indexed by PubMed were not returned by the literature search.

The final search syntax used to query PubMed on May 19, 2021 returned 6,932 results. The references were imported into DistillerSR and duplicated references were removed, eliminating one duplicate. The title and abstract of the remaining 6,931 references were screened as described in the following section.

2.2 Title/Abstract Literature Screening

Title/abstract screening of each reference was performed by two reviewers using DistillerSR. Each reference was screened based on set inclusion and exclusion criteria and subsequently was categorized. Due to the large number of references in the search results, DistillerSR's Artificial Intelligence (AI) model was used. During screening, AI text mining automatically ranked and prioritized unreviewed references. Manually reviewed references served as a training set for the AI screener. Once an appropriate threshold of expected included studies was met (i.e., 99%), the AI screener was used to exclude irrelevant results based on the

manually reviewed training set. For quality control, 10% of the excluded references were then reviewed to confirm accuracy of the AI model.

The inclusion and exclusion criteria used in the SEM are shown in Table 2.

Table 2. Inclusion and exclusion criteria used in the systematic map of 16 PFAS compounds

Category	Include	Exclude
Population	• Human (epidemiological or biomonitoring) • *In vivo* experimental animal • *In vitro*/mechanistic	• Models irrelevant to human health outcomes (e.g., models of ecotoxicity)
Exposure	• Investigates at least one of the PFAS in Table 1 (also refer to Table A- 1. List of sixteen PFAS compounds evaluated)	• No PFAS compound of interest (other PFAS compounds investigated but not one of the 16 of interest) • No chemical of interest (no PFAS compounds investigated) • PFAS detection in other media (e.g., sewage sludge or wastewater) • Studies on treatment following PFAS exposure, unless PFAS was also tested alone
Outcomes	• Health outcome (e.g., apical outcomes) • Toxicokinetics: ADME, PK, PBPK • Mechanistic • Point of departure: NOAEL, NOEL, LOAEL, LOEL, BMD • *In silico*/computational model • Biomonitoring	• Phytotoxicity • Ecotoxicity
Reference type	• Primary references • Reviews: includes relevant risk assessments, meta-analyses, systematic reviews	• Opinion pieces, commentaries, letters to the editor, etc.

Abbreviations: ADME, absorption, distribution, metabolism and excretion; BMD, benchmark dose; NOAEL, no-observed-adverse-effect level; NOEL, no-observed-effect level; LOAEL, lowest-observed-adverse-effect level; LOEL, lowest-observed-effect level; PBPK, physiologically-based pharmacokinetic; PK, pharmacokinetics

At the title/abstract screening stage, the following question was answered with one of four options using a form in DistillerSR:

Is this study potentially relevant for categorization?

- Yes
- No
- No abstract
- Unclear

Beyond the studies not meeting inclusion criteria above, as modified from Schaefer and Myers (2017), studies published in languages other than English also were excluded. When no abstract was available and the title was in English, references were designated as "No abstract" and also were excluded. Corrections and responses to articles were also excluded but were retrieved as needed once the final set of papers was determined.

During the SEM title/abstract screening, if a study was considered relevant by one reviewer and irrelevant by the other reviewer (i.e., an inclusion/exclusion conflict), an attempt at resolution was made by discussion between the two reviewers. If the two screeners did not come to a resolution, a third reviewer was added to resolve the conflict in order to reach a final conclusion. During the title/abstract screening, 4,543 references were excluded, and another 90 references had no abstract, and thus, were not advanced to the next step. The remaining 2,298 references that were included based on the criteria in Table 2 were then further categorized.

2.3 Categorization of References

Included references were further categorized by species, outcomes, duration, and route. If information was missing or the abstract language was difficult to interpret, then "unclear" may have been selected during categorization. For every study type, all PFAS compounds of interest reported in the reference were recorded. The following categories were specified in DistillerSR forms for included references:

Species:

- Human
- Experimental animals
    - Rat
    - Mouse
    - Other mammal (e.g., rabbit, guinea pig, hamster, dog, pig, primate)
    - Non-mammalian (e.g., zebrafish, chicken)

Human studies were categorized further as follows:

- Epidemiology (observational studies)
- Clinical (controlled trials)
- Biomonitoring

- Mechanistic
- Toxicokinetics
- *In silico*/computational model

In vivo experimental animal studies were categorized further by study duration and route of exposure as follows:

Study duration (duration of administration of test chemical):

- Acute (24 hours or less)
- Subacute (1 – 30 days)
- Subchronic (1 – 3 months; e.g., 90-day study)
- Chronic (greater than 3 months; e.g., 2-year bioassay)

Route of exposure:

- Oral/gavage
- Dermal
- Inhalation
- Intraperitoneal/intravenous/subcutaneous
- Immersion

Conflicts of categorization of study types were prioritized based on the goal of development of toxicity factors; therefore, conflicting categorizations in the evidence stream and *in vivo* study information categories were prioritized. If conflicts occurred for categories of lower importance for developing toxicity factors (e.g., *in vitro* studies), both categories were included in the map.

Figure 1 illustrates the literature search and screening process for the SEM.

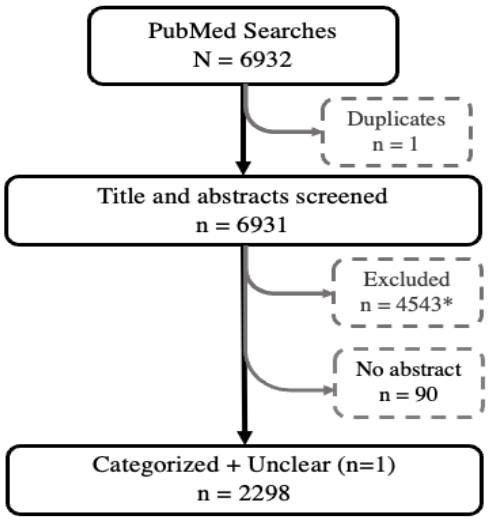

Figure 1 Flow chart of the literature search and screening process for the systematic evidence map.

Solid black lines and boxes indicate literature advancing to the next stage, gray dashed lines indicate literature that was excluded and not advanced to the next stage. * Of the excluded references, 1,187 were screened out by artificial intelligence (AI) with no second reviewer. For quality control, 10% of these were reviewed to confirm the accuracy of the AI model.

2.3.1 PFAS Compounds

After categorization, the greatest numbers of references were available for PFOA (N = 1,538) and PFOS (N = 1,446), followed by PFHxS (N = 528), PFNA (N = 515), and PFDA (N = 365). The fewest number of references (N = 10) were for PFDS. There were 588 references that were categorized as "other" because the authors made generic statements in the title and/or abstract on which compound was evaluated (e.g., 6 PFAS compounds), or PFAS compounds outside of the sixteen of interest were included in the analysis. Table 3 shows the number of references in which a relevant study was reported for each PFAS.

Table 3. Numbers of references reporting a relevant investigation for each PFAS

PFAS Compound Name	PFAS Acronym used by TCEQ	Number of References
Perfluorobutanoic acid	PFBA	61
Perfluorobutanesulfonic acid	PFBS	129
Perfluoropentanoic acid	PFPeA	30
Perfluorohexanoic acid	PFHxA	102
Perfluorohexanesulfonic acid	PFHxS	528
Perfluoroheptanoic acid	PFHpA	76
Perfluorooctanoic acid	PFOA	1,538

PFAS Compound Name	PFAS Acronym used by TCEQ	Number of References
Perfluorooctanesulfonic acid	PFOS	1,446
Perfluorooctanesulfonamide	PFOSA	79
Perfluorononanoic acid	PFNA	515
Perfluorodecanoic acid	PFDA	365
Perfluorodecane sulfonate	PFDS	10
Perfluoroundecanoic acid	PFUnA	172
Perfluorododecanoic acid	PFDoA	88
Perfluorotridecanoic acid	PFTrDA	35
Perfluorotetradecanoic acid	PFTeDA	12
Other PFAS compounds	---	588

2.3.2 Evidence Stream

The references also were categorized according to evidence stream. Human studies were the most frequently identified (N = 1,047). Of interest to the potential development of toxicity values, 727 *in vivo* experimental animal studies were identified across all 16 PFAS compounds. Thirty-three of these *in vivo* studies reported points of departure. There were 507 *in vitro* studies. Review studies accounted for 130 references. Five studies (*ex vivo* or *in ovo*) were identified as "other" studies.

Types of information (i.e., mechanistic data, point of departure) also were categorized (Table 4). The majority of the human studies reported biomonitoring data (N = 896, 85.5%). Epidemiological studies made up 70% (N = 736) of the human study references. As most of the epidemiological studies included biomonitoring data as a part of their analysis, there is significant overlap between these two fields. Specifically, 614 studies reported both epidemiological and biomonitoring data. Clinical studies were the least frequent human study type (N = 12).

Table 4. Type of information reported by *in vitro*, *in vivo*, and human studies

	***In vitro* only**	***In vivo* only**	***In vitro* and *in vivo***	**Human**
Health outcome	6	373	56	N/A [a]
Mechanistic	372	428	86	34
Toxicokinetics/ADME	54	149	43	153
Point-of-departure	1	32	1	0
In silico/computational	28	21	5	59

	***In vitro* only**	***In vivo* only**	***In vitro* and *in vivo***	**Human**
Epidemiology	N/A	N/A	N/A	736 [b]
Clinical	N/A	N/A	N/A	12
Biomonitoring	N/A	N/A	N/A	896

Abbreviations: ADME, absorption, distribution, metabolism, and excretion; N/A, not applicable
a: The human studies did not include a separate category entitled "health outcome" because they generally reported health outcomes and were subsequently reviewed during the full text review.
b: The majority of epidemiology studies included biomonitoring data (N = 614), and therefore, also were captured under the "biomonitoring " study category.

In vivo and *in vitro* studies also were categorized for species studied. For *in vitro* studies, human cell lines were the most frequently evaluated, followed by rat and mouse. *In vitro* models categorized as "other mammal" or "non-mammalian" included studies conducted in Chinese hamster ovary cells and *Salmonella typhimurium*, respectively. Several publications (N = 83) did not provide enough detail in the title or abstract to determine the species of the cell line or model and, therefore, were designated as "unclear".

For *in vivo* studies rats were the most common species evaluated in the relevant PFAS literature (N = 301), followed by mice (N = 277), non-mammals (N = 158), and "other mammal" (N = 28). As with *in vitro* studies, "other mammal" and "non-mammalian" categories only included species potentially relevant to human health endpoints, such as rabbits and zebrafish, respectively.

When adequately described in the title and/or abstract, *in vivo* studies also were categorized according to study duration and route of administration. Short-term studies (> 1 day but < 1 month in duration) were the most frequently reported duration (N = 390). Acute studies (≤ 24 h) accounted for 95 studies, while subchronic (1 – 3 months) and chronic (> 3 months) accounted for 72 and 60 studies, respectively.

Oral studies accounted for 340 references, while inhalation studies were described in five references. Other routes of exposure studied included immersion (N = 109), intraperitoneal/intravenous/subcutaneous (N = 81), dermal (N = 7), and other (N = 66). Studies frequently categorized as "other" included *in utero* and lactational exposures in developmental studies; these were commonly identified as oral exposure and it was also documented whether the dams were evaluated for adverse effects.

When the title and/or abstract did not clearly describe the study duration or route of administration, the publication was designated as "unclear" (N = 256 and N = 268 for study duration and route of administration, respectively).

Chapter 3 Systematic Review

3.1 Literature Identification

Literature identification was based on title and abstract mapping reported by the SEM with application of the Population, Exposure, Comparator, Outcome (PECO) statement to the categories depicted in Table 5 and subsequently advanced to the SR.

Table 5. Category labels assigned during the SEM process required to advance studies in each evidence stream to the SR

Evidence Stream	Study Type	Species	Data Type	Route
Experimental Animal	In vivo	Rat OR Mouse OR Other mammal	Health Outcome	Oral/Gavage OR Inhalation OR Unclear OR other
Human	Epidemiology OR Clinical	N/A	N/A	N/A

Abbreviations: N/A, not applicable; SEM, systematic evidence map; SR, systematic review

To ensure that a comprehensive review was conducted, screening results that were advanced to the SR were compared to publications identified in other sources including USEPA (USEPA 2016a, USEPA 2016b, USEPA 2021a, USEPA 2021b, USEPA 2021c), the Agency for Toxic Substances and Disease Registry (ATSDR 2021), and a PFAS-Tox Database (Pelch et al. 2021). The SEM yielded 1,048 references that were advanced to a full text review. Following the literature search and identification, two critical National Toxicology Program (NTP) reports (Toxicity Reports 96 and 97) that were first published in August 2019 were revised and the final versions published in July 2022. The body of literature was updated to include the revised, finalized versions for both reports.

3.2 Full Text Screening

Because the literature was identified within the SEM, it was assumed that all 1,048 references met the inclusion criteria at the title and abstract stage. Therefore, references were evaluated for relevance at the full-text level by two reviewers in DistillerSR, using project-specific forms developed and piloted during problem formulation. The inclusion/exclusion criteria used for this evaluation are provided in Table 6. The form used for this evaluation is shown in Appendix C.

Table 6. Inclusion and exclusion criteria used in the SR of 16 PFAS compounds

Category	Include	Exclude
Population	• Human • Mammalian experimental animals	• Non-mammalian models and mammalian models irrelevant to human health outcomes (e.g., models of ecotoxicity, livestock)

Exposure	• Investigates at least one of the 16 PFAS listed in Table 1 or salts/ions thereof
	• Oral or inhalation routes of exposure
	• Exposure is based on external dose/exposure
	• Exposure involves quantitative measurement or estimation informed by a measurement
	  ○ Observational human studies: exposure estimates informing exposure-response relationship must have some based on reported quantitative measurements
	• If internal exposure dose metric (e.g., serum PFAS) is used to evaluate outcome association, it is characterized over the duration leading up to outcome observation (e.g., career or lifetime cumulative serum PFAS level)
	  ○ E.g., cumulative serum PFAS estimates modeled on worker biomonitoring data and job exposure matrix (JEM) or some other historical understanding of exposure activities
		• No PFAS compound of interest (other PFAS compounds investigated but not one of the 16 of interest)
		• No chemical of interest (no PFAS compounds investigated)
		• Inadequate exposure data for risk assessment for quantitative characterization of exposure response relationship
		  ○ E.g., observational human studies that do not include measurement of exposure, such as qualitative estimates based solely on job history without environmental or internal exposure measurements
		• Exposure estimates based solely on a single collection of biomonitoring samples, such as serum measurements
		• PFAS concentration data (e.g., detection in media such as sewage sludge or wastewater)
		• Exposure to more than one chemical/compound unless one study group was PFAS only exposure

Outcomes	• Adverse health outcomes (e.g., apical outcomes) as defined in Section 3.6 of TCEQ guidance [a]	• Mechanistic endpoints (e.g., gene expression data, enzyme activity) • Human vaccine studies [b] – effects of PFAS exposure on antibody concentration
Reference type	• Primary references	• Opinion pieces, commentaries, letters to the editor, etc. • Review: includes risk assessments, meta-analyses, systematic reviews

Category	Include	Exclude
Study model/design	• Epidemiological study designs such as cohort studies, case-control studies • *In vivo* study designs	• Case studies or case series • Ecological studies since there are no individual data (i.e., rely on population-level exposure or outcome data and report associations between exposure and outcome at the population level) • Cross-sectional studies since temporal association cannot be established (e.g., National Health and Nutrition Examination Survey [NHANES]) • Clinical trials • Human biomonitoring studies • *In vitro* study models
Additional criteria	• English translation available • Quantitative data (dose response or pairwise significance) available for at least one outcome of interest	

Abbreviations: PFAS, per- and polyfluoroalkyl substances; SR, systematic review
a: The focus of the review is on adverse effects per TCEQ Guidelines to Develop Toxicity Factors (TCEQ 2015). For outcomes where potential adversity is less obvious or unclear, the project team was more inclusive than not.
b: There is debate within the scientific community about whether the endpoint represented an adverse effect. It was decided to set these studies aside and to evaluate them separately.

For studies excluded at full text screening, one or more exclusion reasons were documented. Each full text was reviewed by two screeners. When screening results were considered relevant by one screener and irrelevant by the other screener (i.e., an inclusion/exclusion conflict), an attempt was made for resolution by discussion between the two screeners. If the screeners did not come to a resolution, a third reviewer was added to resolve the conflict in order to reach a conclusion.

Of the 1,048 references subjected to full text screening, 753 references were excluded as they did not meet the inclusion criteria in Table 6. The remaining 295 references were brought forward to the follow-up study evaluation step.

3.3 Follow-up Study Evaluation

Each experimental animal and epidemiology study that met the inclusion criteria was evaluated to determine whether any follow-up or linked studies were available. All linked/follow-up studies were evaluated by two reviewers to determine if they met inclusion criteria. If both studies met the inclusion criteria, both initial and follow-up/linked studies were evaluated together to determine whether articles represented stand-alone datasets that should be extracted separately and articles that contained overlapping or interrelated data from the same

study that should be extracted as a single study. When multiple articles were evaluated as a single study, the study quality was evaluated for the study itself, rather than according to the individual publication.

Some publications contained more than one PFAS studied. If separate studies were conducted and reported in the same publication (e.g., the 28-day oral gavage studies sponsored by the National Toxicology Program, Toxicity Reports 96 and 97), separate data extractions were done for each study.

3.4 Data Extraction

Data extraction was focused on gathering information pertinent to derivation of toxicity factors as outlined in the TCEQ guidelines (TCEQ 2015). For each reference that met inclusion criteria at the full-text review stage, details of study information, experimental design, and results were compiled using project-specific form in DistillerSR. The data extracted depended upon the evidence stream reported by the study. The following information was included in each row of the DistillerSR extraction table:

1) PFAS compound
2) Study duration and species (experimental animal studies)
3) Cohort or sub-group (epidemiology studies)
4) Endpoint/outcome category (experimental animal studies)
5) A single lowest-observed-adverse-effect level (LOAEL) and no-observed-adverse-effect level (NOAEL)

Studies with no statistically significant findings were not captured in the data extraction table, unless there was compelling evidence that the findings were of biological significance (e.g., histopathology findings denoted as significant but where statistical significance of the finding was not quantified; 10% decrease in body weight for adult animals; 5% decrease in body weight for juvenile animals). Mechanistic data were not extracted, but information regarding mechanistic data could be included in the reviewer notes. A total of 15 references (7 human studies, 8 experimental animal studies) did not report statistically significant findings and, therefore, data extractions were not performed for these studies.

General information regarding the data extraction is included in Table 7. Details for data extraction for experimental animal studies and epidemiology studies are in Table 8 and Table 9, respectively. Each data extraction was performed by a single reviewer, which was subsequently checked for quality control (QC) by a second reviewer, and any necessary corrections or revisions were made.

Table 7. Information compiled at the data extraction phase for experimental animal and human studies

Information category	Experimental animal	Human
General	• Study objective • Good laboratory practice (GLP)/guideline compliance	• Study objective • Study design (e.g., cohort, case-control, cross-sectional) • Study date
Population	• Species/strain • Number per group, sex, age	• Name of study/cohort • Population size (N) • Population description (e.g., age, location, pregnancy status, etc.)
Exposure	• PFAS compound • Route of exposure • Dose concentration(s) and frequency • Exposure duration	• PFAS compound • Exposure scenario (e.g., self-reported, serum concentration) • Exposure concentrations
Outcome	• Outcome category • Endpoint information (effect description, direction of change, etc.)	• Outcome category • Endpoint
Values	• No-observed-adverse-effect level (NOAEL)	• Risk ratio, odds ratio, hazard ratio, etc.

