EPA/635/R-23/167
IRIS Assessment Protocol
www.epa.gov/iris

Protocol for the Vanadium and Compounds (Inhalation) IRIS Assessment (Preliminary Assessment Materials)

August 2023

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

DISCLAIMER

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

CONTENTS

TABLES

FIGURES

ABBREVIATIONS

ADME	absorption, distribution, metabolism, and excretion
AOP	adverse outcome pathway
ATSDR	Agency for Toxic Substances and Disease Registry
BMD	benchmark dose
BMDL	benchmark dose lower confidence limit
BMDS	Benchmark Dose Software
CAS	Chemical Abstracts Service
CASRN	Chemical Abstracts Service registry number
COI	conflict of interest
CPAD	Chemical and Pollutant Assessment Division
CPHEA	Center for Public Health and Environmental Assessment
DNA	deoxyribonucleic acid
EPA	Environmental Protection Agency
HAWC	Health Assessment Workspace Collaborative
HEC	human equivalent concentration
HERO	Health and Environmental Research Online
IAP	IRIS Assessment Plan
IARC	International Agency for Research on Cancer
IRIS	Integrated Risk Information System
IUR	inhalation unit risk
LOAEL	lowest-observed-adverse-effect level
LOEL	lowest-observed-effect level
MeSH	Medical Subject Headings
MOA	mode of action
NOAEL	no-observed-adverse-effect level
NOEL	no-observed-effect level
NTP	National Toxicology Program
OAR	Office of Air and Radiation
OECD	Organisation for Economic Co-operation and Development
ORD	Office of Research and Development
OSF	oral slope factor
PBPK	physiologically based pharmacokinetic
PECO	populations, exposures, comparators, and outcomes
PK	pharmacokinetic
PM	particulate matter
POD	point of departure
RfC	inhalation reference concentration
RfD	oral reference dose
ROBINS I	Risk of Bias in Nonrandomized Studies of Interventions
UF	uncertainty factor
UF_H	human variation uncertainty factor
WOS	Web of Science

AUTHORS | CONTRIBUTORS | REVIEWERS

Assessment Managers

Suryanarayana Vulimiri, Ph.D.	EPA/ORD/CPHEA
Kathleen Newhouse, M.S.	

Assessment Team

Francesca Branch, DrPH	EPA/ORD/CPHEA
David Farrar, Ph.D.	
Elizabeth Radke, Ph.D.	
Erin Yost, Ph.D.	
John Stanek, Ph.D.	
Andre Weaver, Ph.D.	
Michelle Angrish, Ph.D.	
Bevin Blake, Ph.D.	
Michele Taylor, Ph.D.	
Martha Powers, Ph.D.	
Dustin Kapraun, Ph.D.	
Bidya Prasad, Ph.D. (formerly EPA)	

Executive Direction

Wayne E. Cascio, M.D. (CPHEA Director)	EPA/ORD/CPHEA
V. Kay Holt, M.S. (CPHEA Deputy Director)	
Samantha Jones, Ph.D. (CPHEA Associate Director)	
Kristina Thayer, Ph.D. (CPAD Director)	
Andrew Kraft, Ph.D. (CPAD Associate Director)	
Paul White, Ph.D. (CPAD Senior Science Advisor)	
Ravi Subramaniam, Ph.D. (CPAD Senior Science Advisor)	
Janice Lee, Ph.D. (Branch Chief)	
Elizabeth Radke-Farabaugh, Ph.D. (Branch Chief)	
Viktor Morozov, Ph.D. (Branch Chief)	
Glenn Rice, Ph.D. (Branch Chief)	
Deborah Segal, M.S. (Acting Branch Chief)	
Vicki Soto, B.S. (Branch Chief)	

Production Team

Maureen Johnson	EPA/ORD/CPHEA
Ryan Jones	
Dahnish Shams	
Avanti Shirke	
Jessica Soto-Hernandez	
Samuel Thacker	
Garland Waleko	
Brittany Schulz	Student Services Contractor, Oak Ridge Associated Universities (ORAU)

1. INTRODUCTION

The Integrated Risk Information System (IRIS) Program is undertaking an assessment of the health effects from inhalation exposure to vanadium and compounds.[1] IRIS assessments provide high quality, publicly available hazard identification and dose-response analyses on chemicals to which the public might be exposed. These assessments are not regulations but provide an important source of toxicity information used by the Environmental Protection Agency (EPA), state and local health agencies, tribes, other federal agencies, and international health organizations.

A draft IRIS Assessment Plan (IAP) for Vanadium and Compounds (Inhalation Exposure) was presented at a public science meeting on July 14, 2021 (see https://www.epa.gov/iris/iris-public-science-meeting-jul-2021) to seek input on the problem formulation components of the assessment plan (U.S. EPA, 2021a). The 2021 IAP specified why vanadium and compounds were selected for evaluation, described the objectives and specific aims of the assessment, provided draft populations, exposures, comparators, and outcomes (PECO) criteria, and identified key areas of scientific complexity. This assessment is being developed at the request of EPA's Office of Air and Radiation (OAR). It may also be used to support actions in other EPA Program and Regional Offices and can inform efforts to address inhalation exposure to vanadium by tribes, states, and international health agencies.

This protocol document includes the IAP content, revised based on public input, and updated EPA scoping needs, and presents the methods for conducting the systematic review and dose-response analysis for the assessment. While the IAP describes *what* the assessment covers, this protocol describes *how* the assessment is conducted (see Figure 1-1). The methods described in this protocol are based on the *Office of Research and Development (ORD) Staff Handbook for Developing Integrated Risk Information System (IRIS) Assessments* (referred to as the "IRIS Handbook") (U.S. EPA, 2022).

The IRIS Program posts assessment protocols on its website. Public input received is considered during preparation of the draft assessment.

[1]An assessment of oral exposure to vanadium and compounds was initiated prior to the inhalation assessment and is being performed separately (https://cfpub.epa.gov/ncea/iris_drafts/recordisplay.cfm?deid=348792).

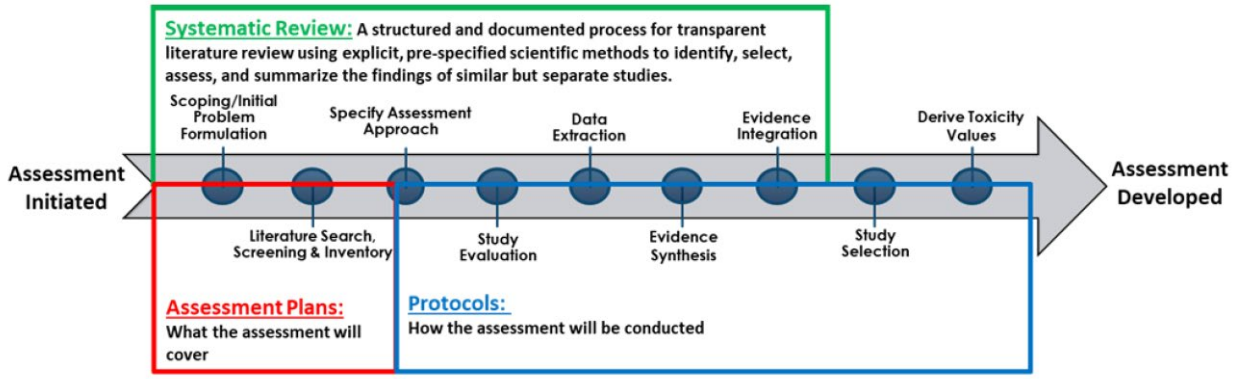


Figure 1-1. IRIS systematic review problem formulation and method documents.

2. SCOPING AND INITIAL PROBLEM FORMULATION

2.1. BACKGROUND

The IRIS Program finalized an assessment of vanadium and compounds in 1987 that included a reference dose (RfD) but did not derive inhalation toxicity values due to lack of inhalation data (U.S. EPA, 1987). Since then, several relevant studies on vanadium inhalation toxicity have been completed, including a two-year inhalation study conducted by the National Toxicology Program (NTP, 2002). The focus of this document is on inhalation exposure to vanadium and compounds and its potential impacts on human health. In this assessment "vanadium" refers to the element vanadium as part of environmentally occurring compounds. Vanadium as a pure metal is not found in the environment since it is unstable. Vanadium alloys (such as ferrovanadium) are not considered within scope because they are mixtures (see Table 5.1). Oral exposure to vanadium and compounds is currently under evaluation in a separate assessment (https://cfpub.epa.gov/ncea/iris_drafts/recordisplay.cfm?deid=348792). Section 2.1 provides a brief overview of aspects of the physiochemical properties, human exposure, and environmental fate characteristics of vanadium. This overview provides a summary of background information for contextual purposes only as it falls outside the scope of a human health toxicity assessment. This overview is not intended to provide a comprehensive description of the available information on these topics and is not recommended for use in decision-making. The reader is encouraged to refer to the source materials cited below, more recent publications on these topics, and authoritative reviews or assessments focused on these topics.

2.1.1. Physical and Chemical Properties

Vanadium has a complex chemistry, existing in the environment with three common oxidation states (+3, +4, +5) (Gustafsson, 2019). Pure elemental vanadium does not exist naturally (Rehder, 2015; ATSDR, 2012). Burning of fossil fuels containing vanadium results in the production of vanadium as oxides, including VO, V2O3, VO2, V2O5, which are emitted as fly ash into the atmosphere (Sturini et al., 2010; Crans et al., 1998; Mamane and Pirrone, 1998). In general, specific vanadium compounds relevant to environmental inhalation have not been well characterized in the literature, and few methods are available to speciate vanadium in particulate matter (PM) (Shafer et al., 2012; Sturini et al., 2010). Vanadium pentoxide (V2O5), a +5 vanadium species, is the most common compound of vanadium used for industrial applications such as metal alloy production and catalytic processes. In crude oils, vanadium is present as an organometallic complex, and upon burning in boilers or furnaces, vanadium is left behind as vanadium oxides in residual oil fly ash (IPCS, 2001). Residual oil fly ash is a mixture of different vanadium compounds and other metals and components of PM (Hauser et al., 1995). Most studies of human inhalation of vanadium, such as

those which evaluate vanadium as a component of PM, do not identify the specific vanadium compound, but instead report total vanadium in air or biological matrix. Evidence from one study suggests that pentavalent vanadium, in particular vanadium pentoxide, can be among the vanadium compounds present in PM emitted from diesel engines and in PM in urban atmospheric aerosols (Shafer et al., 2012).

When dissolved in water, vanadium speciation is a complex function of factors such as pH, redox potential, and vanadium concentration. Vanadate species (+5) predominate under oxic conditions and high pH, while vanadyl (+4) occurs under suboxic conditions and low pH and trivalent vanadium (+3) occurs under anoxic conditions (Gustafsson, 2019; Huang et al., 2015). Vanadium pentoxide undergoes hydrolytic reactions in water generating vanadate solutions (Cohen, 2007; Crans et al., 1998).

Table 2-1 lists the chemical identity of elemental vanadium (for reference only) and inorganic vanadium compounds that have been used in inhalation toxicology studies.

2.1.2. Sources, Production, and Uses

Vanadium is a transition metal that occurs naturally in Earth's crust and is a component of various minerals and most ores, tars, coal, and petroleum crude oils with heavy oils and bitumen from tar sands being especially rich in vanadium (WHO, 1988). Natural sources of vanadium in the air include continental dust, marine aerosol, and volcanic emissions (ATSDR, 2012). Vanadium has been reported to have natural background concentrations in the air ranging from tenths of a nanogram to a few nanograms (WHO, 2000). Vanadium is produced worldwide through mining or recycling residues and waste materials. In the US, the main method of production of vanadium is through reclamation (U.S. Department of Commerce, 2021).

The use of vanadium in industrial applications (e.g., steel production, vanadium redox-flow batteries, and catalytic converters) could contribute to the release of vanadium into the environment (Schlesinger et al., 2017; ATSDR, 2012). However, fossil fuel combustion is thought to be the major anthropogenic source of vanadium to the atmosphere, with vanadium found adsorbed onto PM as a result (Schlesinger et al., 2017; ATSDR, 2012).

2.1.3. Environmental Fate and Transport

As noted above, industrial processes, primarily burning of vanadium rich fuel, are reported to be the major source of vanadium in the atmosphere. Vanadium oxides generated during such combustion combine into particulate fly ash, also called residual oil fly ash (ROFA). ROFA is a mixture of different vanadium compounds and other metals and components of particulate matter (Hauser et al., 1995). At high temperatures encountered in combustion stacks (l00-500°C), lower oxides of vanadium will ultimately oxidize to varius vanadium oxides including V_2O_5 (U.S. EPA, 1985). Deposition is likely the only sink for atmospheric vanadium (Tullar and Suffet, 1975).

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

The ambient air concentrations of vanadium in the United States varies widely depending on factors including urban or rural location, seasonality, and geography (as reviewed by (ATSDR, 2012)). Generally, populations in cities located in the northeastern states have higher vanadium concentrations occuring during winter months when more fuel oil is burnt for heating and electricity (IARC, 2006). In addition, populations near port cities have greater exposure to vanadium due to higher concentrations of vanadium in marine vessel fuel and emissions (Spada et al., 2018); (Agrawal et al., 2009); (Peltier and Lippmann, 2010). Relatively recent publications, using air monitoring between 2007 and 2009, have reported average vanadium concentrations in New York City of approximately 5 ng/m^3 and ranging from approximately 2-15 ng/m^3 which were closely associated with ship traffic (Ito et al., 2016); (Peltier and Lippmann, 2010). Air monitoring in 2011 near the Seattle and Tacoma ports measured median vanadium concentrations of approximately 6-8 ng/m^3 (Spada et al., 2018). However, recent regulations limiting sulfur content of marine fuel oil have resulted in decreased vanadium emissions near ports, due to increased fuel refinement (Kodros et al., 2022); (Tao et al., 2013); (Spada et al., 2018).

Occupational exposure to vanadium in humans can occur through the inhalation of dust generated during vanadium processing and through the inhalation of residual oil fly ash (ROFA) during cleaning of oil-burning boilers and furnaces. Other occupational exposure occurs through cleaning of oil boilers, vanadium pentoxide production, and metallurgical processes (e.g., ferroalloy and V_2O_5 production facilities) associated with production of vanadium-containing vapors that would condense forming respirable aerosols (Kučera and Sabbioni, 1998).

Table 2-1. Chemical identity of vanadium compounds with repeat dose inhalation toxicity data (short term, subchronic, and chronic)

Name	Elemental vanadium[a]	Vanadium pentoxide	Bismuth orthovanadate	Sodium orthovanadate	Sodium metavanadate	Ammonium metavanadate	Vanadium dioxide
CASRN	7440-62-2	1314-62-1	14059-33-7	13721-39-6	13718-26-8	7803-55-6	12036-21-4
DTXSIDb	2040282	2023806	20893971	2037269	3044336	1052533	5065194
Empirical formula	V	V_2O_5	BiO_4V	Na_3VO_4	$NaVO_3$	NH_4VO_3	VO_2
Molecular mass (g/mol)	50.942	181.878	323.918	183.907	121.928	116.978	82.94
Oxidation state	0	+5	+5	+5	+5	+5	+4
Selected synonym(s)	Vanadium	Vanadium oxide; mu-oxido[tetrakis(oxido)] divanadium; divanadium pentoxide; vanadic anhydride; vanadin(V) oxide; vanadium(V) oxide	Bismuth vanadate(V) ($BiVO_4$); bismuth($^{3+}$) tetraoxidovanadate($^{3-}$); bismuth vanadium oxide; vanadic acid; bismuth vanadate ($BiVO_4$); bismuth vanadate yellow	Trisodium tetraoxidovanadate($^{3-}$); sodium vanadium oxide; trisodium vanadate; sodium vanadate(V); vanadic acid, trisodium salt	Sodium vanadate; sodium trioxidovanadate($^{1-}$); sodium vanadium oxide; sodium vanadium trioxide; vanadic acid, monosodium salt; sodium vanadate(V)	Ammonium trioxovanadate($^{1-}$); ammonium trisoxidovanadate($^{1-}$); ammonium monovanadate; ammonium vanadate(V); vanadic acid, ammonium salt; ammonium vanadium oxide; ammonium vanadium trioxide	Bisoxidovanadium; dioxido de vanadio; dioxyde de vanadium; divanadium tetraoxide; divanadium tetroxide; vanadium (IV) oxide; vanadium dioxide; vanadium (IV) oxide
Solubility (g/100 ml)	Insoluble	0.8 (20°C)	-	-	-	-	-
Melting point (◦C)	1.9×10^3 [c]	690[c]	-	858[c]	630[c]	200[d]	-
Boiling point (◦C)	3.0×10^3 [c]	1.75×10^3 [c]	-	-	-	-	-

[a]Elemental vanadium included for reference only.

[b]DTXSIDs are unique substance identifiers used for curation by the EPA's Distributed Structure Searchable Toxicity (DSSTox) project.

[c]Experimental average values for physicochemical properties are provided in EPA's CompTox Chemicals Dashboard at https://comptox.epa.gov/dashboard/. If no experimental values are available on EPA's CompTox Chemicals Dashboard, "-"is shown.

[d]ATSDR (2012)

2.2. SCOPING SUMMARY

During scoping, the IRIS Program met with EPA program and regional offices that had interest in an IRIS assessment for inhalation exposure to vanadium and compounds to discuss specific assessment needs. Table 2-2 summarizes input from this outreach. EPA's Office of Transportation and Air Quality within the Office of Air and Radiation (OAR) nominated vanadium compounds (including vanadium pentoxide) for an inhalation health assessment (both cancer and noncancer) under the IRIS Program. Vanadium pentoxide has been used as a catalyst to control emissions from diesel engines employed in mobile sources such as on-highway heavy-duty trucks, nonroad equipment, and marine vessels.[2] Under certain conditions, the use of vanadium in diesel engine emission control devices can result in the potential for exposures to vanadium compounds, such as vanadium pentoxide. A vanadium (inhalation) assessment could therefore help inform decisions about potential health risks from increased vanadium in the atmosphere.

Table 2-2. EPA program and regional office interest in the assessment of inhalation exposure to vanadium and compounds

EPA program or regional office	Oral[3]	Inhalation	Statute/Regulation	Anticipated uses/Interest
Office of Air and Radiation		✓	Clean Air Act	Vanadium and compounds (including vanadium pentoxide) are mobile source air toxics. Toxicological information developed for this assessment may be used to inform risk management decisions.

2.3. PROBLEM FORMULATION

The IRIS Program published an assessment of vanadium and compounds in 1987 that included a reference dose (RfD) for vanadium pentoxide, but no inhalation toxicity values (U.S. EPA, 1987). A draft IRIS assessment addressing inhaled vanadium pentoxide was released for public comment and external peer review in 2011 (U.S. EPA, 2011b), but was not finalized due to recognition of a cross-Agency need for an assessment with broader consideration of vanadium compounds (potentially aiding in the evaluation of toxic effects and helping to inform data gaps) (see: https://cfpub.epa.gov/ncea/iris2/chemicalLanding.cfm?substance_nmbr=125). EPA's

[2]OAR has issued an Advance Notice regarding plans for a new rulemaking that would establish new emission standards for oxides of nitrogen (NOx) and other pollutants for highway heavy-duty engines (https://www.epa.gov/regulations-emissions-vehicles-and-engines/advance-notice-proposed-rule-control-air-pollution-new).

[3]The IRIS Program is conducting a separate assessment of oral exposure to vanadium and compounds (U.S. EPA, 2021b).

Provisional Peer-Reviewed Toxicity Values (PPRTV) program developed a 2008 assessment on vanadium pentoxide which identified respiratory inflammation in female rats as the most sensitive endpoint for inhaled vanadium pentoxide in a two-year exposure study conducted by the National Toxicology Program (NTP, 2002). Later, in 2009, the PPRTV program also finalized an assessment that included soluble inorganic vanadium compounds other than vanadium pentoxide (U.S. EPA, 2009). However, this assessment found the evidence base was inadequate to support the derivation of chronic or subchronic inhalation toxicity values for soluble inorganic vanadium compounds other than vanadium pentoxide.

The IAP for Vanadium and Compounds (Inhalation) was released in May 2021 and presented at a public science meeting on July 14, 2021 (see https://www.epa.gov/iris/iris-public-science-meeting-jul-2021) to seek input on the problem formulation components of the assessment plan and key science issues (U.S. EPA, 2021a). This protocol considers input received on the 2021 IAP.

2.4. KEY SCIENCE ISSUES

The 2021 IAP for inhalation exposure to vanadium and compounds (inhalation) identified several key science issues based on the preliminary literature survey results and review of past assessments on inhalation exposure to vanadium (see Section 2.3).

- Key Science Issue #1: Consideration of vanadium speciation and oxidation state.

  Considering oxidation status could be important as preliminary results from oral exposure studies in rodents indicates increased toxicity of vanadium in the +5 oxidation state compared to vanadium +4 (Roberts et al., 2016). As noted in Section 2, vanadium in solution can convert between oxidation states and will form different species as a function of factors including pH, concentration, and redox potential. Study evaluations for the available inhalation studies, to the extent possible, will consider factors that could affect vanadium oxidation state and speciation [e.g., study methods that involved aerosolizing vanadium pentoxide (or other vanadium compounds) from solution, e.g., González-Villalva et al. (2011), rather than exposure to vanadium as a dust, e.g., NTP (2002)]. In addition, data to inform potential conversion between vanadium species and oxidation states in the body also will be evaluated and discussed in the assessment.

- Key Science Issue #2: Interpretation of data on noncancer respiratory responses to vanadium pentoxide.

  The two year NTP (2002) study reports increasing incidences of nonneoplastic lesions in the upper and lower respiratory tract of rats and mice (both sexes) with increasing vanadium pentoxide concentrations. Responses in all vanadium pentoxide exposure groups were highly elevated compared to controls. Information on the biology underlying these findings will aid interpretation of their use for hazard identification. Depending on the hazard identification decisions, methods for low-dose extrapolation and the associated uncertainties with any such approaches also would need to be explored and justified.

- Key Science Issue #3: Interpretation of data on rodent tumor responses.

  The NTP (2002) study also reports that tumor responses (alveolar/bronchiolar neoplasms) in male and female mice were highly elevated at all concentrations of vanadium pentoxide exposure: 70–80% increased incidence at the lowest tested vanadium concentration; control incidence in male mice was high (44%), but background incidence in females was very low (2%). Tumor incidence in male rats was elevated slightly but not statistically significant compared to controls. Previous reviews analyzed this tumor incidence against concurrent controls as well as historical controls, which will be useful in interpreting these data as they are considered in the assessment. In summary, aspects of the rodent tumor data noted above, and the uncertainties will be considered in the assessment.

- Key Science Issue #4. Cancer MOA for alveolar/bronchiolar neoplasms.

  There is some support for both a mutagenic MOA and an MOA dependent on cytotoxicity and reparative regeneration (and potentially other undetermined mechanisms) as suggested in the EPA PPRTV assessment (U.S. EPA, 2008). A similar lack of clearly delineated MOA(s) for alveolar/bronchiolar lung tumors with vanadium pentoxide exposure was proposed in the unfinalized draft IRIS Assessment of Vanadium Pentoxide (U.S. EPA, 2011b). As reported in these reviews, mutagenicity tests for vanadium pentoxide appear generally negative, but there is evidence of DNA strand breaks, aneuploidy, cytotoxicity, and cell proliferation. A focused evaluation of the available evidence regarding cancer MOA(s) for alveolar/bronchiolar neoplasms, including judgments regarding human relevance, is expected to be a key component of the vanadium (inhalation) IRIS assessment.

3. OVERALL OBJECTIVES AND SPECIFIC AIMS

3.1. OBJECTIVES

The overall objective of this assessment is to identify adverse health effects and characterize exposure-response relationships for these effects to support development of inhalation toxicity values. This assessment will use systematic review methods to evaluate the epidemiological and toxicological literature for vanadium compounds, including consideration of relevant mechanistic evidence. The evaluation conducted in this assessment will be consistent with relevant EPA guidelines.[4]

3.2. SPECIFIC AIMS

- Develop a systematic evidence map (SEM) to identify epidemiological (i.e., human), toxicological (i.e., experimental animal), and supplemental literature pertinent to characterizing the health effects of inhalation exposure to vanadium. The PECO criteria used to develop the SEM (referred to as "problem formulation PECO") is intended to identify the amount and type of evidence available to address a particular topic and is a useful scoping tool for health effects assessments (Thayer et al., 2022; NASEM, 2021; Wolffe et al., 2019).
    - Supplemental material content includes: mechanistic studies, including in vivo, in vitro, ex vivo, or in silico models; non-mammalian model systems; pharmacokinetic and absorption, distribution, metabolism, and excretion (ADME) studies; PM studies (as measured via air pollution monitoring stations); exposure characteristics (with no health outcome); data pertinent to identify susceptible populations; mixture studies; non-PECO routes of exposure; case studies (of 1-3 individuals); records with no original data; conference abstracts, and errata.
- Use the results of the SEM to (1) develop PECO criteria for the assessment (referred to as "assessment PECO"); (2) define the unit(s) of analysis at the level of endpoint or health outcome for hazard characterization; and (3) identify priority analyses of supplemental material to address the specific aims, uncertainties in hazard characterization, susceptibility, and dose-response analysis.
- Conduct study evaluations (risk of bias and sensitivity) for individual epidemiological and toxicological studies that meet assessment PECO criteria.
- Conduct a scientific and technical review for PBPK models considered for use in the assessment, if available. If a PBPK or PK model is selected for use, the most reliable dose

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

metric will be applied based on analyses of the available dose metrics and the outcomes to which they are being applied.

- Conduct data extraction (summarizing study methods and results) from epidemiological and animal toxicological studies that meet the assessment PECO criteria.

- For each evidence stream, and for each unit of analysis, use a structured framework to develop and describe the certainty of evidence across studies and the supporting rationale ("evidence synthesis"). Depending on the specific health endpoint or outcome, mechanistic information and precursor events may be included in a unit of analysis.

- For each health effect category, use a structured framework to develop and describe weight of evidence judgments across evidence streams and the supporting rationale for those judgments ("evidence integration"). The evidence integration analysis presents inferences and conclusions on human relevance of findings in animals, cross-evidence stream coherence, potentially susceptible populations and lifestages, biological plausibility, and other critical inferences supported by mechanistic, ADME, or PK/PBPK analyses.

- For each health effect category, summarize evidence synthesis (certainty of evidence) and evidence integration (weight of evidence) conclusions in an evidence profile table.

- As supported by the currently available evidence, derive chronic and subchronic inhalation reference concentrations (RfCs) and organ- or system-specific RfCs. Apply pharmacokinetic and dosimetry modeling (possibly including PBPK modeling) to account for interspecies differences, as appropriate. Derive an inhalation unit risk (IUR) as appropriate. Characterize confidence in any toxicity values that are derived.

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

4. LITERATURE SEARCH, SCREENING, AND INVENTORY

The literature search and screening processes described in this section were used to develop a SEM (here) using the problem formulation PECO (see Section 4.1) and supplemental screening criteria (see Section 4.2) to guide the inclusion of studies. The resulting inventory of studies identified in the SEM was used to develop assessment PECO criteria and identify priority analyses of supplemental material (described in Section 5). The initial literature search as well as all subsequent literature search updates are conducted using the processes described in this chapter. The literature inventory is continually updated with new studies as the assessment progresses.

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

PECO criteria are used to focus the assessment question(s), search terms, and inclusion criteria. To meet the PECO criteria a study must meet all PECO elements. The problem formulation PECO criteria used to develop the SEM presented in the IAP are presented in Table 4-1. The SEM PECO criteria were intentionally broad to identify all the available evidence in humans and animal models. As part of problem formulation, the SEM PECO is refined to develop the assessment PECO and these refinements are presented in Section 5.1.

