



National Toxicology Program

U.S. Department of Health and Human Services

**Handbook for Conducting a Literature-Based Health
Assessment Using OHAT Approach for Systematic Review and
Evidence Integration**

March 4, 2019

Office of Health Assessment and Translation (OHAT)

Division of the National Toxicology Program

National Institute of Environmental Health Sciences

TABLE OF CONTENTS

NOTE: page numbers and headings reflect the 2019 OHAT Handbook revisions.

Table of Contents.....ii

2019 Updates and Clarificationsv

Preface1

OHAT Evaluation Process, Systematic Review, and Evidence Integration2

 OHAT Evaluation Process..... 2

 OHAT Systematic Review and Evidence Integration..... 5

Step 1: Formulate Problem and Develop Protocol8

 OHAT Process for Identifying Topics and Formulating the Study Question 8

 Nominations 8

 Scoping, Problem Formulation, and Development of Draft PECO Statement 8

 Develop Protocol 11

 Protocol Format for Step 1 12

Step 2: Search For and Select Studies for Inclusion16

 Evidence Selection Criteria 17

 Database Searches 20

 Literature Search Strategy..... 20

 Databases 21

 Reviews, Letters, Commentaries, or Other Non-Research Articles 22

 Treatment of Special Content Types..... 22

 Non-English Studies 23

 Unpublished Data 23

 Database Content 23

 Conference Abstracts, Grant Awards, and Theses/Dissertations 24

 Identifying Evidence from Other Sources 24

 References and Citations of Included Studies 24

 Grey Literature 24

 Public Input..... 25

 Screening Process 25

 Title/Abstract Review 25

 Full-Text Review 26

 Study Flow Diagram..... 27

Step 3: Extract Data from Studies28

 Data Extraction Process and Data Warehousing 28

 Missing Data 29

 Data Extraction Elements..... 29

Step 4: Assess Internal Validity of Individual Studies33

 Internal Validity (“Risk of Bias”)..... 33

OHAT Handbook (January 9, 2015 REVISION: March 4, 2019)

Excluding or Analyzing Studies Based on Aspects of Study Quality	38
Consideration of Funding Source and Disclosure of Conflict of Interest	40
Consideration of Timing and Duration of Exposure and Route of Administration	40
Risk of Bias Assessment Process	40
Missing Information for Risk of Bias Assessment.....	41
Exposure Assessment	41
Step 5: Synthesize Evidence and Rate Confidence in Body of Evidence.....	43
Considering and Conducting a Meta-Analysis	43
Sensitivity Analysis and Meta-Regression	45
Confidence Rating: Assessment of Body of Evidence	45
Initial Confidence Based on Study Design.....	50
Domains That Can Reduce Confidence.....	50
Risk of Bias Across Studies.....	51
Unexplained Inconsistency.....	53
Directness and Applicability	57
Imprecision	58
Publication Bias	59
Domains That Can Increase Confidence	60
Large Magnitude of Association or Effect	60
Dose Response	61
Residual Confounding or Other Related Factors That Would Increase Confidence in the Estimated Effect	62
Cross-Species/Population/Study Consistency	63
Other.....	63
Combine Confidence Conclusions for All Study Types and Multiple Outcomes	63
Step 6: Translate Confidence Ratings into Level of Evidence for Health Effect	64
Step 7: Integrate Evidence to Develop Hazard Identification Conclusions.....	66
Integration of Human and Animal Evidence.....	66
Consideration of Mechanistic Data.....	68
About the Protocol.....	72
Contributors.....	72
Evaluation Team	72
Technical Advisors	72
Sources of Support.....	72
Protocol History and Revisions	72
Data Display and Software	73
Data Display	73
Software.....	73
Time and Cost Estimates	74
Handbook Peer review and Updates.....	76
Peer Reviewers (January 9, 2015 Release)	76

Future Considerations.....	76
References.....	Error! Bookmark not defined.
Typical protocol Appendices.....	84
Appendix 1: Database-Specific Search Strategies.....	84
Appendix 2: Example of Quick Reference Instructions for Risk of Bias.....	85
Appendix 3: Example of an Evidence Profile Table: PFOS/PFOA and Functional Antibody Response.....	89
Appendix 4: Template Options for Tabular Data Summary.....	90
Human Studies.....	90
Animal Studies.....	92
In Vitro Studies.....	94
Appendix 5: Template Options for Graphical Data Display.....	95
Human Studies.....	95
Animal Studies.....	96
In Vitro Studies.....	97