Table 8. Guidance for data extraction of experimental animal studies

Field	Guidance

N/sex/age	• For non-reproductive and developmental studies:
	  ○ N = number of animals evaluated per endpoint
	  ○ Sex = sex of animals evaluated per endpoint
	  ○ Age at study initiation
	• For reproductive and developmental studies
	  ○ N = number of litters evaluated per endpoint category (may include F_1 generation)
	  ▪ If number of litters (or dams) evaluated are not provided by the authors, record NR and add the number of pups evaluated in the notes
	  ○ Sex = sex of animal exposed (probably the F_0 generation)
	  ○ Age = age of animal at study initiation (i.e., age at start of exposure)
	• If the number of animals evaluated per endpoint was not reported (but the number exposed/tested was provided in the study methods), put "NR" in this column and note the number tested in the notes field [a]
	• Age of animals at initiation of study (i.e., when dosing/exposure was initiated, which may be after a period of acclimatization)
	• For reproductive and developmental study designs where exposure occurred prior to birth, focus was on the parental exposures
	  ○ e.g., N = number of dams, sex = F (mated), age of dams if reported (or "adult" if stated, "NR" if not reported)

PFAS compounds	• include salts/ions of PFAS compounds ○ example: potassium PFOS
Dose/concentration, frequency of dosing and units	• list all doses evaluated in the study including the control, not just doses at which findings occurred, along with the metric unit (e.g., mg/kg body weight-d) • include the frequency of dosing • in studies that included recovery or satellite groups, report findings together with the main dose groups unless an endpoint was significantly affected during the recovery period, but the main group was not. If an endpoint was significantly affected during recovery (but not during the main exposure period), include a separate row with only the dose levels of the satellite groups.
Exposure duration and units	• provide duration of exposure for the entire study
Endpoint category	• select a single, most appropriate endpoint category • if no category was available, then "other" was selected and a new category was provided • select 1 endpoint category per table row • select "developmental" for all endpoints measured in offspring exposed via parents

Field	Guidance

Endpoint	• list statistically and/or biologically significant endpoint(s) per endpoint category [b]. This may include significance determined by two-way comparisons between control and dose group, or a trend test.
	  ○ Examples for endpoint categories:
	  ▪ General toxicity: Body weight, mortality, clinical observations, food/water consumption, gross observations
	  ▪ Developmental: Fetal variations and malformations, developmental markers (vaginal opening, preputial separation, first estrus, eye opening; includes all endpoints measured in offspring exposed via parents, with the exception of litter characteristics)
	  ▪ Litter characteristics: Litter size, viability, pup survival, sex per litter, litter weight (pup information prior to weaning)
	  ▪ Reproductive: % mated or littered, gestation length, andrology, estrus cyclicity, fertility, pregnancy outcome and reproductive endpoints, placental endpoints
	• Provide direction of change
	• Include all endpoints with same direction of change in one row
	• Example:
	  ○ <u>Row 1</u>
	  Endpoint group: organ weight
	  Endpoint: decreased thyroid, thymus weight
	  ○ <u>Row 2</u>
	  Endpoint group: organ weight
	  Endpoint: increased ovary weight

NOAEL/LOAEL	• Include NOAEL and/or LOAEL as defined by the data • Include relevant information that defined the value (e.g., sex, tissue) • When no NOAEL/LOAEL are available, “NA” was entered ○ This includes scenarios where only modeled results are available (e.g., benchmark dose modeling results) and effect levels are not reported • Units are not required (available in the dose column) • Used “unclear” when statistical significance was evaluated but the point of departure value was not apparent • When statistical significance was reported but there was a no/non-monotonic dose response, note the lowest dose where there is a statistically significant differences as the “LOAEL” and include a reviewer note about characterizing the dose response (e.g., lack of dose response, non-monotonic dose response, etc.)
Notes	• Add comments relating to the data captured in the table row • Adding information to this field is not required
QC Agreement with Extraction	• Yes – second reviewer has checked the data row and agrees • No - second reviewer has checked the data row and disagrees – this should trigger a discussion between the two reviewers that will eventually resolve this response to “yes”

a: The number of animals evaluated for an endpoint may differ from the number exposed/tested at the start of the study for several reasons, including the study design itself (e.g., splitting dose groups for assaying multiple endpoints) and unanticipated events (e.g., animal infection/loss).
b: A note may have been added when there was a statistically significant difference with apparent biological significance (e.g., significance for a single endpoint in the low-dose group only). Expert judgment was applied when including endpoints with biological but not statistical significance (e.g., close to, but not, statistically significant or a histopathology finding identified as key but without statistical significance and severity).

Table 9. Guidance for data extraction of epidemiology studies

Field	Guidance
Name of study/cohort	• List one cohort per row • List sub-groups only if statistically and/or biologically significant • Do not include the whole cohort if only the sub-group has statistical significance; in these instances, list the sub-group only and provide a note (per row) to state that the whole group was not significant
PFAS compound(s)	• List 1 PFAS compound per row
Exposure concentration	• Capture exposure concentration quartiles • Include mean exposure value as a note if desired
Endpoint category	• Select a single, most appropriate endpoint category • If no category is available, "other" was selected and a new category was provided • Select 1 endpoint category per table row
Endpoint	• List statistically and/or biologically significant endpoint(s) per endpoint category [a] • Provide direction of change • Include all endpoints with same direction of change in one row • Provide a brief description of the endpoint

a: A note may have been added when there was a statistically significant difference with apparent biological significance (e.g., significance for a single endpoint in the low-dose group only). Expert judgment was applied when including endpoints with biological but not statistical significance (e.g., close to, but not, statistically significant or an outcome that is deemed important).

The formats for data presentation for data extraction of experimental animal studies and human epidemiology studies are shown in Appendix D.

Table 10 shows the number of references for which data were extracted for each PFAS.

Table 10. Numbers of references with data extractions for each PFAS

PFAS Compound Name	PFAS Acronym used by TCEQ	Experimental Animal Studies	Human Studies
Perfluorobutanoic acid	PFBA	2	0
Perfluorobutanesulfonic acid	PFBS	5	0
Perfluoropentanoic acid	PFPeA	0	0
Perfluorohexanoic acid	PFHxA	5	0
Perfluorohexanesulfonic acid	PFHxS	8	1
Perfluoroheptanoic acid	PFHpA	1	0
Perfluorooctanoic acid	PFOA	110	16
Perfluorooctanesulfonic acid	PFOS	106	1
Perfluorooctanesulfonamide	PFOSA	0	0

PFAS Compound Name	PFAS Acronym used by TCEQ	Experimental Animal Studies	Human Studies
Perfluorononanoic acid	PFNA	20	2
Perfluorodecanoic acid	PFDA	10	1
Perfluorodecane sulfonate	PFDS	0	0
Perfluoroundecanoic acid	PFUnA	3	0
Perfluorododecanoic acid	PFDoA	9	0
Perfluorotridecanoic acid	PFTrDA	1	0
Perfluorotetradecanoic acid	PFTeDA	0	0

3.5 Study Quality Evaluation

Studies that met the inclusion criteria with at least one statistically significant outcome were evaluated for study quality and risk of bias using the study evaluation tool developed by USEPA and available on USEPA's Health Assessment Workspace Collaboration (HAWC) tool (Shapiro et al. 2018). A total of 279 references that had data extracted were then evaluated for study quality.

Study evaluation domains and related metrics were scored by a reviewer as Good (++), Adequate (+), Deficient (-), or Critically Deficient, with justification for the category. Following the categorization of each domain and metric, an overall study confidence rating of High (++), Medium (+), or Low (-) was determined. Each assessment was performed by a single reviewer followed by QC by a second reviewer.

The domains and metrics for study quality evaluation are shown in Appendix E.

Figure 2 illustrates the full text literature screening process and outcome.

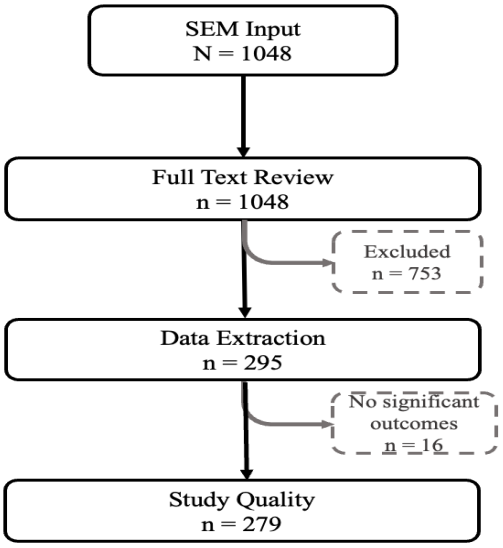

Figure 2 Flow chart of the literature search and screening process at the reference level

Chapter 4 Evidence Integration

Following the study quality evaluation, an Excel file was prepared that included the data extractions and study quality evaluations for all references brought forward to the data extraction and study quality evaluation steps.

Subsequently, the data extractions were sorted by PFAS and reviewed for the potential development of toxicity factors. Evidence integration was performed for each PFAS based on the TCEQ's Guidelines to Develop Toxicity Factors (2015) in the context of development of inhalation toxicity factors (reference values [ReVs], when inhalation studies were available) and oral toxicity factors (RfDs, oral slope factors [SFos]). Details of evidence integration are provided in the DSDs or other documentation for each PFAS. Additional information regarding evidence integration is described in the following sections.

For some PFAS, toxicity factors developed by other agencies (e.g., USEPA) were adopted by the TCEQ. These have been documented (Per- and Poly-fluoroalkyl Substances [PFAS], February 14, 2023; available at https://www.tceq.texas.gov/downloads/toxicology/pfc/pfcs.pdf). For example, the RfDs derived by USEPA for PFBA (USEPA, 2022) and PFHxA (USEPA, 2023) were adopted by the TCEQ. Also, there are several PFAS that do not have relevant experimental animal and/or human studies to support toxicity factor development. Therefore, the toxicity factors for these PFAS may be based on other PFAS, for which there were sufficient data to derive toxicity factors. For example, no relevant experimental animal or human studies are currently available for PFPeA; therefore, the RfD for PFPeA (a 5 carbon perfluorinated carboxylate) was surrogated to PFHxA (a 6 carbon perfluorinated carboxylate). For the other PFAS listed in Table 10 for which there are no relevant studies available, the toxicity factors will be surrogated to appropriate PFAS for which toxicity factors are derived based on relevant

studies, or may be derived independently if relevant experimental animal or human studies become available.

If relevant studies become available outside of the SR, these may be reviewed and if appropriate, used to inform derivation of toxicity factors, with associated documentation.

4.1 Oral Toxicity Studies Conducted in Animals

The majority of toxicology studies conducted in animals were oral studies. For some oral studies, oral doses were not included in the data extraction. These included oral studies in which the compound was admixed into the diet and the diet concentrations were included in the data extraction, or studies in which the compound was dissolved in water and the concentrations of the compound in water were included in the data extraction. Each reference was reviewed to see if the oral doses (mg/kg-d) were included in the publication. If the oral doses were included in the publication, the dose information was verified by a second toxicologist and then added to a copy of the Excel file. If the oral doses were not included in the publication, then the doses were calculated based on data in the publication, if available, or based on default values from USEPA (USEPA 1988), the National Research Council (NRC 1995), or from websites from which the animals were sourced (e.g., Charles River Laboratories website). The calculated oral doses (mg/kg-d) were verified by a second toxicologist, any necessary corrections were made, and then the doses were added to a copy of the Excel file.

4.2 Revised References

One of the references (NTP 2020) was revised in 2023. Although the overall conclusions did not differ from the original report, the revised report was used and cited in the relevant DSD (PFOA).

4.3 Acknowledgements

The TCEQ acknowledges the expertise, experience, and assistance of ToxStrategies in the protocol development and production of the systematic evidence map and systematic review of the literature for the 16 PFAS.

Chapter 5 References

Agency for Toxic Substances and Disease Registry (ATSDR). 2021. Toxicological profile for perfluoroalkyls. Available at: https://www.atsdr.cdc.gov/ToxProfiles/tp200.pdf

National Research Council. 1995. Nutrient requirements of laboratory animals. 4th ed. National Academies Press. Available from: https://nap.nationalacademies.org/catalog/4758/nutrient-requirements-of-laboratory-animals-fourth-revised-edition-1995

National Toxicology Program (NTP). 2019 and revised 2022. NTP Technical report on the toxicity studies of perfluoroalkyl sulfonates (perfluorobutane sulfonic acid, perfluorohexane sulfonate potassium last, and perfluorooctane sulfonic acid) administered by gavage to Sprague Dawley (Hsd:Sprague Dawley SD) rats (revised) Research Triangle Park, NC: National Toxicology Program. Toxicity Report 96.

National Toxicology Program (NTP). 2019 and revised 2022. NTP Technical report on the toxicity studies of perfluoroalkyl carboxylates (perfluorohexanoic acid, perfluorooctanoic acid, perfluorononanoic acid, and perfluorodecanoic acid) administered by gavage to Sprague Dawley (Hsd:Sprague Dawley SD) rats (revised) Research Triangle Park, NC: National Toxicology Program. Toxicity Report 97.

National Toxicology Program. (NTP). 2020 and revised 2023. NTP Technical report on the toxicology and carcinogenesis studies of perfluorooctanoic acid (CASRN 335-67-1) administered in feed to Sprague Dawley (Hsd: Sprague Dawley® SD®) rats (revised). Research Triangle Park, NC: National Toxicology Program. Technical Report 598.

Pelch K, A Reade, T Wolffe, C Kwiatkowski. 2019. PFAS health effects database: protocol for a systematic evidence map. Environ Int. doi: 10.1016/j.envint.2019.05.045

Schaefer HR, JL Myers. 2017. Guidelines for performing systematic reviews in the development of toxicity factors. Regul Toxicol Pharmacol. 91:124–141. doi: 10.1016/j.yrtph.2017.10.008.

Shapiro AJ, S Antoni, KZ Guyton, RM Lunn, D Loomis, I Rusyn, GD Jahnke, PJ Schwingl, SS Mehta, J Addington, N Guha. 2018. Software tools to facilitate systematic review used for cancer hazard identification. Environ Health Perspect. 126:104501. doi: 10.1289/EHP4224

TCEQ. 2015. Guidelines to develop toxicity factors. RG-442: Texas Commission on Environmental Quality (TCEQ). https://www.tceq.texas.gov/toxicology/esl/guidelines

United States Environmental Protection Agency (USEPA). 1988. Recommendations for and documentation of biological values for use in risk assessment. Office of Research and Development. EPA/600/6-87/008

United States Environmental Protection Agency (USEPA). 2016a. Health effects support document for perfluorooctanoic acid (PFOA). Office of Water, Health and Ecological Criteria Division. EPA 822-R-16-003.

United States Environmental Protection Agency (USEPA). 2016b. Health effects support document for perfluorooctane sulfonate (PFOS). Office of Water, Health and Ecological Criteria Division. EPA 822-R-16-002.

United States Environmental Protection Agency (USEPA). 2020. Systematic review protocol for the PFBA, PFHxA, PFHxS, PFNA, and PFDA (anionic and acid forms) IRIS assessments. Supplemental Information-Appendix A. IRIS Assessments Protocol. EPA/635/R-20/131.

United States Environmental Protection Agency (USEPA). 2021a. Human health toxicity values for perfluorobutane sulfonic acid (CASRN 375-73-5) and related compound potassium perfluorobutane sulfonate (CASRN 29420-49-3). EPA/600/R-20/345F.Office of Research and Development, Washington, D.C.

United States Environmental Protection Agency (USEPA). 2021b. Proposed Approaches to the Derivation of a Draft Maximum Contaminant Level Goal for Perfluorooctane Sulfonic Acid (PFOS) (CASRN 1763-23-1) in Drinking Water. External Peer Review Draft. EPA Document No. 822-D-21-002 Office of Water, Washington D.C.

United States Environmental Protection Agency (USEPA). 2021c. Proposed Approaches to the Derivation of a Draft Maximum Contaminant Level Goal for Perfluorooctanoic Acid (PFOA) (CASRN 335-67-1) in Drinking Water. External Peer Review Draft. EPA Document No. 822-D-21-001. Office of Water, Washington, D.C.

United States Environmental Protection Agency (USEPA). 2022. IRIS Toxicological review of perfluorobutanoic acid (PFBA, CASRN 375-22-4) and related salts. EPA/635/R-22/277Fa. US Environmental Protection Agency. Washington D.C.

United States Environmental Protection Agency (USEPA). 2023. IRIS Toxicological review of perfluorohexanoic acid (PFHxA, CASRN 307-24-4) and related salts. EPA/635/R-23/027Fa. US Environmental Protection Agency. Washington D.C.

Appendix A Systematic Evidence Map Protocol

Note: The protocol has been reformatted to allow accessibility. The content is the same as that in the original protocol.

Page numbers in the protocol do not reflect the original page numbers in the protocol but rather are a continuation of the pagination of this overall systematic review and evidence integration document.

A.1 Introduction

Per- and polyfluoroalkyl substances (PFAS) are a large family of fluorinated organic compounds, many of which have been manufactured and used globally since the 1950s. They have unique chemical properties (e.g., repel oil, grease, and water, and resist heat degradation) that have made them important components in the manufacturing of consumer products and fire-fighting foams. Due to their widespread use and environmental stability, PFAS compounds are ubiquitous environmental contaminants, being detected in air, soil, water, and biota. Human biomonitoring studies have indicated that the most commonly detected PFAS are the long-chain legacy PFAS compounds—in particular, perfluorooctanoic acid and perfluorooctane sulfonate (PFOA and PFOS, respectively).

In Texas, PFAS compounds have been found in both drinking water and at contaminated sites. There are currently 25 remediation sites with PFAS contamination in Texas, of which two are former manufacturers, six used PFAS for fire suppression or are fire training facilities, and 17 are military bases. In the last ten years, TCEQ has notified approximately 400 well owners and/or well users about nearby PFAS contamination in five counties. In July of 2011 (updated January 2016), TCEQ released an initial assessment for a set of 16 PFAS compounds (Table A- 1).

Table A- 1. List of sixteen PFAS compounds evaluated

List	PFAS Compound Name	CASRN	PFAS Acronym used by TCEQ
1	Perfluorobutanesulfonic acid	375-73-5	PFBS
2	Perfluorodecanoic acid	335-76-2	PFDA
3	Perfluorododecanoic acid	307-55-1	PFDoA
4	Perfluorohexanoic acid	307-24-4	PFHxA
5	Perfluorononanoic acid	375-95-1	PFNA
6	Perfluorooctanesulfonic acid	1763-23-1	PFOS
7	Perfluorooctanoic acid	335-67-1	PFOA
8	Perfluoroheptanoic acid	375-85-9	PFHpA
9	Perfluorohexanesulfonic acid	355-46-4	PFHxS
10	Perfluoroundecanoic acid	2058-94-8	PFUnA
11	Perfluorobutanoic acid	375-22-4	PFBA
12	Perfluorotetradecanoic acid	376-06-7	PFTeDA
13	Perfluoropentanoic acid	2706-90-3	PFPeA
14	Perfluorotridecanoic acid	72629-94-8	PFTrDA

List	PFAS Compound Name	CASRN	PFAS Acronym used by TCEQ
15	Perfluorooctanesulfonamide	754-91-6	PFOSA
16	Perfluorodecane sulfonate	67906-42-7	PFDS

TCEQ plans to re-evaluate these 16 PFAS and, as a first step, will need to identify the body of toxicology and epidemiology literature for these 16 compounds. To define the current body of literature reporting the toxicology and epidemiology of these 16 PFAS compounds, ToxStrategies will work with TCEQ to develop and conduct a systematic literature search and screen of the peer-reviewed literature. On completion of this effort, TCEQ will be provided with a full systematic evidence map to support the development of potential toxicity values for the 16 PFAS compounds, where data are appropriate to do so. This effort will also provide insight for current data gaps and knowledge clusters.