Table 4-1. Populations, exposures, comparators, and outcomes (PECO) criteria

PECO element	Evidence
Populations	**Human:** Any population and life stage (occupational or general population, including children and other potentially sensitive populations). **Animal:** Nonhuman mammalian animal species (whole organism) at any life stage (including preconception, in utero, lactation, peripubertal, and adult stages). Studies of transgenic animals will be tracked as mechanistic studies under "potentially relevant supplemental material."
Exposures	**Relevant forms:** Any forms of vanadium. **Human:** Any exposure to vanadium compound(s) via the inhalation route, either explicitly stated or considered plausible based on exposure assessment. Exposure can be based on administered concentration, biomonitoring data (e.g., urine, blood, or other specimens), environmental or occupational measurements (e.g., air concentration), or job title or residence. Studies will be included if biomarkers of vanadium exposure are evaluated but the

PECO element	Evidence
	exposure route is unclear. Other exposure routes including oral will be tagged as "potentially relevant supplemental information." **Animal:** Any exposure to vanadium compound(s) via the inhalation route. Studies involving exposures to mixtures will be included only if they include an arm with exposure to a singular vanadium compound alone, otherwise, they will be tagged as "potentially relevant supplemental information." Other exposure routes, including intratracheal instillation, intranasal or oropharyngeal administration, oral, dermal, or injection, will be tagged as "potentially relevant supplemental information."
Comparators	**Human:** A comparison or referent population exposed to lower levels (or no exposure/exposure below detection limits) to vanadium compounds, or exposure for shorter periods of time, or cases versus controls. However, worker surveillance studies are considered to meet PECO criteria even if no referent group is presented. Case reports or case series of >3 people will be considered to meet PECO criteria, while case reports describing findings in 1–3 people in nonoccupational or occupational settings will be tagged as "potentially relevant supplemental information." **Animal:** A concurrent control group exposed to vehicle-only treatment, untreated control, or other treatment group with a different exposure duration.
Outcomes	All health outcomes (both cancer and noncancer). In general, endpoints related to clinical diagnostic criteria, disease outcomes, histopathological examination, or other apical/phenotypic outcomes are considered to meet PECO criteria and are prioritized for evidence synthesis over outcomes such as biochemical measures.
PK or PBPK models	Studies describing pharmacokinetic (PK) or physiologically based pharmacokinetic (PBPK) models for any form of vanadium will be included. **Classical Pharmacokinetic (PK) or Dosimetry Model Studies:** Classical PK or dosimetry modeling usually divides the body into just one or two compartments, which are not specified by physiology, where movement of a chemical into, between, and out of the compartments is quantified empirically by fitting model parameters to ADME (absorption, distribution, metabolism, and excretion) data. This category is for papers that provide detailed descriptions of PK models, that are not a PBPK model. Note: ADME studies often report classical PK parameters, such as bioavailability (fraction of an inhalation concentration absorbed), volume of distribution, clearance rate, or half-life or half-lives. If a paper only provides such results in tables with minimal description of the underlying model or software (i.e., uses standard PK software without elaboration), including "noncompartmental analysis," it should be listed only as a supplemental material ADME study. **Physiologically Based Pharmacokinetic (PBPK) or Mechanistic Dosimetry Model Studies:** PBPK models represent the body as various compartments (e.g., liver, lung, slowly perfused tissue, richly perfused tissue) to quantify the movement of chemicals or particles into and out of the body (compartments) by defined routes of exposure, metabolism, and elimination, and thereby estimate concentrations in blood or target tissues.

4.2. SUPPLEMENTAL CONTENT SCREENING CRITERIA

During the literature screening process, studies containing information that may be potentially relevant to the specific aims of the assessment are tagged as supplemental material by category. Some studies could emerge as being critically important to the assessment and may need to be evaluated and summarized at the individual study level (e.g., certain cancer MOA or ADME

studies), or might be helpful to provide context (e.g., provide hazard evidence from other routes or durations of exposure not meeting the assessment PECO), or might not be cited at all in the assessment (e.g., individual studies that contribute to a well-established scientific conclusion). Because it is often difficult to assess the impact of individual studies tagged as supplemental material on assessment conclusions at the screening stage, the tagging structure, allows for easy retrieval later in the assessment process. Table 4-2 presents the supplemental tagging structure presented in the July 2021 IAP. This structure was slightly refined to align with the IRIS Handbook methods(U.S. EPA, 2022) and the updated supplemental material tagging structure used in the draft assessment is presented in Section 5.2.

Table 4-2. Categories of Potentially Relevant Supplemental Material

Category	Evidence
Mechanistic studies	Studies reporting measurements related to a health outcome that inform the biological or chemical events associated with phenotypic effects, in both mammalian and nonmammalian model systems, including in vitro, in vivo (by any routes of exposure, includes transgenic models), ex vivo, and in silico studies. Genotoxicity tests are considered "mechanistic." Studies in which the chemical is used as a laboratory reagent generally do not need to be tagged (e.g., as a chemical probe used to measure antibody response).
Nonmammalian model systems	Studies in nonmammalian model systems, e.g., zebrafish, birds, *C. elegans*.
Non-inhalation route of administration	Studies in which humans or animals (whole organism) were exposed via a non-inhalation route (e.g., oral, injection, or dermal) and intratracheal, intranasal, or oropharyngeal routes of exposure. This categorization generally does not apply to epidemiological studies in which the exposure route may be unclear; such studies are considered to meet PECO criteria when inhalation exposure is plausible (further review of these studies will include consideration of whether route attribution can be inferred). Studies evaluating oral exposure to vanadium compounds are also under evaluation in a separate IRIS assessment (https://cfpub.epa.gov/ncea/iris_drafts/recordisplay.cfm?deid=348792).

Category	Evidence
Toxicokinetic (ADME)	Toxicokinetic (ADME) studies are primarily controlled experiments, where defined exposures usually occur by intravenous, oral, inhalation, or dermal routes, and the concentration of particles, a chemical, or its metabolites in blood or serum, other body tissues, or excreta are then measured. These data are used to estimate the amount absorbed (A), distributed (D), metabolized (M), or excreted/eliminated (E) through urine, breath, feces. The most informative studies are by the inhalation route and involve measurements over time such that the initial increase and subsequent concentration decline is observed, preferably at multiple exposure levels. Data collected from multiple tissues or excreta at a single timepoint, however, also inform distribution. ADME data also can be collected from human subjects who have had environmental or workplace exposures that are not quantified or fully defined. To be useful, however, such data must involve either repeated measurements over a period when exposure is known (e.g., is zero because previous exposure ended) ***or*** time- and subject-matched tissue or excreta concentrations (e.g., plasma and urine, or maternal and cord blood). ADME data, especially metabolism and tissue partition-coefficient information, can be generated using in vitro model systems. Although in vitro data may not be as definitive as in vivo data, these studies should also be tracked as ADME. For large evidence bases, separately tracking the in vitro ADME studies may be appropriate. *Studies describing environmental fate and transport or metabolism in bacteria are not tagged as ADME.
Exposure characteristics (no health outcome assessment)	Exposure characteristic studies include data that are unrelated to toxicological endpoints, but which provide information on exposure sources or measurement properties of the environmental agent (e.g., demonstrate a biomarker of exposure).
Mixture studies	Mixture studies that are not considered to meet PECO criteria because they do not contain an exposure or treatment group assessing only the chemical of interest. This categorization generally does not apply to epidemiological studies.
Case reports	Case reports of fewer than three subjects that describe health outcomes after exposure.
Records with no original data	Records that do not contain original data, such as other agency assessments, informative scientific literature reviews, editorials, or commentaries.
Conference abstracts/abstract only	Records that do not contain sufficient documentation to support study evaluation and data extraction.

4.3. USE OF EXISTING ASSESSMENTS

The ATSDR *Toxicological Profile for Vanadium* (ATSDR, 2012) was selected as the starting point for the literature search because it is the most recent review of health effects of vanadium and

compounds published by a U.S. federal government agency that has undergone public comment and external peer review. In addition, reference lists from existing assessments (final or publicly available drafts) were manually screened. References were identified from: PPRTV assessment of vanadium pentoxide (U.S. EPA, 2008), PPRTV assessment of vanadium and its soluble compounds other than vanadium pentoxide (U.S. EPA, 2009), IRIS External Review Draft assessment of vanadium pentoxide (U.S. EPA, 2011b), International Agency for Research on Cancer (IARC) document on vanadium pentoxide (IARC, 2006) as well as references pertinent to vanadium from the most recent Integrated Science Assessment for Particulate Matter (U.S. EPA, 2019c). All references from the 2012 ATSDR *Toxicological Profile for Vanadium*, literature searches, and other relevant assessments were extracted by an EPA information specialist and stored in the Health and Environmental Research Online (HERO) database.[5]

4.4. LITERATURE SEARCH STRATEGIES

4.4.1. Core Database Searches

Database searches were conducted to identify records that had been published since development of the 2012 ATSDR *Toxicological Profile for Vanadium*. The sources listed below were searched for records published between 2010 and 2021. The start date of 2010 was selected to ensure records published near the time of release of the ATSDR document were captured.

- PubMed (National Library of Medicine)
- Web of Science (Thomson Reuters)
- Toxline (National Library of Medicine)[6]

The database searches focused only on the chemical name (and synonyms or trade names) with no additional limits. The search terms were based on previous vanadium review efforts by IRIS and were reviewed carefully to ensure that a wide array of vanadium compounds were encompassed. Because each database has its own search architecture, the resulting search strategy was tailored to account for each database's unique search functionality. The detailed search strategies are presented in Appendix A. Literature searches are conducted using EPA's Health and Environmental Research Online (HERO) database.

The database searches will be updated throughout the assessment's development and review process to identify newly published literature. The last full literature search update is conducted several months prior to the planned release of the IRIS draft assessment for public

[5]Health and Environmental Research Online: https://hero.epa.gov/hero/.
[6]The Toxline database was migrated to PubMed prior to the March 2020 literature search update, so the Toxline search was conducted only in March 2019.

comment. The results returned (i.e., the number of references from each electronic database or other literature source described in 4.4.2 and 4.4.3), including the results of any literature search updates, are documented in the literature flow diagrams (see Section 4.4.2), which also reflect the literature screening decisions. The IRIS Program takes extra steps to ensure identification of pertinent studies by encouraging the scientific community and the public to identify additional studies and ongoing research and by considering late breaking studies that would impact the credibility of the conclusions, even during the review process. Studies identified after peer review begins are considered for inclusion only if they meet the PECO criteria and could fundamentally alter the assessment's primary conclusions or address key uncertainties (U.S. EPA, 2022).

4.4.2. Searching Other Sources

The literature search strategies described above are designed to be broad, but like in any search strategy, studies can be missed [e.g., cases where the specific chemical is not mentioned in title, abstract, or keyword content; ability to capture "gray" literature (studies not reported in the peer-reviewed literature) that is not indexed in the databases listed above]. Thus, in addition to the database searches, the sources below are used to identify studies that may have been missed in the database search. Records that appear to meet the problem formulation PECO criteria are uploaded into the screening software, annotated with respect to source of the record, and screened using the methods described in Appendix B. The list of other sources consulted includes:

- Manual review (at the title level) of reference list in studies screened as meeting problem formulation PECO after full-text review.

- Manual review (at the title level) of the reference list from publicly available final or draft assessments from EPA (e.g., IRIS and PPRTV) and other non-EPA Agencies (e.g., IARC [International Agency for Research on Cancer]) or published journal review specifically focused on human health.

- References from EPA's Toxicity Values database (ToxValDB), accessed via EPA's CompTox Chemicals Dashboard (https://comptox.epa.gov/dashboard/), to identify studies or assessments that present point of departure (POD) information. ToxValDB collates publicly available toxicity dose-effect related summary values typically used in risk assessments, many of which are from "gray literature" and are not available in databases such as Pub Med or Web of Science. These include POD data collected from data sources within EPA's ACToR (Aggregated Computational Toxicology Resource) and ToxRefDB (Toxicity Reference Database), and no-observed and lowest-observed (adverse) effect levels (NOEL, NOAEL, LOEL, LOAEL) data extracted from repeated dose toxicity studies submitted under European Union (EU) REACH regulation (Registration, Evaluation, Authorisation and Restriction of Chemicals). Also included are RfDs from EPA's IRIS and dose descriptors from EPA's PPRTV documents. Acute toxicity information is extracted from several different sources, including OECD eChemPortal, ECHA (European Chemicals Agency), NLM (National Library of Medicine) HSDB (Hazardous Substances Data Bank), ChemIDplus via EPA TEST (Toxicity Estimation Software Tool), and the EU JRC (Joint Research Centre) AcutoxBase. Data from the EU COSMOS project (Integrated in Silico Models for the Prediction of Human

Repeated Dose Toxicity of COSMetics to Optimise Safety) have also been included in ToxValDB. Although many of the resources included in the "Other Sources Consulted" list are represented in ToxValDB, they are also manually searched because most of the ToxValDB entries have not undergone quality control to ensure accuracy or completeness and might not include recent studies.

- European Chemicals Agency (ECHA) registration dossiers to identify data submitted by registrants http://echa.europa.eu/information-on-chemicals/information-from-existing-substances-regulation.

- EPA ChemView database (U.S. EPA, 2019a) to identify unpublished studies, information submitted to EPA under Toxic Substances Control Act (TSCA) Section 4 (chemical testing results), Section 8(d) (health and safety studies), Section 8(e) (substantial risk of injury to health or the environment notices), and For Your Information (FYI; voluntary documents). Other databases accessible via ChemView include EPA's High Production Volume (HPV) Challenge database (https://sor.epa.gov/sor_internet/registry/substreg/list/details.do?listId=74) and the Toxic Release Inventory database.

- National Toxicology Program (NTP) Chemical Effects in Biological Systems (CEBS) database of study results and research projects.

- The Organisation for Economic Co-operation and Development (OECD) Screening Information DataSet (SIDS) High Production Volume Chemicals http://www.inchem.org/pages/sids.html.

- Review of the list of references in the ECOTOX database for the chemical(s) of interest.

- The EPA CompTox (Computational Toxicology Program) Chemical Dashboard (U.S. EPA, 2019b) to retrieve a summary of any ToxCast or Tox21 high-throughput screening information. This data can be used to generate mechanistic insight, predict outcomes using appropriate models, and potentially inform dose-response modeling. The data's importance for outcome prediction and dose-response modeling depends on the context, size, and quality of retrieved results and the lack of availability of other data typically used for these purposes.

- References identified by the nominating program office, during public comment periods, by technical consultants, and during peer review.

4.4.3. Non-Peer Reviewed Data

IRIS assessments rely mainly on publicly accessible, peer-reviewed studies. However, it is possible that unpublished data directly relevant to the PECO may be identified during assessment development. In these instances, the EPA will try to get permission to make the data publicly available (e.g., in HERO); data that cannot be made publicly available are not used in IRIS assessments. In addition, on rare occasions where unpublished data would be used to support key assessment decisions (e.g., deriving a toxicity value). EPA may obtain external peer review if the owners of the data are willing to have the study details and results made publicly accessible, or if an

unpublished report is publicly accessible (or submitted to EPA in a nonconfidential manner)(U.S. EPA, 2015).This independent, contractor-driven peer review includes an evaluation of the study similar to that for peer review of a journal publication. The contractor would identify and typically select two to three scientists knowledgeable in scientific disciplines relevant to the topic as potential peer reviewers. Persons invited to serve as peer reviewers would be screened for conflict of interest. In most instances, the peer review would be conducted by letter review. The study authors are informed of the outcome of the peer review and given an opportunity to clarify issues or provide missing details. The study and its related information, if used in the IRIS assessment, would become publicly available. In the assessment, EPA would acknowledge that the document underwent external peer review managed by the EPA, and the names of the peer reviewers would be identified. In certain cases, IRIS will assess the utility of a data analysis of accessible raw data (with descriptive methods) that has undergone rigorous quality assurance/quality control review (e.g., ToxCast/Tox21 data, results of NTP studies not yet published) but that have not yet undergone external peer review.

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

4.5. LITERATURE SCREENING STRATEGY

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

4.5.1. Title and abstract-level screening

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

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

4.5.2. Full-text-level screening

Full-text references are sought through EPA's HERO database for studies screened as meeting problem formulation PECO criteria, or "unclear" based on TIAB screening. Full-text screening occurs in DistillerSR. Full-text copies of these citations are retrieved, stored in the HERO database, and independently assessed by two screeners using a structured form in DistillerSR to confirm eligibility. Screening conflicts are resolved by discussion among the primary screeners with consultation by a third reviewer or technical advisor (as needed to resolve any remaining disagreements). Rationales for excluding citations are documented, e.g., study did not meet problem formulation PECO, full-text not available. Approaches for language translation include online translation tools or engagement of a native speaker. Fee-based translation services for non-English studies are typically reserved for studies that are anticipated as being useful for toxicity value derivation. Conflicts between screeners in applying the supplemental material tags are resolved similarly, erring on the side of over tagging. Note that more granular sub-tagging of supplemental material occurs during preparation of the literature inventory as described in Table 4.2.

4.5.3. Multiple Citations with the Same Data

When there are multiple citations using the same or overlapping data, all citations are included, with one selected for use as the primary citation; the others are considered as secondary publications with annotation in HAWC and HERO indicating their relationship to the primary citation during data extraction. For epidemiology studies, the primary citation is generally the one with the longest follow-up, the largest number of cases, or the most recent publication date. For animal studies, the primary citation is typically the one with the longest duration of exposure, the largest sample size, or with the outcome(s) most informative to the PECO. For both epidemiology and animal studies, the assessment includes relevant data from all citations of the study; although, if the same outcome is reported in more than one citation, the data are extracted only once. For

corrections, retractions, and other companion documents to the included citations, a similar approach to annotation is taken and the most recently published data are incorporated into the assessments.

4.5.4. Literature Flow Diagram

The results of the screening process are posted on the project page for the assessment in the HERO database (https://heronet.epa.gov/heronet/index.cfm/project/page/project_id/2952) and studies have been "tagged" with appropriate category descriptors (e.g., included, excluded, potentially relevant supplemental material). Results for SEM screening against the problem formulation PECO are also summarized in a literature flow diagram (see Figure 4-1).

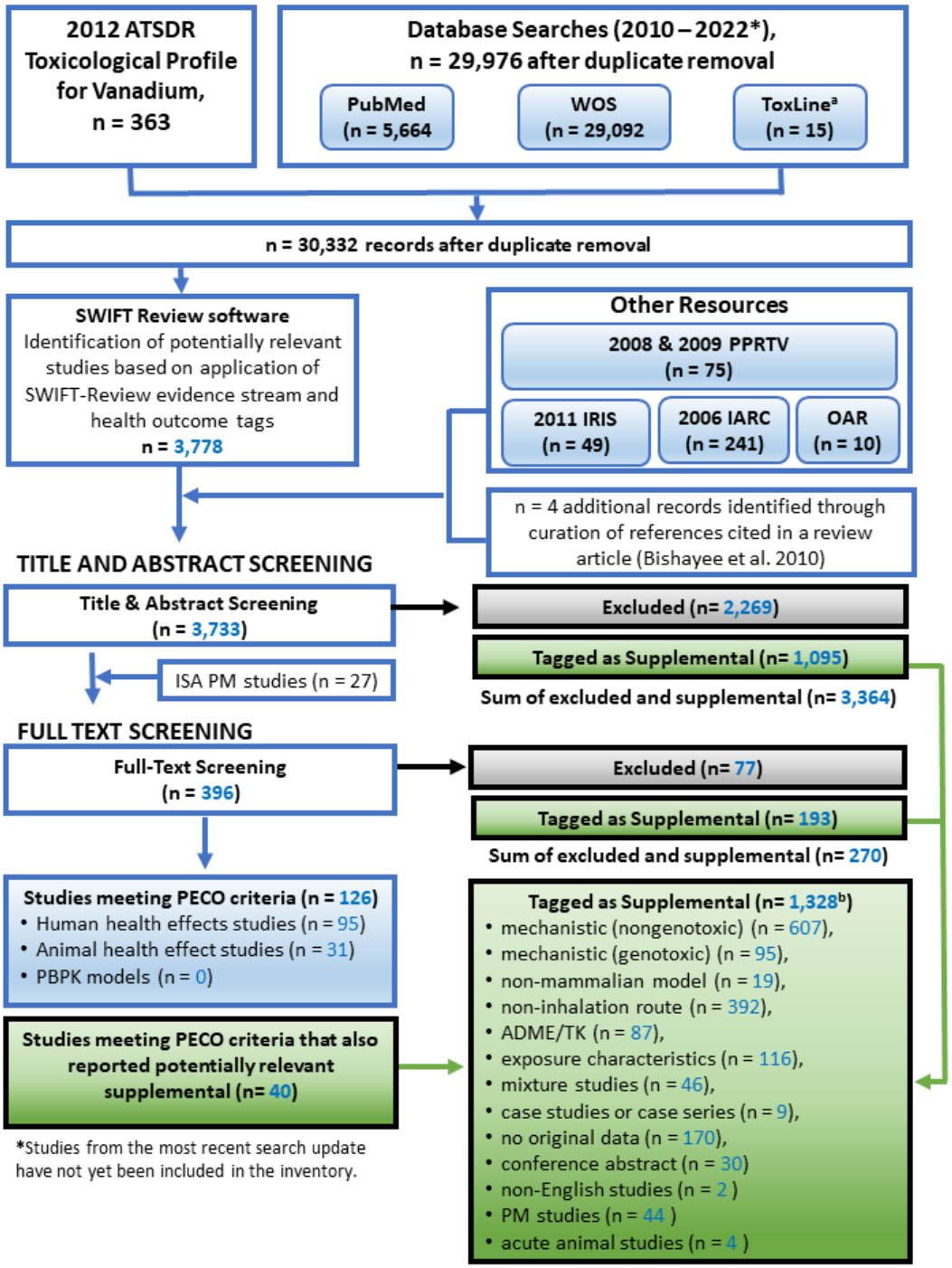


Figure 4-1. Literature search flow diagram for vanadium and compounds.

[a]The Toxline database was migrated to PubMed after the 2019 literature search update, thus it was not included in subsequent literature search updates.

[b]These numbers represent the total number of unique citations that were identified; because some citations are given multiple tags, the sum of the individual material tags is greater than the total number of citations.

4.6. LITERATURE INVENTORY

During title/abstract or full-text-level screening, citations that meet problem formulation PECO criteria are categorized by evidence type (human or animal) or category of supplemental information (e.g., mechanistic, PBPK, ADME). Next, study design details for citations that meet problem formulation PECO criteria are summarized as described in Section 4.6.1. A more granular tagging of supplemental material may be conducted as described in Section 4.6.2. The results of this categorization and tagging are referred to as the literature inventory and is the key analysis output of the SEM.

4.6.1. Studies That Meet Problem Formulation PECO Criteria

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

4.6.2. Organizational Approach for Supplemental Material

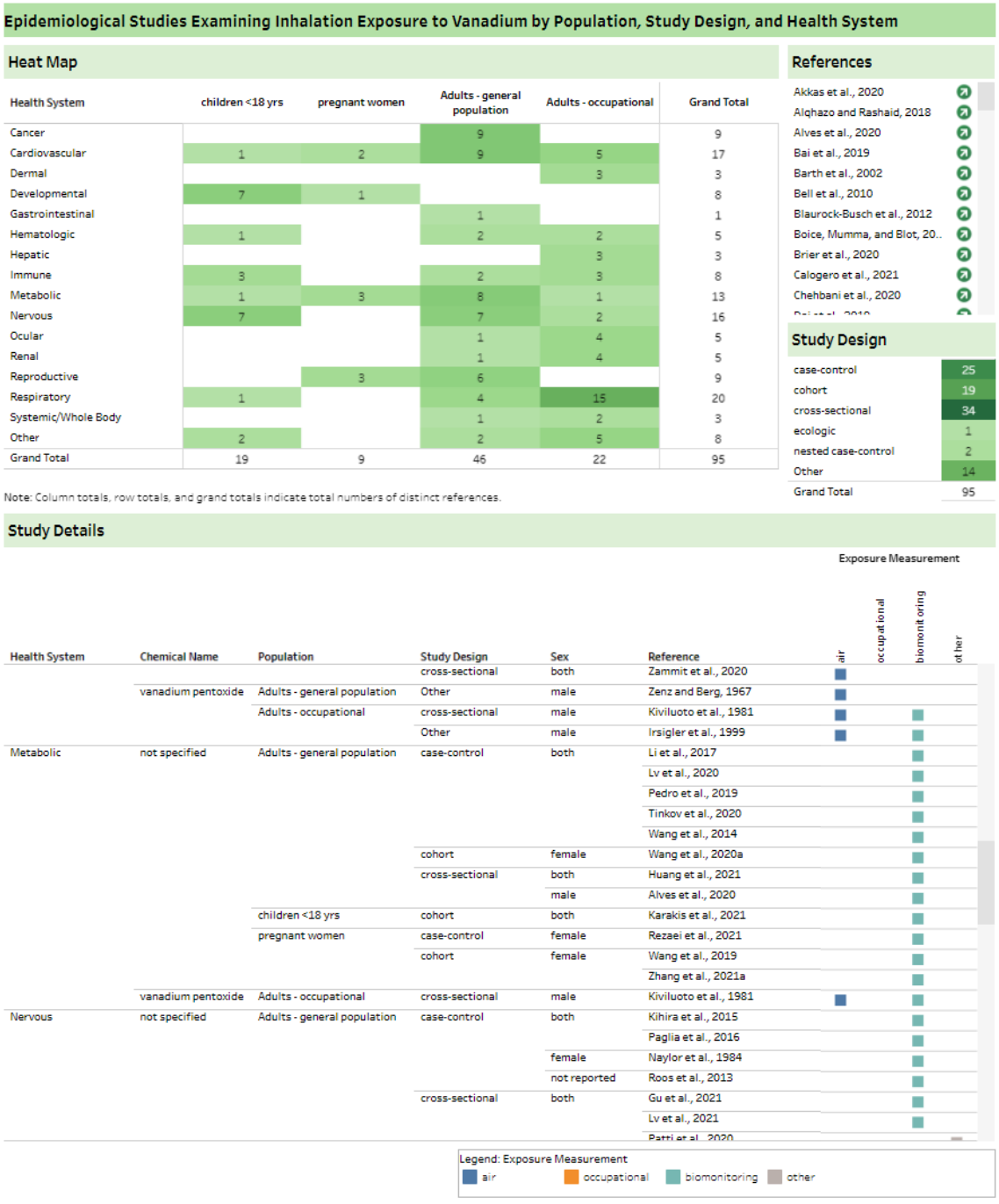
Inventories may also be created for other categories of studies that were tagged as "potentially relevant supplemental material" during screening, including mechanistic studies (e.g., in vitro or in silico models), ADME studies, and other studies that do not meet the specific PECO criteria but that may still be relevant to the research question(s). Here, the objective is to create an inventory of studies that can be tracked and further summarized as needed—for example, by model system, key characteristic [e.g., of carcinogens, (Smith et al., 2016)] mechanistic endpoint, or key event—to support analyses of critical questions that arise at various stages of the systematic review. The analysis of biological processes underlying vanadium-induced respiratory lesions, including lung tumor formation, was identified as a key science issue during problem formulation (see Section 2.4). Studies tagged as containing mechanistic information are inventoried to identify and organize data that can be used to evaluate the MOA(s) for vanadium induced respiratory effects.

4.7. INITIAL LITERATURE INVENTORIES FOR VANADIUM (INHALATION)

Literature inventories for PECO-relevant citations were created to develop summary-level, sortable lists that include some basic study design information (e.g., study population, exposure

information such as doses administered or biomarkers analyzed, age/life stage[7] of exposure, endpoints examined). These literature inventories facilitate subsequent review of individual studies or sets of studies by topic-specific experts. The literature inventory of studies meeting the problem formulation PECO criteria are presented in Figures 4-2 and 4-3 for human and animal studies, respectively. An interactive version of these figures, including additional study design details and a high-level summary of the results is available here.