During problem formulation and exploratory searching, previous efforts to develop a systematic map of the PFAS literature were identified. The PFAS-Tox Database[1] developed by Pelch et al. (2019) provides a systematic evidence map of health and toxicology literature reported for 29 PFAS compounds. While some of these overlap with the set of 16 compounds provided in Table A- 1, the database did not evaluate literature of PFOA, PFOS, or PFOSA. While this database was not used in developing the systematic map described herein, we anticipate that it will be used to cross-check for agreement in the final work product. The full protocol and details of the database are available on the Center for Open Science's Open Science Framework (OSF).[2]

The objective of this exercise is to develop a systematic map at the title/abstract (TiAb) level for use as a tool for characterizing the types of human, animal, and mechanistic studies available for each of the relevant PFAS compounds. Although the review and extraction of relevant full-text articles is not within the scope of this effort, this systematic map will support those activities should they be pursued. It is anticipated that full-text review and extraction will proceed following the completion of the systematic map at the TiAb level.

This protocol describes the framework for the TiAb review and categorization of PFAS toxicology and epidemiology literature. Further, this protocol serves as documentation of study design decisions. It should be noted that the formal protocol (i.e., this document) was completed following the initiation of piloting and screening. However, the approach described herein was developed and written into a Standard Operating

[1] https://pfastoxdatabase.org/

[2] https://osf.io/fpbka/

Procedure (SOP) *a priori*. Any deviations from this protocol, or from the initial SOP, will be documented in a summary report at the conclusion of the systematic map.

A.2 Methods

A.2.1 Literature Search

A.2.1.1 Search Syntax Development

ToxStrategies will develop search syntax based on a list of 16 PFAS compounds of interest to TCEQ. This syntax will be developed to query the PubMed citation database.[3] Initial drafting of syntax using PFAS compound names, acronyms, and CAS numbers, as listed in Table A- 1, will be undertaken by ToxStrategies.

Additional synonyms and acronyms for PFBA, PFHxA, PFHxS, PFNA, and PFDA will be added to the search syntax based on USEPA's *Systematic Review Protocol for the PFBA, PFHxA, PFHxS, PFNA, and PFDA (anionic and acid forms) IRIS Assessments* (USEPA 2020). For the remaining compounds not assessed by USEPA, PubChem[4] will be consulted, and any available synonyms and acronyms will be added to the search syntax. In order to perform a broad and comprehensive search, no additional terms restricting the query will be added (e.g., outcome, species, route). The draft PubMed search syntax for each chemical will be combined into a concatenated string and sent to TCEQ for review and approval (Appendix A.A Identification of Additional PFAS Synonyms and Acronyms).

A pilot query using the concatenated search string was conducted in PubMed. This search generated >3,000,000 results, because PubMed failed to recognize several PFAS synonyms that were included as quoted phrases. As a result, the search syntax was adjusted by removing several synonyms (Appendix A.B.1). The finalized search syntax is provided in Appendix A.B.2.

A.2.1.2 Search Validation

Following the PubMed query, the search results will be validated by comparing them to ~50 primary publications identified in previous USEPA assessments of PFAS (USEPA 2016a, 2016b, 2021). All publications used in validation that are also indexed in PubMed will be identified by the final syntax. References cited by USEPA that are unpublished or not indexed by PubMed will not be returned by the literature search.

A.2.2 Literature Screening

To facilitate the screening and selection process, project-specific DistillerSR forms will be developed based on inclusion and exclusion criteria. Each reference will be screened for

[3] https://pubmed.ncbi.nlm.nih.gov

[4] https://pubchem.ncbi.nlm.nih.gov

inclusion/exclusion and categorized by two reviewers. Prior to initiating the literature screening, reviewer calibration exercises with the project team (i.e., piloting) will be performed to ensure consistency and allow for documentation of a reproducible workflow. For this effort, only TiAb screening will be performed. Full-text screening and data extraction are outside the scope of this effort; however, the resulting systematic map will allow for these subsequent efforts if desired.

Based on the number of anticipated search results, DistillerSR's Artificial Intelligence (AI) model will be utilized where possible. During screening, AI text mining will automatically rank and prioritize unreviewed references. These manually reviewed references will serve as the training set for the AI screener. Once an appropriate threshold of expected included studies has been met (i.e., 95%–99%), the AI screener will be used to exclude irrelevant results based on the manually reviewed training set.

A.2.2.1 Inclusion/Exclusion Criteria

The following inclusion and exclusion criteria were developed based on the objective of the systematic map:

Table A- 2. Inclusion and exclusion criteria implemented in the systematic map of sixteen PFAS compounds

	Include	**Exclude**
Population	• Human (epidemiological or biomonitoring) • *In vivo* experimental animal • *In vitro*/mechanistic	• Models irrelevant to human health outcomes (e.g., models of ecotoxicity)
Exposure	• Investigates at least one of the PFAS in Table A- 1 (Appendix A.A)	• No PFAS compound of interest (other PFAS compounds investigated but not one of the 16 of interest) • No chemical of interest (no PFAS compounds investigated) • PFAS detection in other media • Studies on treatment following PFAS exposure, unless PFAS was also tested alone

	Include	**Exclude**
Outcomes	• Health outcome (e.g., apical outcomes) • Toxicokinetics: ADME, PK, PBPK • Mechanistic • Point of departure: NOAEL, NOEL, LOAEL, LOEL, BMD • *In silico*/computational model • Biomonitoring	• Phytotoxicity • Ecotoxicity
Reference type	• Primary references • Reviews: includes relevant risk assessments, meta-analyses, systematic reviews	• Opinion pieces, commentaries, letters to the editor, etc.

If a study is deemed relevant based on the inclusion criteria above, it will be categorized further by species, outcomes, duration, and route. Reviewers may also select "unclear" if the information is missing or the abstract language is difficult to interpret. For every study type, all PFAS compounds of interest that were reported in the reference will be recorded. The following categories were specified in the DistillerSR forms for included references:

Species:
- Human
- Experimental animals
  - Rat
  - Mouse
  - Other mammal (e.g., rabbit, guinea pig, hamster, dog, pig, primate)
  - Non-mammalian (e.g., zebrafish, chicken)

Human studies were categorized further as follows:
- Epidemiology (observational studies)
- Clinical (controlled trials)
- Biomonitoring
- Mechanistic
- Toxicokinetics
- *In silico*/computational model

In vivo experimental animal studies will be categorized further by study duration and route of exposure, as follows:

Study duration (duration of administration of test chemical):
- Acute (24 hours or less)

- Subacute (1 – 30 days)
- Subchronic (1–3 months; e.g., 90-day study)
- Chronic (greater than 3 months; e.g., 2-year bioassay)

Route of Exposure:
- Oral/gavage
- Dermal
- Inhalation
- Intraperitoneal (IP)/intravenous (IV)/subcutaneous (SC)
- Immersion
- Other (e.g., *in utero*)

Beyond the studies not meeting inclusion criteria above, as modified from Schaefer and Myers (2017), studies published in languages other than English will also be excluded. When no abstract is available and the title is in English, references will be designated as "No abstract" and will be excluded. Corrections and responses to articles will also be excluded but may be retrieved as needed once the final set of papers is determined.

A.2.2.2 Piloting and Reviewer Calibration

Reviewers will pilot references selected at random by a ToxStrategies Scientist in DistillerSR. This exercise will also serve as a pilot for the workflow and calibration of reviewer responses. Feedback on the structure of the DistillerSR form will be discussed among reviewers and the project facilitation team, and the form will be revised as needed. Additionally, differing responses between reviewers will be discussed, and the form and/or inclusion criteria may be revised as necessary for clarity.

A.2.2.3 Conflict Resolution

Each TiAb will be reviewed and categorized by two screeners; therefore, it is anticipated that conflicting screening and categorization results may arise. When screening results in a TiAb review are deemed relevant by one screener and irrelevant by another (i.e., an inclusion/exclusion conflict), this will be resolved by discussion between the two screeners. If the two screeners cannot come to a resolution, a third reviewer will review the TiAb to reach a final conclusion.

Conflicts in the categorization of study types will be prioritized based on the TCEQ project goals (i.e., development of toxicity values). Therefore, conflicting categorizations in the evidence stream and *in vivo* study information categories will be prioritized. Should conflicts occur for categorizations of lower importance for developing a toxicity value (e.g., *in vitro* studies), both categories will be included in the map. Reviewer conflicts regarding the reason for study exclusion will also remain, due to the minor importance of such categorizations.

A.2.3 Title/Abstract Data Map and Reporting

Title and abstract categorizations will be exported from DistillerSR and organized into Microsoft Excel spreadsheets. It is anticipated that multiple spreadsheets will be created, and the references and respective data will be sorted into inclusion and exclusion categorizations. Based on the volume of literature, further breakdown of the TiAb categorizations may occur, such as creating separate lists by evidence stream (i.e., human, *in vitro*, *in vivo* experimental animal).

An accompanying narrative report will also be provided, summarizing methods, deviations from this protocol, and selected visualizations. Visualizations will include a literature flow chart depicting the origins and categorizations of the included literature. Additional data summaries of evidence streams, PFAS investigated, route of exposure, exposure duration, and measured endpoints/outcomes will also be developed as warranted by the data.

A.3 References

Pelch K, Reade A, Wolffe T, Kwiatkowski, C. 2019. PFAS health effects database: Protocol for a systematic evidence map. Environ Int, doi: 10.1016/j.envint.2019.05.045

Schaefer HR, Myers JL. 2017. Guidelines for performing systematic reviews in the development of toxicity factors. Regul Toxicol Pharmacol 91:124–141, doi: 10.1016/j.yrtph.2017.10.008.

USEPA (United States Environmental Protection Agency). 2016a. Health effects support document for perfluorooctanoic acid (PFOA). Office of Water, Health and Ecological Criteria Division. EPA 822-R-16-003.

USEPA (United States Environmental Protection Agency). 2016b. Health effects support document for perfluorooctane sulfonate (PFOS). Office of Water, Health and Ecological Criteria Division. EPA 822-R-16-002.

USEPA (United States Environmental Protection Agency). 2020. Systematic review protocol for the PFBA, PFHxA, PFHxS, PFNA, and PFDA (anionic and acid forms) IRIS assessments. Supplemental Information-Appendix A. IRIS Assessments Protocol. EPA/635/R-20/131.

USEPA (United States Environmental Protection Agency). 2021. Human health toxicity values for perfluorobutane sulfonic acid (CASRN 375-73-5) and related compound potassium perfluorobutane sulfonate (CASRN 29420-49-3). Office of Research and Development, Center for Public Health and Environmental Assessment. EPA/600/R-20/345F.

Appendix A.A Identification of Additional PFAS Synonyms and Acronyms

Table A.A.1. Table of additional PFAS synonyms and acronyms

TCEQ PFAS Names and Acronyms			EPA Assessment Synonyms/Acronyms	PubChem Synonyms/Acronyms	Pubchem direct link	Syntax for all compounds and synonyms	Final Syntax for all PFAS compounds w/ synonyms and CAS nos.	Number of results in PubMed per PFAS compound
PFAS CASRN	PFAS Name	PFAS Acronym						
375-22-4	Perfluorobutyric acid	PFBA	Perfluorobutanoic acid Heptafluoro-1-butanoic acid Heptafluorobutanoic acid Heptafluorobutyric acid Kyseinа heptafluormaselna Perfluoropropanecarboxylic acid Butanoic acid, heptafluoro- Perfluoro-n-butanoic acid Perfluorobutanoate			"Perfluorobutanoic acid" OR "Heptafluorobutanoic acid" OR "Heptafluorobutyric acid" OR "Kyseinа heptafluormaselna" OR "Perfluoropropanecarboxylic acid" OR Perfluorobutanoate OR "Perfluorobutyric acid" OR PFBA OR 375-22-4[EC/RN Number]	"Perfluorobutanoic acid" OR "Heptafluorobutanoic acid" OR "Heptafluorobutyric acid" OR Perfluorobutanoate OR "Perfluorobutyric acid" OR PFBA OR 375-22-4[EC/RN Number]	637
375-73-5	Perfluorobutane sulfonate, Perfluorobutane sulfonic acid	PFBuS	PFBS nonafluorobutane-1-sulfonic acid Perfluorobutanesulfonic acid Monafluoro-1-butanesulfonic acid Monafluorobutanesulfonic acid Perfluorobutanesulfonic acid Perfluorobutanesulfonate Perfluorobutylsulfonate			"Perfluorobutanesulfonic acid" OR "Monafluorobutanesulfonic acid" OR "Perfluorobutanesulfonic acid" OR "Perfluorobutanesulfonate" OR Perfluorobutylsulfonate OR "Perfluorobutane sulfonate" OR "Perfluorobutane sulfonic acid" OR PFBuS OR PFBS OR 375-73-5[EC/RN Number]	"Perfluorobutanesulfonic acid" OR "Perfluorobutanesulfonic acid" OR "Perfluorobutanesulfonic acid" OR Perfluorobutylsulfonate OR "Perfluorobutane sulfonate" OR "Perfluorobutane sulfonic acid" OR PFBuS OR PFBS OR 375-73-5[EC/RN Number]	417
2706-90-3	Perfluoropentanoic acid	PFPeA		Perfluorovaleric acid Monafluorovaleric acid Monafluoropentanoic acid	https://pubchem.ncbi.nlm.nih.gov/compound/79921	"perfluoropentanoic acid" OR "Perfluorovaleric acid" OR "Monafluorovaleric acid" OR "monafluoropentanoic acid" OR PFPeA OR 2706-90-3[EC/RN Number]	"Perfluoropentanoic acid" OR "monafluoropentanoic acid" OR PFPeA OR 2706-90-3[EC/RN Number]	188
355-46-4	Perfluorohexane sulfonate, Perfluorohexane sulfonic acid	PFHxS	Perfluorohexanesulfonic acid Perfluoro-1-hexanesulfonate Perfluorohexane sulfonic acid Perfluorohexane-1-sulphonic acid Perfluorohexanesulfonate Perfluorohexanesulfonic acid Perfluorohexylsulfonate Tridecafluorohexanesulfonic acid tridecafluoro-1-Hexanesulfonic acid			"Perfluorohexanesulfonic acid" OR "Perfluorohexane sulfonic acid" OR Perfluorohexanesulfonate OR "Perfluorohexanesulfonic acid" OR Perfluorohexylsulfonate OR "Tridecafluorohexanesulfonic acid" OR "Perfluorohexane sulfonate" OR "Perfluorohexane sulfonic acid" OR PFHxS OR 355-46-4[EC/RN Number]	"Perfluorohexanesulfonic acid" OR "Perfluorohexane sulfonic acid" OR Perfluorohexylsulfonate OR "Perfluorohexane sulfonate" OR "Perfluorohexane sulfonic acid" OR PFHxS OR 355-46-4[EC/RN Number]	843
307-24-4	Perfluorohexanoic acid	PFHxA	Undecafluorohexanoic acid hexanoic acid, undecafluoro- perfluoro-1-pentanecarboxylic acid perfluorocaproic acid perfluorohexanoate undecafluorocaproic acid undecafluorohexanoic acid			"Undecafluorohexanoic acid" OR "perfluorocaproic acid" OR perfluorohexanoate OR "undecafluorocaproic acid" OR "undecafluorohexanoic acid" OR "Perfluorohexanoic acid" OR PFHxA OR 307-24-4[EC/RN Number]	perfluorohexanoate OR "Perfluorohexanoic acid" OR PFHxA OR 307-24-4[EC/RN Number]	370
375-85-9	Perfluoroheptanoic acid	PFHpA		Tridecafluoroheptanoic acid Perfluoroenanthic Acid Tridecafluoroenanthic Acid	https://pubchem.ncbi.nlm.nih.gov/compound/Perfluoroheptanoic-acid	"Tridecafluoroheptanoic acid" OR "Perfluoroenanthic Acid" OR "Tridecafluoroenanthic Acid" OR "Perfluoroheptanoic acid" OR PFHpA OR 375-85-9[EC/RN Number]	"Tridecafluoroheptanoic acid" OR "Perfluoroheptanoic acid" OR PFHpA OR 375-85-9[EC/RN Number]	257
1763-23-1	Perfluorooctanoic sulfonate, Perfluorooctane sulfonic acid	PFOS		Perfluorooctane sulfonic acid Perfluorooctanesulfonic acid Perfluorooctylsulfonic acid heptadecafluorooctane sulfonic acid heptadecafluorooctane sulphonic acid	https://pubchem.ncbi.nlm.nih.gov/compound/74483	"Perfluorooctane sulfonic acid" OR "Perfluorooctanesulfonic acid" OR "Perfluorooctylsulfonic acid" OR "heptadecafluorooctane sulfonic acid" OR "heptadecafluorooctane sulphonic acid" OR "Perfluorooctanoic sulfonate" OR "Perfluorooctanoic sulfonate" OR PFOS OR 1763-23-1[EC/RN Number]	"Perfluorooctane sulfonic acid" OR "Perfluorooctanesulfonic acid" OR "heptadecafluorooctane sulfonic acid" OR "Perfluorooctanoic sulfonate" OR PFOS OR 1763-23-1[EC/RN Number]	3730
335-67-1	Perfluorooctanoic acid, Perfluorooctanoate	PFOA		Pentadecafluorooctanoic acid Perfluorocaprylic acid Perfluorooctanoic acid Perfluoroheptanecarboxylic acid perfluorooctylcarboxylic acid	https://pubchem.ncbi.nlm.nih.gov/compound/9554#section=Other-Identifiers	"Pentadecafluorooctanoic acid" OR "Perfluorocaprylic acid" OR "Perfluorooctanoic acid" OR "Perfluoroheptanecarboxylic acid" OR "perfluorooctylcarboxylic acid" OR "Perfluorooctanoate" OR PFOA OR 335-67-1[EC/RN Number]	"Pentadecafluorooctanoic acid" OR "Perfluorooctanoic acid" OR "Perfluorooctanoate" OR PFOA OR 335-67-1[EC/RN Number]	3639
754-91-6	Perfluorooctane sulfonamide	PFOSA		Perfluorooctanesulfonamide Perfluorooctane sulfonamide Perfluorooctylsulfonamide Heptadecafluorooctanesulphonamide Perfluorooctanesulphonamide	https://pubchem.ncbi.nlm.nih.gov/compound/69778#section=Synonyms	Perfluorooctanesulfonamide OR "Perfluorooctane sulfonamide" OR Perfluorooctylsulfonamide OR Heptadecafluorooctanesulphonamide OR Perfluorooctanesulphonamide OR "Perfluorooctane sulfonamide" OR PFOSA OR 754-91-6[EC/RN Number]	Perfluorooctanesulfonamide OR "Perfluorooctane sulfonamide" OR Perfluorooctylsulfonamide OR Heptadecafluorooctanesulphonamide OR Perfluorooctanesulphonamide OR "Perfluorooctane sulfonamide" OR PFOSA OR 754-91-6[EC/RN Number]	340
375-95-1	Perfluorononanoic acid	PFNA	heptadecafluorononanoic acid Perfluoro-n-nonanoic acid Perfluorononanoate Perfluoropelargonic acid			"Heptadecafluorononanoic acid" OR Perfluorononanoate OR "Perfluoropelargonic acid" OR "Perfluorononanoic acid" OR PFNA OR 375-95-1[EC/RN Number]	"Heptadecafluorononanoic acid" OR Perfluorononanoate OR "Perfluorononanoic acid" OR PFNA OR 375-95-1[EC/RN Number]	1265

Table A.A.1. Table of additional PFAS synonyms and acronyms (continued)

TCEQ PFAS Names and Acronyms			EPA Assessment Synonyms/Acronyms	PubChem Synonyms/Acronyms	Pubchem direct link	Syntax for all compounds and synonyms	Final Syntax for all PFAS compounds w/ synonyms and CAS nos.	Number of results in PubMed per PFAS compound
PFAS CASRN	PFAS Name	PFAS Acronym						
335-76-2	Perfluorodecanoic acid	PFDeA	PFDA Nonadecafluoro-n-decanoic acid Nonadecafluorodecanoic acid Perfluoro-n-decanoic acid Perfluorodecanoate PFDcA			"Nonadecafluorodecanoic acid" OR Perfluorodecanoate OR "Perfluorodecanoic acid" OR PFDcA OR PFDA OR PFDeA OR 335-76-2[EC/RN Number]	"Nonadecafluorodecanoic acid" OR Perfluorodecanoate OR "Perfluorodecanoic acid" OR PFDcA OR PFDA OR PFDeA OR 335-76-2[EC/RN Number]	676
67906-42-7	Perfluorodecane sulfonate	PFDS		Ammonium henicosafluorodecanesulphonate Perfluorodecanesulfonic acid ammonium salt ammonium perfluorodecanesulfonate	https://pubchem.ncbi.nlm.nih.gov/compound/Ammonium-henicosafluorodecanesulphonate	"Ammonium henicosafluorodecanesulphonate" OR "Perfluorodecanesulfonic acid ammonium salt" OR "ammonium perfluorodecanesulfonate" OR "Perfluorodecane sulfonate" OR PFDS OR 67906-42-7[EC/RN Number]	"Perfluorodecane sulfonate" OR PFDS OR 67906-42-7[EC/RN Number]	300
2058-94-8	Perfluoroundecanoic acid	PFUA		Perfluoroundecanoic acid Henicosafluoroundecanoic acid heneicosafluoroundecanoic acid Eicosafluorondecanoic acid	https://pubchem.ncbi.nlm.nih.gov/compound/Perfluoroundecanoic-acid	"Perfluoroundecanoic acid" OR "Henicosafluoroundecanoic acid" OR "heneicosafluoroundecanoic acid" OR "Eicosafluorondecanoic acid" OR PFUA OR 2058-94-8[EC/RN Number]	"Perfluoroundecanoic acid" OR PFUA OR 2058-94-8[EC/RN Number]	221
307-55-1	Perfluorododecanoic acid	PFDoA		Perfluorododecanoic acid Tricosafluorododecanoic acid Perfluorolauric acid Tricosafluorolauric Acid	https://pubchem.ncbi.nlm.nih.gov/compound/Perfluorododecanoic-acid	"Perfluorododecanoic acid" OR "Tricosafluorododecanoic acid" OR "Perfluorolauric acid" OR "Tricosafluorolauric Acid" OR PFDoA OR 307-55-1[EC/RN Number]	"Perfluorododecanoic acid" OR PFDoA OR 307-55-1[EC/RN Number]	166
72629-94-8	Perfluorotridecanoic acid	PFTrDA		Perfluorotridecanoic acid Pentacosafluorotridecanoic acid	https://pubchem.ncbi.nlm.nih.gov/compound/Perfluorotridecanoic-acid	"Perfluorotridecanoic acid" OR "Pentacosafluorotridecanoic acid" OR PFTrDA OR 72629-94-8[EC/RN Number]	"Perfluorotridecanoic acid" OR PFTrDA OR 72629-94-8[EC/RN Number]	74
376-06-7	Perfluorotetradecanoic acid	PFTeDA		Perfluorotetradecanoic acid Perfluoromyristic acid Heptacosafluorotetradecanoic acid	https://pubchem.ncbi.nlm.nih.gov/compound/Perfluorotetradecanoic-acid	"Perfluorotetradecanoic acid" OR "Perfluoromyristic acid" OR "Heptacosafluorotetradecanoic acid" OR PFTeDA OR 376-06-7[EC/RN Number]	"Perfluorotetradecanoic acid" OR PFTeDA OR 376-06-7[EC/RN Number]	67

Appendix A.B Final Search Syntax

Appendix A.B.1 List of PFAS synonyms not recognized by PubMed and removed from search syntax

"Kyselina heptafluormaselna", "Perfluoropropanecarboxylic acid", "Nonafluorobutanesulfonic acid", "Perfluorovaleric acid", "Nonafluorovaleric acid", "Tridecafluorohexanesulfonic acid", "Undecafluorohexanoic acid", "perfluorocaproic acid", "undecafluorocaproic acid", "undecafluorohexanoic acid", "Perfluoroenanthic Acid", "Tridecafluoroenanthic Acid", "Perfluorooctylsulfonic acid", "heptadecafluorooctane sulphonic acid", "Perfluorocaprylic acid", "Perfluoroheptanecarboxylic acid", "perfluorooctylcarboxylic acid", "Perfluoropelargonic acid", "Ammonium henicosafluorodecanesulphonate", "Perfluorodecanesulfonic acid ammonium salt", "ammonium perfluorodecanesulfonate", "Henicosafluoroundecanoic acid", "heneicosafluoroundecanoic acid", "Eicosafluorondecanoic acid", "Tricosafluorododecanoic acid", "Perfluorolauric acid", "Tricosafluorolauric Acid", "Pentacosafluorotridecanoic acid", "Perfluoromyristic acid", "Heptacosafluorotetradecanoic acid"

Appendix A.B.2 Final syntax for literature search conducted on May 19, 2021, in PubMed

("Perfluorobutanoic acid" OR "Heptafluorobutanoic acid" OR "Heptafluorobutyric acid" OR Perfluorobutanoate OR "Perfluorobutyric acid" OR PFBA OR 375-22-4[EC/RN Number]

OR

"Perfluorobutanesulfonic acid" OR "Perfluorobutanesulfonic acid" OR "Perfluorobutanesulfonate" OR Perfluorobutylsulfonate OR "Perfluorobutane sulfonate" OR "Perfluorobutane sulfonic acid" OR PFBuS OR PFBS OR 375-73-5[EC/RN Number]

OR

"Perfluoropentanoic acid" OR "nonafluoropentanoic acid" OR PFPeA OR 2706-90-3[EC/RN Number]

OR

"Perfluorohexanesulfonic acid" OR "Perfluorohexane sulfonic acid" OR Perfluorohexanesulfonate OR "Perfluorohexanesulfonic acid" OR Perfluorohexylsulfonate OR "Perfluorohexane sulfonate" OR "Perfluorohexane sulfonic acid" OR PFHxS OR 355-46-4[EC/RN Number]

OR

perfluorohexanoate OR "Perfluorohexanoic acid" OR PFHxA OR 307-24-4[EC/RN Number]

OR

"Tridecafluoroheptanoic acid" OR "Perfluoroheptanoic acid" OR PFHpA OR 375-85-9[EC/RN Number]

OR

"Perfluorooctane sulfonic acid" OR "Perfluorooctanesulfonic acid" OR "heptadecafluorooctane sulfonic acid" OR "Perfluorooctanoic sulfonate" OR "Perfluorooctanoic sulfonate" OR PFOS OR 1763-23-1[EC/RN Number]

OR

"Pentadecafluorooctanoic acid" OR "Perfluoroctanoic acid" OR "Perfluorooctanoate" OR PFOA OR 335-67-1[EC/RN Number]

OR
Perfluorooctanesulfonamide OR "Perfluorooctane sulfonamide" OR Perfluoroctylsulfonamide OR Heptadecafluorooctanesulphonamide OR Perfluorooctanesulphonamide OR "Perfluorooctane sulfonamide" OR PFOSA OR 754-91-6[EC/RN Number]

OR

"Heptadecafluorononanoic acid" OR Perfluorononanoate OR "Perfluorononanoic acid" OR PFNA OR 375-95-1[EC/RN Number]

OR

"Nonadecafluorodecanoic acid" OR Perfluorodecanoate OR "Perfluorodecanoic acid" OR PFDcA OR PFDA OR PFDeA OR 335-76-2[EC/RN Number]

OR

"Perfluorodecane sulfonate" OR PFDS OR 67906-42-7[EC/RN Number]

OR

"Perfluoroundecanoic acid" OR PFUA OR 2058-94-8[EC/RN Number]

OR

"Perfluorododecanoic acid" OR PFDoA OR 307-55-1[EC/RN Number]

OR

"Perfluorotridecanoic acid" OR PFTrDA OR 72629-94-8[EC/RN Number]

OR

"Perfluorotetradecanoic acid" OR PFTeDA OR 376-06-7[EC/RN Number])

Appendix B Systematic Review Protocol

Note: The protocol has been reformatted to allow accessibility. The content is the same as that in the original protocol.

Page numbers in the protocol do not reflect the original page numbers in the protocol but rather are a continuation of the pagination of this overall systematic review and evidence integration document.

This Appendix was copied from the original protocol. The content is the same as in the protocol, but there may be some minor differences in formatting (i.e., font color, size, etc.). Also, the links to outside documents were removed from Appendix B.A.

B.1 Introduction

Per- and polyfluoroalkyl substances (PFAS) are a large family of fluorinated organic compounds, many of which have been manufactured and used globally since the 1950s. They have unique chemical properties (e.g., repel oil, grease, and water, and resist heat degradation) that have made them important components in the manufacturing of consumer products and fire-fighting foams. Due to their widespread use and environmental stability, PFAS compounds are ubiquitous environmental contaminants, being detected in air, soil, water, and biota. Human biomonitoring studies have indicated that the most commonly detected PFAS are the long-chain legacy PFAS compounds—in particular, perfluorooctanoic acid and perfluorooctane sulfonate (PFOA and PFOS, respectively).

In Texas, PFAS compounds have been found in both drinking water and at contaminated sites. There are currently 18 remediation sites with PFAS contamination in Texas, of which two are former manufacturers, two are fire training facilities, and 14 are military bases. In the last ten years, the Texas Commission of Environmental Quality (TCEQ) has notified approximately 400 well owners and/or well users about nearby PFAS contamination in five counties. In July of 2011 (updated January 2016), TCEQ released an initial assessment for a set of 16 PFAS compounds (Table B- 1).

Table B- 1. List of sixteen PFAS compounds evaluated

List	PFAS Compound Name	CASRN	PFAS Acronym used by TCEQ
1	Perfluorobutanesulfonic acid	375-73-5	PFBS
2	Perfluorodecanoic acid	335-76-2	PFDA
3	Perfluorododecanoic acid	307-55-1	PFDoA
4	Perfluorohexanoic acid	307-24-4	PFHxA
5	Perfluorononanoic acid	375-95-1	PFNA
6	Perfluorooctanesulfonic acid	1763-23-1	PFOS
7	Perfluorooctanoic acid	335-67-1	PFOA
8	Perfluoroheptanoic acid	375-85-9	PFHpA
9	Perfluorohexanesulfonic acid	355-46-4	PFHxS
10	Perfluoroundecanoic acid	2058-94-8	PFUnA
11	Perfluorobutanoic acid	375-22-4	PFBA
12	Perfluorotetradecanoic acid	376-06-7	PFTeDA
13	Perfluoropentanoic acid	2706-90-3	PFPeA
14	Perfluorotridecanoic acid	72629-94-8	PFTrDA

List	PFAS Compound Name	CASRN	PFAS Acronym used by TCEQ
15	Perfluorooctanesulfonamide	754-91-6	PFOSA
16	Perfluorodecane sulfonate	67906-42-7	PFDS

TCEQ plans to re-evaluate these 16 PFAS in context of a development support document and, as a first step, defined the body of toxicology and epidemiology literature in 2021 by executing a systematic evidence map (SEM). This effort resulted in a full SEM to support the development of potential toxicity values for the 16 PFAS compounds, where data are appropriate to do so.

The objective of this exercise is to perform a systematic review (SR) of the toxicology and epidemiology literature identified in the systematic evidence map. Specifically, the findings of this SR will inform the development of toxicity values in a TCEQ development support document. The protocol contained herein was developed under the direction of TCEQ and describes the framework of this review. Further, this protocol serves as documentation of study design decisions. Any deviations from this protocol will be documented in a summary report at the conclusion of the SR effort.

B.2 Methods

B.2.1 Literature Identification

B.2.1.1 SEM literature

Literature identified by title/abstract in the SEM will be advanced to the SR based on its categorization within such. Table B- 2 below displays the required category labels to advance, along with the estimated numbers associated with each evidence stream.

Table B- 2. Category labels assigned during the SEM process required to advance studies in each evidence stream to the SR

Evidence Stream	SEM labels	N
Experimental animal	"In vivo" AND "Rat" OR "Mouse" OR "Other mammal" AND "Health outcome" AND "Oral/gavage" OR "Inhalation" OR "Unclear" OR "Other"	309
Human	"Epidemiology" OR "Clinical"	739

B.2.1.2 Validation of results

To ensure a comprehensive review, results will be validated by comparing them to publications identified in a variety of secondary sources. Specifically, relevant citations from the following recent assessments will be used to cross-check those advanced to the SR:

- Health effects support document for perfluorooctanoic acid (PFOA) (USEPA, 2016a)
- Health effects support document for perfluorooctane sulfonate (PFOS) (USEPA, 2016b)
- Proposed Approaches to the Derivation of a Draft Maximum Contaminant Level Goal for Perfluorooctane Sulfonic Acid (PFOS) (CASRN 1763-23-1) in Drinking Water (USEPA, 2021a)
- Proposed Approaches to the Derivation of a Draft Maximum Contaminant Level Goal for Perfluorooctanoic Acid (PFOA) (CASRN 335-67-1) in Drinking Water (USEPA, 2021b)
- Human health toxicity values for perfluorobutane sulfonic acid (CASRN 375-73-5) and related compound potassium perfluorobutane sulfonate (CASRN 29420-49-3) (USEPA, 2021c)
- Toxicological Profile for Perfluoroalkyls (ATSDR, 2021)
- PFAS-Tox Database (Pelch et al., 2021)

B.2.2 Literature Screening

To facilitate the screening and selection process, project specific DistillerSR forms will be developed based on inclusion and exclusion criteria. As literature has already been reviewed and categorized at the TiAb level, each reference will be screened at the full text level for inclusion/exclusion and categorized by two reviewers. Two web-based literature review software platforms (DistillerSR and HAWC) were considered. Given prior familiarity and demonstrated availability of technical support, it was decided that all workflows would be executed through DistillerSR. Prior to initiating the literature screening, reviewer calibration exercises with the project team (i.e., piloting) will be performed to ensure consistency and allow for documentation of a reproducible workflow.

B.2.2.1 Inclusion/Exclusion Criteria

Inclusion and exclusion criteria were developed by TCEQ based on the objective of the SR (Table B- 3). As stated above, each publication will be reviewed in the context of these criteria to determine whether it will be included in the SR. For studies excluded at full text screening, one or more exclusion reasons will be documented.

Table B- 3. Inclusion and exclusion criteria to be utilized during full text review of the SR

Category	Include	Exclude
Population	• Human • Mammalian experimental animals	• Non-mammalian models and mammalian models irrelevant to human health outcomes (e.g., models of ecotoxicity, livestock)

Exposure	• Investigates at least one of the 16 PFAS listed in Table B- 1 or salts/ions thereof
	• Oral or inhalation routes of exposure
	• Exposure is based on external dose/exposure
	• Exposure involves quantitative measurement or estimation informed by a measurement
	  ○ Observational human studies: exposure estimates informing exposure-response relationship must have some based on reported quantitative measurements
	• If internal exposure dose metric (e.g., serum PFAS) is used to evaluate outcome association, it is characterized over the duration leading up to outcome observation (e.g., career or lifetime cumulative serum PFAS level)
	  ○ E.g., cumulative serum PFAS estimates modeled on worker biomonitoring data and job exposure matrix (JEM) or some other historical understanding of exposure activities
		• No PFAS compound of interest (other PFAS compounds investigated but not one of the 16 of interest)
		• No chemical of interest (no PFAS compounds investigated)
		• Inadequate exposure data for risk assessment for quantitative characterization of exposure response relationship
		  ○ E.g., observational human studies that do not include measurement of exposure, such as qualitative estimates based solely on job history without environmental or internal exposure measurements
		• Exposure estimates based solely on a single collection of biomonitoring samples, such as serum measurements
		• PFAS concentration data (e.g., detection in media such as sewage sludge or wastewater)
		• Exposure to more than one chemical/compound unless one study group was PFAS only exposure

Outcomes	• Adverse health outcomes (e.g., apical outcomes) as defined in Section 3.6 of TCEQ guidance [a]	• Mechanistic endpoints (e.g., gene expression data, enzyme activity) • Human vaccine studies [b] – effects of PFAS exposure on antibody concentration
Reference type	• Primary references	• Opinion pieces, commentaries, letters to the editor, etc. • Review: includes risk assessments, meta-analyses, systematic reviews

Category	Include	Exclude
Study model/ design	• Epidemiological study designs such as cohort studies, case-control studies • *In vivo* study designs	• Case studies or case series • Ecological studies since there are no individual data (i.e., rely on population-level exposure or outcome data and report associations between exposure and outcome at the population level) • Cross-sectional studies since temporal association cannot be established (e.g., National Health and Nutrition Examination Survey [NHANES]) • Clinical trials • Human biomonitoring studies • *In vitro* study models
Additional criteria	• English translation available • Quantitative data (dose response or pairwise significance) available for at least one outcome of interest	

a: The focus of the review is on adverse effects per TCEQ Guidelines to Develop Toxicity Factors (TCEQ 2015). For outcomes where potential adversity is less obvious or unclear, TCEQ directed reviewers to be more inclusive than not.
b: The decision to exclude was communicated by TCEQ staff to ToxStrategies during a call on April 28, 2022. (and reiterated during a call on May 12, 2022). The TCEQ staff were aware that there was a debate within the scientific community about whether the endpoint represented an adverse effect. TCEQ decided to set these studies aside from the current effort and evaluate them separately.

B.2.2.2 Piloting and Reviewer Calibration

Reviewers will pilot a random selection of references produced by the literature search, by applying the inclusion and exclusion criteria above. This exercise will also serve as a pilot for the workflow and calibration of reviewer responses. Feedback on the structure of the form will be discussed among reviewers and the project facilitation team, and the form will be revised as needed. Additionally, differing responses between reviewers will be discussed, and the form and/or inclusion criteria may be revised as necessary for clarity. It is anticipated that this pilot will include 10 full text papers (5 experimental animal and 5 epidemiological studies), and at least 5 reviewers. All stages of this SR will be included in pilot and reviewer calibration exercises, i.e., full text screening, data extraction, and study quality.

B.2.2.3 Conflict Resolution

Each full text will be reviewed for inclusion by two screeners. When screening results in a are deemed relevant by one screener and irrelevant by another (i.e., an inclusion/exclusion conflict), this will be resolved by discussion between the two

screeners. If the two screeners cannot come to a resolution, a third reviewer will review the full text to reach a conclusion.

B.2.3 Follow-up Study Evaluation

Each experimental animal and epidemiology study that meets inclusion criteria will be evaluated to determine whether any follow-up studies were available. The review team will evaluate all included articles and document linked citations in Excel. Initial and follow-up studies will be evaluated together to determine which should be included / excluded. Following evaluation, QC of all linked studies will be performed to ensure accuracy. Full text screen exclusion criteria will be updated in Distiller for all excluded articles.

B.2.4 Data Extraction

Data extraction will focus on gathering information pertinent to the derivation of toxicological values as outlined by TCEQ's Guidelines to Develop Toxicity Factors (2015). The review team will extract details of study information, experimental design, and results in a project-specific form using DistillerSR. As previously stated, following the development of a draft extraction form, the review team will pilot extraction to ensure reviewer consistency and to identify any additional data determined to be useful to informing the research question. The pilot exercises may result in iterative refinement of the form. Following data extraction, QC of the data extraction will be performed to ensure accuracy. All updates made during QC will be documented using Distiller's audit log only.

As a result of piloting and reviewer calibration, and in discussion with TCEQ, the following criteria were established to ensure consistent reviewer responses (Table B- 4 and Table B- 5). Each row in the Distiller data extraction table will be defined by:

1. PFAS compound
2. Study duration and species (experimental animal)
3. Cohort or sub-group (epidemiology)
4. Endpoint/Outcome category (experimental animal)
5. A single LOAEL / NOAEL value

Studies with no statistically significant findings will not be captured in the data extraction table, unless there is compelling evidence that the findings are biologically significant; however, lack of statistically significant findings will be noted within the DistillerSR record. Examples of biological significance: histopathology findings denoted by the author as significant but where statistical significance of the finding incidence is not quantified; a 10% decrease in body weight (for adults) or 5% (for pups) is biologically significant. Additionally, results from mechanistic experimental endpoints (e.g., results from transgenic animals, gene expression data, etc.) will not be extracted, but a brief description of the reported model/endpoint will be included in the reviewer notes for

relevant studies (e.g., humanized PPARα mouse model tested; oxidative stress data reported; gene expression data reported).