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



Health System	children <18 yrs	pregnant women	Adults - general population	Adults - occupational	Grand Total
Cancer			9		9
Cardiovascular	1	2	9	5	17
Dermal				3	3
Developmental	7	1			8
Gastrointestinal			1		1
Hematologic	1		2	2	5
Hepatic				3	3
Immune	3		2	3	8
Metabolic	1	3	8	1	13
Nervous	7		7	2	16
Ocular			1	4	5
Renal			1	4	5
Reproductive		3	6		9
Respiratory	1		4	15	20
Systemic/Whole Body			1	2	3
Other	2		2	5	8
Grand Total	19	9	46	22	95

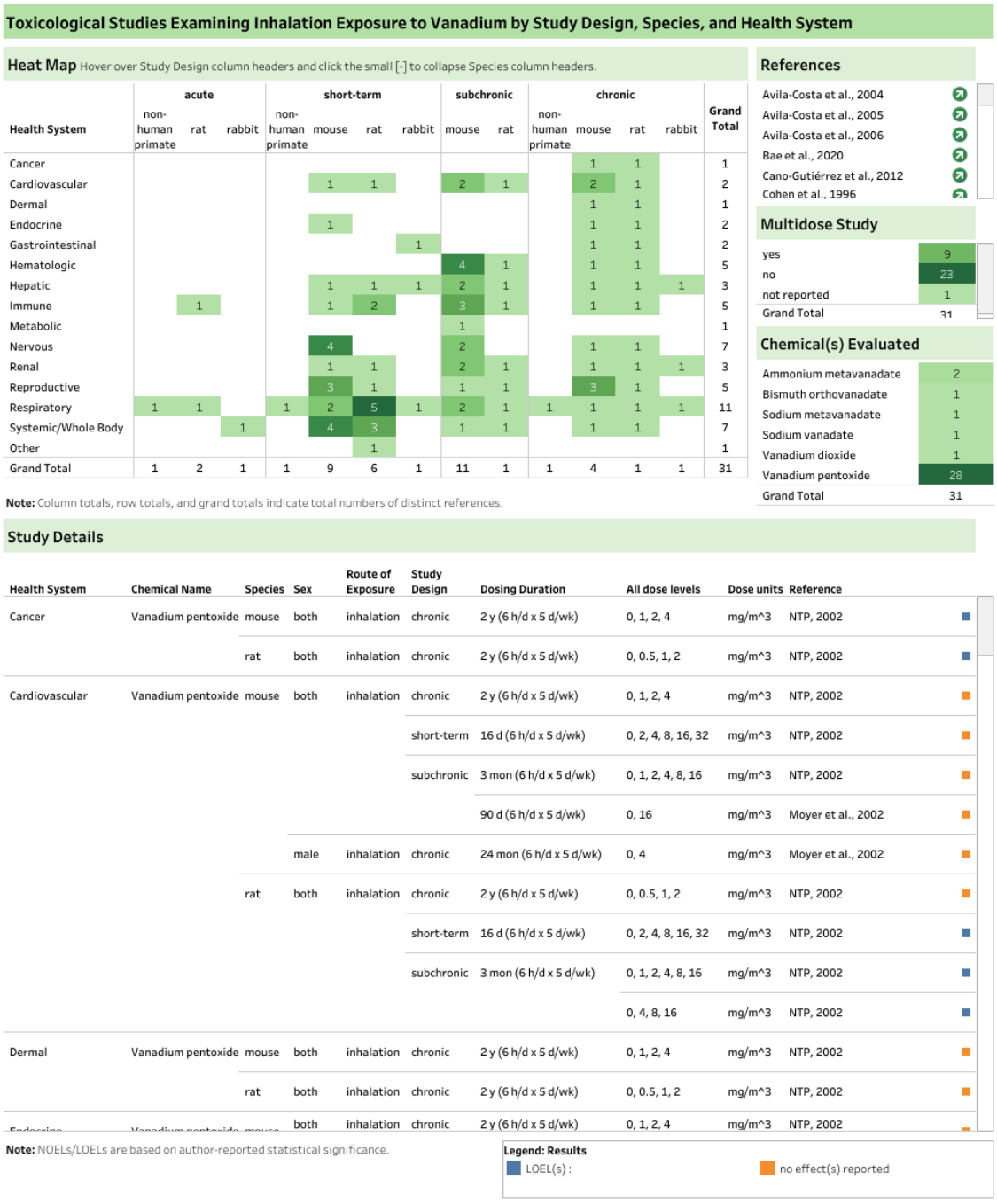
Note: Column totals, row totals, and grand totals indicate total numbers of distinct references.

References: Akkas et al., 2020; Alqhazo and Rashaid, 2018; Alves et al., 2020; Bai et al., 2019; Barth et al., 2002; Bell et al., 2010; Blaurock-Busch et al., 2012; Boice, Mumma, and Blot, 20..; Brier et al., 2020; Calogero et al., 2021; Chehbani et al., 2020

Study Design	
case-control	25
cohort	19
cross-sectional	34
ecologic	1
nested case-control	2
Other	14
Grand Total	95

Health System	Chemical Name	Population	Study Design	Sex	Reference	air	occupational	biomonitoring	other
			cross-sectional	both	Zammit et al., 2020	■			
	vanadium pentoxide	Adults - general population	Other	male	Zenz and Berg, 1967	■			
		Adults - occupational	cross-sectional	male	Kiviluoto et al., 1981	■		■	
			Other	male	Irsigler et al., 1999	■		■	
Metabolic	not specified	Adults - general population	case-control	both	Li et al., 2017			■	
					Lv et al., 2020			■	
					Pedro et al., 2019			■	
					Tinkov et al., 2020			■	
					Wang et al., 2014			■	
			cohort	female	Wang et al., 2020a			■	
			cross-sectional	both	Huang et al., 2021			■	
				male	Alves et al., 2020			■	
		children <18 yrs	cohort	both	Karakis et al., 2021			■	
		pregnant women	case-control	female	Rezaei et al., 2021			■	
			cohort	female	Wang et al., 2019			■	
					Zhang et al., 2021a			■	
	vanadium pentoxide	Adults - occupational	cross-sectional	male	Kiviluoto et al., 1981	■		■	
Nervous	not specified	Adults - general population	case-control	both	Kihira et al., 2015			■	
					Paglia et al., 2016			■	
				female	Naylor et al., 1984			■	
				not reported	Roos et al., 2013			■	
			cross-sectional	both	Gu et al., 2021			■	
					Lv et al., 2021			■	
					Patri et al., 2020			■	

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



Toxicological Studies Examining Inhalation Exposure to Vanadium by Study Design, Species, and Health System

Heat Map Hover over Study Design column headers and click the small [-] to collapse Species column headers.

Health System	acute			short-term				subchronic		chronic				Grand Total
	non-human primate	rat	rabbit	non-human primate	mouse	rat	rabbit	mouse	rat	non-human primate	mouse	rat	rabbit	
Cancer											1	1		1
Cardiovascular					1	1		2	1		2	1		2
Dermal											1	1		1
Endocrine					1						1	1		2
Gastrointestinal							1				1	1		2
Hematologic								4	1		1	1		5
Hepatic					1	1	1	2	1		1	1	1	3
Immune		1			1	2		3	1		1	1		5
Metabolic								1						1
Nervous					4			2			1	1		7
Renal					1	1		2	1		1	1	1	3
Reproductive					3	1		1	1		3	1		5
Respiratory	1	1		1	2	5	1	2	1	1	1	1	1	11
Systemic/Whole Body			1		4	3		1	1		1	1		7
Other						1								1
Grand Total	1	2	1	1	9	6	1	11	1	1	4	1	1	31

Note: Column totals, row totals, and grand totals indicate total numbers of distinct references.

References

Avila-Costa et al., 2004	
Avila-Costa et al., 2005	
Avila-Costa et al., 2006	
Bae et al., 2020	
Cano-Gutiérrez et al., 2012	
Cohen et al., 1996	

Multidose Study

yes	9
no	23
not reported	1
Grand Total	31

Chemical(s) Evaluated

Ammonium metavanadate	2
Bismuth orthovanadate	1
Sodium metavanadate	1
Sodium vanadate	1
Vanadium dioxide	1
Vanadium pentoxide	28
Grand Total	31

Study Details

Health System	Chemical Name	Species	Sex	Route of Exposure	Study Design	Dosing Duration	All dose levels	Dose units	Reference
Cancer	Vanadium pentoxide	mouse	both	inhalation	chronic	2 y (6 h/d x 5 d/wk)	0, 1, 2, 4	mg/m^3	NTP, 2002
		rat	both	inhalation	chronic	2 y (6 h/d x 5 d/wk)	0, 0.5, 1, 2	mg/m^3	NTP, 2002
Cardiovascular	Vanadium pentoxide	mouse	both	inhalation	chronic	2 y (6 h/d x 5 d/wk)	0, 1, 2, 4	mg/m^3	NTP, 2002
					short-term	16 d (6 h/d x 5 d/wk)	0, 2, 4, 8, 16, 32	mg/m^3	NTP, 2002
					subchronic	3 mon (6 h/d x 5 d/wk)	0, 1, 2, 4, 8, 16	mg/m^3	NTP, 2002
						90 d (6 h/d x 5 d/wk)	0, 16	mg/m^3	Moyer et al., 2002
			male	inhalation	chronic	24 mon (6 h/d x 5 d/wk)	0, 4	mg/m^3	Moyer et al., 2002
		rat	both	inhalation	chronic	2 y (6 h/d x 5 d/wk)	0, 0.5, 1, 2	mg/m^3	NTP, 2002
					short-term	16 d (6 h/d x 5 d/wk)	0, 2, 4, 8, 16, 32	mg/m^3	NTP, 2002
					subchronic	3 mon (6 h/d x 5 d/wk)	0, 1, 2, 4, 8, 16	mg/m^3	NTP, 2002
							0, 4, 8, 16	mg/m^3	NTP, 2002
Dermal	Vanadium pentoxide	mouse	both	inhalation	chronic	2 y (6 h/d x 5 d/wk)	0, 1, 2, 4	mg/m^3	NTP, 2002
		rat	both	inhalation	chronic	2 y (6 h/d x 5 d/wk)	0, 0.5, 1, 2	mg/m^3	NTP, 2002
[illegible]	[illegible]	[illegible]	both	inhalation	chronic	2 y (6 h/d x 5 d/wk)	0, 1, 2, 4	mg/m^3	NTP, 2002

Note: NOELs/LOELs are based on author-reported statistical significance.

Legend: Results
LOEL(s) : no effect(s) reported

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

5. REFINE PROBLEM FORMULATION AND SPECIFY ASSESSMENT APPROACH

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

5.1. REFINEMENTS TO PECO CRITERIA

Refinements to the problem formulation PECO criteria were made based on the creation of initial literature inventories, which are presented in here. The assessment PECO criteria (see Table 5-1) reflect the subset of studies that will be the focus of the systematic review and will move forward for study evaluation and evidence synthesis.

The vanadium and compounds (inhalation) IRIS assessment will focus on the health outcome categories identified in the literature inventory that appear to have sufficient information available to support hazard identification, i.e., respiratory, immune, reproductive, developmental, hepatic, renal, cardiometabolic, hematologic, nervous, and cancer. It is clear that in the absence of additional inhalation studies there will not be sufficient evidence to draw conclusions about gastrointestinal, dermal, ocular, or endocrine effects. Thus, unless more evidence becomes available, studies on these health outcomes will not undergo study evaluation or evidence synthesis to inform hazard characterization, and the information relating to those effects will be briefly summarized at the literature inventory level. Animal toxicological studies reporting effects tagged as "Systemic/Whole Body" (body weight, food/water consumption, mortality) that do not evaluate any other health systems also will not undergo study evaluation or evidence synthesis but can be considered to help interpret findings for other outcomes and will be summarized at the literature inventory level.

Many observational epidemiological studies evaluated health outcomes in relation to internal biomarkers of vanadium exposure (e.g., total vanadium in blood, urine, nails). However, the primary route of vanadium exposure is unclear. This body of studies will be evaluated and

synthesized as part of both the vanadium and compounds oral and vanadium and compounds inhalation assessments and may contribute to hazard identification. Studies where total vanadium was measured only as a component of PM, as measured via stationary monitoring sites were tagged as "potentially relevant supplemental material." Across these studies, there is considerable potential for exposure misclassification due to spatial and temporal heterogeneity of vanadium exposures [PM mass has less heterogeneity than the individual components (U.S. EPA, 2019c)]; thus, exposure estimates from stationary monitoring stations may not adequately distinguish between individuals. In addition, there are concerns for confounding by other PM components, which are difficult to disentangle given their high correlations. Even when multi-pollutant modeling is performed, there is potential for amplification bias from highly correlated co-exposures (Weisskopf et al., 2018). Because of these limitations, these studies are expected to be of lower confidence overall and are unlikely to provide sufficient evidence of an association with a health effect on their own. These studies will only undergo study evaluation and inclusion in evidence synthesis if they can inform the evaluation of an outcome that has evidence of adversity based on other epidemiology or animal toxicology data.

Table 5-1. Assessment PECO for the vanadium and compounds (inhalation) assessment

PECO element	Evidence
Populations	**Human:** Any population and life stage (occupational or general population, including children and other potentially sensitive populations). **Animal:** Nonhuman mammalian animal species (whole organism) at any life stage (including preconception, in utero, lactation, peripubertal, and adult stages). Studies of transgenic animals will be tracked as mechanistic studies under "potentially relevant supplemental material."
Exposures	**Relevant forms:** All forms of vanadium except alloys. Vanadium alloys will be tracked as supplemental. **Human:** Exposure to vanadium compound(s) via the inhalation route, either explicitly stated or considered plausible based on exposure assessment. Exposure can be based on administered concentration, biomonitoring data (e.g., urine, blood, or other specimens), environmental or occupational measurements (e.g., air concentration), or job title or residence. Studies will be included if biomarkers of vanadium exposure are evaluated but the exposure route is unclear. Other exposure routes including oral will be tagged as "potentially relevant supplemental information." Human inhalation exposure to vanadium (as a component of PM) measured via population stationary air monitoring sites will be tagged as "potentially relevant supplemental material" due to high potential for exposure misclassification (see Section 6.2.1). **Animal:** Any exposure to vanadium compound(s) via the inhalation route. Studies involving exposures to mixtures will be included only if they include an arm with exposure to a singular vanadium compound alone, otherwise, they will be tagged as "potentially relevant supplemental information." Other exposure routes, including intratracheal instillation, intranasal or oropharyngeal administration, oral, dermal, or injection, will be tagged as "potentially relevant supplemental information." Acute studies (<24 hours) will be included in the literature inventory as they can be helpful to interpret findings from studies more directly

PECO element	Evidence
	informative for developing a chronic toxicity value; however, these studies will be tagged as potentially relevant supplemental material and will not undergo study evaluation or full data extraction.
Comparators	**Human:** A comparison or referent population exposed to lower levels (or no exposure/exposure below detection limits) to vanadium, or exposure for shorter periods of time, or cases versus controls. However, worker surveillance studies are considered to meet PECO criteria even if no referent group is presented. Case reports or case series of >3 people will be considered to meet PECO criteria, while case reports describing findings in 1–3 people in nonoccupational or occupational settings will be tagged as "potentially relevant supplemental information." **Animal:** A concurrent control group exposed to vehicle only treatment, untreated control, or other treatment group with a different exposure duration.
Outcomes	Health outcomes: respiratory, immune, reproductive, developmental, hepatic, renal, cardiometabolic, hematologic, nervous, and cancer. In general, endpoints related to clinical diagnostic criteria, disease outcomes, histopathological examination, or other apical/phenotypic outcomes are considered to meet PECO criteria and are prioritized for evidence synthesis over outcomes such as biochemical measures.
PK or PBPK models	Studies describing pharmacokinetic (PK) or physiologically based pharmacokinetic (PBPK) models for any form of vanadium will be included. Classical Pharmacokinetic (PK) or Dosimetry Model Studies: Classical PK or dosimetry modeling usually divides the body into just one or two compartments, which are not specified by physiology, where movement of a chemical into, between, and out of the compartments is quantified empirically by fitting model parameters to ADME (absorption, distribution, metabolism, and excretion) data. This category is for papers that provide detailed descriptions of PK models, that are not a PBPK model. Note: ADME studies often report classical PK parameters, such as bioavailability (fraction of an inhalation concentration absorbed), volume of distribution, clearance rate, or half-life or half-lives. If a paper only provides such results in tables with minimal description of the underlying model or software (i.e., uses standard PK software without elaboration), including "noncompartmental analysis," it should be listed only as a supplemental material ADME study. Physiologically Based Pharmacokinetic (PBPK) or Mechanistic Dosimetry Model Studies: PBPK models represent the body as various compartments (e.g., liver, lung, slowly perfused tissue, richly perfused tissue) to quantify the movement of chemicals or particles into and out of the body (compartments) by defined routes of exposure, metabolism, and elimination, and thereby estimate concentrations in blood or target tissues.

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

5.1.1. Other Exclusions Based on Full-Text Content

In addition to failure to meet PECO criteria (described above), epidemiological and toxicological studies may be excluded at the full-text level due to critical reporting limitations. Reporting limitations can be identified during full-text screening but are more commonly identified during subsequent phases of the assessment (e.g., literature inventory, study evaluation). Regardless of when the limitation is identified, exclusions based on full-text content are

documented at the level of full-text exclusions in literature flow diagrams with a rationale of "critical reporting limitation."

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

Critical reporting information for different study types are summarized below:

Epidemiology studies

- Sample size
- Exposure characterization and/or measurement method
- Outcome ascertainment method
- Study design

Animal studies

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

In vitro studies prioritized for focused analysis

- Cell/tissue type(s) or test system
- Test article name
- Concentration and duration of treatment
- Quantitative or qualitative results for at least one endpoint of interest

5.2. REFINEMENTS TO SUPPLEMENTAL CONTENT SCREENING CRITERIA

As noted in the refinements to PECO criteria (see Section 5.1), studies evaluating exposure to vanadium as a component of PM and animal studies with acute (<24 hour) exposure durations will be considered as potentially relevant supplemental material in the assessment. A revised list of supplemental content screening criteria that includes these two categories is presented in Table 5-2.

Table 5-2. Categories of potentially relevant supplemental material

Category (Tag)	Description	Typical Assessment Use
Pharmacokinetics Data Potentially Informative to Assessment Analyses		
Pharmacokinetic (ADME)	Pharmacokinetic (ADME) studies are primarily controlled experiments, where defined exposures usually occur by intravenous, oral, inhalation, or dermal routes, and the concentration of particles, a chemical, or its metabolites in blood or serum, other body tissues, or excreta are then measured. These data are used to estimate the amount absorbed (A), distributed (D), metabolized (M), or excreted/eliminated (E) through urine, breath, feces. • The most informative studies are by the inhalation route and involve measurements over time such that the initial increase and subsequent concentration decline is observed, preferably at multiple exposure levels. Data collected from multiple tissues or excreta at a single timepoint, however, also inform distribution. • ADME data also can be collected from human subjects who have had environmental or workplace exposures that are not quantified or fully defined. To be useful, however, such data must involve either repeated measurements over a period when exposure is known (e.g., is zero because previous exposure ended) ***or*** time- and subject-matched tissue or excreta concentrations (e.g., plasma and urine, or maternal and cord blood). • ADME data, especially metabolism and tissue partition-coefficient information, can be generated using in vitro model systems. Although in vitro data may not be as definitive as in vivo data, these studies should also be tracked as ADME. For large evidence bases, separately tracking the in vitro ADME studies may be appropriate. *Studies describing environmental fate and transport or metabolism in bacteria are not tagged as ADME.	ADME studies are inventoried and prioritized for possible inclusion in an ADME synthesis section on the chemical's PK properties and for conducting quantitative adjustments or extrapolations (e.g., animal-to-human). Specialized expertise in PK is necessary for inventory and prioritization. Standard operating procedures for PBPK/PK model evaluation and the identification, organization, and evaluation of ADME studies are outlined in *An Umbrella Quality Assurance Project Plan (QAPP) for PBPK models (*U.S. EPA, 2018b*).*

Category (Tag)	Description	Typical Assessment Use
Supplemental Evidence Potentially Informative to Assessment Analyses		
Mechanistic studies	Studies reporting measurements related to a health outcome that inform the biological or chemical events associated with phenotypic effects, in either mammalian and nonmammalian model systems, including in vitro, in vivo (by any routes of exposure, includes transgenic models), ex vivo, and in silico studies. Genotoxicity tests are considered "mechanistic." Studies in which the chemical is used as a laboratory reagent generally do not need to be tagged (e.g., as a chemical probe used to measure antibody response).	Prioritized studies of mechanistic endpoints are described in the mechanistic synthesis sections; subsets of the most informative studies may become part of the units of analysis. Mechanistic evidence can provide support for the relevance of animal effects to humans and biological plausibility for evidence integration judgments (including MOA analyses, e.g., using the MOA framework in the US EPA Cancer Guidelines (U.S. EPA, 2005a)).
Nonmammalian model systems	Studies in nonmammalian model systems, e.g., zebrafish, birds, *C. elegans*.	Studies of non-PECO animal models can be summarized to inform evaluations of consistency (e.g., across species), coherence, or adversity; subsets of the most informative studies may be included in the unit of analysis. These studies may also be used to inform evidence integration judgments of biological plausibility and/or MOA analyses and thus may be summarized as part of the mechanistic evidence synthesis.
Non-inhalation route of administration	Studies in which humans or animals (whole organism) were exposed via a non-inhalation route (e.g., oral, injection, or dermal) and intratracheal, intranasal, or oropharyngeal routes of exposure. This categorization generally does not apply to epidemiological studies in which the exposure route may be unclear; such studies are considered to meet PECO criteria when inhalation exposure is plausible (further review of these studies will include consideration of whether route attribution can be inferred). Studies evaluating oral exposure to vanadium and compounds are also under evaluation in a separate IRIS assessment (https://cfpub.epa.gov/ncea/iris_drafts/recordisplay.cfm?deid=348792).	Routes of exposure which fall outside of the PECO may be summarized to inform evidence synthesis and integration judgments, and/or MOA analyses.

Category (Tag)	Description	Typical Assessment Use
Susceptible population	Studies that help to identify potentially susceptible subgroups, including studies on the influence of intrinsic factors such as sex, lifestage (including pregnancy and developmental period), or genotype, as well as some other factors (e.g., health status). These are often co-tagged with other supplemental material categories, such as mechanistic or ADME. Studies meeting PECO criteria that also address susceptibility should be co-tagged as supplemental. **Susceptibility based on most extrinsic factors, such as increased risk for exposure due to residential proximity to exposure sources, is not considered an indicator of susceptible populations for the purposes of IRIS assessments.*	Provides information on factors that might predispose sensitive populations or lifestages to a higher risk of adverse health effects following exposure to the chemical. This information is summarized during evidence integration for each health effect and is considered during dose-response, where it can directly impact modeling decisions.
PM studies	Human studies evaluating health outcomes associated with the vanadium component of particulate matter (as measured via air pollution monitoring stations). No study evaluation will be done on these studies.	Studies which provide more detailed exposure monitoring (e.g., personal air sampling, or occupational exposure measurements) will be considered PECO relevant and will undergo study evaluation.
Acute studies	Animal studies with acute exposure durations (defined as less than 24 hours) that otherwise meet PECO criteria.	Acute animal studies are retained in the literature inventory since they can be helpful in interpreting studies used in developing chronic toxicity values.
Non-English language studies	Records in foreign language with the abstracts in English.	For non-English language studies online translation tools (e.g., Google translator) or engagement with a native speaker can be used to summarize studies at the level of the literature inventory. Fee-based translation services for non-English studies are typically reserved for studies considered potentially informative for dose response, a consideration that occurs after preparation of the initial literature inventory during draft assessment development.
Background Information Potentially Useful to Problem Formulation and Protocol Development (These studies fall outside the scope of IRIS assessment analyses)		

Category (Tag)	Description	Typical Assessment Use
Exposure characteristics (no health outcome assessment)	Exposure characteristic studies include data that are unrelated to toxicological endpoints, but which provide information on exposure sources or measurement properties of the environmental agent (e.g., demonstrate a biomarker of exposure).	This information may be useful for developing exposure criteria for study evaluation or refining problem formulation decisions. Notably, providing an assessment of typical human exposures (e.g., sources, levels) falls outside the scope of an IRIS assessment.
Mixture studies	Animal studies which included co-exposure to multiple chemicals are not considered to meet PECO criteria because they do not contain an exposure or treatment group assessing only the chemical of interest.	Mixture studies are tracked to help inform cumulative risk analyses, which may provide useful context for risk assessment but fall outside the scope of an IRIS assessment.
Case studies or case series	Case reports of 1-3 subjects that describe health outcomes after exposure.	Tracking case studies can facilitate awareness of potential human health issues missed by other types of studies during problem formulation.
Reference Materials		
Records with no original data	Records that do not contain original data, such as other agency assessments, informative scientific literature reviews, editorials, or commentaries.	Studies that are tracked for potential use in identifying missing studies, background information, or current scientific opinions (e.g., hypothesized MOAs).
Posters or conference abstracts	Records that do not contain sufficient documentation to support study evaluation and data extraction.	

Underlined text show modifications in the categories of potentially relevant supplemental material since release of the IAP.

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

The planned units of analysis based on outcomes identified in the assessment PECO are summarized in Table 5-2. General considerations for defining the units of analysis are presented in the IRIS Handbook (U.S. EPA, 2022). Each unit of analysis is initially synthesized and judged separately within an evidence stream (see Section 8.1)

Evidence integration judgments focus on the stronger within evidence stream synthesis conclusions when multiple units of analysis are synthesized. The evidence synthesis judgments are used alongside other key considerations (i.e., human relevance of findings in animal evidence, coherence across evidence streams, information on susceptible populations or lifestages, and other critical inferences that draw on mechanistic evidence) to draw an overall evidence integration judgment for each health effect category or more granular health outcome grouping (see Section 8.2). As new evidence to inform potential vanadium-associated health hazards become available, updates to the units of analysis will be considered as appropriate.

Table 5-3. Human and animal endpoint grouping categories.

Health Effect Categories for Evidence Integration	Units of Analysis for Evidence Synthesis Integration (Each bullet represents a unit of analysis)	
	Human	**Animals**
Cancer	• Tumors and precancerous lesions	• Tumors and precancerous lesions
Cardiometabolic	• Cardiovascular outcomes (e.g., CVD mortality, hospital admissions) • Clinical effects (e.g., blood pressure, pulse) • Serum lipids, glucose, A1C	• Histopathology • Heart weight • Serum lipids, glucose
Developmental	• Fetal viability/pregnancy outcomes • Fetal structural alterations	• Offspring mortality/survival • Offspring growth • Developmental milestones (e.g., eye opening, incisor eruption, etc) • Structural alterations (e.g., external, skeletal, and soft tissue findings)

Health Effect Categories for Evidence Integration	Units of Analysis for Evidence Synthesis Integration (Each bullet represents a unit of analysis)	
	Human	**Animals**
Hematologic	• Red blood cell count and size, hematocrit or hemoglobin, platelets	• Red blood cell count and size, hematocrit or hemoglobin, platelets
Hepatic	• Clinical chemistry (e.g., ALT, AST, ALP, GGT)	• Histopathology • Liver weight • Clinical Chemistry (e.g., ALT, AST, ALP, GGT)
Immune	• Immune cell counts • Functional immune assays • Autoimmune response	• Immune organ weight histopathology • Immune cell counts • Functional immune assays
Nervous	• Neurodevelopmental/ neurobehavioral disorders • Neurological or sensory symptoms	• Histopathology • Brain weight • Functional observational battery, including motor activity and reflex responses • Learning and memory
Renal/Urinary	• Clinical Chemistry (e.g., BUN, CREA, KIM1, NGAL) • Urinalysis (protein, glucose) • Renal function (e.g., GFR)	• Histopathology • Organ weight • Clinical Chemistry (e.g., BUN, CREA, KIM1, NGAL) • Urinalysis (e.g., protein, glucose) • Renal function (e.g., GFR)
Reproductive	• Fertility • Pregnancy outcomes • Menstrual disorders	• Fertility and pregnancy outcomes • Histopathology • Reproductive organ weight • Reproductive hormones • Dam body weight/body weight gain

Health Effect Categories for Evidence Integration	Units of Analysis for Evidence Synthesis Integration (Each bullet represents a unit of analysis)	
	Human	**Animals**
Respiratory	• Pulmonary function (e.g., FEV, FVC, MEF) • Asthma incidence/severity • Respiratory symptoms (e.g., wheezing, irritation, shortness of breath)	• Lung weight • Histopathology • Pulmonary function

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

6. STUDY EVALUATION (RISK OF BIAS AND SENSITIVITY)

The general approach for evaluating the primary health effect studies meeting PECO criteria for all study types is described in Section 6.1. Instructional and informational materials for study evaluations are available at https://hawcprd.epa.gov/assessment/100000039/. The approach is conceptually the same for epidemiology, animal toxicology, and controlled human exposure, but the application specifics differ; thus, they are described separately in Sections 6.2, 6.3, and 6.4, respectively. Any physiologically based PBPK models used in the assessment are evaluated using methods described in the Quality Assurance Project Plan for PBPK models(U.S. EPA, 2018b) (see Section 6.5).