Table B- 4. Guidance for Reviewers Completing the Experimental Animal Data Extraction Table

Field	Guidance

N/sex/age	• For non-reproductive and developmental studies:
	  ○ N = number of animals evaluated per endpoint
	  ○ Sex = sex of animals evaluated per endpoint
	  ○ Age at study initiation
	• For reproductive and developmental studies
	  ○ N = number of litters evaluated per endpoint category (may include F_1 generation)
	  ▪ If number of litters (or dams) evaluated are not provided by the authors, record NR and add the number of pups evaluated in the notes
	  ○ Sex = sex of animal exposed (probably the F_0 generation)
	  ○ Age = age of animal at study initiation (i.e., age at start of exposure)
	• If the number of animals evaluated per endpoint was not reported (but the number exposed/tested was provided in the study methods), put "NR" in this column and note the number tested in the notes field [a]
	• Age of animals at initiation of study (i.e., when treatment was initiated, which may be after a period of acclimatization)
	• For reproductive and developmental study designs where exposure occurred prior to birth, focus should be on the parental exposures
	  ○ e.g., N = number of dams, sex = F (mated), age of dams if reported (or "adult" if stated, "NR" if not reported)

PFAS compounds	• include salts/ions of PFAS compounds ○ example: potassium PFOS
Dose/concentration, frequency of dosing and units	• list all doses evaluated in the study including the control, not just doses at which findings occurred, along with the metric unit (e.g., mg/kg body weight-d) • include the frequency of dosing • in studies that included recovery or satellite groups, report findings together with the main dose groups unless an endpoint was significantly affected during the recovery period, but the main group was not. If an endpoint was significantly affected during recovery (but not during the main exposure period), include a separate row with only the dose levels of the satellite groups.
Exposure duration and units	• provide duration of exposure for the entire study
Endpoint category	• select a single, most appropriate endpoint category • if no category was available, then "other" was selected and a new category was provided • select 1 endpoint category per table row • select "developmental" for all endpoints measured in offspring exposed via parents

Field	Guidance

Endpoint	• list statistically and/or biologically significant endpoint(s) per endpoint category [b]. This may include significance determined by two-way comparisons between control and dose group, or a trend test.
	  ○ Examples for endpoint categories:
	  ▪ General toxicity: Body weight, mortality, clinical observations, food/water consumption, gross observations
	  ▪ Developmental: Fetal defects, developmental markers (e.g., vaginal opening, preputial separation, first estrus, eye opening; includes all endpoints measured in offspring exposed via parents, with the exception of litter characteristics)
	  ▪ Litter characteristics: Litter size, viability, pup survival, sex per litter, litter weight (pup information prior to weaning)
	  ▪ Reproductive: % mated or littered, gestation length, andrology, estrus cyclicity, fertility, pregnancy outcome and reproductive endpoints, placental endpoints
	• Provide direction of change
	• Include all endpoints with same direction of change in one row
	• Example:
	  ○ <u>Row 1</u>
	  Endpoint group: organ weight
	  Endpoint: decreased thyroid, thymus weight
	  ○ <u>Row 2</u>
	  Endpoint group: organ weight
	  Endpoint: increased ovary weight

NOAEL/LOAEL	• Include NOAEL and/or LOAEL as defined by the data • Include relevant information that defined the value (e.g., sex, tissue) • When no NOAEL/LOAEL are available, reviewers should enter "NA" ○ This includes scenarios where only modeled results are available (e.g., BMDU/BMDL) and effect levels are not reported • Units are not required (available in the dose column) • Use "unclear" when statistical significance was evaluated but the point of departure value is not apparent • When statistical significance is reported but there is a no/non-monotonic dose response, note the lowest dose where there is a statistically significant differences as the "LOAEL" and include a reviewer note about characterizing the dose response (e.g., lack of dose response, non-monotonic dose response, etc.)
Notes	• Add comments relating to the data captured in the table row • Adding information to this field is not required
QC Agreement with Extraction	• Yes – second reviewer has checked the data row and agrees • No - second reviewer has checked the data row and disagrees – this should trigger a discussion between the two reviewers that will eventually resolve this response to "yes"

[a] *The number of animals "evaluated" for an endpoint can differ from the number "exposed/tested" at the start of the study for several reasons, including the study design itself (e.g., splitting exposure groups for assaying multiple endpoints) and unanticipated events (e.g., animal infection/loss).*

[b] *Reviewers will add a note (in the "Notes" field of the review form) to indicate statistical significance without apparent biological significance (e.g., significance for a single endpoint, low-dose group only).*

Reviewers should apply expert judgement when including endpoints with biological but not statistical significance (e.g., close to (but not) statistically significant or a histopathology finding identified as key but without statistical incidence and severity). Endpoints/outcomes may be raised with the group for discussion as needed.

Table B- 5. Guidance for Reviewers Completing the Human Epidemiology Data Extraction Table

Field	Guidance
Name of study/cohort	• List one cohort per row • List sub-groups only if statistically and/or biologically significant • Do not include the whole cohort if only the sub-group has statistical significance; in these instances, list the sub-group only and provide a note (per row) to state that the whole group was not significant
PFAS compound(s)	• List 1 PFAS compound per row
Exposure concentration	• Capture exposure concentration quartiles • Include mean exposure value as a note if desired
Endpoint category	• Select a single, most appropriate endpoint category • If no category is available, select "other" and provide a new category • Select 1 endpoint category per table row
Endpoint	• List statistically and/or biologically significant endpoint(s) per endpoint category [a] • Provide direction of change • Include all endpoints with same direction of change in one row • Provide a brief description of the endpoint

[a] *Reviewers will add a note (in the "Notes" field of the review form) to indicate statistical significance without apparent biological significance (e.g., significance for a single endpoint, low-dose group only). Reviewers should apply expert judgement when including endpoints with biological but not statistical significance (e.g., close to (but not) statistically significant or an outcome the reviewers deems important). Endpoints/outcomes maybe raised with the group for discussion as needed.*

B.2.5 Study Quality

Each experimental animal and epidemiology study that meets inclusion criteria and had significant adverse outcome data to extract will be evaluated for study quality and risk of bias (RoB). Where needed, study quality questions will be refined to target specific aspects of PFAS literature. This process will undergo user reviewer piloting and reviewer calibration, which may result in iterative refinement of the form. Study quality forms, containing reviewer guidance, are presented in Appendix B.A. For experimental animal studies, the quality assessment will be conducted at the study design level rather than the outcome or publication levels. In the hypothetical publication below each study would have a separate study quality form.

- Study 1: PFOA was given to pregnant mice by gavage once daily from GD 1 through GD 17; samples taken for teratological analysis
- Study 2: PFOA was given to adult mice by oral gavage daily for 90 days; samples taken for histology on day 90; samples taken for clinical chemistry on days 45 and 90.

- Study 3: PFOA was given to adult mice by oral gavage daily for 90 days, the mice allowed to recover for 90 days; samples taken for histology on day 90; samples taken for clinical chemistry on days 45, 90 and 180.

Study quality will be conducted for all included studies where health effects were evaluated and significant outcome data were extracted; study quality will not be evaluated for mechanistic endpoints evaluated in a health effects study, nor for studies in which no significant adverse outcome data were reported. In this case, the reviewer will notate such by adding "N/A" to the Overall Study Quality field. For each metric, reviewers should provide sufficient information to justify their quality categorization (copy from publication if possible).

For epidemiology (human) studies, study quality and RoB will be conducted at the study design level rather than the outcome/s evaluated in the study. This assessment will be adapted from study evaluation domains available on USEPA's Health Assessment Workspace Collaboration (HAWC) (Shapiro et al., 2018). These domains are arranged into metrics which provide both core and prompting questions to aid the reviewer in assessing the study's reporting, RoB and study sensitivity on an outcome-specific basis. Using the provided guidance, reviewers will independently score each domain as Good (++), Adequate (+), Deficient (-), or Critically Deficient; each reviewer will provide sufficient information to justify their selection of the category as free text. Following domain scoring, each study will be given an overall confidence rating of High (++), Medium (+), or Low (-) based on these domain scores.

At the onset of the review, the study quality evaluation workflow included two independent reviewers. Conflicts on the overall confidence rating were resolved by discussion between the two reviewers. If the two reviewers could not come to a resolution, a third reviewer gave input to make a final determination. Due to time and resource constraints, this workflow was revised during the implementation of the SR to include a first reviewer performing the full study quality evaluation, followed by a second reviewer performing QC.

B.2.6 Evidence Integration and Synthesis

Evidence integration and synthesis phases of the SR will not be executed by ToxStrategies, and thus are not included in this protocol.

B.3 References

Pelch K, Reade A, Wolffe T, Kwiatkowski, C. 2019. PFAS health effects database: Protocol for a systematic evidence map. Environ Int, doi: 10.1016/j.envint.2019.05.045

Schaefer HR, Myers JL. 2017. Guidelines for performing systematic reviews in the development of toxicity factors. Regul Toxicol Pharmacol 91:124–141, doi: 10.1016/j.yrtph.2017.10.008.

Shapiro AJ, Antoni S, Guyton KZ, Lunn RM, Loomis D, Rusyn I, Jahnke GD, Schwingl PJ, Mehta SS, Addington J, Guha N. 2018. Software Tools to Facilitate Systematic Review Used for Cancer Hazard Identification. Environ Health Perspect. 126(10):104501. DOI: 10.1289/EHP4224

TCEQ (Texas Commission on Environmental Quality). 2015. TCEQ Guidelines to Develop Toxicity Factors. Available at: Guidelines to Develop Inhalation and Oral Cancer and Non-Cancer Toxicity Factors (texas.gov)

USEPA (United States Environmental Protection Agency). 2016a. Health effects support document for perfluorooctanoic acid (PFOA). Office of Water, Health and Ecological Criteria Division. EPA 822-R-16-003.

USEPA (United States Environmental Protection Agency). 2016b. Health effects support document for perfluorooctane sulfonate (PFOS). Office of Water, Health and Ecological Criteria Division. EPA 822-R-16-002.

USEPA (United States Environmental Protection Agency). 2020. Systematic review protocol for the PFBA, PFHxA, PFHxS, PFNA, and PFDA (anionic and acid forms) IRIS assessments. Supplemental Information-Appendix A. IRIS Assessments Protocol. EPA/635/R-20/131.

USEPA (United States Environmental Protection Agency). 2021a. Proposed Approaches to the Derivation of a Draft Maximum Contaminant Level Goal for Perfluorooctane Sulfonic Acid (PFOS) (CASRN 1763-23-1) in Drinking Water. External Peer Review Draft. EPA Document No. 822D21002

USEPA (United States Environmental Protection Agency). 2021b. Proposed Approaches to the Derivation of a Draft Maximum Contaminant Level Goal for Perfluorooctanoic Acid (PFOA) (CASRN 335-67-1) in Drinking Water. External Peer Review Draft. EPA Document No. 822D21001

USEPA (United States Environmental Protection Agency). 2021c. Human health toxicity values for perfluorobutane sulfonic acid (CASRN 375-73-5) and related compound potassium perfluorobutane sulfonate (CASRN 29420-49-3). Office of Research and Development, Center for Public Health and Environmental Assessment. EPA/600/R-20/345F.

B.4 Change Log

Document Version Number	Release Phase	Update Date	Change(s)
v1.0	Original document	08 April 2022	
v2.0	Post pre-pilot 1	11 May, 2022	Per client request, protocol updated: • Study inclusion / exclusion criteria for human studies updated (quantifiable exposure, study design) • Define which data and how to categorize for data extraction • A follow-up/ follow-on study evaluation protocol added
v3.0	Post pre-pilot 2	12 May 2022	Per client request, protocol updated: • Add frequency of dosing • Consolidate endpoint categories (animal studies) and add endpoint alignment to categories • Definition of data row (single NOAEL/LOAEL) • Clarification of human exposure inclusion/exclusion criteria • Definition of a study unit for study quality evaluation
v4.0	Post animal pilot	18 May 2022	Updated based on outcome and comments from the pilot: • Endpoints / endpoint categories • Designation of salts / ions • Example of biological significance • Additional guidance for data extraction fields
V4.1	Post pilot client meeting	02 June 2022	Updated based on outcome and comments from the pilot: • Epi studies – exposure concentration by quartile • QC tracking performed using Distiller audit log
V4.2	Client requested updates	15 July 2022	• Data extraction table updates (route and age) • GLP compliance
V5.0	Production revisions	15 Sept 2022	• Addition of study quality criteria (as previously provided within DistillerSR only) • Direction on data extraction of non-monotonic dose-response data • Clarification of selected data extraction guidance • Removal of reference to preliminary outcome list
V5.1	Production revisions	18 Nov 2022	• Clarification on data extraction of developmental and reproductive toxicity studies • Direction on data extraction of studies that do not provide raw data (e.g., only BMDL/BMDU reported) • Direction on data extraction of studies with a recovery or satellite group • Clarification on study quality evaluation
V5.2	Final revisions	17 Feb 2023	• Minor revisions/formatting

Appendix B.A Study Quality Evaluation Guidance

B.A.1 EPA HAWC Study Evaluation Metrics and Guidance for Experimental Animal Studies

Table B.A- 1. Domain #1: Reporting Quality

Domain #1	**Name: Reporting Quality** Description:
Metric #1	**Name:** Reporting quality • **Short name:** Reporting • **Required animal** **Description:** **CORE QUESTION** Does the study report information for evaluating the design and conduct of the study for the endpoint(s)/outcome(s) of interest? **PROMPTING QUESTIONS** Does the study report the following? **Critical information** necessary to perform study evaluation: • Species; test article name; levels and duration of exposure; route (e.g., oral; inhalation); qualitative or quantitative results for at least one endpoint of interest **Important information** for evaluating the study methods: • Test animal: strain, sex, source, and general husbandry procedures • Exposure methods: source, purity, method of administration • Experimental design: frequency of exposure, animal age and lifestage during exposure and at endpoint/outcome evaluation • Endpoint evaluation methods: assays or procedures used to measure the endpoints/outcomes of interest *NOTES:* • *Reviewers should reach out to authors to obtain missing information when studies are considered key for hazard evaluation and/or dose-response.* • *This domain is limited to reporting. Other aspects of the exposure methods, experimental design, and endpoint evaluation methods are evaluated using the domains related to risk of bias and study sensitivity.* **BASIC CONSIDERATIONS** These considerations typically do not need to be refined by assessment teams, although in some instances the **important information** may be refined depending on the endpoints/outcomes of interest or the chemical under investigation. A judgment and rationale for this domain should be given for the study. Typically, these will not change regardless of the endpoints/outcomes

Domain #1	**Name: Reporting Quality** Description:
Metric #1 (cont.)	investigated by the study. **In the rationale, reviewers should indicate whether the study adhered to GLP, OECD, or other testing guidelines.** ***Good***: All **critical and important information** is reported or inferable for the endpoints/outcomes of interest. ***Adequate***: All **critical information** is reported but some **important information** is missing. However, the missing information is not expected to significantly impact the study evaluation. ***Deficient***: All **critical information** is reported but **important information** is missing that is expected to significantly reduce the ability to evaluate the study. ***Critically Deficient***: Study report is missing any pieces of **critical information**. Studies that are Critically Deficient for reporting are Uninformative for the overall rating and not considered further for evidence synthesis and integration. **EXAMPLE RATING** **Study** - ***Good*** **-** Important information is provided for test species, strain, sex, age, exposure methods, experimental design, endpoint evaluations and the presentation of results. The authors report that 'the study was conducted in compliance with the OECD guidelines for Good Laboratory Practice [c(81) 30 (Final)]. Follow link to see attachments that contain example answers to the animal study evaluation domains.

Table B.A- 2. Domain #2: Selection and Performance

Domain #2	**Name: Selection and Performance** **Description:**
Metric #1	• **Name:** Allocation • **Short name:** Allocation • **Required animal** **Description:** **CORE QUESTION:** Were animals assigned to experimental groups using a method that minimizes selection bias? **PROMPTING QUESTIONS** For each study: • Did each animal or litter have an equal chance of being assigned to any experimental group (i.e., random allocation)? • Is the allocation method described? • Aside from randomization, were any steps taken to balance variables across experimental groups during allocation? **BASIC CONSIDERATIONS**

Domain #2	**Name: Selection and Performance** **Description:**
Metric #1 (Cont.)	These considerations typically do not need to be refined by assessment teams. A judgment and rationale for this domain should be given for each cohort or experiment in the study. ***Good***: Experimental groups were randomized and any specific randomization procedure was described or inferable (e.g., computer-generated scheme). [Note that normalization is not the same as randomization (see response for 'Adequate').] ***Adequate***: Authors report that groups were randomized but do not describe the specific procedure used (e.g., 'animals were randomized'). Alternatively, authors used a non-random method to control for important modifying factors across experimental groups (e.g., body weight normalization). ***Not Reported*** **(interpreted as Deficient)**: No indication of randomization of groups or other methods (e.g., normalization) to control for important modifying factors across experimental groups. ***Critically Deficient***: Bias in the animal allocations was reported or inferable. **EXAMPLE RATING** **All Cohorts/Experiments -** ***Good*** **-** The study authors report that 'Fifty males and fifty females were randomly assigned to groups by a computer-generated weight-ordered distribution such that individual body weights did not exceed + 20% of the mean weight for each sex.' Follow **link** to see attachments that contain example answers to the animal study evaluation domains.
Metric #2	**Name:** Observational bias/blinding • **Short name:** Blinding • **Required animal** **Description:** **CORE QUESTION** Did the study implement measures to reduce observational bias? **PROMPTING QUESTIONS** For each endpoint/outcome or grouping of endpoints/outcomes in a study: • Does the study report blinding or other methods/procedures for reducing observational bias? • If not, did the study use a design or approach for which such procedures can be inferred? What is the expected impact of failure to implement (or report implementation) of these methods/procedures on results? **BASIC CONSIDERATIONS**

Domain #2	**Name: Selection and Performance** **Description:**
Metric #2 (cont.)	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[1]). ***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. • ***(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 could have impacted results [1] For non-targeted or screening-level histopathology outcomes often used in guideline studies, blinding during the initial evaluation of tissues is generally not recommended as masked evaluation can make 'the task of separating treatment-related changes from normal variation more difficult' and 'there is concern that masked review during the initial evaluation may result in missing subtle lesions.' Generally, blinded evaluations are recommended for targeted secondary review of specific tissues or in instances when there is a pre-defined set of outcomes that is known or predicted to occur (Crissman 2004). **EXAMPLE RATINGS** **Histopathology - *Good*** - Although the study did not indicate blinding, blinding during the initial evaluation of tissues for initial or non-targeted evaluations is generally not recommended as masked evaluation can make the task of separating treatment-related changes from normal variation more difficult and may result in subtle lesions being overlooked (Crissman 2004). The study did include a secondary evaluation by a pathology working group (PWG) review on coded pathology slides which minimized the potential for observational bias. **Organ weights, functional observational battery, motor activity, swim maze and histopathology - *Good* -** Authors reported that the

Domain #2	**Name: Selection and Performance** **Description:**
Metric #2 (cont.)	investigators were blinded to the animal treatment group during evaluation for all outcome measures. Although blinding is not recommended for initial or non-targeted evaluations (Crissman 2004), this study evaluated prespecified outcomes in targeted evaluations for which blinding is appropriate (cell counts in the CA3 region of the hippocampus). Follow **link** to see attachments that contain example answers to the animal study evaluation domains.