6.1. STUDY EVALUATION OVERVIEW FOR HEALTH EFFECT STUDIES

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

(a) Individual evaluation domains

Epidemiology	Animal	In vitro
• Exposure measurement • Outcome ascertainment • Participant selection • Confounding • Analysis • Selective reporting • Sensitivity	• Allocation • Observational bias/blinding • Confounding • Attrition • Chemical administration and characterization • Endpoint measurement • Results presentation • Selective reporting • Sensitivity	• Observational bias/blinding • Variable control • Selective reporting • Chemical administration and characterization • Endpoint measurement • Results presentation • Sensitivity

(b) Domain level judgments and overall study rating

Domain judgments

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

Overall study rating for an outcome

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

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

To calibrate the assessment specific considerations, the study evaluation process includes a pilot phase to assess and refine the evaluation process. Following this pilot, at least two reviewers independently evaluate studies to identify characteristics that bear on the informativeness (i.e., validity and sensitivity) of the results. The independent reviewers use structured web-forms for study evaluation housed within the EPA's version of HAWC (https://hawcprd.epa.gov/assessment/100000039/) to record separate judgments for each domain and the overall study for each outcome and unit of analysis, to reach consensus between reviewers, and when necessary, resolve differences by discussion between the reviewers or consultation with additional independent reviewers. As reviewers examine a group of studies, additional chemical specific knowledge or methodological concerns could emerge, and a second pass of all pertinent studies might become necessary.

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

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

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

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

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

- *Adequat*e indicates a judgment that methodological limitations related to the evaluation domain are (or are likely to be) present, but those limitations are unlikely to be severe or to notably impact the study results or interpretation of the study findings.

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

- *Not reported* indicates the information necessary to evaluate the domain question was not available in the study. Depending on the expected impact, the domain may be interpreted as *adequate* or *deficient* for the purposes of the study confidence rating.

- *Critically deficient* reflects a judgment that the study conduct relating to the evaluation domain introduced a serious flaw that is interpreted to be the primary driver of any observed effect(s) or makes the study uninterpretable. Studies with *critically deficient* judgments in any evaluation domain are almost always classified as overall *uninformative* for the relevant outcome(s).

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

- *High* confidence: No notable deficiencies or concerns were identified; the potential for bias is unlikely or minimal, and the study used sensitive methodology. *High* confidence studies generally reflect judgments of *good* across all or most evaluation domains.

- *Medium* confidence: Possible deficiencies or concerns were identified, but the limitations are unlikely to have a significant impact on the study results or their interpretation. Generally, *medium* confidence studies include *adequate* or *good* judgments across most domains, with the impact of any identified limitation not being judged as severe.

- *Low* confidence: Deficiencies or concerns are identified, and the potential for bias or inadequate sensitivity is expected to have a significant impact on the study results or their interpretation. Typically, *low* confidence studies have a *deficient* evaluation for one or more domains, although some *medium* confidence studies might have a *deficient* rating in domain(s) considered to have less influence on the magnitude or direction of effect estimates. *Low* confidence results are given less weight compared to *high* or m*edium* confidence results during evidence synthesis and integration (see Sections 7 and 8) and are generally not used as the primary sources of information for hazard identification or derivation of toxicity values, unless they are the only studies available (in which case, this significant uncertainty would be

emphasized during dose-response analysis). Studies rated *low* confidence only because of sensitivity concerns are asterisked or otherwise noted because they often require additional consideration during evidence synthesis. Effects observed in studies that are biased toward the null may increase confidence in the results, assuming the study is otherwise well conducted (see Section 8).

- *Uninformative*: Serious flaw(s) are judged to make the study results uninterpretable for use in the assessment. Studies with *critically deficient* judgments in any evaluation domain are almost always rated *uninformative*. Studies with multiple *deficient* judgments across domains may also be considered *uninformative*. Given that the findings of interest are considered uninterpretable based on the identified flaws (see above definition of *critically deficient*) and do not provide information of use to assessment interpretations, these studies have no impact on evidence synthesis or integration judgments and are not useable for dose-response analyses but may be used to highlight research gaps.

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

6.2. EPIDEMIOLOGY STUDY EVALUATION

Evaluation of epidemiology studies of health effects to assess risk of bias and study sensitivity are conducted for the following domains: exposure measurement, outcome ascertainment, participant selection, potential confounding, analysis, study sensitivity, and selective reporting. Bias can result in false positives and negatives, whereas study sensitivity is typically concerned with identifying the latter.

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

reviews. Exposure and confounding domain considerations specific to vanadium are presented in Sections 6.2.1.

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

Domain	Types of information that may need to be collected or are important for evaluating the domain
Exposure measurement	Source(s) of exposure (e.g., consumer products, occupational, an industrial accident) and source(s) of exposure data, blinding to outcome, level of detail for job history data, when measurements were taken, type of biomarker(s), assay information, reliability data from repeated-measures studies, validation studies.
Outcome ascertainment	Source of outcome (effect) measure, blinding to exposure status or level, how measured/classified, incident vs. prevalent disease, evidence from validation studies, prevalence (or distribution summary statistics for continuous measures).
Participant selection	Study design, where and when was the study conducted, and who was included? Recruitment process, exclusion and inclusion criteria, type of controls, total eligible, comparison between participants and nonparticipants (or followed and not followed), and final analysis group. Does the study include potential susceptible populations or life stages (see discussion in Table 8.6)?
Confounding	Background research on key confounders for specific populations or settings; participant characteristic data, by group; strategy/approach for consideration of potential confounding; strength of associations between exposure and potential confounders and between potential confounders and outcome; and degree of exposure to the confounder in the population.
Analysis	Extent (and if applicable, treatment) of missing data for exposure, outcome, and confounders; approach to modeling; classification of exposure and outcome variables (continuous vs. categorical); testing of assumptions; sample size for specific analyses; and relevant sensitivity analyses.
Sensitivity	What are the ages of participants (e.g., not too young in studies of pubertal development)? What is the length of follow-up (for outcomes with long latency periods)? Choice of referent group, the exposure range, and the level of exposure contrast between groups (i.e., the extent to which the "unexposed group" is truly unexposed, and the prevalence of exposure in the group designated as "exposed").
Selective reporting	Are results presented with adequate detail for all the endpoints and exposure measures reported in the methods section, and are they relevant to the PECO? Are results presented for the full sample as well as for specified subgroups? Were stratified analyses (effect modification) motivated by a specific hypothesis?

Table 6-2. Domains, questions, and general considerations to guide the evaluation of epidemiology studies

Domain and core question	Prompting questions	Follow-up questions	Criteria that apply to most exposures and outcomes
Exposure measurement Does the exposure measure reliably distinguish between levels of exposure in a time window considered most relevant for a causal effect with respect to the development of the outcome?	For all: • Does the exposure measure capture the variability in exposure among the participants, considering intensity, frequency, and duration of exposure? • Does the exposure measure reflect a relevant time window? If not, can the relationship between measures in this time and the relevant time window be estimated reliably? • Was the exposure measurement likely to be affected by 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, period, and use of specific materials? For biomarkers of exposure, general population: • Is a standard assay used? What are the intra- and inter-assay coefficients of variation? Is the assay likely to be affected by contamination? Are values less than the limit of detection dealt with adequately? • What exposure period is reflected by the biomarker? If the half-life is short, what is the correlation between serial measurements of exposure?	Is the degree of exposure misclassification likely to vary by exposure level? If the correlation between exposure measurements is moderate, is there an adequate statistical approach to ameliorate variability in measurements? If potential for bias is a concern, is the predicted direction or distortion of the bias on the effect estimate (if there is enough information)?	**Good** • Valid exposure assessment methods used, which represent the etiologically relevant period of interest. • Exposure misclassification is expected to be minimal. **Adequate** • Valid exposure assessment methods used, which represent the etiologically relevant period of interest. • Exposure misclassification could exist but is not expected to greatly change the effect estimate. **Deficient** • Valid exposure assessment methods used, which represent the etiologically relevant time period of interest. Specific knowledge about the exposure and outcome raises concerns about reverse causality, but whether it is influencing the effect estimate is uncertain. • Exposed groups are expected to contain a notable proportion of unexposed or minimally exposed individuals, the method did not capture important temporal or spatial variation, or other evidence of exposure misclassification would be expected to notably change the effect estimate. **Critically deficient** • Exposure measurement does not characterize the etiologically relevant period of exposure or is not valid. • Evidence exists that reverse causality is very likely to account for the observed association. • Exposure measurement was not independent of outcome status.

Domain and core question	Prompting questions	Follow-up questions	Criteria that apply to most exposures and outcomes
<u>Outcome ascertainment</u> Does the outcome measure reliably distinguish the presence or absence (or degree of severity) of the outcome?	For all: • Is outcome ascertainment likely affected by knowledge, or presence, of exposure (e.g., consider access to health care, if based on self-reported history of diagnosis)? For case-control studies: • Is the comparison group without the outcome (e.g., controls in a case-control study) based on objective criteria with little or no likelihood of inclusion of people with the disease? For mortality measures: • How well does cause-of-death data reflect occurrence of the disease in an individual? How well do mortality data reflect incidence of the disease? For diagnosis of disease measures: • Is the diagnosis based on standard clinical criteria? If it is based on self-report of the diagnosis, what is the validity of this measure? For laboratory-based measures (e.g., hormone levels): • Is a standard assay used? Does the assay have an acceptable level of inter-assay variability? Is the sensitivity of the assay appropriate for the outcome measure in this study population?	Is there a concern that any outcome misclassification is nondifferential, differential, or both? What is the predicted direction or distortion of the bias on the effect estimate (if there is enough information)?	**Good** • High certainty in the outcome definition (i.e., specificity and sensitivity), minimal concerns with respect to misclassification. • Assessment instrument was validated in a population comparable to the one from which the study group was selected. **Adequate** • Moderate confidence that outcome definition was specific and sensitive, some uncertainty with respect to misclassification but not expected to greatly change the effect estimate. • Assessment instrument was validated but not necessarily in a population comparable to the study group. **Deficient** • Outcome definition was not specific or sensitive. • Uncertainty regarding validity of assessment instrument. **Critically deficient** • Invalid/insensitive marker of outcome. • Outcome ascertainment is very likely to be affected by knowledge of, or presence of, exposure. Note: Lack of blinding should not be automatically construed to be *critically deficient*.

Participant selection Is there evidence that selection into or out of the study (or analysis sample) was jointly related to exposure and to outcome?	For longitudinal cohort: • Did participants volunteer for the cohort on the basis of knowledge of exposure or preclinical disease symptoms? Was entry into, 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 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?	Were differences in participant enrollment and follow-up evaluated to assess bias? If potential for bias is a concern, what is the predicted direction or distortion of the bias on the effect estimate (if there is enough information)? Were appropriate analyses performed to address changing exposures over time relative to symptoms? Is there a comparison of participants and nonparticipants to address whether differential selection or study retention/continuation is likely?	**Good** • Minimal concern for selection bias based on description of recruitment process and follow-up (e.g., selection of comparison population, population-based random sample selection, recruitment from sampling frame including current and previous employees). • Exclusion and inclusion criteria specified and would not induce bias. • Participation rate is reported at all steps of study (e.g., initial enrollment, follow-up, selection into analysis sample). If rate is not high, appropriate rationale is given for why it is unlikely to be related to exposure (e.g., comparison between participants and nonparticipants or other available information indicates differential selection is not likely). **Adequate** • Enough of a description of the recruitment process to be comfortable that there is no serious risk of bias. • Inclusion and exclusion criteria specified and would not induce bias. • Participation rate is incompletely reported but available information indicates participation is unlikely to be related to exposure. **Deficient** • Little information on recruitment process, selection strategy, sampling framework, and participation OR aspects of these processes raises the potential for bias (e.g., healthy worker effect, survivor bias). **Critically deficient** • Aspects of the processes for recruitment, selection strategy, sampling framework, or participation result in concern that selection bias is likely to have had a large impact on effect estimates (e.g., convenience sample with no information about recruitment and selection, cases and controls are recruited from different sources with different likelihood of

Domain and core question	Prompting questions	Follow-up questions	Criteria that apply to most exposures and outcomes
			exposure, recruitment materials stated outcome of interest and potential participants are aware of or are concerned about specific exposures).
<u>Confounding</u> Is confounding of the effect of the exposure likely?	Is confounding adequately addressed by considerations in: • Participant selection (matching or restriction)? • Accurate information on potential confounders and statistical adjustment procedures? • Lack of association between confounder and outcome, or confounder and exposure in the study? • Information from other sources? Is the assessment of confounders based on a thoughtful review of published literature, potential relationships (e.g., as can be gained through directed acyclic graphing), and minimizing potential overcontrol (e.g., inclusion of a variable on the pathway between exposure and outcome)?	If potential for bias is a concern, what is the predicted direction or distortion of the bias on the effect estimate (if there is enough information)?	**Good** • Conveys strategy for identifying key confounders, including co-exposures. This may include a priori biological consideration, published literature, causal diagrams, or statistical analyses, with the recognition that not all "risk factors" are confounders. • Inclusion of potential confounders in statistical models not based solely on statistical significance criteria (e.g., $p < 0.05$ from stepwise regression). • Does not include variables in the models likely to be influential colliders or intermediates on the causal pathway. Key confounders are evaluated appropriately and considered unlikely sources of substantial confounding. This often will include: ○ Presenting the distribution of potential confounders by levels of the exposure of interest or the outcomes of interest (with amount of missing data noted); ○ Consideration that potential confounders were rare among the study population, or were expected to be poorly correlated with exposure of interest; ○ Consideration of the most relevant functional forms of potential confounders; ○ Examination of the potential impact of measurement error or missing data on confounder adjustment; or ○ Presenting a progression of model results with adjustments for different potential confounders, if warranted.

Domain and core question	Prompting questions	Follow-up questions	Criteria that apply to most exposures and outcomes
			Adequate • Similar to good but might not have included all key confounders, or less detail might be available on the evaluation of confounders (e.g., sub bullets in good). That residual confounding could explain part of the observed effect is possible, but concern is minimal. **Deficient** • Does not include variables in the models shown to be influential colliders or intermediates on the causal pathway. • And any of the following: o The potential for bias to explain some results is high based on an inability to rule out residual confounding, such as a lack of demonstration that key confounders of the exposure-outcome relationships were considered. o Descriptive information on key confounders (e.g., their relationship relative to the outcomes and exposure levels) are not presented; or o Strategy of evaluating confounding is unclear or is not recommended (e.g., only based on statistical significance criteria or stepwise regression [forward or backward elimination]). **Critically deficient** • Includes variables in the models that are colliders or intermediates in the causal pathway, indicating that substantial bias is likely from this adjustment; or • Confounding is likely present and not accounted for, indicating that all results were most likely due to bias.
Analysis Does the analysis strategy and presentation convey	• Are missing outcome, exposure, and covariate data recognized, and if necessary, accounted for in the analysis?	If potential for bias is a concern, what is the predicted direction or distortion of the bias on	**Good**

Domain and core question	Prompting questions	Follow-up questions	Criteria that apply to most exposures and outcomes
the necessary familiarity with the data and assumptions?	• Does the analysis appropriately consider variable distributions and modeling assumptions? • Does the analysis appropriately consider subgroups or lifestages 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)?	the effect estimate (if there is enough information)?	• Use of an optimal characterization of the outcome variable, including presentation of subgroup- or lifestage-specific comparisons (as appropriate for the outcome). • Quantitative results presented (effect estimates and confidence limits or variability in estimates) (i.e., not presented only as a p-value or "significant"/"not significant"). • Descriptive information about outcome and exposure provided (where applicable). • Amount of missing data noted and addressed appropriately (discussion of selection issues—missing at random vs. differential). • Where applicable, for exposure, includes LOD (and percentage below the LOD), and decision to use log transformation. • Includes analyses that address robustness of findings, e.g., examination of exposure-response (explicit consideration of nonlinear possibilities, quadratic, spline, or threshold/ceiling effects included, when feasible); relevant sensitivity analyses; effect modification examined based only on a priori rationale with sufficient numbers. • No deficiencies in analysis evident. Discussion of some details might be absent (e.g., examination of outliers). **Adequate** • Same as 'Good', except: • Descriptive information about exposure provided (where applicable) but might be incomplete; might not have discussed missing data, cut-points, or shape of distribution(s).

Domain and core question	Prompting questions	Follow-up questions	Criteria that apply to most exposures and outcomes
			• Includes analyses that address robustness of findings (examples in 'Good'), but some important analyses are not performed. **Deficient** • Does not conduct analysis using optimal characterization of the outcome variable. • Descriptive information about exposure levels not provided (where applicable). • Effect estimate and p-value presented, without standard error or confidence interval. • Results presented as statistically "significant"/ "not significant." **Critically deficient** • Analysis methods are not appropriate for design or data of the study.
Selective reporting Is there reason to be concerned about selective reporting?	• Were results provided for all the primary analyses described in the methods section? • Is appropriate justification given for restricting the amount and type of results shown? • Are only statistically significant results presented?	If potential for bias is a concern, what is the predicted direction or distortion of the bias on the effect estimate (if there is enough information)?	**Good** • The results reported by study authors are consistent with the primary and secondary analyses described in a registered protocol or methods paper. **Adequate** • The authors described their primary (and secondary) analyses in the methods section and results were reported for all primary analyses. **Deficient** • Concerns were raised based on previous publications, a methods paper, or a registered protocol indicating that analyses were planned or conducted that were not reported, or that hypotheses originally considered to be secondary were represented as primary in the reviewed paper.

Domain and core question	Prompting questions	Follow-up questions	Criteria that apply to most exposures and outcomes
			• Only subgroup analyses were reported, suggesting that results for the entire group were omitted. • Only statistically significant results were reported.
Sensitivity Is there a concern that sensitivity of the study is not adequate to detect an effect?	• Is the exposure contrast adequate to detect associations and exposure-response relationships? • Was the appropriate population or lifestage 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? • Do other aspects related to risk of bias or otherwise raise concerns about sensitivity?		**Good** • There is sufficient variability/contrast in exposure to evaluate primary hypotheses. • The study population was sensitive to the development of the outcomes of interest (e.g., ages, lifestage, sex). • The timing of outcome ascertainment was appropriate given expected latency for outcome development (i.e., adequate follow-up interval). • The study was adequately powered to observe an effect. • No other concerns raised regarding study sensitivity. **Adequate** Same considerations as *Good*, except: • There may be issues identified that could reduce sensitivity, but they are considered unlikely to substantially impact the overall findings of the study. **Deficient** • Concerns were raised about the considerations described for Good that are expected to notably decrease the sensitivity of the study to detect associations for the outcome. **Critically deficient** • Severe concerns were raised about the sensitivity of the study such that any observed associations are likely to be explained by bias.

6.2.1. Epidemiology Study Evaluation Considerations Specific to Vanadium

6.2.1.1 Exposure measurement

Exposure to vanadium by the inhalation route may be measured based on occupational exposure (e.g., job duties), air sampling, or biomonitoring data, or a combination of these. Criteria for evaluating each of these information types are summarized in Table 6-3, with some additional considerations described below.

Biomarker measurements of total vanadium may represent exposure via any route. Where possible, evaluations will indicate the likely predominant route; studies where exposure is likely to be primarily via inhalation will be given more weight. Measurements from urine, blood, hair, or toenails will be considered to be relevant to either acute or long-term continuous exposure. Metal concentrations in hair or toenails may reflect exposures during the previous several months based on their rate of growth, although the precise exposure window has not been investigated for vanadium (Gutiérrez-González et al., 2019). Toenail vanadium was strongly correlated with vanadium in hair (r = 0.61) in a study of 26 adults (primarily workers) (Raińska et al., 2005). Validated reference values are available for hair, blood, plasma, and urine using Inductively Coupled Plasma Mass Spectrometry (ICP-MS) (Goullè et al., 2005). Quality control procedures include the use of certified reference material generated by individual laboratories, recovery analysis, procedural blanks, duplicate samples, or spike samples. Sample mass has been associated with concentrations measured in toenails; therefore, correction methods are necessary.

Well-established and sensitive methods for measurement of total vanadium concentrations include measurement using graphite furnace atomic absorption (GF-AAS) (with a preconcentration procedure), isotope dilution mass spectrometry (ID-MS), inductively coupled plasma mass spectrometry (ICP-MS), and neutron activation analysis (NAA) with radiochemical separation. Detection limits of these methods have been summarized previously (ATSDR, 2012). Because toxic properties of vanadium species differ, measurements that report vanadium species are preferred to measurements of total vanadium. If only total vanadium were measured in the sample media used for internal biomarker measurement, and there were no other serious limitations in the measurement of exposure (e.g., invalid measure, inappropriate timing of measurement, inadequate detail of analysis including quality control), the exposure measurement domain would be rated *adequate* rather than *good* to reflect the reduced sensitivity resulting from combining the effects of vanadium species.

Occupational exposure to vanadium compounds can occur in a variety of occupational settings, including mining and/or processing of vanadium ore, maintenance of oil-burning boilers, and some steel production processes. In most occupational studies of vanadium inhalation exposures, the vanadium compound is identified in the indicated job category. It is preferred if these categories are validated by a quantitative measure such as personal sampling of air or biomonitoring matrices in at least some participants.

Vanadium in air is generally measured via regular monitoring for particulate matter (PM) and its components. However, the spatial distribution for the monitoring of PM in air is limited; in the U.S., only a few hundred monitoring stations capture data that allow for analysis of individual PM components (from the Chemical Speciation Network (CSN) and the Interagency Monitoring of Protection Visual Environments (IMPROVE) systems). Given that vanadium levels are more spatially and temporally variable than PM mass (U.S. EPA, 2019c), there is considerable uncertainty as to whether monitoring measurements can provide accurate estimates of individual exposure. In addition, some studies used land use regression (LUR) models supplemented by a short period of air monitoring that predicted levels of vanadium and other PM components using variables such as land density, population density, altitude, traffic intensity, and road network; at this time, none of these models have been validated for prediction of vanadium and thus their ability to distinguish exposures is uncertain. A small number of studies measured exposure using personal or home air samples which may provide more specific estimates of individual exposure when samples are collected over enough time to capture variability.

Table 6-3 Criteria for evaluating exposure measurement in epidemiology studies of vanadium

Rating	Criteria
Good	Biomarker measures: • Evidence that exposure was consistently assessed using well-established analytical methods. Well-established and sensitive methods include measurement of total vanadium using GF-AAS (with a preconcentration procedure), ID-MS, ICP-MS, and NAA with radiochemical separation. Occupational measures: • For a specific job site(s): Evidence that measurement of current/recent exposure is based on personal samples (air or biomonitoring). Ideally, this would cover all workers or randomly selected workers within specific areas/jobs/tasks for at least one full shift, allowing for examination of variation in exposure among workers at a particular worksite, but this is not required (i.e., categorization by job duties with validation using personal samples in a sample of workers is acceptable). OR for long-term exposure, monitoring data covering a substantial portion of the time period of interest specific to work locations, job titles, and tasks with information provided on changes in exposure conditions over time; job histories are available for a substantial period of employment in exposed jobs. Air measures: • Personal/Home samples: Integrated personal measurements using passive monitors over multiple 24-hour periods, or time-weighted summary concentrations incorporating concentrations in residence and school/workplace. OR Area measurements in home using passive or active monitors, with an average of measurements in one or more rooms (average over longer periods is preferred, with multiple seasons if estimating annual

Rating	Criteria
	average). For either type of measure, sampling details are provided including type and placement of samplers, sampling periods, and chemical analysis methods. And all of the following (where relevant): • Measurement of vanadium included species or exposure was to a specific species. • Exposure was assessed in a relevant time-window (i.e., temporality is established, and sufficient latency occurred prior to disease onset) for development of the outcome. • There is evidence that a sufficient number of the exposure data measurements are above the limit of quantification for the assay. • Details on quality control provided include measures to avoid contamination in sampling, sample handling and storage of blood and urine samples, sample mass (minimum 10 mg with adjustment for mass (Gutiérrez-González et al., 2019)) for toenails. QA statistics on precision and accuracy reported. • There is sufficient specificity/sensitivity and range or variation in exposure measurements that would minimize potential for exposure measurement error and misclassification by allowing exposure classifications to be differentiated (i.e., can reliably categorize participants into groups such as high vs. low exposure).
Adequate	Biomarker measures • Evidence that exposure was consistently assessed using methods described in *Good,* but there were some concerns about quality control measures or other potential for non-differential misclassification. Occupational measures • For a specific job site(s): With known exposure to vanadium at the site, evidence that current/recent exposure based on job duties alone (without personal samples) are used with a comparison group where exposure levels are known to be low (i.e., similar to background levels in the general population) or monitoring data is less comprehensive, raising the possibility of nondifferential misclassification. OR for long-term exposure, monitoring data is less comprehensive with regard to time, work site, job title or tasks, or job history data are less complete than described in *Good.* • For population-based occupational studies: Job exposure matrix that incorporates industry, time period, tasks, and material used, and has validation data confirming its ability to differentiate between exposure levels. Air measures • Personal/Home samples: sampling occurs over a shorter period than described in *Good*, or some details on sampling and analysis are not provided but appear appropriate. And all of the following (where relevant): • Exposure was assessed in a relevant time-window for development of the outcome. • There is evidence that a sufficient number of the exposure data measurements are above the limit of quantification for the assay. • The laboratory analysis included some data on standard quality control measures with demonstrated precision and accuracy.

Rating	Criteria
	• There is sufficient specificity/sensitivity and range or variation in exposure measurements that would minimize potential for exposure measurement error and misclassification by allowing exposure classifications to be differentiated (i.e., can reliably categorize participants into groups such as high vs. low exposure) Measurement of total vanadium will reduce the rating from *Good* to *Adequate* unless the exposure is known to be a specific vanadium species.
Deficient	Any of the following: • There is a lack of detail on the sampling or analytical methods that reduces the ability to assess exposure misclassification. • There is some concern, but no direct evidence, that the exposure was assessed using methods that have not been validated or empirically shown to be consistent with methods that directly measure exposure. • Exposure was assessed in a relevant time-window(s) for development of the outcome, but there could be some concern about the potential for bias due to reverse causation between exposure and outcome, but there is no direct evidence that it is present. • There is some concern over insufficient specificity/sensitivity and range or variation in exposure measurements that may result in considerable exposure measurement error and misclassification when exposure classifications are compared (i.e., data do not lend themselves to reliably categorize participants into groups such as high vs. low exposure, and/or there is considerable uncertainty in exposure values which do not allow for confidence in the examination of small per unit changes in continuous exposures)
Critically deficient	Any of the following: • Exposure was assessed in a time-window that is unknown or not relevant for development of the outcome. This could be due to clear evidence of bias due to reverse causation between exposure and outcome, or other concerns such as the lack of temporal ordering of exposure and disease onset, insufficient latency, or having exposure measurements that are not reliable measures of exposure during the etiologic window(s). • Direct evidence that bias was likely since the exposure was assessed using methods with poor validity. • Evidence of differential exposure misclassification (e.g., differential recall of self-reported exposure). • There is evidence that an insufficient number of the exposure data measurements were above the limit of quantification for the assay.