Table B.A- 3. Domain #3: Confounding/Variable Control

Domain #3	**Name: Confounding/Variable Control** **Description:**
Metric #1	• **Name:** Confounding/variable control • **Short name:** Confounding/Variable Control • **Required animal** **Description:** **CORE QUESTION** Are variables with the potential to confound or modify results controlled for and consistent across all experimental groups? **PROMPTING QUESTIONS** For each study: • Are there differences across the treatment groups (e.g., co-exposures, vehicle, diet, palatability, husbandry, health status, etc.) that could bias the results? • If differences are identified, to what extent are they expected to impact the results? **BASIC CONSIDERATIONS** 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.

Domain #3	**Name: Confounding/Variable Control** **Description:**
Metric #1 (cont.)	***Adequate***: Some concern that variables that were likely to confound or modify results were uncontrolled or inconsistent across groups, but are expected to have a minimal impact on the results. ***Deficient***: Notable concern that potentially confounding variables were uncontrolled or inconsistent across groups, and are expected to substantially impact the results. ***Critically deficient***: Confounding variables were presumed to be uncontrolled or inconsistent across groups, and are expected to be a primary driver of the results. **EXAMPLE RATING** **All Cohorts/Experiments/Endpoints - *Good* -** Based on the study report, vehicle (deionized water with 2% tween 80) and husbandry practices were inferred to be the same in controls and treatment groups. The experimental conditions described provided no indication of concern for uncontrolled variables or different practices across groups. Follow **link** to see attachments that contain example answers to the animal study evaluation domains.

Table B.A- 4. Domain #4: Selective Reporting/Attrition

Domain #4	**Name: Selective Reporting/Attrition** **Description:**
Metric #1	• **Name:** Selective reporting and attrition • **Short name:** Selective Reporting/Attrition • **Required animal** **Description:** **CORE QUESTION** Did the study report results for all prespecified outcomes and tested animals? **PROMPTING QUESTIONS** For each study: *Selective reporting bias:* • Are all results presented for endpoints/outcomes described in the methods (see note)? *Attrition bias:* • Are all animals accounted for in the results? • If there are discrepancies, do authors provide an explanation (e.g., death or unscheduled sacrifice during the study)? • If unexplained results omissions and/or attrition are identified, what is the expected impact on the interpretation of the results? *NOTE: This domain does **not** consider the appropriateness of the analysis/results presentation. This aspect of study quality is evaluated in another domain.*

Domain #4	**Name: Selective Reporting/Attrition** **Description:**
Metric #1 (cont.)	**BASIC CONSIDERATIONS** These considerations typically do not need to be refined by assessment teams. A judgment and rationale for this domain should be given for each cohort or experiment in the study. ***Good***: Quantitative or qualitative results were reported for all prespecified outcomes (explicitly stated or inferred), exposure groups and evaluation timepoints. Data not reported in the primary article is available from supplemental material. If results omissions or animal attrition are identified, the authors provide an explanation and these are not expected to impact the interpretation of the results. ***Adequate***: Quantitative or qualitative results are reported for most prespecified outcomes (explicitly stated or inferred), exposure groups and evaluation timepoints. Omissions and/or attrition are not explained, but are not expected to significantly impact the interpretation of the results. ***Deficient***: Quantitative or qualitative results are missing for many prespecified outcomes (explicitly stated or inferred), exposure groups and evaluation timepoints and/or high animal attrition; omissions and/or attrition are not explained and may significantly impact the interpretation of the results. ***Critically Deficient***: Extensive results omission and/or animal attrition are identified and prevents comparisons of results across treatment groups. **EXAMPLE RATING** **Inhalation study - *Good* -** Animal loss was reported (the authors treated 10 rats/sex/dose group and noted one death in a high-dose male rat at day 85 of study). All endpoints described in methods were reported qualitatively or quantitatively. Follow **link** to see attachments that contain example answers to the animal study evaluation domains.

Table B.A- 5. Domain #5: Exposure Methods

Domain #5	Name: Exposure Methods Description:
Metric #1	• **Name:** Chemical administration and characterization • **Short name:** Exposure Characterization • **Required animal** **Description:** **CORE QUESTION** Did the study adequately characterize exposure to the chemical of interest and the exposure administration methods? **PROMPTING QUESTIONS** For each study: • Does the study report the source and purity and/or composition (e.g., identity and percent distribution of different isomers) of the chemical? If not, can the purity and/or composition be obtained from the supplier (e.g., as reported on the website) • Was independent analytical verification of the test article purity and composition performed? • Did the authors take steps to ensure the reported exposure levels were accurate? • For inhalation studies: were target concentrations confirmed using reliable analytical measurements in chamber air? • For oral studies: if necessary based on consideration of chemical-specific knowledge (e.g., instability in solution; volatility) and/or exposure design (e.g., the frequency and duration of exposure), were chemical concentrations in the dosing solutions or diet analytically confirmed? Are there concerns about the methods used to administer the chemical (e.g., inhalation chamber type, gavage volume, etc.)? *NOTE: Consideration of the appropriateness of the route of exposure is not evaluated at the individual study level. Relevance and utility of the routes of exposure are considered in the PECO criteria for study inclusion and during evidence synthesis.* **BASIC CONSIDERATIONS** 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 cohort or experiment in the study. ***Good***: Chemical administration and characterization is complete (i.e., source, purity, and analytical verification of the test article are provided). There are no concerns about the composition, stability, or purity of the administered chemical, or the specific methods of administration. For

Domain #5	**Name: Exposure Methods** **Description:**
Metric #1 (cont.)	inhalation studies, chemical concentrations in the exposure chambers are verified using reliable analytical methods. ***Adequate***: Some uncertainties in the chemical administration and characterization are identified but these are expected to have minimal impact on interpretation of the results (e.g., source and vendor- reported purity are presented, but not independently verified; purity of the test article is sub-optimal but not concerning; For inhalation studies, actual exposure concentrations are missing or verified with less reliable 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; levels of impurities are substantial or concerning; deficient administration methods, such as use of static inhalation chambers or a gavage volume considered too large for the species and/or lifestage at exposure). ***Critically Deficient***: Uncertainties in the exposure characterization are identified and there is reasonable certainty that the results are largely attributable to factors other than exposure to the chemical of interest (e.g., identified impurities are expected to be a primary driver of the results). **EXAMPLE RATINGS** **Oral study - *Good* -** Source (3M) and purity (98%) are described, and the authors provided verification using analytical methods (GC/MS). Addressing concerns about known instability in solution for this chemical, the authors verified the dosing solutions twice weekly over the course of the experiment. Animals were exposed via gavage with all dose groups receiving the same volume. **Inhalation study - *Good* -** Source (3M) and purity (98%) of the test article are described. All animals were transferred to dynamic inhalation exposure chambers for the exposures. The concentration of the test chemical in the air was continuously monitored from the animals' breathing zone throughout the 6-hour exposure periods and mean daily average concentrations and variability were reported. Follow **link** to see attachments that contain example answers to the animal study evaluation domains.

Domain #5	Name: Exposure Methods Description:
Metric #2	**Name:** Exposure timing, frequency and duration • **Short name:** Study Design Applicability • **Required animal** **Description:** **CORE QUESTION** Was the timing, frequency, and duration of exposure sensitive for the endpoint(s)/outcome(s) of interest? **PROMPTING QUESTIONS** For each endpoint/outcome or grouping of endpoints/outcomes in a study: • Does the exposure period include the critical window of sensitivity? • Was the duration and frequency of exposure sensitive for detecting the endpoint of interest? **BASIC CONSIDERATIONS** Considerations for this domain are highly variable depending on the endpoint(s)/outcome(s) of interest and must be refined by assessment teams. A judgment and rationale for this domain should be given for each endpoint/outcome or group of endpoints/outcomes investigated in the study. ***Good***: The duration and frequency of the exposure was sensitive and the exposure included the critical window of sensitivity (if known). ***Adequate***: The duration and frequency of the exposure was sensitive and the exposure covered most of the critical window of sensitivity (if known). ***Deficient***: The duration and/or frequency of the exposure is not sensitive and did not include the majority of the critical window of sensitivity (if known). These limitations are expected to bias the results towards the null. ***Critically deficient***: The exposure design was not sensitive and is expected to strongly bias the results towards the null. The rationale should indicate the specific concern(s). **EXAMPLE RATINGS** **All Endpoints/Outcomes - Good -** Study uses a standard OECD short-term (28-day) study design to examine toxicological effects that are routinely evaluated in this testing guideline. Developmental and Male Reproductive effects - Good - The experimental design and exposure period were appropriate for evaluation of potential male reproductive and developmental effects. The experiment was designed to evaluate reproductive and developmental outcomes and followed recommendations in OECD 416 and EPA OPPT 870.3800 guidelines.

Domain #5	Name: Exposure Methods Description:
Metric #2 (cont.)	Follow link to see attachments that contain example answers to the animal study evaluation domains.

Table B.A- 6. Domain #6: Outcome Methods/Results Presentation

Domain #6	Name: Outcome Methods/Results Presentation Description:
Metric #1	• **Name:** Outcome Assessment • **Short name:** Outcome Assessment • **Required animal** **Description:** **CORE QUESTION** Are the procedures sensitive and specific for evaluating the endpoint(s)/outcome(s) of interest? **PROMPTING QUESTIONS** For each endpoint/outcome or grouping of endpoints/outcomes in a study: • Are there concerns regarding the specificity and validity of the protocols? • Are there serious concerns regarding the sample size (see note)? • Are there concerns regarding the timing of the endpoint assessment? *NOTE: Sample size alone is not a reason to conclude an individual study is critically deficient.* **BASIC CONSIDERATIONS** Considerations for this domain are highly variable depending on the endpoint(s)/outcome(s) of interest and must be refined by assessment teams. A judgment and rationale for this domain should be given for each endpoint/outcome or group of endpoints/outcomes investigated in the study. Examples of potential concerns include: • Selection of protocols that are insensitive or non-specific for the endpoint of interest • Use of unreliable methods to assess the outcome • Assessment of endpoints at inappropriate or insensitive ages, or without addressing known endpoint variation (e.g., due to circadian rhythms, estrous cyclicity, etc.). • Decreased specificity or sensitivity of the response due to the timing of endpoint evaluation, as compared to exposure (e.g., short-acting depressant or irritant effects of chemicals; insensitivity due to prolonged period of non-exposure prior to testing). **EXAMPLE RATING**

Domain #6	Name: Outcome Methods/Results Presentation Description:
Metric #1 (cont.)	**Organ weight, body weights, and hormone measures - *Good* -** No concerns regarding the specificity and validity of the protocols and measures were identified. Study authors used standard methodology for evaluating organ and body weights. Thyroid hormones were measured using commercial electrochemiluminescence-immunoassay methods, and the known diurnal variation in these measures was accounted for during blood collection. Follow **link** to see attachments that contain example answers to the animal study evaluation domains.
Metric #2	• **Name:** Results presentation • **Short name:** Results Presentation • **Required animal** **Description:** **CORE QUESTION** Are the results presented in a way that makes the data usable and transparent? **PROMPTING QUESTIONS** For each endpoint/outcome or grouping of endpoints/outcomes in a study: • Does the level of detail allow for an informed interpretation of the results? • Are the data analyzed, compared, or presented in a way that is inappropriate or misleading? BASIC CONSIDERATIONS Considerations for this domain are highly variable depending on the outcomes of interest and must be refined by assessment teams. A judgment and rationale for this domain should be given for each endpoint/outcome or group of endpoints/outcomes investigated in the study. Examples of potential concerns include: • Non-preferred presentation, such as developmental toxicity data averaged across pups in a treatment group, when litter responses are more appropriate • Failing to present quantitative results • Pooling data when responses are known or expected to differ substantially (e.g., across sexes or ages) • Failing to report on or address overt toxicity when exposure levels are known or expected to be highly toxic • Lack of full presentation of the data (e.g., presentation of mean without variance data; concurrent control data are not presented) **EXAMPLE RATING**

Domain #6	Name: Outcome Methods/Results Presentation Description:
Metric #2 (cont.)	All Endpoints/Outcomes - Good - There are no notable concerns about the way the results are analyzed or presented. Follow link to see attachments that contain example answers to the animal study evaluation domains.

Table B.A- 7. Domain #11: Overall Study Confidence

Domain #11	Name: Overall Study Confidence Description:
Metric #1	**Name:** Overall confidence (animal) • **Short name:** Overall confidence • **Required animal** **Description:** **CORE QUESTION** Considering the identified strengths and limitations, what is the overall confidence rating for the endpoint(s)/outcome(s) of interest? **PROMPTING QUESTIONS** For each endpoint/outcome or grouping of endpoints/outcomes in a study: • Were concerns (i.e., limitations or uncertainties) related to the reporting quality, risk of bias, or sensitivity identified? • If yes, what is their expected impact on the overall interpretation of the reliability and validity of the study results, including (when possible) interpretations of impacts on the magnitude or direction of the reported effects? *NOTE: Reviewers should mark studies that are rated lower than high confidence only due to low sensitivity (i.e., bias towards the null) for additional consideration during evidence synthesis. If the study is otherwise well-conducted and an effect is observed, the confidence may be increased.* **BASIC CONSIDERATIONS** The overall confidence rating considers the likely impact of the noted concerns (i.e., limitations or uncertainties) in reporting, bias and sensitivity on the results. A confidence rating and rationale should be given for each endpoint/outcome or group of endpoints/outcomes investigated in the study. ***High confidence***: No notable concerns are identified (e.g. most or all domains rated Good). ***Medium confidence***: Some concerns are identified, but expected to have minimal impact on the interpretation of the results. (e.g., most domains rated Adequate or Good; may include studies with Deficient ratings if concerns are not expected to strongly impact the magnitude or direction

Domain #11	Name: Overall Study Confidence Description:
Metric #1 (cont.)	of the results). Any important concerns should be carried forward to evidence synthesis. ***Low confidence***: Identified concerns are expected to significantly impact on the study results or their interpretation (e.g., generally, Deficient ratings for one or more domains). The concerns leading to this confidence judgment must be carried forward to evidence synthesis (see note). ***Uninformative***: Serious flaw(s) that make the study results unusable for informing hazard identification (e.g., generally, Critically Deficient rating in any domain; many Deficient ratings). Uninformative studies are not considered further in the synthesis and integration of evidence. **EXAMPLE RATINGS** **Reproductive and developmental effects other than behavior - *High Confidence*** - The study was well-designed for the evaluation reproductive and developmental toxicity induced by chemical exposure. The study applied established approaches, recommendations, and best practices, and employed an appropriate exposure design for these endpoints. Evidence was presented clearly and transparently. **Behavioral measures - *Low Confidence* -** The cursory cage-side observations of activity are considered to be insensitive and non-specific methods for detecting motor effects, with a strong bias towards the null. Follow **link** to see attachments that contain example answers to the animal study evaluation domains

B.A.2 EPA HAWC Study Evaluation Metrics and Guidance for Epidemiology Studies

Table B.A- 8. Domain #2: Selection and Performance

Domain #2	Name: Selection and Performance Description:
Metric #3	• **Name:** Participant selection • **Short name:** Participant • **Required epi** QUESTION: Is there evidence that selection into or out of the study (or analysis sample) was jointly related to exposure and to outcome? **EXAMPLE TEXT: Adequate.** Nested case-control design in Mexico City birth cohort with 30 cases of preterm birth and 30 controls selected randomly from same population of woman who were recruited during prenatal visits at one of four clinics (serving low to moderate income population). Recruitment and eligibility criteria (inclusion/exclusion criteria) discussed. Little discussion of participants versus nonparticipants but the available information indicates that differential selection is

Domain #2	Name: Selection and Performance Description:
Metric #3 (cont.)	possible but not likely. Participation rate reported to be low (36%). Evaluates the vulnerable population of low-moderate income pregnant women. Add other concerns or limitations. Add impact and direction to effect estimate, if applicable. **RATING GUIDANCE**: Is there evidence that selection into or out of the study (or analysis sample) was jointly related to exposure and to outcome? 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), final analysis group. Does the study include potential vulnerable/susceptible groups or lifestages? Follow **link** to see attachments that contain example prompting and follow-up questions for epidemiological studies.

Table B.A- 9. Domain #5: Exposure Methods

Domain #5	Name: Exposure Methods Description:
Metric #3	**Name:** Exposure measures • **Short name:** Measures • **Required epi** **QUESTION: 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?** **EXAMPLE TEXT: Poor** for long-chained (DEHP, DiNP) and adequate for short-chained (DEP, DBP, DiBP) phthalate metabolites based on number of samples. A single spot (second morning void) urine sample was collected from each woman during a third-trimester visit to the project's research center; third trimester sample is relevant to later term preterm births. Analytical approach described and appropriate. High percent >LOD. Add other concerns or limitations. Add impact and direction to effect estimate, if applicable. **RATING GUIDANCE**: 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? Follow **link** to see attachments that contain example prompting and follow-up questions for epidemiological studies.

Table B.A- 10. Domain #6: Outcome Methods/Results Presentation

Domain #6	**Name: Outcome Methods/Results Presentation** **Description:**
Metric #3	**Name:** Outcome measures • **Short name:** Outcome • **Required epi** **Description:** QUESTION: Does the outcome measure reliably distinguish the presence or absence (or degree of severity) of the outcome? EXAMPLE TEXT: Adequate. Preterm birth defined by length of gestation (< 37 weeks), a standard measure of birth outcome, estimated by maternal recall of the date of last menstrual period, rather than the preferred early ultrasound. Potential misclassification of preterm cases due to maternal recall of last menstrual period to estimate gestational age which may be nondifferential with respect to exposure; however, differential misclassification is still possible but unlikely. Add other concerns or limitations. Add impact and direction to effect estimate, if applicable. **RATING GUIDANCE**: Does the outcome measure reliably distinguish the presence or absence (or degree of severity) of the outcome? Source of outcome (effect) measure, blinding to exposure status or level, how measured/classified, incident versus prevalent disease, evidence from validation studies, prevalence (or distribution summary statistics for continuous measures). Follow **link** to see attachments that contain example prompting and follow-up questions for epidemiological studies.

Table B.A- 11. Domain #7: Confounding

Domain #7	**Name: Confounding** **Description:**
Metric #1	• **Name:** Confounding • **Short name:** Confounding • **Required epi** **QUESTION:** Is confounding of the effect of the exposure likely? **EXAMPLE TEXT: Adequate**. Information on key confounders was collected through questionnaire. The strategy for evaluating confounding and the process for retaining variables in the models was described. Rationale for selecting confounders not provided. Inclusion in model not solely based on statistical significance. Adjustment for relative co-exposures. Add other concerns or limitations. Add impact and direction to effect estimate, if applicable. **RATING GUIDANCE: Is confounding of the effect of the exposure unlikely?**

Domain #7	Name: Confounding Description:
Metric #1 (cont.)	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; degree of exposure to the confounder in the population. Follow **link** to see attachments that contain example prompting and follow-up questions for epidemiological studies

Table B.A- 12. Domain #8: Analysis

Domain #8	Name: Analysis Description:
Metric #1	• **Name:** Does the analysis strategy and presentation convey the necessary familiarity with the data and assumptions? • **Short name:** Analysis • **Required epi** **QUESTION:** Does the analysis strategy and presentation convey the necessary familiarity with the data and assumptions? **EXAMPLE TEXT: Adequate.** Multivariable (multivariate) logistic regression used to take into account potential confounding variables; quantitative results presented (ORs and 95% CIs with ORs adjusted for confounders). Imputation techniques used when phthalate metabolite concentrations were below the LOD (filling in data where there wasn't); Amount of missing data not noted; Dichotomous exposure (reduced sensitivity) and use of median as the cut-off adjusted for urine creatinine and specific gravity to assess effect of method used. Add other concerns or limitations. Add impact and direction to effect estimate, if applicable. **RATING GUIDANCE: Does the analysis strategy and presentation convey the necessary familiarity with the data and assumptions?** Extent (and if applicable, treatment) of missing data for exposure, outcome, and confounders, approach to modeling, classification of exposure and outcome variables (continuous versus categorical), testing of assumptions, sample size for specific analyses, relevant sensitivity analyses. An ideal study would convey a thoughtful and thorough description of the analytical approach, and descriptive data for key variables (e.g., exposure measures, outcome measures), including the amount of missing data (or proportion less than the limit of detection [LOD]). The ideal analysis would use an appropriate and well thought out modeling approach for the study design (e.g., logistic regression for case-control data) and

Domain #8	Name: Analysis Description:
Metric #1 (cont.)	specify the covariates used in the final model; the methods should be described in enough detail such that they could be applied to the data from another study. In addition, the results should be presented with sufficient detail to enable estimation of effect estimates and precision of the estimates (e.g., standard error [SE] or confidence interval [CI] Follow **link** to see attachments that contain example prompting and follow-up questions for epidemiological studies.