6.2.1.2 Confounding by co-exposures

Exposure to vanadium via the inhalation route is typically co-occurring with other chemical exposures. In the general population, overall PM mass and other individual PM components are highly correlated with vanadium (U.S. EPA, 2019c), while in occupational studies, co-exposures depend on the specific job duties. These co-exposures represent potentially important confounders when estimating the effect of an individual component from a larger mixture. The likelihood of

confounding by co-exposures will be considered during study evaluation. In order for confounding to occur, the co-pollutant would need to be associated with both vanadium and the outcome of interest, and not act as an intermediate in the causal pathway. Thus, where correlations across exposures are reported, they will be reviewed to assess the likelihood that confounding could explain the observed results. In addition, many studies, particularly those published recently, may also have performed multipollutant modeling to adjust for co-exposures. These analyses can provide additional context, but even when they are available, it is often not possible to fully disentangle the associations due to high correlations. This stems from the potential for amplification bias that can occur following adjustment of highly correlated exposures (Weisskopf et al., 2018). Thus, in most studies, there may be some residual uncertainty about the risk of confounding by co-exposures. A *Good* rating for the confounding domain will be reserved for situations where there is minimal concern for substantial confounding across co-exposures as well as other sources of confounding. This could occur in studies where there are robust results following multipollutant modeling (i.e., minimal change between single- and multi-pollutant models), which would also indicate minimal concern for amplification bias. Potential confounding by co-exposures may result in a *Deficient* rating if there is considerable concern that the observed effect could be explained by correlated co-exposures.

Because of the challenge in evaluating individual studies for confounding by co-exposures, this issue will also be assessed across studies during the evidence synthesis phase, primarily when there is support for an association with adverse health effects in the epidemiology evidence (i.e., *moderate,* or *robust* evidence in humans, as described below). Analyses may include comparison of results across studies in populations with different exposure mixture profiles (e.g., general population vs. occupational) and considering results of multi-pollutant models across studies when available. In situations where there is considerable uncertainty regarding the impact of residual confounding by co-exposures, this will be captured as a factor that decreases the overall strength of evidence (see Section 10.1).

6.3. EXPERIMENTAL ANIMAL STUDY EVALUATION

Using the principles described in Section 6.1, the animal studies of health effects to assess risk of bias and sensitivity are evaluated for the following domains: allocation, observational bias/blinding, confounding, selective reporting, attrition, chemical administration and characterization, endpoint measurement and validity, results presentation and comparisons, and sensitivity (see Table 6-4).

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

Vanadium speciation chemistry in animal toxicological studies will be considered in the *Exposure methods sensitivity* domain. The highest confidence will be placed in studies that report the vanadium compound that was used and have analytical chemistry data indicating the vanadium species present. Considering oxidation status could be important as results from some oral exposure studies in rodents suggest increased toxicity of vanadium in the +5 oxidation state compared to vanadium +4 (Roberts et al., 2016); (National Toxicology Program (NTP)). Study evaluations for the available inhalation studies, to the extent possible, will consider factors that could affect vanadium oxidation state and speciation (e.g., study methods that involved aerosolizing vanadium pentoxide from solution, rather than exposure to vanadium pentoxide as a dust).

Table 6-4. Domains, questions, and general considerations to guide the evaluation of animal toxicology studies

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

Domain and core question	Prompting questions	General considerations
		• **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 impacted the results.
Confounding Are variables with the potential to confound or modify results controlled for and consistent across experimental groups? *Note:* *Consideration of overt toxicity (possibly masking more specific effects) is addressed under endpoint measurement reliability.*	For each study: • Are there difference across the treatment groups, considering both differences related to the exposure (e.g. co-exposures, vehicle, diet, palatability) and other aspects of the study design or animal groups (e.g., animal source, husbandry, or health status), that could bias the results? • If differences are identified, to what extent are they expected, based on a specific scientific understanding, to impact the results?	These considerations may need to be refined by assessment teams, as the specific variables of concern can vary by experiment or chemical. A judgment and rationale for this domain should be given for each cohort or experiment in the study, noting when the potential for confounding is restricted to specific endpoints/outcomes. • **Good**: Outside of the exposure of interest, variables that are likely to confound or modify results appear to be controlled for and consistent across experimental groups. • **Adequate**: Some concern that variables that were likely to confound or modify results were uncontrolled or inconsistent across groups but are expected to have a minimal impact on the results. • **Deficient**: Notable concern that potentially confounding variables were uncontrolled or inconsistent across groups and are expected based on to substantially impact the results. • **Critically deficient**: Confounding variables were presumed to be uncontrolled or inconsistent across groups and are expected to be a primary driver of the results.
Attrition Did the study report results for all tested animals?	For each study: • Are all animals accounted for in the results?	These considerations typically do not need to be refined by assessment teams. A judgment and rationale for this domain should be given for each cohort or experiment in the study.

Domain and core question	Prompting questions	General considerations
	• If there is attrition, do authors provide an explanation (e.g., death or unscheduled sacrifice during the study)? • If unexplained attrition of animals for outcome assessment is identified, what is the expected impact on the interpretation of the results?	• **Good**: Results were reported for all animals. If animal attrition is identified, the authors provide an explanation, and these are not expected to impact the interpretation of the results. • **Adequate**: Results are reported for most animals. Attrition is not explained but this is not expected to significantly impact the interpretation of the results. • **Deficient**: Moderate to high level of animal attrition that is not explained and may significantly impact the interpretation of the results. • **Critically deficient**: Extensive animal attrition that prevents comparisons of results across treatment groups.
Chemical administration and characterization Did the study adequately characterize exposure to the chemical of interest and the exposure administration methods? *Note:* *Consideration of the appropriateness of the route of exposure (not the administration method) is not a risk of bias consideration. Relevance and utility of the routes of exposure are considered in the PECO criteria for study inclusion and during evidence synthesis.* *Relatedly, consideration of exposure level selection (e.g., were levels sufficiently high to elicit effects) is addressed during evidence synthesis and*	For each study: • Are there concerns [specific to this chemical] regarding the source and purity and/or composition (e.g., identity and percent distribution of different isomers) of the chemical? • Was independent analytical verification of the test article (e.g., composition, homogeneity, and purity) performed? • Were nominal exposure levels verified analytically? Are there concerns about the methods used to administer the chemical (e.g., inhalation chamber type, gavage volume)?	It is essential that these considerations are considered, and potentially refined, by assessment teams, as the specific variables of concern can vary by chemical (e.g., stability may be an issue for one chemical but not another). A judgment and rationale for this domain should be given for each cohort or experiment in the study. • **Good**: Chemical administration and characterization is complete (i.e., source and purity are provided or can be obtained from the supplier and test article is analytically verified). There are no notable concerns about the composition, stability, or purity of the administered chemical, or the specific methods of administration. Exposure levels are verified using reliable analytical methods. • **Adequate**: Some uncertainties in the chemical administration and characterization are identified but these are expected to have minimal impact on interpretation of the results (e.g., purity of the test article is suboptimal but interpreted as unlikely to have a significant impact; analytical verification of exposure levels is not reported or verified with non-preferred methods). • **Deficient**: Uncertainties in the exposure characterization are identified and expected to substantially impact the results (e.g., source of the test article is not reported, and composition is not independently verified; impurities are substantial or concerning; administration methods are considered likely to introduce

Domain and core question	Prompting questions	General considerations
is not a risk of bias consideration.		confounders, such as use of static inhalation chambers or a gavage volume considered too large for the species or lifestage at exposure). • **Critically deficient**: Uncertainties in the exposure characterization are identified and there is reasonable certainty that the study results are largely attributable to factors other than exposure to the chemical of interest (e.g., identified impurities are expected to be a primary driver of the results).
Endpoint measurement Are the selected procedures, protocols, and animal models adequately described and appropriate for the endpoint(s)/outcome(s) of interest? *Notes:* *Considerations related to the sensitivity of the animal model and timing of endpoint measurement are evaluated under Sensitivity.* *Considerations related to adjustments/corrections to endpoint measurements (e.g., organ weight corrected for body weight) are addressed under results presentation.*	For each endpoint/outcome or grouping of endpoints/outcomes in a study: • Are the evaluation methods and animal model adequately described and appropriate? • Are there concerns regarding the methodology selected for endpoint evaluation? • Are there concerns about the specificity of the experimental design? • Are there serious concerns regarding the sample size or how endpoints were sampled? • Are appropriate control groups for the study/assay type included?	Considerations for this domain are highly variable depending on the endpoint(s)/outcome(s) of interest and typically must be refined by assessment teams. A judgment and rationale for this domain should be given for each endpoint/outcome or group of endpoints/outcomes investigated in the study. Some considerations include the following: • **Good**: Adequate description of methods and animal models. Use of generally accepted and reliable endpoint methods. Sample sizes are generally considered adequate for the assay or protocol of interest and there are no notable concerns about sampling in the context of the endpoint protocol (e.g., sampling procedures for histological analysis).Includes appropriate control groups and any use of nonconcurrent or historical control data (e.g., for evaluation of rare tumors) is justified (e.g., authors or evaluators considered the similarity between current experimental animals and laboratory conditions to historical controls). Ratings of **Adequate, Deficient, and Critically Deficient** are generally defined as follows: • **Adequate**: Issues are identified that may affect endpoint measurement but are considered unlikely to substantially impact the overall findings or the ability to reliably interpret those findings. • **Deficient**: Concerns are raised that are expected to notably affect endpoint measurement and reduce the reliability of the study findings. • **Critically deficient**: Severe concerns are raised about endpoint measurement and any findings are likely to be largely explained by these limitations. The following specific examples of relevant concerns are typically associated with a **Deficient** rating, but **Adequate** or **Critically Deficient** might be applied depending on the expected impact of limitations on the reliability and interpretation of the results:

Domain and core question	Prompting questions	General considerations
		• Study report lacks important details that are necessary to evaluate the appropriateness of the study design (e.g., description of the assays or protocols; information on the strain, sex, or lifestage of the animals) • Selection of protocols that are nonpreferred or lack specificity for investigating the endpoint of interest. This includes omission of additional experimental criteria (e.g., inclusion of a positive control or dosing up to levels causing minimal toxicity) when required by specific testing guidelines/protocols. * • Overt toxicity (e.g., mortality, extreme weight loss) is observed or expected based on findings from similarly designed studies and may mask interpretation of outcome(s) of interest. • Sample sizes are smaller than is generally considered adequate for the assay or protocol of interest. Inadequate sampling can also be raised within the context of the endpoint protocol (e.g., in a pathology study, bias that is introduced by only sampling a single tissue depth or an inadequate number of slides per animal). ** • Control groups are not included, considered inappropriate, or comparisons to non-concurrent or historical controls are not adequately justified. *These limitations typically also raise a concern for insensitivity ** Sample size alone is not a reason to conclude an individual study is critically deficient.
Results presentation Are the results presented and compared in a way that is appropriate and transparent?	For each endpoint/outcome or grouping of endpoints/outcomes in a study: • Does the level of detail allow for an informed interpretation of the results? • Are the data compared, or presented, in a way that is inappropriate or misleading?	Considerations for this domain are highly variable depending on the outcomes of interest and typically must be refined by assessment teams. A judgment and rationale for this domain should be given for each endpoint/outcome or group of endpoints/outcomes investigated in the study. Some considerations include the following: • **Good**: No concerns with how the data are presented. Results are quantified or otherwise presented in a manner that allows for an independent consideration of the data (assessments do not rely on author interpretations). No concerns with completeness of the results reporting.* Ratings of **Adequate**, **Deficient**, and **Critically Deficient** are generally defined as follows: • **Adequate**: Concerns are identified that may affect results presentation but are considered unlikely to substantially impact the overall findings or the ability to reliably interpret those findings.

Domain and core question	Prompting questions	General considerations
		• **Deficient**: Concerns with results presentation are identified and expected to substantially impact results interpretation and reduce the reliability of the study findings. • **Critically deficient**: Severe concerns about results presentation were identified and study findings are likely to be largely explained by these limitations. The following specific examples of relevant concerns are typically associated with a **Deficient** rating but **Adequate** or **Critically Deficient** might be applied depending on expected impact of limitations on the reliability and interpretation of the results: • Nonpreferred presentation of data (e.g., developmental toxicity data averaged across pups in a treatment group, when litter responses are more appropriate; presentation of only absolute organ weight data when relative weights are more appropriate). • Pooling data when responses are known or expected to differ substantially (e.g., across sexes or ages). • Incomplete presentation of the data* (e.g., presentation of mean without variance data; concurrent control data are not presented; dichotomizing or truncating continuous data). *Failure to describe *any* findings for assessed outcomes (i.e., report lacks any qualitative or quantitative description of the results in tables, figures, or text) is addressed under Selective Reporting.
Selective reporting Did the study report results for all prespecified outcomes? *Note:* *This domain does not consider the appropriateness of the analysis/results presentation. This aspect of study quality is evaluated in another domain.*	For each study: • Are results presented for all endpoints/outcomes described in the methods (see note)? • If unexplained results omissions are identified, what is the expected impact on the interpretation of the results?	These considerations typically do not need to be refined by assessment teams. A judgment and rationale for this domain should be given for each cohort or experiment in the study. • **Good**: Quantitative or qualitative results were reported for all prespecified outcomes (explicitly stated or inferred), exposure groups and evaluation time points. Data not reported in the primary article is available from supplemental material. If results omissions 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) and evaluation time points. Omissions and

Domain and core question	Prompting questions	General considerations
		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), omissions are not explained and may significantly impact the interpretation of the results. • **Critically deficient**: Extensive results omission is identified and prevents comparisons of results across treatment groups.
Sensitivity Are there concerns that sensitivity in the study is not adequate to detect an effect? *Note:* *Consideration of exposure level selection (e.g., were levels sufficiently high to elicit effects) is addressed during evidence synthesis and is not a study sensitivity consideration.*	• Was the exposure period, timing (e.g., lifestage), frequency, and duration sensitive for the outcome(s) of interest? • Based on knowledge of the health hazard of concern, did the selection of species, strain, and/or sex of the animal model reduce study sensitivity? • Are there concerns regarding the timing (e.g., lifestage) of the outcome evaluation? • Are there aspects related to risk of bias domains that raise concerns about insensitivity (e.g., selection of protocols that are known to be insensitive or nonspecific for the outcome(s) of interest)	These considerations may require customization to the specific exposure and outcomes. Some study design features that affect study sensitivity may have already been included in the other evaluation domains; these should be noted in this domain, along with any features that have not been addressed elsewhere. Some considerations include: • **Good:** The experimental design (considering exposure period, timing, frequency, and duration) is appropriate and sensitive for evaluating the outcome(s) of interest. The selected animal model (considering species, strain, sex, and/or lifestage) is known or assumed to be appropriate and sensitive for evaluating the outcome(s) of interest. No significant concerns with the ability of the experimental design to detect the specific outcome(s) of interest. (e.g., outcomes evaluated at the appropriate lifestage; study designed to address known endpoint variability that is unrelated to treatment, such as estrous cyclicity or time of day). Timing of endpoint measurement in relation to the chemical exposure is appropriate and sensitive (e.g., behavioral testing is not performed during a transient period of test chemical-induced depressant or irritant effects; endpoint testing does not occur only after a prolonged period, such as weeks or months, of nonexposure). Potential sources of bias towards the null are not a substantial concern. • **Adequate:** Same considerations as *Good*, except: The duration and frequency of the exposure was appropriate, and the exposure covered most of the critical window (if known) for the outcome(s) of interest. Potential issues are identified that could reduce sensitivity, but they are unlikely to impact the overall findings of the study.

Domain and core question	Prompting questions	General considerations
		• **Deficient**: Concerns were raised about the considerations described for *Good* or *Adequate* that are expected to notably decrease the sensitivity of the study to detect a response in the exposed group(s). • **Critically deficient**: Severe concerns were raised about the sensitivity of the study and experimental design such that any observed associations are likely to be explained by bias. The rationale should indicate the specific concern(s).
Overall confidence Considering the identified strengths and limitations, what is the overall confidence rating for the endpoint(s)/outcome(s) of interest?	For each endpoint/outcome or grouping of endpoints/outcomes in a study: • Were concerns (i.e., limitations or uncertainties) related to the risk of bias or sensitivity identified? • If yes, what is their expected impact on the overall interpretation of the reliability and validity of the study results, including (when possible) interpretations of impacts on the magnitude or direction of the reported effects?	The overall confidence rating considers the likely impact of the noted concerns (i.e., limitations or uncertainties) in reporting, bias, and sensitivity on the results. Reviewers should mark studies that are rated lower than high confidence only due to low sensitivity (i.e., bias towards the null) for additional consideration during evidence synthesis. If the study is otherwise well conducted and an effect is observed, it may increase the strength of evidence judgment. A confidence rating and rationale should be given for each endpoint/outcome or group of endpoints/outcomes investigated in the study. Confidence ratings are described above (see Section 6.1.1).

6.4. CONTROLLED HUMAN EXPOSURE STUDY EVALUATION

This study design involves human volunteers to test specific hypotheses about short-term exposures and biological responses that inform potential mechanisms and understanding of exposure-response patterns. The exposures are generated in the laboratory to achieve predetermined concentrations for periods of minutes to hours. For study evaluation, a process incorporating aspects of the approaches used for epidemiology studies and experimental animal studies, as well as the ROBINS-I tool discussed in Section 6.2 (Sterne et al., 2016), are used to evaluate controlled exposure studies in humans. Controlled human exposure studies are evaluated for important attributes of experimental studies, including randomization of exposure assignments, blinding of subjects and investigators, exposure generation, inclusion of a clean air control exposure (if applicable), study sensitivity, and other aspects of the exposure protocol. Evaluation will also include confirmation that the study protocol was approved by an institutional review board.

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

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

No PBPK models for vanadium and compounds were identified in the survey of the literature. If the comprehensive literature search or updates to that initial search identify any PBPK models, they will be evaluated in accordance with the Quality Assurance Project Plan for PBPK models (U.S. EPA, 2020b, 2018b).

6.6. IN VITRO AND OTHER MECHANISTIC STUDY EVALUATION

As described in Section 4.4, the initial literature screening identifies sets of other potentially informative studies, including mechanistic studies, as "potentially relevant supplemental information." Mechanistic information includes any experimental measurement related to a health outcome that informs the biological or chemical events associated with phenotypic effects. These measurements can improve understanding of the mechanisms involved in the biological effects following exposure to a chemical but are not generally considered by themselves adverse outcomes. Mechanistic data are reported in a diverse array of observational and experimental studies across

species, model systems, and exposure paradigms, including in vitro, in vivo (by various routes of exposure), ex vivo, and in silico studies.

Individual study-level evaluations of mechanistic endpoints are pursued only in some select cases. For some chemical assessments, it may be necessary to identify assay-specific considerations for study endpoint evaluations, on a case-by-case basis, to provide a more detailed summary and evaluation for the most relevant individual studies. This may be done, for example, when the scientific understanding of a critical mechanistic event or MOA is less established or lacks scientific consensus, when the reported findings on a mechanistic endpoint are conflicting, when the available mechanistic evidence addresses a complex and influential aspect of the assessment, or when in vitro or in silico data make up the bulk of the evidence base and there is little or no evidence from epidemiological studies or animal bioassays.

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

The evaluation of mechanistic studies applies similar principles as those described above for the evaluation of experimental animal studies. Table 6-5 provides the standard domains and core questions for evaluating studies conducted in in vitro test systems, along with some basic considerations for guiding the evaluation. The evaluation process focuses on assessing aspects of the study design and conduct through three broad types of evaluations: reporting quality, risk of bias, and study sensitivity. Some domain considerations are tailored to the chemical, as well as the assay(s) and/or endpoint(s) being evaluated. Assessment teams work with subject-matter experts to develop specific considerations. These specific considerations are determined before performing the study evaluation, although they may be refined as the study evaluation proceeds (e.g., during pilot testing). Assessment-specific and/or assay-specific considerations are documented and made publicly available in the assessment.

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

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

Domain and core question	Prompting questions	General considerations
controlled for and consistent across experimental groups?	cell passages, variations in reagent production lots, mycoplasma infections) that could bias the results? If differences are identified, to what extent are they expected to impact the results? • Did the study address feature inherent to the physicochemical properties of the test substance(s) that have the potential to bias the results away from the null? For example, could the test article interfere with a given assay (e.g., auto-fluoresces or inhibits enzymatic processes necessary for assay signals), potentially leading to an erroneous positive signal? *(Note that concerns related to dose are addressed in chemical administration and characterization.)* • Are there known variations in cellular signaling unique to the model system that could influence the possibility of detecting the effect(s) of interest? • Are there concerns regarding the negative (untreated and/or vehicle) controls used? Were negative controls run concurrently?	A judgment and rationale for this domain should be given for each experiment in the study, noting when the potential to affect results is restricted to specific assays or endpoints. • **Good**: Outside of the exposure of interest, variables or features of the test system and/or chemical properties that are likely to impact results appear to be controlled for and consistent across experimental groups. • **Adequate**: Some concern that variables or features of the test system and/or chemical properties that are likely to modify or interfere with results were uncontrolled or inconsistent across groups but are expected to have a minimal impact on the results. • **Deficient**: Notable concern that important study variables and/or features of the test system lacked specificity or were uncontrolled or inconsistent across groups and are expected to substantially impact the results. • **Critically deficient**: Features of the test system are known to be nonspecific for this endpoint, and/or influential study variables were presumed to be uncontrolled or inconsistent across groups and are expected to be a primary driver of the results.
Selective reporting Did the study present results, quantitatively or qualitatively, for all prespecified assays or	For each study: • Are results presented for all endpoints/outcomes described in the methods?	These considerations typically do not need to be refined by assessment teams. A judgment and rationale for this domain should be given for each assay or endpoint in the study.

Domain and core question	Prompting questions	General considerations
endpoints and replicates described in the methods? *Note: The appropriateness of the analysis or results presentation is considered under results presentation.*	• Did the study clearly indicate the number of replicate experiments performed? Were the replicates technical (from the same sample) or independent (from separate, distinct exposures)? • If unexplained results omissions are identified, what is the expected impact on the interpretation of the results?	• **Good**: Quantitative or qualitative results were reported for all prespecified assays or endpoints (explicitly stated or inferred), exposure groups and evaluation timepoints. Data not reported in the primary article is available from supplemental material. If results omissions are identified, the authors provide an explanation, and these are not expected to impact the interpretation of the results. • **Adequate**: Quantitative or qualitative results are reported for most prespecified assays or endpoints (explicitly stated or inferred), exposure groups and evaluation timepoints. Omissions are not explained but are not expected to significantly impact the interpretation of the results. • **Deficient**: Quantitative or qualitative results are missing for many prespecified assays or endpoints (explicitly stated or inferred), exposure groups and evaluation timepoints; omissions are not explained and may significantly impact the interpretation of the results. • **Critically deficient**: Extensive results omissions are identified, preventing comparisons of results across treatment groups.
Chemical administration and characterization Did the study adequately characterize exposure to the chemical of interest and the exposure administration methods?	For each study: • Are there concerns regarding the purity and/or composition (e.g., identity and percent distribution of different isomers) of the test material/chemical? If so, can the purity and/or composition be obtained from the supplier (e.g., as reported on the website)? • Was independent analytical verification of the test article purity and composition performed? If not, is	It is essential that these criteria are considered, and potentially refined, by assessment teams, as the specific variables of concern can vary by chemical (e.g., stability may be an issue for one chemical but not another). A judgment and rationale for this domain should be given for each experiment in the study. • **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.

Domain and core question	Prompting questions	General considerations
	this a significant concern for this substance? • Are there concerns about the stability of the test chemical in the vehicle and/or culture media (e.g., pH, solubility, volatility, adhesion to plastics) that were not corrected for, leading to potential bias away from the null (e.g., observed precipitate formation at high concentrations) or toward the null (e.g., enclosed chambers not used for testing volatile chemicals)? • Are there concerns about the preparation or storage conditions of the test substance? • Are there concerns about the methods used to administer the chemical?	• **Adequate**: Some uncertainties in the chemical administration and characterization are identified but these are expected to have minimal impact on interpretation of the results (e.g., source and vendor-reported purity are presented but not independently verified; purity of the test article is suboptimal but not concerning). • **Deficient**: Uncertainties in the exposure characterization are identified and expected to substantially impact the results (e.g., the source and purity of the test article are not reported, and no independent verification of the test article was conducted; levels of impurities are substantial or concerning; deficient administration methods were used). • **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).
Endpoint measurement Are the selected protocols, procedures, and test systems adequately described and appropriate for evaluating the endpoint(s) of interest? *Notes:* *Considerations related to adjustments or corrections to endpoint measurements are addressed under results presentation.*	For each endpoint or grouping of endpoints in a study: • Are the evaluation methods and test systems adequately described and appropriate? • Are there concerns regarding the methodology selected (e.g., accepted guidelines, established criteria) for endpoint evaluation? • Are there concerns about the specificity of the experimental design? Did the study address feature inherent	Considerations for this domain are highly variable depending on the assay or endpoint(s) of interest and must be refined by assessment teams. A judgment and rationale for this domain should be given for each assay or endpoint or group of endpoints investigated in the study. Some considerations include the following: • **Good**: Adequate description of methods and test system. Use of generally accepted and reliable endpoint methods that are consistent with accepted guidelines or established criteria for the assay(s)/endpoint(s) of interest. Sample sizes are generally considered adequate for the assay or protocol of interest and there are no notable concerns about sampling in the context of the

Domain and core question	Prompting questions	General considerations
Considerations related to the sensitivity of the animal model and timing of endpoint measurement are evaluated under sensitivity.	to the test system or experiment that have the potential to lead to bias away from the null? • Are there serious concerns about the number of replicates or sample size in the study? • Are appropriate control groups for the study/assay type included? Was there a need for the assay to include specific controls to reduce potential sources of underlying bias? • Did the test compound induce cytotoxicity (known, or expected based on other studies of similar design) to a degree that is expected to affect interpretation of results?	endpoint protocol. Includes appropriate control groups (e.g., use of loading controls) and any use of nonconcurrent or historical control data (e.g., for comparison to background levels in negative controls) is justified (e.g., authors or evaluators considered the similarity between current cell cultures and laboratory conditions to historical controls). Ratings of **Adequate**, **Deficient**, and **Critically Deficient** are generally defined as follows: • **Adequate**: Issues are identified that may affect endpoint measurement but are considered unlikely to substantially impact the overall findings or the ability to reliably interpret those findings. • **Deficient**: Concerns are raised that are expected to notably affect endpoint measurement and reduce the reliability of the study findings • **Critically deficient**: Severe concerns are raised about endpoint measurement and any findings are likely to be largely explained by these limitations. The following specific examples of relevant concerns are typically associated with a Deficient rating, but Adequate or Critically Deficient might be applied depending on the expected impact of limitations on the reliability and interpretation of the results: • Study report lacks important details that are necessary to evaluate the appropriateness of the study design (e.g., description of the assays or protocols; information on the cell line, passage number). • Selection of protocols that are nonpreferred or lack specificity for investigating the endpoint of interest. This includes omission of additional experimental criteria (e.g., inclusion of a positive control or dosing up to levels causing minimal toxicity) when required by specific testing guidelines/protocols.*

Domain and core question	Prompting questions	General considerations
		• Cytotoxicity is observed or expected based on findings from similarly designed studies and may mask interpretation of outcome(s) of interest. Sample sizes are smaller than is generally considered adequate for the assay or protocol of interest. Inadequate sampling can also be raised within the context of the endpoint protocol (e.g., in a pathology study, bias that is introduced by only sampling a single tissue depth or an inadequate number of slides per animal)** Controls are not included or considered inappropriate. *These limitations typically also raise a concern for insensitivity **Sample size alone is not a reason to conclude an individual study is critically deficient.
Results presentation Are the results presented and compared in a way that is appropriate and transparent and makes the data usable?	For each assay/endpoint or grouping of endpoints in a study: • Does the level of detail allow for an informed interpretation of the results? • If applicable, was the assay signal normalized to account for nonbiological differences across replicates and exposure groups? • Are the data compared or presented in a way that is inappropriate or misleading (e.g., presenting western blot images without including numerical values for densitometry analysis, or vice versa)? Flag potentially inappropriate statistical comparisons for further review.	Considerations for this domain are highly variable depending on the endpoints of interest and must be refined by assessment teams. A judgment and rationale for this domain should be given for each assay or endpoint or group of endpoints investigated in the study. Some considerations include the following: **Good**: • No concerns with how the data are presented. • Results are quantified or otherwise presented in a manner that allows for an independent consideration of the data (assessments do not rely on author interpretations). • No concerns with completeness of the results reporting.* Ratings of **Adequate**, **Deficient**, and **Critically Deficient** are generally defined as follows: • **Adequate:** Concerns are identified that may affect results presentation but are considered unlikely to substantially impact the overall findings or the ability to reliably interpret those findings.

Domain and core question	Prompting questions	General considerations
		• **Deficient:** Concerns with results presentation are identified and expected to substantially impact results interpretation and reduce the reliability of the study findings. • **Critically deficient:** Severe concerns about results presentation were identified and study findings are likely to be largely explained by these limitations. The following specific examples of relevant concerns are typically associated with a **Deficient** rating but **Adequate** or **Critically Deficient** might be applied depending on expected impact of limitations on the reliability and interpretation of the results: • Nonpreferred presentation of data (e.g., averaging technical replicates rather than independent replicates). • Failure to present quantitative results. • Pooling data when responses are known or expected to differ substantially (e.g., across cell types or passage number). • Incomplete presentation of the data* (e.g., presentation of mean without variance data; concurrent control data are not presented; failure to report or address overt cytotoxicity). *Failure to describe *any* findings for assessed outcomes (i.e., report lacks any qualitative or quantitative description of the results in tables, figures, or text) will result in a critically deficient rating for the outcome(s) of interest for Results Presentation; overall completeness of reporting at the study level is addressed under Selective Reporting.
Sensitivity Are there concerns that sensitivity in the study is not adequate to detect an effect?	• Was the exposure period, timing (i.e., cell passage number, insufficient culture maturity for the adequate expression of mature cell markers; insufficient treatment and/or measurement duration for the production of protein above the level	Are there concerns regarding the need for positive controls (e.g., concerns that the effects of interest may be inhibited or otherwise poorly manifest in the test system, for example due to differences from in vivo biology)? If used, was the selected positive test substance (and dose) reasonable and appropriate and was the intended positive response induced?

Domain and core question	Prompting questions	General considerations
	of detection), frequency, and duration of exposure sensitive for the assay/model system of interest, particularly in the absence of a positive control? • Assay-specific considerations regarding sensitivity, specificity, and validity of the selection of the test methods will be described here (e.g., metabolic competency, antibody specificity) (some of these external considerations may have been applied during prioritization of studies for evaluation). • Are there aspects related to risk of bias domains that raise concerns about insensitivity (e.g., selection of protocols or methods that are known to be insensitive or nonspecific for the outcome(s) of interest)? • Are there concerns regarding the need for positive controls (e.g., concerns that the effects of interest may be inhibited or otherwise poorly manifest in the test system, for example due to differences from in vivo biology)? If used, was the selected positive test substance (and dose) reasonable and appropriate and was the intended positive response induced?	Considerations for this domain are highly variable depending on the specific assay/model system used or endpoint(s) of interest and must be refined by assessment teams. Some study design features that affect study sensitivity may have already been included in the other evaluation domains; these should be noted in this domain, along with any features that have not been addressed elsewhere. Some considerations include: **Good** • The experimental design (considering exposure period, timing, frequency, and duration) is appropriate and sensitive for evaluating the outcome(s) of interest. • The selected test system is appropriate and sensitive for evaluating the outcome(s) of interest (e.g., cell line/cell type is appropriate and routinely used for the selected assay). • No significant concerns with the ability of the experimental design to detect the specific outcome(s) of interest. (e.g., study designed to address known endpoint variability that is unrelated to treatment, such as doubling time or confluency). • Timing of endpoint measurement in relation to the chemical exposure is appropriate and sensitive (e.g., cultures adequately express mature cell markers). • Potential sources of bias toward the null are not a substantial concern. **Adequate** • Potential issues are identified related to the considerations described for *Good* that could reduce sensitivity, but they are unlikely to impact the overall findings of the study. **Deficient**

Domain and core question	Prompting questions	General considerations
		• Concerns were raised about the considerations described for *Good* that are expected to notably decrease the sensitivity of the study to detect a response in the exposed group(s). **Critically deficient** • Severe concerns were raised about the sensitivity of the study and experimental design such that any observed associations are likely to be explained by bias. The rationale should indicate the specific concern(s).
Overall confidence Considering the identified strengths and limitations, what is the overall confidence rating for the assay(s) or endpoint(s) of interest? *Note:* *Reviewers should mark studies for additional consideration during evidence synthesis if, due to low sensitivity only (i.e., bias toward the null), these studies are rated as lower than high confidence. If the study is otherwise well conducted and an effect is observed, the confidence may be increased.*	For each assay or endpoint or grouping of endpoints in a study: • Were concerns (i.e., limitations or uncertainties) related to the risk of bias or sensitivity identified? • If yes, what is their expected impact on the overall interpretation of the reliability and validity of the study results, including (when possible) interpretations of impacts on the magnitude or direction of the reported effects?	The overall confidence rating considers the likely impact of the noted concerns (i.e., limitations or uncertainties) in reporting, bias, and sensitivity on the results. A confidence rating and rationale should be given for each assay or endpoint, or group of endpoints investigated in the study. Confidence rating definitions are described above (see Section 4.1).

1

7. DATA EXTRACTION OF STUDY METHODS AND RESULTS

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

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

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

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

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

For animal studies that met the assessment PECO criteria, HAWC is used for full extraction of study methods and results. For animal studies, compared with the literature inventory forms used to describe studies that meet problem formulation PECO criteria, full data extraction in HAWC includes summarizing more details of study design (e.g., diet, chemical purity) and gathering effect size information. Instructions on how to conduct data extraction in HAWC are available at (https://hawcproject.org/resources/). Over 100 distinct extraction fields are collected for each animal study and endpoint (for list of data extraction fields, see Downloads > Animal Bioassay Data > Complete Export at the HAWC Vanadium and Compounds (Inhalation) Project (https://hawc.epa.gov/assessment/100500286/). An additional resource used to implement use of a consistent vocabulary to summarize endpoints assessed in animal studies is available in the HAWC project "IRIS PPRTV SEM Template Figures and Resources" (see "Attachments," then select the "Environmental Health Vocabulary (EHV)— a recommended terminology for outcomes/endpoints" file).

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

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

For non-English language studies online translation tools (e.g., Google translator) or engagement with a native speaker can be used to summarize studies at the level of the literature inventory. Fee-based translation services for non-English studies are typically reserved for studies considered potentially informative for dose response, a consideration that occurs after preparation of the initial literature inventory during draft assessment development. Digital rulers, such as WebPlotDigitizer (http://arohatgi.info/WebPlotDigitizer/), are used to extract numerical information from figures, and their use is be documented during extraction. For studies that evaluate endpoints at multiple time points (e.g., 7 days, 3 weeks, 3 months) data are generally summarized for the longest duration in the study report, but other durations may be summarized if they provide important contextual information for hazard characterization (e.g., an effect was

present at an interim time point but did not appear to persist or the magnitude of the effect diminished). A free text field is available in HAWC to describe cases when the approach for summarizing results requires explanation.

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

Exposures are standardized to common units when possible. For hazard characterization, exposure levels are typically presented as reported in the study and standardized to common units (e.g., ppm or mg/m^3 for inhalation studies) as an initial phase in evidence synthesis and integration. For inhalation exposures to vanadium, concentration in air will be reported as mg vanadium/m^3.

8. EVIDENCE SYNTHESIS AND INTEGRATION

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

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

- Evidence integration: The animal and human evidence judgments are combined to draw an overall evidence integration judgment(s) that incorporates inferences drawn based on information on the human relevance of the animal evidence, coherence across evidence

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

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

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

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

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

Evidence synthesis (certainty of evidence) judgments **(Note that many factors and judgments require elaboration or evidence-based justification; see IRIS Handbook for details)**					**Evidence integration (Weight of evidence) judgment(s)**
Studies	**Summary of key findings**	**Factors that increase certainty (Applied to each unit of analysis)**	**Factors that decrease certainty (Applied to each unit of analysis)**	**Evidence synthesis judgment(s)**	Describe overall evidence integration judgment(s): ⊕⊕⊕ Evidence demonstrates ⊕⊕⊙ Evidence indicates (likely) ⊕⊙⊙ Evidence suggests ⊙⊙⊙ Evidence inadequate – – – Strong evidence supports no effect Highlight the primary supporting evidence for each integration judgment* Present inferences and conclusions on: • Human relevance of findings in animals* • Cross-stream coherence* • Potential susceptibility* • Biological plausibility* • Other critical inferences (e.g., from ADME or other supplemental information)*
Evidence from human studies					
Unit of analysis #1 Studies considered and study confidence	Description of the primary results	• All/Mostly *medium* or *high* confidence studies • Consistency • Dose-response gradient • Large or concerning magnitude of effect • Coherence*	• All/Mostly *low* confidence studies • Unexplained inconsistency • Imprecision • Concerns about biological significance* • Indirect outcome measures* • Lack of expected coherence*	Judgment reached for each unit of analysis* ⊕⊕⊕ *Robust* ⊕⊕⊙ *Moderate* ⊕⊙⊙ Slight ⊙⊙⊙ *Indeterminate* – – – *Compelling evidence of no effect*	
Evidence from animal studies					
Unit of analysis #1 Studies considered and study confidence	Description of the primary results	• All/Mostly *medium* or *high* confidence studies • Consistency • Dose-response gradient • Large or concerning magnitude of effect • Coherence*	• All/Mostly l*ow* confidence studies • Unexplained inconsistency • Imprecision • Concerns about biological significance* • Indirect outcome measures* • Lack of expected coherence*	Judgment reached for each unit of analysis ⊕⊕⊕ *Robust* ⊕⊕⊙ *Moderate* ⊕⊙⊙ *Slight* ⊙⊙⊙ *Indeterminate* – – – *Compelling evidence of no effect*	

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

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

Mechanistic analyses		
Biological events or pathways (or other relevant evidence grouping)	**Summary of key findings and interpretation**	**Judgment(s) and rationale**
Different analyses may be presented separately, e.g., by exposure route or key uncertainty addressed. Each analysis may include multiple rows separated by biological events or other feature of the approach used for the analysis. • Generally, will cite mechanistic synthesis (e.g., for references, for detailed analysis) • Does not have to be chemical-specific (e.g., read-across)	May include separate summaries, for example by study type (e.g., new approach methods vs. in vivo biomarkers), dose, or design. *Interpretation*: Summary of expert interpretation for the body of evidence and supporting rationale. *Key findings*: Summary of findings across the body of evidence (may focus on or emphasize highly informative designs or findings), including key sources of uncertainty or identified limitations of the study designs tested (e.g., regarding the biological event or pathway being examined)	Overall summary of expert interpretation across the assessed set of biological events, potential mechanisms of toxicity, or other analysis approach (e.g., AOP). • Includes the primary evidence supporting the interpretation(s). • Describes and substantiates the extent to which the evidence influences inferences across evidence streams. • Characterizes the limitations of the evaluation and highlights existing data gaps. • May have overlap with factors summarized for other streams.

1

8.1. EVIDENCE SYNTHESIS

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

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

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

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

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

Consideration	Increased evidence certainty (of the human or animal evidence for hazard[a])	Decreased evidence certainty (of the human or animal evidence for hazard[a])
Risk of bias and sensitivity (across studies)	• An evidence base of mostly (or all) *high* or *medium* confidence studies is interpreted as being only minimally affected by bias and insensitivity. • This factor should not be used if no other factors would increase or decrease the confidence for a given unit of analysis. • In addition, consideration of risk of bias and sensitivity should inform how other factors are evaluated, i.e., can inconsistency be potentially explained by variation in confidence judgments?	• An evidence base of mostly (or all) *low* confidence studies decreases certainty. An exception to this is an evidence base of studies in which the issues resulting in *low* confidence are related to insensitivity. This may increase evidence certainty in cases where an association is identified because the expected impact of study insensitivity is toward the null. • An evidence base of mostly null findings where insensitivity is a serious concern decreases certainty that the evidence is sufficient to support a lack of health effect or association. • Decisions to increase certainty for other considerations in this table should generally not be made if there are serious concerns for risk of bias.
Consistency	• Similarity of findings for a given outcome (e.g., of a similar direction) across independent studies or experiments, especially when *medium* or *high* confidence, increases certainty. The increase in certainty is larger when consistency is observed across populations (e.g., geographical location) or exposure scenarios in human studies, and across laboratories, species, or exposure scenarios (e.g., route; timing) in animal studies. When seemingly inconsistent findings are identified, patterns should be further analyzed to discern if the inconsistencies can potentially be explained based on study confidence, dose or exposure levels, population, or experimental model differences, etc. This factor is typically	• Unexplained inconsistency [i.e., conflicting evidence; see (U.S. EPA, 2005a)] decreases certainty. Generally, certainty should not be decreased if discrepant findings can be reasonably explained by considerations such as study confidence conclusions (including sensitivity); variation in population or species, sex, or lifestage (including understanding of differences in pharmacokinetics); or exposure patterns (e.g., intermittent versus continuous), levels (*low* versus *high*), or duration. Similar to current recommendations in the Cochrane Handbook [(Higgins et al., 2022),see Section 7.8.6], clear conflicts of interest (COI) related to funding source can be considered as a factor to explain apparent inconsistency. For small evidence bases, it may be hard to assess consistency. An evidence base of a single or a few studies where consistency cannot be accurately assessed does not, on its own, increase or decrease evidence certainty. Similarly, a reasonable explanation for inconsistency does not necessarily result in an increase in evidence certainty.

Consideration	Increased evidence certainty (of the human or animal evidence for hazard[a])	Decreased evidence certainty (of the human or animal evidence for hazard[a])
	given the most attention during evidence synthesis.	
Effect magnitude and imprecision	• Evidence of a large or concerning magnitude of effect can increase certainty (generally only when observed in *medium* or *high* confidence studies). • Judgments on effect magnitude and imprecision consider the rarity and severity of the effect.	• Certainty may be decreased if the findings are considered not likely to be biologically significant. Effects that are small in magnitude might not be considered to be biologically significant (adverse[b]) based on information such as historical responses and variability. However, effects that appear to be of small magnitude may be meaningful at the population level (e.g., IQ shifts); in such cases, certainty would not be decreased. • Certainty may also be decreased for imprecision, particularly if there are only a few studies available to evaluate consistency in effect magnitude across studies.
Dose-response	• Evidence of dose-response or exposure-response in *high* or *medium* confidence studies increases certainty. Dose-response may be demonstrated across studies or within studies, and it can be dose- or duration-dependent. It may also not be a monotonic dose-response (monotonicity should not necessarily be expected as different outcomes may be expected at low vs. high doses or long vs. short durations due to factors such as activation of different mechanistic pathways, systemic toxicity at high doses, or tolerance/acclimation). Sometimes, grouping studies by level of exposure is helpful to identify the dose-response pattern. • Decreases in a response (e.g., symptoms of current asthma) after a documented cessation of exposure also may increase certainty in a relationship between exposure and outcome	• A lack of dose-response when expected based on biological understanding can decrease certainty in the evidence. If the data are not adequate to evaluate a dose-response pattern, however, then certainty is neither increased nor decreased. • In some cases, duration-dependent patterns in the dose-response can decrease evidence certainty. Such patterns are generally only observable in experimental studies. Specifically, the magnitude of effects at a given exposure level might decrease with longer exposures (e.g., due to tolerance or acclimation) or, effects might rapidly resolve under certain experimental conditions (e.g., reversibility after removal of exposure). As many reversible and short-lived effects can be of high concern, decisions about whether such patterns decrease evidence certainty depend on considering the pharmacokinetics of the chemical and the conditions of exposure [see(U.S. EPA, 1998)], endpoint severity, judgments regarding the potential for delayed or secondary effects, the underlying mechanism(s) involved, as well as the exposure context focus of the assessment (e.g., addressing intermittent or short-term exposures).

Consideration	Increased evidence certainty (of the human or animal evidence for hazard[a])	Decreased evidence certainty (of the human or animal evidence for hazard[a])
	(this is primarily applicable to epidemiology studies because of their observational nature).	
Directness of outcome/endpoint measures	• Not applicable	• If the evidence base primarily includes outcomes or endpoints that are indirect measures (e.g., biomarkers) of the unit of analysis, certainty (for that unit of analysis) is typically decreased. Judgments to decrease certainty based on indirectness should focus on findings that have an unclear linkage to an apical or clinical (adverse[b]) outcome. Scenarios where the magnitude of the response is not considered to reflect a biologically meaningful level of change (i.e., biological significance; see 'effect magnitude and imprecision' row above) are not considered under indirectness. • Related to indirectness, certainty in the evidence may be decreased when the findings are determined to be nonspecific to the hazard under evaluation. This consideration is generally only applicable to animal evidence and the most common example is effects only with exposures (level, duration) shown to cause excessive toxicity in that species and lifestage (including consideration of maternal toxicity in developmental evaluations). This does not apply when an effect is viewed as secondary to other changes (e.g., effects on pulmonary function because of disrupted immune responses).
Coherence	• Biologically related findings within or across studies, within an organ system or across populations (e.g., sex), increase certainty (generally only when observed in *medium* or *high* confidence studies). Certainty is further increased when a temporal or dose-dependent progression of related effects is observed within or across studies, or when related findings of increasing severity are observed with increasing exposure.	• An observed lack of expected coherent changes (e.g., in well-established biological relationships) within or across biologically related units of analysis typically decrease evidence certainty. This includes mechanistic changes when included in the unit of analysis. However, as described for decisions to increase certainty in the biological relationships between the endpoints being compared, and the sensitivity and specificity of the measures used, need to be carefully examined. The decision to decrease depends on the availability of evidence across multiple related endpoints for which changes would be anticipated, and it considers factors (e.g.,

Consideration	Increased evidence certainty (of the human or animal evidence for hazard[a])	Decreased evidence certainty (of the human or animal evidence for hazard[a])
	• Coherence across findings within a unit of analysis (e.g., consistent changes in disease markers and biological precursors in exposed humans) can increase certainty in the evidence for an effect. • Coherence within or across biologically related units of analysis can also increase certainty for a given (or multiple) unit(s) of analysis. This considers certainty in the biological relationships between the endpoints being compared, and the sensitivity and specificity of the measures used. • Mechanistic support for, or biological understanding of, the relatedness between different endpoints within (or across different) units of analysis, can inform an understanding of coherence.	dose and duration of exposure, strength of expected relationship) across the studies of related changes.
Other factors	• Unusual scenarios that cannot be addressed by the considerations above, e.g., read-across inferences supporting the adversity of observed changes.	• Unusual scenarios that cannot be addressed by the considerations above, e.g., strong evidence of publication bias.[c]

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

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

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

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

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

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

Evidence synthesis judgment	Description
	the certainty of evidence to robust for a set of studies that otherwise would be described as moderate. Such evidence not included in the unit of analysis can also inform evaluations of the coherence of the human evidence, the directness of the outcome measures, and the biological significance of the findings. Causality is inferred for a human evidence base of robust.
Moderate (⊕⊕⊙) …evidence in human studies *(Signal of effect with some uncertainty)*	A set of evidence that does not reach the degree of certainty required for *Robust*, but which includes at least one *high* or *medium* confidence study reporting an association and additional information increasing the certainty of evidence. For multiple studies, there is primarily consistent evidence of an association with reasonable support for adversity, but there may be some uncertainty due to potential chance, bias, or confounding or because of the indirectness of some measures. For a single study, there is a large magnitude or severity of the effect, or a dose-response gradient, or other supporting evidence, and there are no serious residual methodological uncertainties. Supporting evidence could include associations with related endpoints, including mechanistic evidence from exposed humans when included within the unit of analysis. When available and included in the unit of analysis, mechanistic data in humans that address the above considerations may raise the certainty of evidence to *Moderate* for a set of studies that otherwise would be described as *Slight*. In exceptional cases, biological support from mechanistic evidence in exposed humans may support raising the certainty of evidence to *Moderate* for evidence that would otherwise be described as *Indeterminate*.
Slight (⊕⊙⊙) …evidence in human studies *(Signal of effect with large amount of uncertainty)*	One or more studies reporting an association between exposure and the health outcome, but considerable uncertainty exists and supporting coherent evidence is sparse. In general, the evidence is limited to a set of consistent *low* confidence studies, or higher confidence studies with significant unexplained heterogeneity or other serious residual uncertainties. It also applies when one *medium* or *high* confidence study is available without additional information strengthening the likelihood of a causal association (e.g., coherent findings within the same study or from other studies). This category serves primarily to encourage additional study where evidence does exist that might provide some support for an association, but for which the evidence does not reach the degree of confidence required for moderate.
Indeterminate (⊙⊙⊙) …evidence in human studies *(Signal cannot be determined for or against an effect)*	No studies available in humans or situations when the evidence is inconsistent and primarily of *low* confidence. In addition, this may include situations where higher confidence studies exist, but there are major concerns with the evidence base such as unexplained inconsistency, a lack of expected coherence from a stronger set of studies, very small effect magnitude (i.e., major concerns about biological significance), or uncertainties or methodological limitations that result in an inability to discern effects from exposure. It also applies for a single *low* confidence study in the absence of factors that increase certainty. A set of largely null studies could be concluded to be *Indeterminate* if the evidence does not reach the level required for *Compelling evidence of no effect*.
Compelling evidence of no effect (- - -)	A set of *high* confidence studies examining a reasonable spectrum of endpoints showing null results (for example, an odds ratio of 1.0), ruling out alternative explanations including chance, bias, and confounding) with reasonable confidence. Each of the studies should have used an optimal outcome and exposure assessment and adequate sample size (specifically

Evidence synthesis judgment	Description
…in human studies *(Strong signal for lack of an effect with little uncertainty)*	for higher exposure groups and for susceptible populations). The set as a whole should include diverse sampling (across sexes [if applicable] and different populations) and include the full range of levels of exposures that human beings are known to encounter, an evaluation of an exposure-response gradient, and an examination of at-risk populations and lifestages. Mechanistic data in humans that address the above considerations or that provide information supporting the lack of an association between exposure and effect with reasonable confidence may provide additional support for this judgment.

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

Evidence synthesis judgment	Description
Robust (⊕⊕⊕) …evidence in animal studies *(strong signal of effect with very little uncertainty)*	The set of *high* or *medium* confidence, independent experiments (i.e., across laboratories, exposure routes, experimental designs [for example, a subchronic study and a multigenerational study], or species) reporting effects of exposure on the health outcome(s). The set of studies is primarily consistent, with reasonable explanations when results differ (i.e., due to differences in study design, exposure level, animal model, or study confidence), and the findings are considered adverse (i.e., biologically significant and without notable concern for indirectness). At least two of the following additional factors in the set of experiments increase the certainty of evidence: coherent effects across multiple related endpoints (within or across biologically related units of analysis and may include mechanistic endpoints); an unusual magnitude of effect, rarity, age at onset, or severity; a strong dose-response relationship; or consistent observations across animal lifestages, sexes, or strains. Mechanistic evidence from animals included in the unit of analysis or used to assess coherence of findings in the animal evidence may raise the certainty of evidence to *robust* for a set of studies that otherwise would be described as *moderate.*
Moderate (⊕⊕⊙) …evidence in animal studies *(signal of effect with some uncertainty)*	A set of evidence that does not reach the degree of certainty required for *Robust*, but which includes at least one *high* or *medium* confidence study and additional information increasing the certainty of evidence. For multiple studies or a single study, the evidence is primarily consistent or coherent with reasonable support for adversity, but there are notable remaining uncertainties (e.g., difficulty interpreting the findings due to concerns for indirectness of some measures); however, these uncertainties are not sufficient to reduce or discount the level of concern regarding the positive findings and any conflicting findings are from a set of experiments of lower confidence. The set of experiments supporting the effect provide additional information increasing the certainty of evidence, such as consistent effects across laboratories or species; coherent effects across multiple related endpoints (may include mechanistic endpoints within the unit of analysis); an unusual magnitude of effect, rarity, age at onset, or severity; a strong

Evidence synthesis judgment	Description
	dose-response relationship; and/or consistent observations across exposure scenarios (e.g., route, timing, duration), sexes, or animal strains. When available and included in the unit of analysis, mechanistic data in animals that address the above considerations may raise the certainty of evidence to *Moderate* for a set of studies that otherwise would be described as *Slight*. In exceptional cases, strong biological support from mechanistic studies may raise the certainty of evidence to *Moderate* for evidence that would otherwise be described as *Indeterminate*.
Slight (⊕⊙⊙) …evidence in animal studies *(signal of effect with large amount of uncertainty)*	One or more studies reporting an effect on an exposure on the health outcome, but considerable uncertainty exists and supporting coherent evidence is sparse. In general, the evidence is limited to a set of consistent *low* confidence studies, or higher confidence studies with significant unexplained heterogeneity or other serious uncertainties (e.g., concerns about adversity) across studies. It also applies when one *medium* or *high* confidence experiment is available without additional information increasing the certainty of evidence (e.g., coherent findings within the same study or from other studies). Biological evidence from mechanistic studies may also be independently interpreted as *Slight*. This category serves primarily to encourage additional study where evidence does exist that might provide some support for an association, but for which the evidence does not reach the degree of confidence required for *Moderate*.
Indeterminate (⊙⊙⊙) …evidence in animal studies *(signal cannot be determined for or against an effect)*	No studies available in animals or situations when the evidence is inconsistent and primarily of *low* confidence. In addition, this may include situations where higher confidence studies exist, but there are major concerns with the evidence base such as unexplained inconsistency, a lack of expected coherence from a stronger set of studies, very small effect magnitude (i.e., major concerns about biological significance), or uncertainties or methodological limitations that result in an inability to discern effects from exposure. It also applies for a single *low* confidence study in the absence of factors that increase certainty. A set of largely null studies could be concluded to be *Indeterminate* if the evidence does not reach the level required for *Compelling evidence of no effect*.
Compelling evidence of no effect (- - -) …in animal studies *(strong signal for lack of an effect with little uncertainty)*	A set of *high* confidence experiments examining a reasonable spectrum of endpoints that demonstrate a lack of biologically significant effects across multiple species, both sexes, and a broad range of exposure levels. The data are compelling in that the experiments have examined the range of scenarios across which health effects in animals could be observed, and an alternative explanation (e.g., inadequately controlled features of the studies' experimental designs; inadequate sample sizes) for the observed lack of effects is not available. Each of the studies should have used an optimal endpoint and exposure assessment and adequate sample size. The evidence base should represent both sexes and address potentially susceptible populations and lifestages. Mechanistic data in animals that address the above considerations or that provide information supporting the lack of an association between exposure and effect with reasonable confidence may provide additional support for this judgment.