Table B.A- 13. Domain #9: Selective Reporting

Domain #9	Name: Selective Reporting Description:
Metric #1	• **Name:** Selective Reporting • **Short name:** Selective • **Required epi** **Description:** QUESTION: Is there reason to be concerned about selective reporting? Selective Reporting EXAMPLE TEXT: Adequate. No concerns for selective reporting. **RATING GUIDANCE: Is there concern for selective reporting?** Rating should be 2-level - Adequate or Deficient. Are results presented with adequate detail for all the endpoints of interest? Are results presented for the full sample as well as for specified subgroups? Were stratified analyses (effect modification) motivated by a specific hypothesis? Follow **link** to see attachments that contain example prompting and follow-up questions for epidemiological studies.

Table B.A- 14. Domain #10: Sensitivity

Domain #10	**Name: Sensitivity** **Description:**
Metric #1	• **Name:** Are there concerns for study sensitivity • **Short name:** Study sensitivity • **Required epi** **Description:** QUESTION: Is there a concern that sensitivity of the study is not adequate to detect an effect? Sensitivity EXAMPLE TEXT: Deficient. Small sample size/ Potential nondifferential misclassification of outcome and exposure. Low exposure levels. Range of exposure is narrow. Healthy worker effect. Add other concerns or limitations. Add impact and direction to effect estimate, if applicable. **RATING GUIDANCE: Are there concerns for study sensitivity?** What exposure range is spanned in this study? 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 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'). Is the study relevant to the exposure and outcome of interest? Follow **link** to see attachments that contain example prompting and follow-up questions for epidemiological studies.

Table B.A- 15. Domain #11: Overall Study Confidence

Domain #11	**Name: Overall Study Confidence** **Description:**
Metric #2	• **Name:** Overall confidence (epi) • **Short name:** Overall confidence • **Required epi** **Description:** QUESTION: Considering the identified strengths and limitations, what is the overall confidence rating for the endpoint(s)/outcome(s) of interest? **EXAMPLE TEXT: Low confidence.** Give brief rationale for rating. Add other concerns or limitations. Add impact and direction to effect estimate, if applicable. **RATING GUIDANCE:** Once the evaluation domains have been classified, these ratings will be combined to reach an overall study confidence classification of **High**, **Medium**, **Low**, or **Uninformative**.

Domain #11	**Name: Overall Study Confidence** **Description:**
Metric #2 (cont.)	This classification will be based on the classifications in the evaluation domains, and will include consideration of the likely impact of the noted deficiencies in bias and sensitivity on the results. Studies with critical deficiencies in any evaluation domain will be classified as **Uninformative**. Other classifications will generally follow a sorting such that **High Confidence** studies would have the highest evaluation ('Good') for all or most domains; **Low Confidence** studies would have a 'Poor' evaluation for one or more domains (unless the impact of the particular limitation(s) is judged to be unlikely to be severe), and **Medium Confidence** studies are in between these groups (e.g., most domains receiving a mid-level **Adequate** evaluation, with no limitations judged to be severe.) Once initial evaluation has been performed with consensus between reviewers, the classifications will be re-evaluated, looking at the variability 'within' and 'between' levels to ensure that the separation between the levels of confidence are appropriate and that no additional criteria need to be considered. Follow **link** to see attachments that contain example prompting and follow-up questions for epidemiological studies.

Appendix C Form for Full Text Review and Screening

Population:

- Human (epidemiological)
- *In vivo* experimental animal

Exposure:

- Investigates at least one of the 16 PFAS listed (but not a mixture)
- Oral or inhalation routes of exposure

Outcomes:

- Any adverse outcome (e.g., apical outcome) – for outcomes that are not clear please consult the master list provided as an attachment to the protocol

Reference type:

- Primary reference/empirical evidence

Additional criteria:

- English translation available
- Quantitative data available for at least one outcome of interest

Does this publication meet the above publication criteria?

- Yes
- No
- Unclear (state uncertainty in adjacent text box)

Is this publication a follow-up study?

- Yes (Note: Select “yes” if the publication identifies the data as being a follow-up to a previous study. No response required if no follow-up study identified)

Reviewer notes:

Appendix D Format for Data Extraction of Studies

D.1 Format for Extraction of Experimental Animal Studies

Reviewer should extract details of study information, experimental design, and results in the table below.

QC reviewer should confirm extracted details:

- If all details are correct, QC reviewer should select 'Yes' in the QC Agreement with Extraction column
- If the QC reviewer considers changes are needed, these should be discussed with the primary reviewer and updates made
- Once all changes have been made and the QC complete, the QC reviewer should click 'Submit'

1. Objective (as reported by author)
2. Is this study GLP-compliant (as reported by the author)
    - Yes
    - No or not stated
3. Study information. This information will be input using a tabular format.
    - Species
    - Strain
    - Route of exposure
    - N/Sex/Age and units
    - PFAS Compound
    - Dose/concentration, frequency and units
    - Exposure duration and units
    - Endpoint Category
    - Endpoint
    - NOAEL
    - LOAEL
    - Notes
    - QC Agreement on Extraction
4. Reviewer notes

Note that for route of exposure, this was further categorized for oral studies (e.g., oral drinking water, oral dietary admixture, oral gavage, etc.) and inhalation studies (e.g., inhalation nose, inhalation whole body).

Note that for endpoint category, this was categorized depending upon the adverse outcome seen (e.g., hepatotoxicity, thyroid pathway, developmental, general toxicity, other category, etc.)

D.2 Format for Extraction of Human Epidemiology Studies

Reviewer should extract details of study information, experimental design, and results in the table below.

QC reviewer should confirm extracted details:

- If all details are correct, QC reviewer should select 'Yes' in the QC Agreement with Extraction column
- If the QC reviewer considers changes are needed, these should be discussed with the primary reviewer and updates made
- Once all changes have been made and the QC complete, the QC reviewer should click 'Submit'

1. Objective (as reported by author)
2. Study information. This information will be input using a tabular format.

    - Study Design
    - Name of Study/Cohort
    - Population Size (N)
    - Population Description (age, location, pregnancy status, etc.)
    - Study Date
    - PFAS Compound(s)
    - Exposure Scenario (e.g., self-reported, serum concentration)
    - Exposure Concentration
    - Endpoint/Outcome Category
    - Endpoint/Outcome
    - RR, OR, HR, etc.
    - Notes
    - QC Agreement on Extraction

Abbreviations: HR, hazard ratio; OR, odds ratio; RR, risk ratio

3. Reviewer notes

Appendix E Study Quality Evaluation

Note: This appendix replicates the forms that were used for study quality evaluation. The study quality evaluation criteria and risk of bias evaluation were based on the study evaluation tool developed by USEPA and available on USEPA's Health Assessment Workspace Collaboration (HAWC) (Shapiro et al. 2018). The domains and metrics are numbered in the same manner as those in USEPA's HAWC tool. Because not all domains and metrics were used for study quality evaluation, the numbering of the domains and metrics are not sequential.

E.1 Study Quality Evaluation for Experimental Animal Studies

Use this form to detail the study quality assessment for experimental animal studies.

One form should be completed per study design. Example: 90-day repeat dose study and an extended one generation study would be assessed for quality on different forms.

Do not include mechanistic data in the quality assessment.

If there are multiple studies within a publication, select "this form, next instance" to open a new and empty form.

Additional information on how to assign quality criteria can be found in the supporting documents.

Enter short study name (e.g., 90-day oral rat study; two-generation oral mouse study).

E.1.1 Domain #1: Reporting Quality

E.1.1.1 Metric #1: Reporting Quality

E.1.1.1.1 CORE QUESTION

Does the study report information for evaluating the design and conduct of the study for the endpoint(s)/outcome(s) of interest?

Additional guidance: PFAS purity needs to be ≥ 90% to score 'good'.

E.1.1.1.2 PROMPTING QUESTIONS

Does the study report the following?

Critical information necessary to perform study evaluation:

- Species; test article name; levels and duration of exposure; route (e.g., oral; inhalation); qualitative or quantitative results for at least one endpoint of interest.

Important information for evaluating the study methods:

- Test animal: strain, sex, source, and general husbandry procedures.
- Exposure methods: source, purity, method of administration.
- Experimental design: frequency of exposure, animal age and lifestage during exposure and at endpoint/outcome evaluation.
- Endpoint evaluation methods: assays or procedures used to measure the endpoints/outcomes of interest.

NOTES:

This domain is limited to reporting. Other aspects of the exposure methods, experimental design, and endpoint evaluation methods are evaluated using the domains related to risk of bias and study sensitivity.

E.1.1.1.3 BASIC CONSIDERATIONS

These considerations typically do not need to be refined by assessment teams, although in some instances the important information may be refined depending on the endpoints/outcomes of interest or the chemical under investigation.

Typically, these will not change regardless of the endpoints/outcomes investigated by the study. In the rationale, reviewers should indicate whether the study adhered to GLP, OECD, or other testing guidelines.

Good: All critical and important information is reported or inferable for the endpoints/outcomes of interest.

Adequate: All critical information is reported but some important information is missing. However, the missing information is not expected to significantly impact the study evaluation.

Deficient: All critical information is reported but important information is missing that is expected to significantly reduce the ability to evaluate the study.

Critically Deficient: Study report is missing any pieces of critical information. Studies that are Critically Deficient for reporting are Uninformative for the overall rating and not considered further for evidence synthesis and integration.

E.1.1.1.4 EXAMPLE RATING

Study - Good - Important information is provided for test species, strain, sex, age, exposure methods, experimental design, endpoint evaluations and the presentation of results. The authors report that 'the study was conducted in compliance with the OECD guidelines for Good Laboratory Practice [c(81) 30 (Final)].

E.1.2 Domain #2: Selection and Performance

E.1.1.2.1 Metric #1: Allocation

E.1.1.2.1.1 CORE QUESTION

Were animals assigned to experimental groups using a method that minimizes selection bias?

E.1.1.2.1.2 PROMPTING QUESTIONS

For each study:

- Did each animal or litter have an equal chance of being assigned to any experimental group (i.e., random allocation)?
- Is the allocation method described?
- Aside from randomization, were any steps taken to balance variables across experimental groups during allocation?

E.1.1.2.1.3 BASIC CONSIDERATIONS

These considerations typically do not need to be refined by assessment teams.

Good: Experimental groups were randomized and any specific randomization procedure was described or inferable (e.g., computer-generated scheme). [Note that normalization is not the same as randomization (see response for 'Adequate').]

Adequate: Authors report that groups were randomized but do not describe the specific procedure used (e.g., 'animals were randomized'). Alternatively, authors used a non-random method to control for important modifying factors across experimental groups (e.g., body weight normalization).

Not Reported (interpreted as Deficient): No indication of randomization of groups or other methods (e.g., normalization) to control for important modifying factors across experimental groups.

Critically Deficient: Bias in the animal allocations was reported or inferable.

E.1.1.2.1.4 EXAMPLE RATING

All Cohorts/Experiments - Good - The study authors report that “Fifty males and fifty females were randomly assigned to groups by a computer-generated weight-ordered distribution such that individual body weights did not exceed + 20% of the mean weight for each sex.”

E.1.1.2.2 Metric #2: Observational bias/blinding

E.1.1.2.2.1 CORE QUESTION

Did the study implement measures to reduce observational bias?

Additional Guidance:

- For assays reporting numeric measurements e.g., hormone levels, no report of blinding does not automatically = deficient
- When there is risk of subjectivity in the assay measurements lack of blinding may result in a lower score e.g., histopathology

E.1.1.2.2.2 PROMPTING QUESTIONS

For each endpoint/outcome or grouping of endpoints/outcomes in a study:

Does the study report blinding or other methods/procedures for reducing observational bias?

If not, did the study use a design or approach for which such procedures can be inferred?

What is the expected impact of failure to implement (or report implementation) of these methods/procedures on results?

E.1.1.2.2.3 BASIC CONSIDERATIONS

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.]

Good: Measures to reduce observational bias were described (e.g. blinding to conceal treatment groups during endpoint evaluation; consensus-based evaluations of histopathology lesions[1]).

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.

---

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

- (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.
- (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 could have impacted results

E.1.1.2.2.4 EXAMPLE RATINGS

Histopathology - Good - Although the study did not indicate blinding, blinding during the initial evaluation of tissues for initial or non-targeted evaluations is generally not recommended as masked evaluation can make the task of separating treatment-related changes from normal variation more difficult and may result in subtle lesions being overlooked (Crissman 2004). The study did include a secondary evaluation by a pathology working group (PWG) review on coded pathology slides which minimized the potential for observational bias.

Organ weights, functional observational battery, motor activity, swim maze and histopathology - Good - Authors reported that the investigators were blinded to the animal treatment group during evaluation for all outcome measures (i.e.,). Although blinding is not recommended for initial or non-targeted evaluations (Crissman 2004), this study evaluated prespecified outcomes in targeted evaluations for which blinding is appropriate (cell counts in the CA3 region of the hippocampus).

E.1.3 Domain #3: Confounding/Variable Control

E.1.3.1 Metric #1: Confounding/variable control

E.1.3.1.1 CORE QUESTION

Are variables with the potential to confound or modify results controlled for and consistent across all experimental groups?

Additional Guidance:

- Consider significant decrease in body weight or signs of overt toxicity as a potential confounder

E.1.3.1.2 PROMPTING QUESTIONS

For each study:

- Are there differences across the treatment groups (e.g., co-exposures, vehicle, diet, palatability, husbandry, health status, etc.) that could bias the results?

- If differences are identified, to what extent are they expected to impact the results?

E.1.3.1.3 BASIC CONSIDERATIONS

These considerations may need to be refined by assessment teams, as the specific variables of concern can vary by experiment or chemical.

Good: Outside of the exposure of interest, variables that are likely to confound or modify results appear to be controlled for and consistent across experimental groups.

Adequate: Some concern that variables that were likely to confound or modify results were uncontrolled or inconsistent across groups, but are expected to have a minimal impact on the results.

Deficient: Notable concern that potentially confounding variables were uncontrolled or inconsistent across groups, and are expected to substantially impact the results.

Critically deficient: Confounding variables were presumed to be uncontrolled or inconsistent across groups, and are expected to be a primary driver of the results.

E.1.3.1.4 EXAMPLE RATING

All Cohorts/Experiments/Endpoints - Good - Based on the study report, vehicle (deionized water with 2% tween 80) and husbandry practices were inferred to be the same in controls and treatment groups. The experimental conditions described provided no indication of concern for uncontrolled variables or different practices across groups.

E.1.4 Domain #4: Selective Reporting/Attrition

E.1.4.1 Metric #1: Selective reporting and attrition

E.1.4.1.1 CORE QUESTION

Did the study report results for all prespecified outcomes and tested animals?

Additional Guidance

- Consider attrition to be the unexplained loss of animals
- Do not reduce the score if the number of animals per endpoint is not given (captured in Domain 6)

E.1.4.1.2 PROMPTING QUESTIONS

For each study:

Selective reporting bias:

- Are all results presented for endpoints/outcomes described in the methods (see note)?

Attrition bias:

- Are all animals accounted for in the results?
- If there are discrepancies, do authors provide an explanation (e.g., death or unscheduled sacrifice during the study)?
- If unexplained results omissions and/or attrition are identified, what is the expected impact on the interpretation of the results?

NOTE: This domain does not consider the appropriateness of the analysis/results presentation. This aspect of study quality is evaluated in another domain.

E.1.4.1.3 BASIC CONSIDERATIONS

These considerations typically do not need to be refined by assessment teams.

Good: Quantitative or qualitative results were reported for all prespecified outcomes (explicitly stated or inferred), exposure groups and evaluation timepoints. Data not reported in the primary article is available from supplemental material. If results omissions or animal attrition are identified, the authors provide an explanation and these are not expected to impact the interpretation of the results.

Adequate: Quantitative or qualitative results are reported for most prespecified outcomes (explicitly stated or inferred), exposure groups and evaluation timepoints. Omissions and/or attrition are not explained, but are not expected to significantly impact the interpretation of the results.

Deficient: Quantitative or qualitative results are missing for many prespecified outcomes (explicitly stated or inferred), exposure groups and evaluation timepoints and/or high animal attrition; omissions and/or attrition are not explained and may significantly impact the interpretation of the results.

Critically Deficient: Extensive results omission and/or animal attrition are identified and prevents comparisons of results across treatment groups.

E.1.4.1.4 EXAMPLE RATING

Inhalation study - Good - Animal loss was reported (the authors treated 10 rats/sex/dose group and noted one death in a high-dose male rat at day 85 of study). All endpoints described in methods were reported qualitatively or quantitatively.

E.1.5 Domain #5: Exposure Methods

E.1.5.1 Metric #1: Chemical administration and characterization

E.1.5.1.1 CORE QUESTION

Did the study adequately characterize exposure to the chemical of interest and the exposure administration methods?

E.1.5.1.2 PROMPTING QUESTIONS

For each study:

- Does the study report the source and purity and/or composition (e.g., identity and percent distribution of different isomers) of the chemical? If not, can the purity and/or composition be obtained from the supplier (e.g., as reported on the website)
- Was independent analytical verification of the test article purity and composition performed?
- Did the authors take steps to ensure the reported exposure levels were accurate?
- For inhalation studies: were target concentrations confirmed using reliable analytical measurements in chamber air?
- For oral studies: if necessary based on consideration of chemical-specific knowledge (e.g., instability in solution; volatility) and/or exposure design (e.g., the frequency and duration of exposure), were chemical concentrations in the dosing solutions or diet analytically confirmed?
- Are there concerns about the methods used to administer the chemical (e.g., inhalation chamber type, gavage volume, etc.)?

NOTE: Consideration of the appropriateness of the route of exposure is not evaluated at the individual study level. Relevance and utility of the routes of exposure are considered in the PECO criteria for study inclusion and during evidence synthesis.

E.1.5.1.3 BASIC CONSIDERATIONS

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).

Good: Chemical administration and characterization is complete (i.e., source, purity, and analytical verification of the test article are provided). There are no concerns about the composition, stability, or purity of the administered chemical, or the specific methods of administration. For inhalation studies, chemical concentrations in the exposure chambers are verified using reliable analytical methods.

Adequate: Some uncertainties in the chemical administration and characterization are identified but these are expected to have minimal impact on interpretation of the results (e.g., source and vendor- reported purity are presented, but not independently verified; purity of the test article is sub-optimal but not concerning; For inhalation studies, actual exposure concentrations are missing or verified with less reliable methods). Additionally, the full chemical name or the CASRN is reported.

Deficient: Uncertainties in the exposure characterization are identified and expected to substantially impact the results (e.g., source of the test article is not reported; levels of impurities are substantial or concerning; deficient administration methods, such as use of static inhalation chambers or a gavage volume considered too large for the species

and/or lifestage at exposure). Or, only the common name is reported and there is uncertainty in the exact chemical administered.

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).