8.2. EVIDENCE INTEGRATION

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

Table 8-6. Considerations that inform evidence integration judgments

Judgment	Description
Human relevance of findings	• Used to describe and justify the interpretation of the relevance of the animal data to humans. This can include consideration of mechanistic or other supplemental information. When human evidence is lacking or has results that differ from animals, analyses of the mechanisms underlying the animal response in relation to those presumed to operate in humans, and the chemical's pharmacokinetics, can inform the extent to which the animal response is likely to be relevant to humans and potentially strengthen overall confidence in the evidence integration conclusion. Conversely, evidence for a mechanistic pathway that is expected to only occur in animals and not in humans can provide support for a conclusion that the animal evidence for an effect is not relevant to humans. • In the absence of chemical-specific evidence informing human relevance, the evidence integration narrative will briefly describe the interpreted comparability of experimental animal organs/systems to humans based on underlying biological similarity (e.g., thyroid signaling processes are well conserved across rodents and humans). Generally, a high-level systems summary should be possible for most encountered effects. In some cases, however, it may be appropriate to use a statement such as, 'without evidence to the contrary, [health effect described in the table] responses in animals are presumed to be relevant to humans.' As noted in EPA guidelines (U.S. EPA, 2005a), there needs to be evidence or a biological explanation to support an interpreted lack of human relevance for findings in animals, and site concordance is neither expected nor required.
Cross-stream coherence	• Addresses the concordance of findings known to be biologically related across human, animal, and mechanistic studies, considering factors such as exposure timing and levels. Notably, for many health effects (e.g., some nervous system and reproductive effects; cancer), it is not necessary (or expected) that effects manifest in humans are identical to those observed in animals, although this typically provides stronger evidence. For example, tumors in one animal species can be predictive of carcinogenic potential in

Judgment	Description
	humans or other species, but not necessarily at the same site. EPA guidelines and other resources (e.g., OECD guidelines) are consulted when drawing these inferences. • Mechanistic support for, or biological understanding of, the relatedness between different outcomes (and the manner in which they are manifest) in different species can inform an understanding of coherence across evidence streams. Evidence supporting a biologically plausible mechanistic pathway across species adds coherence (see below).
Potential susceptibility Susceptible populations and lifestages	• Used to summarize analyses relating to individual and social factors that may increase susceptibility to exposure-related health effects in certain populations or lifestages, or to highlight the lack of such information. These analyses are based on knowledge about the health outcome or organ system affected and focus primarily on the influence of intrinsic biological factors such as race/ethnicity, genetic variability, sex, lifestage, and pre-existing health conditions (which can also have an extrinsic basis). Information on extrinsic factors potentially influencing susceptibility (e.g., proximity to exposure; certain lifestyle factors including subsistence living) are not considered in evidence integration judgments on potential susceptibility; these exposure-focused factors are considered by risk managers after the human health assessment is complete. Evaluation of potential susceptibility can also include consideration of mechanistic and ADME evidence.
Biological plausibility or MOA understanding	• Support for the biological plausibility of an association between exposure and the health effect increases evidence certainty, particularly when observed across species. This may be provided by data from experimental studies of mechanistic pathways, particularly when support is provided for key events or is conserved across multiple components of the pathway. Mechanisms or biological changes with broad scientific acceptance for their relevance to chemical toxicity or the health effect (e.g., key characteristics, hallmarks of cancer) may be used to organize the chemical-specific evidence and identify key events leading from exposure to the health effect. For each key event and key event relationship, the evidence is considered regarding the consistency of experimental data and the generalizability, or likelihood of similarities (e.g., in presence or function) across species, as well as the strength of the support for the biological mechanism. • Mechanistic evidence from well conducted studies that demonstrates that the health effect is unlikely to occur (i.e., species-specific effects, irrelevant exposure conditions) can support a judgment that the effects from animal or human studies are not biologically relevant, which weakens the summary evidence integration judgment. Such a decision depends on an evaluation of the certainty of the information supporting vs. opposing biological plausibility, as well as the certainty of the health effect specific findings (e.g., stronger health effect data require more certainty in mechanistic evidence opposing plausibility). Importantly, because understanding biological plausibility is dependent on expert knowledge and canonical scientific knowledge, the lack of such understanding does not provide a rationale to decrease the certainty of the evidence for an effect (NTP, 2015); (NRC, 2014). • These analyses are typically conducted separately to establish MOA understanding and referenced in the evidence integration judgment. If sufficiently supported, MOA understanding can serve to increase (e.g., strong support for mutagenicity) or increase

Judgment	Description
	(e.g., critical dependence on a key event not likely to be operant in humans) certainty in the evidence integration judgments.
Other critical inferences (optional)	• Consideration of other evidence or nonchemical-specific information that informs evidence integration judgments (e.g., read-across analyses, ADME understanding used to inform other considerations; judgments on other health effects expected to be linked to the health effect under evaluation; read-across analyses or inferences) may be separately described as "other critical inferences."

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

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

Summary evidence integration judgment[a] in narrative	Evidence integration judgment level	Explanation and example scenarios[b]
The currently available **evidence demonstrates** that [chemical] causes [health effect] in humans[c] given sufficient exposure conditions. This conclusion is based on studies of [humans or animals] that assessed [exposure or dose] levels of [range of concentrations or specific cutoff level concentration[d]].	**Evidence demonstrates**	• A strong evidence base demonstrating that [chemical] exposure causes [health effect] in humans. • This conclusion level is used if there is *robust* human evidence supporting an effect. • This conclusion level could also be used with *moderate* human evidence and *robust* animal evidence if there is strong mechanistic evidence that MOAs and key precursors identified in animals are anticipated to occur and progress in humans.
The currently available **evidence indicates** that [chemical] likely causes [health effect] in humans given sufficient exposure conditions. This conclusion is based on studies of [humans or animals] that assessed [exposure or dose] levels of [range of concentrations or specific cutoff level concentration].	**Evidence indicates (likely[e])**	• An evidence base that indicates that [chemical] exposure likely causes [health effect] in humans, although there may be outstanding questions or limitations that remain, and the evidence is insufficient for the higher conclusion level. • This conclusion level is used if there is *robust* animal evidence supporting an effect and *slight*-to-*indeterminate* human evidence, or with *moderate* human evidence when strong mechanistic evidence is lacking. • This conclusion level could also be used with *moderate* human evidence supporting an effect and *moderate-to-indeterminate* animal evidence, or with *moderate* animal evidence supporting an effect and *moderate-to-indeterminate* human evidence. In these scenarios, any uncertainties in the *moderate* evidence are not sufficient to substantially reduce confidence in the reliability of the evidence, or mechanistic evidence in the *slight* or *indeterminate* evidence base (e.g., precursors) exists to increase confidence in the reliability of the *moderate* evidence.
The currently available **evidence suggests** that [chemical] may cause [health effect] in humans This conclusion is based on studies of [humans or animals] that assessed [exposure or dose] levels of [range of concentrations or specific cutoff level concentration].	**Evidence suggests**	• An evidence base that suggests that [chemical] exposure may cause [health effect] in humans, but there are very few studies that contributed to the evaluation, the evidence is very weak or conflicting, and/or the methodological conduct of the studies is poor. • This conclusion level is used if there is *slight* human evidence and *indeterminate*-to-*slight* animal evidence. • This conclusion level is also used with *slight* animal evidence and *indeterminate*-to-*slight* human evidence.

Summary evidence integration judgment[a] in narrative	Evidence integration judgment level	Explanation and example scenarios[b]
		• This conclusion level could also be used with *moderate* human evidence and *slight* or *indeterminate* animal evidence, or with *moderate* animal evidence and *slight* or *indeterminate* human evidence. In these scenarios, there are outstanding issues or uncertainties regarding the *moderate* evidence (i.e., the synthesis judgment was borderline with *slight*), or mechanistic evidence in the *slight* or *indeterminate* evidence base (e.g., null results in well-conducted evaluations of precursors) exists to decrease confidence in the reliability of the *moderate* evidence. • Exceptionally, when there is general scientific understanding of mechanistic events that result in a health effect, this conclusion level could also be used if there is strong mechanistic evidence that is sufficient to highlight potential human toxicity[f]—in the absence of informative conventional studies in humans or in animals (i.e., *indeterminate* evidence in both).
The currently available **evidence is inadequate** to assess whether [chemical] may cause [health effect] in humans.	**Evidence inadequate**	• This conveys either a lack of information or an inability to interpret the available evidence for [health effect]. On an assessment-specific basis, a single use of this "inadequate" conclusion level might be used to characterize the evidence for multiple health effect categories (i.e., all health effects that were examined and did not support other conclusion levels).[g] • This conclusion level is used if there is *indeterminate* human and animal evidence. • This conclusion level is also used with *slight* animal evidence and *compelling evidence of no effect* human evidence. • This conclusion level could also be used with *slight*-to-*robust* animal evidence and *indeterminate* human evidence if strong mechanistic information indicated that the animal evidence is unlikely to be relevant to humans. A conclusion of **inadequate** is not a determination that the agent does not cause the indicated health effect(s). It simply indicates that the available evidence is insufficient to reach conclusions.
Strong evidence supports no effect in humans. This conclusion is based on studies of [humans or animals] that assessed [exposure or dose] levels of [range of concentrations].	**Strong evidence supports no effect**	• This represents a situation in which extensive evidence across a range of populations and exposure levels has identified no effects/associations. This scenario requires a *high* degree of confidence in the conduct of individual studies, including consideration of study sensitivity, and comprehensive assessments of the endpoints and lifestages of exposure relevant to the heath effect of interest.

Summary evidence integration judgment[a] in narrative	Evidence integration judgment level	Explanation and example scenarios[b]
		• This conclusion level is used if there is compelling evidence of no effect in human studies and compelling evidence of no effect to indeterminate in animals. • This conclusion level is also used if there is *indeterminate* human evidence and *compelling evidence of no effect* animal evidence in models concluded to be relevant to humans. • This conclusion level could also be used with *compelling evidence of no effect* in human studies and *moderate* to *robust* animal evidence if strong mechanistic information indicated that the animal evidence is unlikely to be relevant to humans.

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

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

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

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

[e]For some applications, such as benefit-cost analysis, to better differentiate the categories of "evidence demonstrates" and "evidence indicates," the latter category should be interpreted as evidence that supports an exposure-effect linkage that is likely to be causal.

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

[g]Specific narratives for each of these health effects may also be deemed unnecessary.

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

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

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

9.1. OVERVIEW

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

For IRIS assessments, dose response assessments are typically performed for both noncancer and cancer hazards, and for both oral and inhalation routes of exposure following chronic exposure[11] to the chemical of interest, if supported by existing data. For noncancer hazards, an inhalation reference concentration (RfC) and an oral reference dose (RfD) will be derived. In addition to an RfC and RfD, this assessment will attempt to derive organ- or system-specific toxicity values when the data are sufficiently strong (i.e., noncancer conclusions of evidence demonstrate or evidence indicates [likely]). A reference value may also be derived for cancer effects in cases where a nonlinear MOA is concluded that indicates a key precursor event necessary for carcinogenicity does not occur below a specific exposure level (U.S. EPA, 2005a) (see Section 3.3.4). In addition, when feasible and if the available data are appropriate for doing so, the assessment will derive a less-than-lifetime toxicity value (a "subchronic" reference value) for noncancer hazards. Both less-than-lifetime and hazard-specific values may be useful to EPA risk assessors within specific decision contexts.

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

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

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

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

9.2. SELECTING STUDIES FOR DOSE-RESPONSE ASSESSMENT

9.2.1. Hazard and MOA Considerations for Dose Response

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

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

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

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

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

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

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

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

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

[a]An exposure or other variable that is associated with both exposure and outcome but is not an intermediary between the two.

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

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

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

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

Endpoint	Study reference/ confidence	Exposure route and duration	Human population or test species and strain	Lifestage and sex	POD derivation	Rationale
Endocrine Effects (hazard judgment of evidence indicates [likely])						
Decreased serum free and total T4	(NTP, 2018); high confidence	Gavage, 28 d	S-D rat	Adult female	Yes, ✓	Dose-dependent effects in free and total T4 in females and free T4 in males; large magnitude of effect in both sexes (91% reduction in free T4 in males at low dose where body weight unaffected, and 36%–53% reduction in free and total T4 in females at ≥3.12 mg/kg-d); effects in males were not prioritized due to elevated weight loss at higher doses.

Endpoint	Study reference/ confidence	Exposure route and duration	Human population or test species and strain	Lifestage and sex	POD derivation	Rationale
Endocrine Effects (hazard judgment of evidence indicates [likely])						
	(NTP, 2018); high confidence	Gavage, 28 d	S-D rat	Adult male	No, X	
Add a second endpoint, maybe not modeled due to large insensitivity vs. T4				Adult males and females	No, X	

1

9.3. CONDUCTING DOSE-RESPONSE ASSESSMENTS

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

1) Within the observed dose range, the preferred approach is to use dose-response modeling to incorporate as much of the data set as possible into the analysis for the purpose of deriving a POD, see Section 9.3.1 for more details.

2) Derivation of cancer risk estimates or reference values nearly always involves extrapolation to exposures lower than the POD and is described in more detail in Sections 9.3.2 and 9.3.3, respectively.

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

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

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

9.3.1. Dose-Response Analysis in the Range of Observation

For conducting a dose response assessment, pharmacodynamic ("biologically based") modeling can be used when there are sufficient data to ascertain the mode of action and quantitatively support model parameters that represent rates and other quantities associated with the key precursor events of the modes of action. When pharmacodynamic modeling is not available to assess health effects associated with inhalation exposure to vanadium compounds, empirical dose-response modeling is used to fit the data (on the apical outcomes or a key precursor events) in the ranges of observation. For this purpose of empirical dose-response modeling, EPA has developed a standard set of models (http://www.epa.gov/ncea/bmds) that can be applied to typical dichotomous and continuous data sets, including those that are nonlinear. In situations where there are alternative models with significant biological support, the users of the assessment can be informed by the presentation of these alternatives along with the models' strengths and

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

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

For noncancer (and nonlinear cancer) datasets, EPA recommends (1) application of a preferred set of models that use maximum likelihood estimation (MLE) methods (default models in BMDS) and (2) selection of a POD from a single model based on criteria designed to limit model selection subjectivity (auto implemented in BMDS version 3 and higher). For the linear analysis of cancer datasets, EPA recommends (1) application of the Multistage MLE model; (2) selection of a single Multistage degree; and (3) in c ses where tumors are observed in multiple organ systems, use of a multi-tumor model (i.e., MS-Combo) that appropriately estimates combined tumor risk (both (2) and (3) are available in BMDS).[12]

Version 3.2 and higher of BMD also provides an alternative modeling approach that uses Bayesian model averaging for dichotom ous modeling average (DMA). BMDS also provide a BMA modeling approach for dichotomous data. EPA is in the process of evaluating this approach for use in assessments and may provide supplementary values derived from such modeling.

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

The selection of the response level at which the POD is calculated is guided by the severity of the endpoint. If linear extrapolation is used, selection of a response level corresponding to the POD is not highly influential, so standard values near the low end of the observable range are

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

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

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

- Intermittent study exposures are standardized to a daily average over the duration of exposure. For chronic effects, daily exposures are averaged over the lifespan. Exposures during a critical period, however, are not averaged over a longer duration ((U.S. EPA, 2005a), see Section 3.1.1; (U.S. EPA, 1991), see Section 3.2). Note that this will typically be done after modeling because the conversion is linear.

- Doses are standardized to equivalent human terms to facilitate comparison of results from different species. Oral doses are scaled allometrically using $mg/kg^{3/4}$ day as the equivalent dose metric across species. Allometric scaling pertains to equivalence across species, not across life stages, and is not used to scale doses from adult humans or mature animals to infants or children ((U.S. EPA, 2011a); (U.S. EPA, 2005a), see section 3.1.3). Inhalation exposures are scaled using dosimetry models that apply species-specific physiologic and anatomic factors and consider whether the effect occurs at the site of first contact or after systemic circulation ((U.S. EPA, 1994); (U.S. EPA, 2012a), see Section 3).

- It can be informative to convert doses across exposure routes. If this is done, the assessment describes the underlying data, algorithms, and assumptions ((U.S. EPA, 2005a), see Section 3.1.4).

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

- The preferred approach for dosimetry extrapolation from animals to humans is through PBPK modeling.

- Briefly, PBPK model simulations can be used to estimate internal dose metrics corresponding to the applied doses for each experimental animal bioassay. By simulating the exposure scenario for each toxicity study (e.g., 6 hours/day, 5 days/week inhalation exposure), the resulting internal metric effectively accounts for the difference between the pattern and a nominal 24 hours/day, 7 days/week exposure. The set of internal dose metrics for each toxicity study and endpoint can then be used in dose-response analysis to identify a BMDL or other POD for individual animal toxicity studies. The human version of the PBPK model can then be used to estimate the exposure concentration in air which, given continuous (24 hour/day, 7 day/week) inhalation exposure, would result in a given internal

dose POD. Any remaining uncertainty factors, including the factor of 10 for human inter-individual variability (UF_H), will then be applied for derivation of the human equivalent concentration (HECs).

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

9.3.2. Extrapolation: Slope Factors and Unit Risk

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

9.3.3. Extrapolation: Reference Values

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

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

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

modeling, which may range from detailed chemical-specific to default approaches (U.S. EPA, 2014, 2011a), and a factor of 101/2 (rounded to 3) is applied to account for the remaining uncertainty involving pharmacokinetic and pharmacodynamic differences.

- Human variation: The assessment accounts for variation in susceptibility across the human population and the possibility that the available data may not represent individuals who are most susceptible to the effect, by using a data-based adjustment or UF or a combination of the two. Where appropriate data or models for the effect or for characterizing the internal dose are available, the potential for databased adjustments for pharmacodynamics or pharmacokinetics is considered 9, 10 (U.S. EPA, 2014, 2002). When sufficient data are available, an intraspecies UF either less than or greater than 10fold may be justified (U.S. EPA, 2002). This factor may be reduced if the POD is derived from or adjusted specifically for susceptible individuals [not for a general population that includes both susceptible and nonsusceptible individuals; (see (U.S. EPA, 2002), Section 4.4.5; (U.S. EPA, 1998), Section 4.2; (U.S. EPA, 1996), Section 4; (U.S. EPA, 1994), Section 4.3.9.1;(U.S. EPA, 1991), Section 3.4). When the use of such data or modeling is not supported, an UF with a default value of 10 is considered.

- LOAEL to NOAEL: If a POD is based on a LOAEL, the assessment includes an adjustment to an exposure level where such effects are not expected. This can be a matter of great uncertainty if there is no evidence available at lower exposures. A factor of 3 or 10 is generally applied to extrapolate to a lower exposure expected to be without appreciable effects. A factor other than 10 may be used depending on the magnitude and nature of the response and the shape of the dose-response curve (U.S. EPA, 2002, 1998, 1996, 1994, 1991).

- Subchronic-to-chronic exposure: When using subchronic studies to make inferences about chronic/lifetime exposure, the assessment considers whether lifetime exposure could have effects at lower levels of exposure. A factor of up to 10 may be applied to the POD, depending on the duration of the studies and the nature of the response (U.S. EPA, 2002, 1998, 1994).

- Database deficiencies: In addition to the adjustments above, if database deficiencies raise concern that further studies might identify a more sensitive effect, organ system, or life stage, the assessment may apply a database UF (U.S. EPA, 2002, 1998, 1996, 1994, 1991). The size of the factor depends on the nature of the database deficiency. For example, the EPA typically follows the recommendation that a factor of 10 be applied if both a prenatal toxicity study and a two-generation reproduction study are missing and a factor of $10^{1/2}$ (i.e., 3) if either one or the other is missing ((U.S. EPA, 2002), Section 4.4.5).

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

10. PROTOCOL HISTORY

REFERENCES

Agrawal, H; Eden, R; Zhang, X; Fine, PM; Katzenstein, A; Miller, JW; Ospital, J; Teffera, S; Cocker, DR, III. (2009). Primary particulate matter from ocean-going engines in the Southern California Air Basin. Environ Sci Technol 43: 5398-5402. http://dx.doi.org/10.1021/es8035016.

ATSDR (Agency for Toxic Substances and Disease Registry). (2012). Toxicological profile for vanadium [ATSDR Tox Profile]. Atlanta, GA: U.S. Department of Health and Human Services, Public Health Service. http://www.atsdr.cdc.gov/toxprofiles/tp58.pdf.

Cohen, MD. (2007). Toxicity of Vanadium Compounds: Pulmonary and Immune System Targets. ACS Symp Ser Am Chem Soc 974: 217-239.

Crans, DC; Amin, SS; Keramidas, AD. (1998). Chemistry of relevance to vanadium in the environment. In JO Nriagu (Ed.), Vanadium in the environment Part 1: Chemistry and biochemistry (pp. 73-95). New York: John Wiley & Sons.

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

EFSA (European Food Safety Authority). (2017). Guidance on the use of the weight of evidence approach in scientific assessments. EFSA J 15: 1-69. http://dx.doi.org/10.2903/j.efsa.2017.4971.

González-Villalva, A; Piñón-Zárate, G; De la Peña Díaz, A; Flores-García, M; Bizarro-Nevares, P; Rendón-Huerta, EP; Colín-Barenque, L; Fortoul, TI. (2011). The effect of vanadium on platelet function. Environ Toxicol Pharmacol 32: 447-456. http://dx.doi.org/10.1016/j.etap.2011.08.010.

Goullè, JP; Mahieu, L; Castermant, J; Neveu, N; Bonneau, L; Lainè, G; Bouige, D; Lacroix, C. (2005). Metal and metalloid multi-elementary ICP-MS validation in whole blood, plasma, urine and hair Reference values. Forensic Sci Int 153: 39-44. http://dx.doi.org/10.1016/j.forsciint.2005.04.020.

Gustafsson, JP. (2019). Vanadium geochemistry in the biogeosphere -speciation, solid-solution interactions, and ecotoxicity. Appl Geochem 102: 1-25. http://dx.doi.org/10.1016/j.apgeochem.2018.12.027.

Gutiérrez-González, E; García-Esquinas, E; de Larrea-Baz, NF; Salcedo-Bellido, I; Navas-Acien, A; Lope, V; Gómez-Ariza, JL; Pastor, R; Pollán, M; Pérez-Gómez, B. (2019). Toenails as biomarker of exposure to essential trace metals: A review [Review]. Environ Res 179: 108787. http://dx.doi.org/10.1016/j.envres.2019.108787.

Hauser, R; Elreedy, S; Hoppin, JA; Christiani, DC. (1995). Airway obstruction in boilermakers exposed to fuel oil ash. A prospective investigation. Am J Respir Crit Care Med 152: 1478-1484. http://dx.doi.org/10.1164/ajrccm.152.5.7582280.

Higgins, JPT; Thomas, J; Chandler, J; Cumpston, M; Li, T; Page, MJ; Welch, VA. (2022). Cochrane handbook for systematic reviews of interventions version 6.3. Higgins, JPT; Thomas, J; Chandler, J; Cumpston, M; Li, T; Page, MJ; Welch, VA. http://www.training.cochrane.org/handbook.

Hill, AB. (1965). The environment and disease: Association or causation? Proc R Soc Med 58: 295-300. http://dx.doi.org/10.1177/003591576505800503.

Hoenig, JM; Heisey, DM. (2001). The abuse of power: The pervasive fallacy of power calculations for data analysis. Am Stat 55: 19-24.

Howard, BE; Phillips, J; Miller, K; Tandon, A; Mav, D; Shah, MR; Holmgren, S; Pelch, KE; Walker, V; Rooney, AA; Macleod, M; Shah, RR; Thayer, K. (2016). SWIFT-Review: A text-mining

workbench for systematic review. Syst Rev 5: 87. http://dx.doi.org/10.1186/s13643-016-0263-z.

Huang, J, enHow; Huang, F; Evans, L, es; Glasauer, S. (2015). Vanadium: Global (bio)geochemistry. Chem Geol 417: 68-89. http://dx.doi.org/10.1016/j.chemgeo.2015.09.019.

IARC (International Agency for Research on Cancer). (2006). Vanadium pentoxide. In IARC monographs on the evaluation of the carcinogenic risk of chemicals to humans Vol 86: Cobalt in hard metals and cobalt sulfate gallium arsenide, indium phosphide and vanadium pentoxide. Lyon, France.

IPCS (International Programme on Chemical Safety). (2001). Concise International Chemical Assessment Document 29: Vanadium pentoxide and other inorganic vanadium compounds. (CICAD 29). Geneva, Switzerland: World Health Organization. http://www.inchem.org/documents/cicads/cicads/cicad29.htm.

IPCS (International Programme on Chemical Safety). (2012). Guidance for immunotoxicity risk assessment for chemicals. (Harmonization Project Document No. 10). Geneva, Switzerland: World Health Organization. https://apps.who.int/iris/handle/10665/330098.

Ito, K; Johnson, S; Kheirbek, I; Clougherty, J; Pezeshki, G; Ross, Z; Eisl, H; Matte, TD. (2016). Intraurban variation of fine particle elemental concentrations in New York City. Environ Sci Technol 50: 7517-7526. http://dx.doi.org/10.1021/acs.est.6b00599.

Kodros, JK; Bell, ML; Dominici, F; L'Orange, C; Godri Pollitt, KJ; Weichenthal, S; Wu, X; Volckens, J. (2022). Unequal airborne exposure to toxic metals associated with race, ethnicity, and segregation in the USA. Nat Commun 13: 6329. http://dx.doi.org/10.1038/s41467-022-33372-z.

Kučera, J; Sabbioni, E. (1998). Baseline vanadium levels in human blood, serum, and urine. In JO Nriagu (Ed.), Vanadium in the environment Part 2: Health effects (pp. 75-90). New York, NY: John Wiley & Sons.

Mamane, Y; Pirrone, N. (1998). Vanadium in the atmosphere. In JO Nriagu (Ed.), Vanadium in the environment Part 1: Chemistry and biochemistry (pp. 37-71). New York, NO: John Wiley & Sons.

NASEM (National Academies of Sciences, Engineering, and Medicine). (2021). Review of U.S. EPA's ORD staff handbook for developing IRIS assessments: 2020 version. Washington, DC: National Academies Press. http://dx.doi.org/10.17226/26289.

National Toxicology Program (NTP). NTP Technical Report on the Toxicity Studies of Sodium Metavanadate (CASRN 13718-26-8) and Vanadyl Sulfate (CASRN 27774-13-6) Administered in Drinking Water to Sprague Dawley (Hsd:Sprague Dawley® SD®) Rats and B6C3F1/N Mice. (Toxicology Report 106). Research Triangle Park, NC.

NRC (National Research Council). (2014). Review of EPA's Integrated Risk Information System (IRIS) process. Washington, DC: The National Academies Press. http://dx.doi.org/10.17226/18764.

NTP (National Toxicology Program). (2002). NTP technical report on the toxicology and carcinogenesis studies of vanadium pentoxide (CAS No. 1314-62-1) in F344/N rats and B6C3F1 mice (inhalation studies). (NTP TR 507; NIH Publication No. 03-4441). Research Triangle Park, NC. https://ntp.niehs.nih.gov/publications/reports/tr/500s/tr507.

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

NTP (National Toxicology Program). (2018). 28-day evaluation of the toxicity (C04049) of perfluorononaoic acid (PFNA) (375-95-1) on Harlan Sprague-Dawley rats exposed via gavage [NTP]. http://dx.doi.org/10.22427/NTP-DATA-002-02655-0003-0000-3.

NTP (National Toxicology Program). (2019). Handbook for conducting a literature-based health assessment using OHAT approach for systematic review and evidence integration. Research Triangle, NC: National Institute of Environmental Health Sciences. https://ntp.niehs.nih.gov/ntp/ohat/pubs/handbookmarch2019_508.pdf.

Peltier, RE; Lippmann, M. (2010). Residual oil combustion: 2. Distributions of airborne nickel and vanadium within New York City. J Expo Sci Environ Epidemiol 20: 342–350. http://dx.doi.org/10.1038/jes.2009.28.

Rańska, E; Biziuk, M; Sarbu, C; Szczepaniak, K; Frontasyeva, MV; Culicov, O; Bode, P; Astel, A. (2005). Assessment of phosphatic fertilizer production impact on occupational staff based on NAA of hair, nails, and inhald particles. J Environ Sci Health A Tox Hazard Subst Environ Eng 40: 2137-2152. http://dx.doi.org/10.1080/10934520500234635.

Rehder, D. (2015). The role of vanadium in biology. Metallomics 7: 730-742. http://dx.doi.org/10.1039/c4mt00304g.

Roberts, GK; Stout, MD; Sayers, B; Fallacara, DM; Hejtmancik, MR; Waidyanatha, S; Hooth, MJ. (2016). 14-day toxicity studies of tetravalent and pentavalent vanadium compounds in Harlan Sprague Dawley rats and B6C3F1/N mice via drinking water exposure. Toxicol Rep 3: 531-538. http://dx.doi.org/10.1016/j.toxrep.2016.05.001.

Schlesinger, WH; Klein, EM; Vengosh, A. (2017). Global biogeochemical cycle of vanadium [Review]. Proc Natl Acad Sci USA 114: E11092-E11100. http://dx.doi.org/10.1073/pnas.1715500114.

Schünemann, H; Brożek, J; Guyatt, G; Oxman, A. (2013). GRADE handbook. Available online at https://gdt.gradepro.org/app/handbook/handbook.html (accessed April 22, 2022).

Shafer, MM; Toner, BM; Overdier, JT; Schauer, JJ; Fakra, SC; Hu, S; Herner, JD; Ayala, A. (2012). Chemical Speciation of Vanadium in Particulate Matter Emitted from Diesel Vehicles and Urban Atmospheric Aerosols. Environ Sci Technol 46: 189-195. http://dx.doi.org/10.1021/es200463c.

Smith, MT; Guyton, KZ; Gibbons, CF; Fritz, JM; Portier, CJ; Rusyn, I; Demarini, DM; Caldwell, JC; Kavlock, RJ; Lambert, PF; Hecht, SS; Bucher, J. R.; Stewart, BW; Baan, RA; Cogliano, VJ; Straif, K. (2016). Key characteristics of carcinogens as a basis for organizing data on mechanisms of carcinogenesis [Review]. Environ Health Perspect 124: 713-721. http://dx.doi.org/10.1289/ehp.1509912.

Spada, NJ; Cheng, X; White, WH; Hyslop, NP. (2018). Decreasing Vanadium Footprint of Bunker Fuel Emissions. Environ Sci Technol. http://dx.doi.org/10.1021/acs.est.8b02942.

Sterne, JAC; Hernán, MA; Reeves, BC; Savović, J; Berkman, ND; Viswanathan, M; Henry, D; Altman, DG; Ansari, MT; Boutron, I; Carpenter, JR; Chan, AW; Churchill, R; Deeks, JJ; Hróbjartsson, A; Kirkham, J; Jüni, P; Loke, YK; Pigott, TD; Ramsay, CR; Regidor, D; Rothstein, HR; Sandhu, L; Santaguida, PL; Schünemann, HJ; Shea, B; Shrier, I; Tugwell, P; Turner, L; Valentine, JC; Waddington, H; Waters, E; Wells, GA; Whiting, PF; Higgins, JPT. (2016). ROBINS-I: A tool for assessing risk of bias in non-randomised studies of interventions. BMJ 355: i4919. http://dx.doi.org/10.1136/bmj.i4919.

Sturini, M; Maraschi, F; Cucca, L; Spini, G; Talamini, G; Profumo, A. (2010). Determination of vanadium(V) in the particulate matter of emissions and working areas by sequential dissolution and solid-phase extraction. Anal Bioanal Chem 397: 395-399. http://dx.doi.org/10.1007/s00216-009-3277-8.

Tao, L; Fairley, D; Kleeman, MJ; Harley, RA. (2013). Effects of switching to lower sulfur marine fuel oil on air quality in the San Francisco Bay area. Environ Sci Technol 47: 10171-10178. http://dx.doi.org/10.1021/es401049x.

Thayer, KA; Shaffer, RM; Angrish, M; Arzuaga, X; Carlson, LM; Davis, A; Dishaw, L; Druwe, I; Gibbons, C; Glenn, B; Jones, R; Kaiser, JP; Keshava, C; Keshava, N; Kraft, A; Lizarraga, L; Markey, K; Persad, A; Radke, EG; ... Yost, E. (2022). Use of systematic evidence maps within the US environmental protection agency (EPA) integrated risk information system (IRIS) program: Advancements to date and looking ahead [Comment]. Environ Int 169: 107363. http://dx.doi.org/10.1016/j.envint.2022.107363.

Tullar, IV; Suffet, IH. (1975). The fate of vanadium in an urban air shed: The lower Delaware River Valley. J Air Pollut Control Assoc 25: 282-286.

U.S. Department of Commerce. (2021). The effect of imports of vanadium on the national security: an investigation conducted under section 232 of the trade expansion act of 1962, as amended.

U.S. EPA (U.S. Environmental Protection Agency). (1985). Health and environmental effects profile for vanadium pentoxide [EPA Report]. (EPA/600/X-85/114). Cincinnati, OH. https://ntrl.ntis.gov/NTRL/dashboard/searchResults/titleDetail/PB88182696.xhtml.

U.S. EPA (U.S. Environmental Protection Agency). (1987). Health effects assessment for vanadium and compounds [EPA Report]. (EPA/600/8-88/061). http://nepis.epa.gov/exe/ZyPURL.cgi?Dockey=2000T901.txt.

U.S. EPA (U.S. Environmental Protection Agency). (1988). Recommendations for and documentation of biological values for use in risk assessment [EPA Report]. (EPA600687008). Cincinnati, OH. http://cfpub.epa.gov/ncea/cfm/recordisplay.cfm?deid=34855.

U.S. EPA (U.S. Environmental Protection Agency). (1991). Guidelines for developmental toxicity risk assessment. Fed Reg 56: 63798-63826.

U.S. EPA (U.S. Environmental Protection Agency). (1994). Methods for derivation of inhalation reference concentrations and application of inhalation dosimetry [EPA Report]. (EPA600890066F). Research Triangle Park, NC. https://cfpub.epa.gov/ncea/risk/recordisplay.cfm?deid=71993&CFID=51174829&CFTOKEN=25006317.

U.S. EPA (U.S. Environmental Protection Agency). (1996). Guidelines for reproductive toxicity risk assessment [EPA Report]. (EPA/630/R-96/009). Washington, DC: U.S. Environmental Protection Agency, Risk Assessment Forum. https://nepis.epa.gov/Exe/ZyPURL.cgi?Dockey=30004YQB.txt.

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

U.S. EPA (U.S. Environmental Protection Agency). (2002). A review of the reference dose and reference concentration processes. (EPA630P02002F). Washington, DC. https://www.epa.gov/sites/production/files/2014-12/documents/rfd-final.pdf.

U.S. EPA (U.S. Environmental Protection Agency). (2005a). Guidelines for carcinogen risk assessment [EPA Report]. (EPA630P03001F). Washington, DC. https://www.epa.gov/sites/production/files/2013-09/documents/cancer_guidelines_final_3-25-05.pdf.

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

https://www.epa.gov/risk/supplemental-guidance-assessing-susceptibility-early-life-exposure-carcinogens.

U.S. EPA (U.S. Environmental Protection Agency). (2008). Provisional Peer-Reviewed Toxicity Values for vanadium pentoxide (CASRN 1314-62-1) [EPA Report]. Cincinnati, OH.

U.S. EPA (U.S. Environmental Protection Agency). (2009). Provisional peer-reviewed toxicity values for vanadium and its soluble inorganic compounds other than vanadium pentoxide (CASRN 7440-62-2 and others): Derivation of subchronic and chronic oral RfDs [EPA Report]. (EPA/690/R-09/070F). Cincinnati, OH. https://cfpub.epa.gov/ncea/pprtv/documents/Vanadium.pdf.

U.S. EPA (U.S. Environmental Protection Agency). (2010). Toxicological review of ethylene glycol monobutyl ether (EGBE) (CAS no. 111-76-2) in support of summary information on the integrated risk information system (IRIS), march 2010. (EPA/635/R-08/006F). https://iris.epa.gov/static/pdfs/0500tr.pdf.

U.S. EPA (U.S. Environmental Protection Agency). (2011a). Recommended use of body weight 3/4 as the default method in derivation of the oral reference dose. (EPA100R110001). Washington, DC. https://www.epa.gov/sites/production/files/2013-09/documents/recommended-use-of-bw34.pdf.

U.S. EPA (U.S. Environmental Protection Agency). (2011b). Toxicological review of vanadium Pentoxide (External review draft). (635R11004C). http://cfpub.epa.gov/ncea/iris_drafts/recordisplay.cfm?deid=236587.

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

U.S. EPA (U.S. Environmental Protection Agency). (2012b). Benchmark dose technical guidance [EPA Report]. (EPA100R12001). Washington, DC: U.S. Environmental Protection Agency, Risk Assessment Forum. https://www.epa.gov/risk/benchmark-dose-technical-guidance.

U.S. EPA (U.S. Environmental Protection Agency). (2014). Guidance for applying quantitative data to develop data-derived extrapolation factors for interspecies and intraspecies extrapolation [EPA Report]. (EPA/100/R-14/002F). Washington, DC: Risk Assessment Forum, Office of the Science Advisor. https://www.epa.gov/sites/production/files/2015-01/documents/ddef-final.pdf.

U.S. EPA (U.S. Environmental Protection Agency). (2015). Peer review handbook [EPA Report] (4th ed.). (EPA/100/B-15/001). Washington, DC: U.S. Environmental Protection Agency, Science Policy Council. https://www.epa.gov/osa/peer-review-handbook-4th-edition-2015.

U.S. EPA (U.S. Environmental Protection Agency). (2017). Guidance to assist interested persons in developing and submitting draft risk evaluations under the Toxic Substances Control Act. (EPA/740/R17/001). Washington, DC: U.S Environmental Protection Agency, Office of Chemical Safety and Pollution Prevention. https://www.epa.gov/sites/production/files/2017-06/documents/tsca_ra_guidance_final.pdf.

U.S. EPA (U.S. Environmental Protection Agency). (2018a). Chemistry Dashboard. Washington, DC. Retrieved from https://comptox.epa.gov/dashboard

U.S. EPA (U.S. Environmental Protection Agency). (2018b). An umbrella Quality Assurance Project Plan (QAPP) for PBPK models [EPA Report]. (ORD QAPP ID No: B-0030740-QP-1-1). Research Triangle Park, NC.

U.S. EPA (U.S. Environmental Protection Agency). (2019a). ChemView [Database]. Retrieved from https://chemview.epa.gov/chemview

U.S. EPA (U.S. Environmental Protection Agency). (2019b). CompTox Chemicals Dashboard [Database]. Research Triangle Park, NC. Retrieved from https://comptox.epa.gov/dashboard

U.S. EPA (U.S. Environmental Protection Agency). (2019c). Integrated Science Assessment (ISA) for particulate matter (final report, Dec 2019) [EPA Report]. (EPA/600/R-19/188). Washington, DC. https://cfpub.epa.gov/ncea/isa/recordisplay.cfm?deid=347534.

U.S. EPA (U.S. Environmental Protection Agency). (2020a). ORD staff handbook for developing IRIS assessments (public comment draft) [EPA Report]. (EPA/600/R-20/137). Washington, DC: U.S. Environmental Protection Agency, Office of Research and Development, Center for Public Health and Environmental Assessment. https://cfpub.epa.gov/ncea/iris_drafts/recordisplay.cfm?deid=350086.

U.S. EPA (U.S. Environmental Protection Agency). (2020b). Umbrella quality assurance project plan (QAPP) for dosimetry and mechanism-based models. (EPA QAPP ID Number: L-CPAD-0032188-QP-1-2). Research Triangle Park, NC.

U.S. EPA (U.S. Environmental Protection Agency). (2021a). IRIS assessment plan for inhalation exposure to vanadium and compounds (scoping and problem formulation materials). (EPA/635/R-21/077). https://cfpub.epa.gov/ncea/iris_drafts/recordisplay.cfm?deid=351650.

U.S. EPA (U.S. Environmental Protection Agency). (2021b). Systematic review protocol for the vanadium and compounds (oral exposure) IRIS assessment (preliminary assessment materials). (EPA/635/R-21/047).

U.S. EPA (U.S. Environmental Protection Agency). (2022). ORD staff handbook for developing IRIS assessments [EPA Report]. (EPA 600/R-22/268). Washington, DC: U.S. Environmental Protection Agency, Office of Research and Development, Center for Public Health and Environmental Assessment. https://cfpub.epa.gov/ncea/iris_drafts/recordisplay.cfm?deid=356370.

Weisskopf, MG; Seals, RM; Webster, TF. (2018). Bias amplification in epidemiologic analysis of exposure to mixtures. Environ Health Perspect 126. http://dx.doi.org/10.1289/EHP2450.

WHO (World Health Organization). (1988). Vanadium. Geneva, Switzerland. http://www.inchem.org/documents/ehc/ehc/ehc81.htm.

WHO (World Health Organization). (2000). Air quality guidelines for Europe (2nd ed.). Copenhagen, Denmark: World Health Organization, Regional Office for Europe. http://www.euro.who.int/en/health-topics/environment-and-health/air-quality/publications/pre2009/air-quality-guidelines-for-europe.

Wolffe, TAM; Whaley, P; Halsall, C; Rooney, AA; Walker, VR. (2019). Systematic evidence maps as a novel tool to support evidence-based decision-making in chemicals policy and risk management. Environ Int 130: 104871. http://dx.doi.org/10.1016/j.envint.2019.05.065.

APPENDIX A. ELECTRONIC DATABASE SEARCH STRATEGIES

Table A-1. Database search strategies for vanadium and compounds

Source	Search Strategy	Number of records
WOS 3/28/2019 3/9/2020 6/3/2021	((TS="Ammonium metavanadate" OR TS="Ammonium monovanadate" OR TS="Ammonium trioxovanadate" OR TS="Monosodium trioxovanadate" OR TS="Oxosulfatovanadium pentahydrate" OR TS="Sodium metavanadate" OR TS="Sodium o-vanadate" OR TS="Sodium orthovanadate" OR TS="Sodium pervanadate" OR TS="Sodium tetraoxovanadate" OR TS="Sodium trioxovanadate" OR TS="Sodium vanadate" OR TS="Trisodium orthovanadate" OR TS="Trisodium tetraoxovanadate" OR TS="Trisodium vanadate" OR TS="Vanadic sulfate" OR TS="vanadium" OR TS="Vanadyl sulfate" OR TS="Vanadic" OR TS="Vanadin" OR TS="sodium peroxyvanadate" OR TS="Vanadyl sulfate pentahydrate" OR TS="Ammonium vanadate" OR TS="Divanadium trioxide" OR TS="Sodium hexavanadate") AND PY=(2010-2019)) ((TS="Sodium tetravanadate" OR TS="Sodium vanadite" OR TS="Sulfovanadic acid" OR TS="vanadium salt" OR TS="Tetrachlorovanadium" OR TS="Trichlorooxo vanadium" OR TS="Trichlorooxovanadium" OR TS="Trichlorooxovanadium oxide" OR TS="Vanadic acid" OR TS="Vanadic oxide" OR TS="Vanadious" OR TS="Vanadosulfuric acid" OR TS="Vanadyl chloride" OR TS="Vanadyl trichloride" OR TS="Divanadium pentaoxide" OR TS="Divanadium pentoxide" OR TS="Vanadic acid anhydride" OR TS="Vanadic anhydride" OR TS="Vanadin(V) oxide" OR TS="Vanadium dust" OR TS="Vanadium fume" OR TS="Vanadium oxide" OR TS="Vanadium pentaoxide" OR TS="Vanadium pentoxide") AND PY=(2010-2019)) ((TS="Vanadium" AND (TS="chloride" OR TS="dichloride" OR TS="oxide" OR TS="oxychloride" OR TS="oxytrichloride" OR TS="sesquioxide" OR TS="sulfate" OR TS="sulphate" OR TS="tetrachloride" OR TS="trichloride" OR TS="trioxide")) AND PY=2010-2019)	29,092
PUBMED 3/28/2019 3/9/2020 6/3/2021	(((7440-62-2[rn] OR 00J9J9XKDE[rn] OR 27774-13-6[rn] OR 6DU9Y533FA[rn] OR 13718-26-8[rn] OR 13721-39-6[rn] OR 7803-55-6[rn] OR FL85PX638G[rn] OR 12439-96-2[rn] OR "Ammonium metavanadate"[tw] OR "Ammonium monovanadate"[tw] OR "Ammonium trioxovanadate"[tw] OR "Monosodium trioxovanadate"[tw] OR "Oxosulfatovanadium pentahydrate"[tw] OR "Sodium metavanadate"[tw] OR "Sodium o-vanadate"[tw] OR "Sodium orthovanadate"[tw] OR "Sodium pervanadate"[tw] OR "Sodium tetraoxovanadate"[tw] OR "Sodium trioxovanadate"[tw] OR "Sodium vanadate"[tw] OR "Trisodium orthovanadate"[tw] OR "Trisodium tetraoxovanadate"[tw] OR "Trisodium vanadate"[tw] OR "Vanadic sulfate"[tw] OR vanadium[tw] OR "Vanadyl sulfate"[tw] OR Vanadic[tw] OR Vanadin[tw] OR "sodium peroxyvanadate"[tw] OR "Vanadyl sulfate pentahydrate"[tw] OR 16785-	5,664

Source	Search Strategy	Number of records
	81-2[rn] OR 12436-28-1[rn] OR 12058-74-1[rn] OR 64082-34-4[rn] OR 10580-52-6[rn] OR 7718-98-1[rn] OR 1314-34-7[rn] OR 7632-51-1[rn] OR 11115-67-6[rn] OR 7727-18-6[rn] OR "Ammonium vanadate"[tw] OR "Divanadium trioxide"[tw] OR "Sodium hexavanadate"[tw] OR "Sodium tetravanadate"[tw] OR "Sodium vanadite"[tw] OR "Sulfovanadic acid"[tw] OR "vanadium salt"[tw] OR Tetrachlorovanadium[tw] OR "Trichlorooxo vanadium"[tw] OR Trichlorooxovanadium[tw] OR "Trichlorooxovanadium oxide"[tw] OR "Vanadic acid"[tw] OR "Vanadic oxide"[tw] OR Vanadious[tw] OR "Vanadosulfuric acid"[tw] OR "Vanadyl chloride"[tw] OR "Vanadyl trichloride"[tw] OR 1314-62-1[rn] OR "Divanadium pentaoxide"[tw] OR "Divanadium pentoxide"[tw] OR "Vanadic acid anhydride"[tw] OR "Vanadic anhydride"[tw] OR "Vanadin(V) oxide"[tw] OR "Vanadium dust"[tw] OR "Vanadium fume"[tw] OR "Vanadium oxide"[tw] OR "Vanadium pentaoxide"[tw] OR "Vanadium pentoxide"[tw]) OR (Vanadium[tw] AND (chloride[tw] OR dichloride[tw] OR oxide[tw] OR oxychloride[tw] OR oxytrichloride[tw] OR sesquioxide[tw] OR sulfate[tw] OR sulphate[tw] OR tetrachloride[tw] OR trichloride[tw] OR trioxide[tw]))) AND ("2010"[PDAT] : "3000"[PDAT]))	
TOXLINE 3/28/2019	@SYN0+@AND+@OR+(@TERM+@rn+7440-62-2+@TERM+@rn+27774-13-6+@TERM+@rn+13718-26-8+@TERM+@rn+13721-39-6+@TERM+@rn+7803-55-6+@TERM+@rn+12439-96-2+@TERM+@rn+16785-81-2+@TERM+@rn+12436-28-1+@TERM+@rn+12058-74-1+@TERM+@rn+64082-34-4+@TERM+@rn+10580-52-6+@TERM+@rn+7718-98-1+@TERM+@rn+1314-34-7+@TERM+@rn+7632-51-1+@TERM+@rn+11115-67-6+@TERM+@rn+7727-18-6+@TERM+@rn+1314-62-1)+@RANGE+yr+2010+2019+@NOT+@org+pubmed+pubdart+nih @SYN0+@AND+@OR+(FL85PX638G+6DU9Y533FA+00J9J9XKDE+"Ammonium+metavanadate"+"Ammonium+monovanadate"+"Ammonium+trioxovanadate"+"Monosodium+trioxovanadate"+"Oxosulfatovanadium+pentahydrate"+"Sodium+metavanadate"+"Sodium+o-vanadate"+"Sodium+orthovanadate"+"Sodium+pervanadate"+"Sodium+tetraoxovanadate"+"Sodium+trioxovanadate"+"Sodium+vanadate"+"Trisodium+orthovanadate"+"Trisodium+tetraoxovanadate"+"Trisodium+vanadate"+"Vanadic+sulfate"+vanadium+"Vanadyl+sulfate"+Vanadic+Vanadin+"sodium+peroxyvanadate"+"Vanadyl+sulfate+pentahydrate"+"Ammonium+vanadate"+"Divanadium+trioxide"+"Sodium+hexavanadate"+"Sodium+tetravanadate"+"Sodium+vanadite"+"Sulfovanadic+acid"+"vanadium+salt"+"Trichlorooxo+vanadium"+Tetrachlorovanadium+Trichlorooxovanadium+"Trichlorooxovanadium+oxide"+"Vanadic+acid"+"Vanadium+dust"+"Vanadium+fume"+"Vanadium+oxide"+"Vanadium+pentaoxide"+"Vanadium+pentoxide"+"Vanadic+oxide"+Vanadious+"Vanadosulfuric+acid"+"Vanadyl+chloride"+"Vanadyl+trichloride"+"Divanadium+pentaoxide"+"Divanadium+pentoxide"+"Vanadic+acid+anhydride"+"Vanadic+anhydride"+"Vanadin+V+oxide")+@RANGE+yr+2010+2019+@NOT+@org+pubmed+pubdart+nih @SYN0+@AND+vanadium+@OR+(chloride+dichloride+oxide+oxychloride+oxytrichloride+sesquioxide+sulfate+sulphate+tetrachloride+trichloride+trioxide)+@RANGE+yr+2010+2019+@NOT+@org+pubmed+pubdart+nih	15

Source	Search Strategy	Number of records
ATSDR Toxicological Profile for Vanadium (2012)	References pulled from ATSDR document	363
2008 & 2009 PPRTV Assessments	References pulled from PPRTV documents	75
2011 IRIS External Review Draft	References pulled from V_2O_5 IRIS document	49
2006 IARC Document	References pulled from IARC document	241
2019 PM Integrated Science Assessment	References pulled from the Integrated Science Assessment for Particulate Matter	27
OAR	References provided by Office of Air and Radiation (OAR)	10
TOTAL	25,988 unique items were discovered using this search strategy.	30,332

1

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

As noted in Section 4, reference lists from existing assessments (final or publicly available draft) were manually screened. References were identified from: PPRTV assessment of vanadium pentoxide (U.S. EPA, 2008), PPRTV assessment of vanadium and its soluble compounds other than vanadium pentoxide (U.S. EPA, 2009), IRIS External Review Draft assessment of vanadium pentoxide (U.S. EPA, 2011b), International Agency for Research on Cancer (IARC) document on vanadium pentoxide (IARC, 2006) as well as references pertinent to vanadium from the most recent Integrated Science Assessment for Particulate Matter (U.S. EPA, 2019c). In addition, references suggested by the Office of Air and Radiation (OAR) were included for screening. References were annotated with respect to the source of the record and screened using the same methods applied to the rest of the literature inventory.

Review of the citation reference lists is typically done manually because they are not available in a file format (e.g., IRIS) that permits uploading into screening software applications. Manual review entails scanning the title, study summary, or study details as presented in the resource for those that appear to meet the PECO criteria. Any records identified that were not already identified from the other sources are formatted in an RIS file format, imported into DistillerSR, annotated with respect to source, and screened as outlined in Section 4.5. For tracking assessments or reviews, the name of the source citation and the number of records imported into DistillerSR are noted. The reference list of any study included in the literature inventory is reviewed manually to identify titles that appear relevant to the PECO criteria. These citations are tracked in a spreadsheet, compared against the literature base to determine whether they are unique to the project, and then added to DistillerSR to be screened at the title and abstract stage for PECO relevance.

B.1. EPA COMPTOX CHEMICALS DASHBOARD (TOXVAL)

ToxVal is searched in the EPA CompTox Chemicals Dashboard (U.S. EPA, 2018a), and data available from the "Hazard" tab is exported from the CompTox File Transfer Protocol site. Using both the human health POD summary file and the Record Source file, citations are identified that apply to human health PODs. A citation for each referenced study is generated in HERO and verified that it is not already identified from the database search (or searches of "other sources consulted") prior to moving forward to screening in DistillerSR. Full texts are retrieved where possible; if full

texts are not available, data from the ToxVal dashboard are entered and the citation is annotated accordingly for Tableau and HAWC visualizations by adding "(ToxVal)" to the citation.

B.2. EUROPEAN CHEMICALS AGENCY (ECHA)

A search of the ECHA registered substances database is conducted using the CASRN. The registration dossier associated with the CASRN is retrieved by navigating to and clicking the eye-shaped view icon displayed in the chemical summary panel. The general information page and all subpages included under the Toxicological Information tab are downloaded in Portable Document Format (PDF), including all nested reports having unique URLs. In addition, the data are extracted from each dossier page and used to populate an Excel tracking sheet. Extracted fields include data from the general information page regarding the registration type and publication dates, and on a typical study summary page the primary fields reported in the administrative data, data source, and effect levels sections. Each study summary results in more than one row in the tracking sheet if more than one data source or effect level is reported.

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

B.3. EPA CHEMVIEW

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

- High Production Volume Challenge Database (HPVIS)
- Human Health studies (Substantial Risk Reports)
- Monitoring (includes environmental, occupational, and general entries)
- TSCA Section 4 (chemical testing results)
- TSCA Section 8(d) (health and safety studies)

- TSCA Section 8(e) (substantial risk)
- FYI (voluntary documents)

All records for ecotoxicology and physical and chemical property entries are excluded. When results are available, extractors navigate into each record until a substantial risk report link is identified and saved as a PDF file. If the report cannot be saved, due to file corruption or broken links, the record is excluded during full-text review as "unable to obtain record." Most substantial risk reports contain multiple document IDs, so citations are derived by concatenating the unique report numbers such as the (formerly) Office of Toxic Substances (OTS); TSCA Section 8(e) submission (8EHQ Num); Document Control Number (DCN); Toxic Substances Control Act Test Submissions (TSCATS RefID); and Chemical Information System (CIS) associated with each document, along with the typical author organization, year, and title. Once a citation is generated, the study moves forward to DistillerSR where it is screened according to PECO and supplemental material criteria.

B.4. NTP CHEMICAL EFFECTS IN BIOLOGICAL SYSTEMS

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

B.5. OECD ECHEMPORTAL

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

- OECD HPV
- OECD SIDS IUCLID
- SIDS United Nations Environment Programme (UNEP)

B.6. ECOTOX DATABASE

EPA's ECOTOX Knowledgebase (https://cfpub.epa.gov/ecotox/search.cfm) is searched using the CASRN. Results are refined to terrestrial mammalian studies by selecting the terrestrial tab at the top of the search page and sorting the results by species group. A citation for each referenced study is generated in HERO and verified that it is not already identified from the

database search (or searches of "other sources consulted") search prior to moving forward to screening in DistillerSR.

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

Version 3.0.9 of the CompTox Chemicals Dashboard (U.S. EPA, 2019b)is accessed for high-throughput screening (HTS) data by searching the Dashboard by CASRN. Next, the "Bioactivity" section is selected and the availability of ToxCast/Tox21 HTS data for active and inactive assays is examined in the "TOXCAST: Summary" tab. If active assays are reported, the figure is copied for presentation in the systematic evidence map. This figure presents (1) a scatterplot of scaled assay responses versus AC50 values for each active assay endpoint and (2) a cytotoxicity limit as a vertical line. More detailed information on the results of ToxCast and Tox21 assays are available in the CompTox Chemicals Dashboard section "ToxCast/Tox21," which includes chemical analysis data, dose-response data and model fits, and "flags" assigned by an automated analysis, which might suggest false positivity/negativity or indicate other anomalies in the data. This information is not summarized further for the purposes of the systematic evidence map, which is focused on identifying the extent of available evidence.