E.1.5.1.4 EXAMPLE RATINGS

Oral study - Good - Source (3M) and purity (98%) are described, and the authors provided verification using analytical methods (GC/MS). Addressing concerns about known instability in solution for this chemical, the authors verified the dosing solutions twice weekly over the course of the experiment. Animals were exposed via gavage with all dose groups receiving the same volume.

Inhalation study - Good - Source (3M) and purity (98%) of the test article are described. All animals were transferred to dynamic inhalation exposure chambers for the exposures. The concentration of the test chemical in the air was continuously monitored from the animals' breathing zone throughout the 6-hour exposure periods and mean daily average concentrations and variability were reported.

E.1.5.2 Metric #2: Exposure timing, frequency, and duration

E.1.5.2.1 CORE QUESTION

Was the timing, frequency, and duration of exposure sensitive for the endpoint(s)/outcome(s) of interest?

E.1.5.2.2 PROMPTING QUESTIONS

For each endpoint/outcome or grouping of endpoints/outcomes in a study:

- Does the exposure period include the critical window of sensitivity?
- Was the duration and frequency of exposure sensitive for detecting the endpoint of interest?

E.1.5.2.3 BASIC CONSIDERATIONS

Considerations for this domain are highly variable depending on the endpoint(s)/outcome(s) of interest and must be refined by assessment teams.

Good: The duration and frequency of the exposure was sensitive and the exposure included the critical window of sensitivity (if known).

Adequate: The duration and frequency of the exposure was sensitive and the exposure covered most of the critical window of sensitivity (if known).

Deficient: The duration and/or frequency of the exposure is not sensitive and did not include the majority of the critical window of sensitivity (if known). These limitations are expected to bias the results towards the null.

Critically deficient: The exposure design was not sensitive and is expected to strongly bias the results towards the null. The rationale should indicate the specific concern(s).

E.1.5.2.4 EXAMPLE RATINGS

All Endpoints/Outcomes - Good - Study uses a standard OECD short-term (28-day) study design to examine toxicological effects that are routinely evaluated in this testing guideline.

Developmental and Male Reproductive effects - Good - The experimental design and exposure period were appropriate for evaluation of potential male reproductive and developmental effects. The experiment was designed to evaluate reproductive and developmental outcomes and followed recommendations in OECD 416 and EPA OPPT 870.3800 guidelines.

E.1.6 Domain #6: Outcome Methods/Results Presentation

E.1.6.1 Metric #1: Outcome Assessment

E.1.6.1.1 CORE QUESTION

Are the procedures sensitive and specific for evaluating the endpoint(s)/outcome(s) of interest?

E.1.6.1.2 PROMPTING QUESTIONS

For each endpoint/outcome or grouping of endpoints/outcomes in a study:

- Are there concerns regarding the specificity and validity of the protocols?
- Are there serious concerns regarding the sample size (see note)?
- Are there concerns regarding the timing of the endpoint assessment?

NOTE: Sample size alone is not a reason to conclude an individual study is critically deficient.

E.1.6.1.3 BASIC CONSIDERATIONS

Considerations for this domain are highly variable depending on the endpoint(s)/outcome(s) of interest and must be refined by assessment teams.

Examples of potential concerns include:

- Selection of protocols that are insensitive or non-specific for the endpoint of interest
- Use of unreliable methods to assess the outcome

- Assessment of endpoints at inappropriate or insensitive ages, or without addressing known endpoint variation (e.g., due to circadian rhythms, estrous cyclicity, etc.).
- Decreased specificity or sensitivity of the response due to the timing of endpoint evaluation, as compared to exposure (e.g., short-acting depressant or irritant effects of chemicals; insensitivity due to prolonged period of non-exposure prior to testing).

E.1.6.1.4 EXAMPLE RATING

Organ weight, body weights, and hormone measures - Good - No concerns regarding the specificity and validity of the protocols and measures were identified. Study authors used standard methodology for evaluating organ and body weights. Thyroid hormones were measured using commercial electrochemiluminescence-immunoassay methods, and the known diurnal variation in these measures was accounted for during blood collection.

E.1.6.2 Metric #2: Results presentation

E.1.6.2.1 CORE QUESTION

Are the results presented in a way that makes the data usable and transparent?

Additional Guidance:

- Consider whether sample size is provided (dose group size or number of animals evaluated/ endpoint)
- Consider whether the N was appropriate for the analysis
- Failure to mention sample size for each endpoint does NOT automatically reduce the quality category

Histopathology data:

- Good studies will provide an indication of the incidence and severity of a finding
- Adequate studies will provide indication of the incidence or severity. If findings are reported for a single animal from the group then a statement that the findings were representative of the entire group is required.
- Deficient studies provide results from a single animal without mention of it being representative of the entire group or without mention of incidence or severity (1 or a group of animals).

E.1.6.2.2 PROMPTING QUESTIONS

For each endpoint/outcome or grouping of endpoints/outcomes in a study:

- Does the level of detail allow for an informed interpretation of the results?
- Are the data analyzed, compared, or presented in a way that is inappropriate or misleading?

E.1.6.2.3 BASIC CONSIDERATIONS

Considerations for this domain are highly variable depending on the outcomes of interest and must be refined by assessment teams.

Examples of potential concerns include:

- Non-preferred presentation, such as developmental toxicity data averaged across pups in a treatment group, when litter responses are more appropriate
- Failing to present quantitative results
- Pooling data when responses are known or expected to differ substantially (e.g., across sexes or ages)
- Failing to report on or address overt toxicity when exposure levels are known or expected to be highly toxic
- Lack of full presentation of the data (e.g., presentation of mean without variance data; concurrent control data are not presented)

E.1.6.2.4 EXAMPLE RATING

All Endpoints/Outcomes - Good - There are no notable concerns about the way the results are analyzed or presented.

E.1.7 Domain #11: Overall Study Confidence

E.1.7.1 Metric #1: Overall Confidence

E.1.7.1.1 CORE QUESTION

Considering the identified strengths and limitations, what is the overall confidence rating for the endpoint(s)/outcome(s) of interest?

Additional Guidance:

- High confidence- at least 4 of 7 domains are good and the remaining are adequate but not deficient or critically deficient
- Medium confidence-there is a mix of good and adequate
- Low confidence- mix of good, adequate, and deficient
- Uninformative-critically deficient in 1 or more domains

E.1.7.1.2 PROMPTING QUESTIONS

For each endpoint/outcome or grouping of endpoints/outcomes in a study:

- Were concerns (i.e., limitations or uncertainties) related to the reporting quality, risk of bias, or sensitivity identified?

- If yes, what is their expected impact on the overall interpretation of the reliability and validity of the study results, including (when possible) interpretations of impacts on the magnitude or direction of the reported effects?

NOTE: Reviewers should mark studies that are rated lower than high confidence only due to low sensitivity (i.e., bias towards the null) for additional consideration during evidence synthesis. If the study is otherwise well-conducted and an effect is observed, the confidence may be increased.

E.1.7.1.3 BASIC CONSIDERATIONS

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.

High confidence: No notable concerns are identified (e.g. most or all domains rated Good).

Medium confidence: Some concerns are identified, but expected to have minimal impact on the interpretation of the results. (e.g., most domains rated Adequate or Good; may include studies with Deficient ratings if concerns are not expected to strongly impact the magnitude or direction of the results). Any important concerns should be carried forward to evidence synthesis.

Low confidence: Identified concerns are expected to significantly impact on the study results or their interpretation (e.g., generally, Deficient ratings for one or more domains). The concerns leading to this confidence judgment must be carried forward to evidence synthesis (see note).

Uninformative: Serious flaw(s) that make the study results unusable for informing hazard identification (e.g., generally, Critically Deficient rating in any domain; many Deficient ratings). Uninformative studies are not considered further in the synthesis and integration of evidence.

E.1.7.1.4 EXAMPLE RATINGS

Reproductive and developmental effects other than behavior - High Confidence - The study was well-designed for the evaluation reproductive and developmental toxicity induced by chemical exposure. The study applied established approaches, recommendations, and best practices, and employed an appropriate exposure design for these endpoints. Evidence was presented clearly and transparently.

Behavioral measures - Low Confidence - The cursory cage-side observations of activity are considered to be insensitive and non-specific methods for detecting motor effects, with a strong bias .

E.2 Study Quality Evaluation for Human Epidemiology Studies

Use this form to detail the study quality assessment for human studies.

One form should be completed per study design.

Do not include mechanistic data in the quality assessment.

If there are multiple studies within a publication, select "this form, next instance" to open a new and empty form.

Additional information on how to assign quality criteria can be found in the supporting documents.

E.2.1 Domain #2: Selection and Performance

E.2.1.1 Metric #3: Participant Selection

E.2.1.1.1 CORE QUESTION

Is there evidence that selection into or out of the study (or analysis sample) was jointly related to exposure and to outcome?

E.2.1.1.2 PROMPTING QUESTIONS

For longitudinal cohort:

- Did participants volunteer for the cohort based on knowledge of exposure and/or preclinical disease symptoms? Was entry into the cohort or continuation in the cohort related to exposure and outcome?

For occupational cohort:

- Did entry into the cohort begin with the start of the exposure?
- Was follow-up or outcome assessment incomplete, and if so, was follow-up related to both exposure and outcome status?
- Could exposure produce symptoms that would result in a change in work assignment/work status (“healthy worker survivor effect”)?

For case-control study:

- Were controls representative of population and time periods from which cases were drawn?
- Are hospital controls selected from a group whose reason for admission is independent of exposure?
- Could recruitment strategies, eligibility criteria, or participation rates result in differential participation relating to both disease and exposure?

For population-based survey:

- Was recruitment based on advertisement to people with knowledge of exposure, outcome, and hypothesis?

E.2.1.1.3 RATING GUIDANCE

Is there evidence that selection into or out of the study (or analysis sample) was jointly related to exposure and to outcome?

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), final analysis group. Does the study include potential vulnerable/susceptible groups or lifestages?

Add other concerns or limitations.

Add impact and direction to effect estimate, if applicable.

E.2.1.1.4 EXAMPLE TEXT

Adequate. Nested case-control design in Mexico City birth cohort with 30 cases of preterm birth and 30 controls selected randomly from same population of woman who were recruited during prenatal visits at one of four clinics (serving low to moderate income population). Recruitment and eligibility criteria (inclusion/exclusion criteria) discussed. Little discussion of participants versus nonparticipants but the available information indicates that differential selection is possible but not likely. Participation rate reported to be low (36%). Evaluates the vulnerable population of low-moderate income pregnant women.

E.2.2 Domain #5: Exposure Methods

E.2.2.1 Metric #3: Exposure measures

E.2.2.1.1 CORE QUESTION

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?

E.2.2.1.2 PROMPTING QUESTIONS

For all:

- Does the exposure measure capture the variability in exposure among the participants, considering intensity, frequency, and duration of exposure?
- Does the exposure measure reflect a relevant time window? If not, can the relationship between measures in this time and the relevant time window be estimated reliably?

- Was the exposure measurement likely to be affected by a knowledge of the outcome?
- Was the exposure measurement likely to be affected by the presence of the outcome (i.e., reverse causality)?

For case-control studies of occupational exposures:

- Is exposure based on a comprehensive job history describing tasks, setting, time period, and use of specific materials?

For biomarkers of exposure, general population:

- Is a standard assay used? What are the intra- and interassay coefficients of variation? Is the assay likely to be affected by contamination? Are values less than the limit of detection dealt with adequately?
- What exposure time-period is reflected by the biomarker? If the half-life is short, what is the correlation between serial measurements of exposure?

E.2.2.1.3 RATING GUIDANCE

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?

Source(s) of exposure (consumer products, occupational, an industrial accident) and source(s) of exposure data, blinding to outcome, level of detail for job history data, when measurements were taken, type of biomarker(s), assay information, reliability data from repeat measures studies, validation studies.

Add other concerns or limitations.

Add impact and direction to effect estimate, if applicable.

E.2.2.1.4 EXAMPLE TEXT

Poor for long-chained (DEHP, DiNP) and adequate for short-chained (DEP, DBP, DiBP) phthalate metabolites based on number of samples. A single spot (second morning void) urine sample was collected from each woman during a third-trimester visit to the project's research center; third trimester sample is relevant to later term preterm births. Analytical approach described and appropriate. High percent >LOD.

E.2.3 Domain #6: Outcome Methods/Results Presentation

E.2.3.1 Metric #3: Outcome measures

E.2.3.1.1 CORE QUESTION

Does the outcome measure reliably distinguish the presence or absence (or degree of severity) of the outcome?

E.2.3.1.2 PROMPTING QUESTIONS

For all:

- Is outcome ascertainment likely to be affected by knowledge of, or presence of, exposure (e.g. consider access to health care, if based on self-reported history of diagnosis)?

For case-control studies:

- Is the comparison group without the outcome (e.g., controls in a case-control study) based on objective criteria with little or no likelihood of inclusion of people with the disease?

For mortality measures:

- How well does cause of death data reflect occurrence of the disease in an individual? How well do mortality data reflect incidence of the disease?

For diagnosis of disease measures:

- Is diagnosis based on standard clinical criteria? If based on self-report of diagnosis, what is the validity of this measure?

For laboratory-based measures (e.g., hormone levels):

- Is a standard assay used? Does the assay have an acceptable level of interassay variability? Is the sensitivity of the assay appropriate for the outcome measure in this study population?

E.2.3.1.3 RATING GUIDANCE

Does the outcome measure reliably distinguish the presence or absence (or degree of severity) of the outcome?

Source of outcome (effect) measure, blinding to exposure status or level, how measured/classified, incident versus prevalent disease, evidence from validation studies, prevalence (or distribution summary statistics for continuous measures).

Add other concerns or limitations.

Add impact and direction to effect estimate, if applicable.

E.2.3.1.4 EXAMPLE TEXT

Adequate. Preterm birth defined by length of gestation (< 37 weeks), a standard measure of birth outcome, estimated by maternal recall of the date of last menstrual period, rather than the preferred early ultrasound. Potential misclassification of preterm cases due to maternal recall of last menstrual period to estimate gestational age which may be nondifferential with respect to exposure; however, differential misclassification is still possible but unlikely.

E.2.4 Domain #7: Confounding

E.2.4.1 Metric #1: Confounding

E.2.4.1.1 CORE QUESTION

Is confounding of the effect of the exposure likely?

E.1.4.1.2 PROMPTING QUESTIONS

Is confounding adequately addressed by considerations in:

- participant selection (matching or restriction)?
- accurate information on potential confounders, and statistical adjustment procedures?
- lack of association between confounder and outcome, or confounder and exposure in the study?
- information from other sources?

Is the assessment of confounders based on a thoughtful review of published literature, potential relationships (e.g., as can be gained through directed acyclic graphing), minimizing potential overcontrol (e.g., inclusion of a variable on the pathway between exposure and outcome)?

E.1.4.1.3 RATING GUIDANCE

Is confounding of the effect of the exposure unlikely?

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; degree of exposure to the confounder in the population.

Add other concerns or limitations.

Add impact and direction to effect estimate, if applicable.

E.1.4.1.4 EXAMPLE TEXT

Adequate. Information on key confounders was collected through questionnaire. The strategy for evaluating confounding and the process for retaining variables in the models was described. Rationale for selecting confounders not provided. Inclusion in model not solely based on statistical significance. Adjustment for relative co-exposures.

E.2.5 Domain #8: Analysis

E.2.5.1 Metric #1: Analysis

E.2.5.1.1 CORE QUESTION

Does the analysis strategy and presentation convey the necessary familiarity with the data and assumptions?

E.2.5.1.2 PROMPTING QUESTIONS

- 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)?

E.2.5.1.3 RATING GUIDANCE

Does the analysis strategy and presentation convey the necessary familiarity with the data and assumptions?

Extent (and if applicable, treatment) of missing data for exposure, outcome, and confounders, approach to modeling, classification of exposure and outcome variables (continuous versus categorical), testing of assumptions, sample size for specific analyses, relevant sensitivity analyses.

An ideal study would convey a thoughtful and thorough description of the analytical approach, and descriptive data for key variables (e.g., exposure measures, outcome measures), including the amount of missing data (or proportion less than the limit of detection [LOD]). The ideal analysis would use an appropriate and well thought out modeling approach for the study design (e.g., logistic regression for case-control data) and specify the covariates used in the final model; the methods should be described in enough detail such that they could be applied to the data from another study. In addition, the results should be presented with sufficient detail to enable estimation of

effect estimates and precision of the estimates (e.g., standard error [SE] or confidence interval [CI]

Add other concerns or limitations.

Add impact and direction to effect estimate, if applicable.

E.2.5.1.4 EXAMPLE TEXT

Adequate. Multivariable (multivariate) logistic regression used to take into account potential confounding variables; quantitative results presented (ORs and 95% CIs with ORs adjusted for confounders). Imputation techniques used when phthalate metabolite concentrations were below the LOD (filling in data where there wasn't); Amount of missing data not noted; Dichotomous exposure (reduced sensitivity) and use of median as the cut-off adjusted for urine creatinine and specific gravity to assess effect of method used.

E.2.6 Domain #9: Selective Reporting

E.2.6.1 Metric #1: Selective reporting

E.2.6.1.1 CORE QUESTION

Is there reason to be concerned about selective reporting?

E.2.6.1.2 PROMPTING QUESTIONS

Are only statistically significant results presented?

E.2.6.1.3 EXAMPLE TEXT

Adequate. No concerns for selective reporting.

E.2.6.1.4 RATING GUIDANCE

Is there concern for selective reporting?

Rating should be 2-level - Adequate or Deficient.

Are results presented with adequate detail for all the endpoints of interest? Are results presented for the full sample as well as for specified subgroups? Were stratified analyses (effect modification) motivated by a specific hypothesis?

E.2.7 Domain #10: Sensitivity

E.2.7.1 Metric #1: Study Sensitivity

E.2.7.1.1 CORE QUESTION

Is there a concern that sensitivity of the study is not adequate to detect an effect?

E.2.7.1.2 PROMPTING QUESTIONS

- Is the exposure range adequate?
- Was the appropriate population included?
- Was the length of follow-up adequate? Is the time/age of outcome ascertainment optimal given the interval of exposure and the health outcome?
- Are there other aspects related to risk of bias or otherwise that raise concerns about sensitivity?

E.2.7.1.3 RATING GUIDANCE

Are there concerns for study sensitivity?

What exposure range is spanned in this study? 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 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'). Is the study relevant to the exposure and outcome of interest?

Add other concerns or limitations.

Add impact and direction to effect estimate, if applicable.

E.2.7.1.4 EXAMPLE TEXT

Deficient. Small sample size/ Potential nondifferential misclassification of outcome and exposure. Low exposure levels. Range of exposure is narrow. Healthy worker effect.

E.2.8 Domain #11: Overall Study Confidence

E.2.8.1 Metric #2: Overall confidence

E.2.8.1.1 RATING GUIDANCE

Once the evaluation domains have been classified, these ratings will be combined to reach an overall study confidence classification of High, Medium, Low, or Uninformative.

This classification will be based on the classifications in the evaluation domains, and will include consideration of the likely impact of the noted deficiencies in bias and sensitivity on the results. Studies with critical deficiencies in any evaluation domain will be classified as Uninformative. Other classifications will generally follow a sorting such that High Confidence studies would have the highest evaluation ('Good') for all or most domains; Low Confidence studies would have a 'Poor' evaluation for one or more domains (unless the impact of the particular limitation(s) is judged to be unlikely to be severe), and Medium Confidence studies are in between these groups (e.g., most domains receiving a mid-level Adequate evaluation, with no limitations judged to be severe.) Once initial evaluation has been performed with consensus between reviewers,

the classifications will be re-evaluated, looking at the variability 'within' and 'between' levels to ensure that the separation between the levels of confidence is appropriate and that no additional criteria need to be considered.

Add other concerns or limitations.

Add impact and direction to effect estimate, if applicable.

E.2.8.1.2 EXAMPLE TEXT

Low confidence. Give brief rationale for rating.