REVISION: 2019 UPDATES AND CLARIFICATIONS

REVISION: The Office of Health Assessment and Translation (OHAT) Handbook was first published online in 2015 (NTP 2015) to outline the standard operating procedures for systematic review and evidence integration for conducting OHAT literature-based assessments. As outlined in this handbook, the procedures are a living document with the expectation that approaches will be updated as methodological practices are refined and strategies identified that improve the ease and efficiency of conducting a systematic review.

REVISION: Consistent with the expectation for updates, the updates and clarifications in the 2019 OHAT Handbook address two topics that were identified during the conduct of evidence evaluations: 1) the process for reaching hazard conclusions from human health data alone (i.e., in the absence of animal data or when there is low confidence in the available animal data); and 2) the process for developing confidence conclusions in the overall body of evidence across multiple outcomes, study types, or exposures.

REVISION: To transparently document the updates and clarifications, the 2019 OHAT Handbook includes the following documents:

- **REVISION: 2019 OHAT Handbook Update and Clarification Summary Document**
- **REVISION: 2019 OHAT Handbook**
- **REVISION: Track changes version of the 2019 OHAT Handbook documenting the changes as follows:**
 - **REVISION:** Updated language or new text is indicated with the word “REVISION:” and is formatted in bold character font, and
 - **REVISION:** Text that has been modified or deleted is formatted with strikethrough character font.

Handbook for Conducting a Literature-Based Health Assessment Using OHAT Approach for Systematic Review and Evidence Integration

PREFACE

The National Toxicology Program (NTP) and the National Institute of Environmental Health Sciences established the NTP Office of Health Assessment and Translation (OHAT) to serve as an environmental health resource to the public and to regulatory and health agencies (Bucher *et al.* 2011). This office conducts evaluations to assess the evidence that environmental chemicals, physical substances, or mixtures (collectively referred to as "substances") cause adverse health effects and provides opinions on whether these substances may be of concern, given what is known about current human exposure levels. The opinions are referred to as NTP Level of Concern (LoC) conclusions. OHAT also organizes workshops or state-of-the-science evaluations to address issues of importance in environmental health sciences. OHAT assessments are typically published as OHAT monographs, meeting reports, and/or peer-reviewed journal publications.

In 2011, OHAT began exploring systematic-review methodology as a means to enhance transparency, foster greater consistency in methods, and increase efficiency in summarizing and synthesizing findings for literature-based health assessments of environmental substances (NTP 2012b, NTP 2012a, Birnbaum *et al.* 2013, NTP 2013b). A systematic review uses an explicit, pre-specified approach to identify, select, assess, and synthesize the data from studies in order to address a specific scientific or public health question (Higgins and Green 2011, Institute of Medicine 2011). On the basis of the systematic review, a structured framework is applied to reach conclusions on the evidence following a defined and transparent decision making process (Guyatt *et al.* 2011a, U.S. Preventive Services Task Force (USPSTF) 2011, Agency for Healthcare Research and Quality (AHRQ) 2012a). Although these methods were originally developed for evaluating the efficacy of healthcare interventions, over the past decade methods have been adapted and applied to a range of health-related activities, including diagnostic testing, treatment efficacy in preclinical studies, and health questions in animal husbandry. Systematic review methodology and structured frameworks are increasingly recommended by a wide range of agencies and institutions to address environmental health questions (European Food Safety Authority (EFSA) 2010, Agency for Toxic Substances and Disease Registry (ATSDR) 2012, Silbergeld and Scherer 2013, Johnson *et al.* 2014b, Koustas *et al.* 2014, Lam *et al.* 2014, Mandrioli *et al.* 2014, Murray and Thayer 2014, National Research Council (NRC) 2014b, NRC 2014a, Woodruff and Sutton 2014).

Systematic review methods do not supplant the role of expert scientific judgment, public participation, or other existing processes used by OHAT and NTP in the evaluation of environmental substances. However, the systematic review methods outlined here are a major part of evidence-based decision making in terms of ensuring the collection of the most complete and reliable evidence to form the basis for decisions or conclusions. Knowledge of the quality and confidence in the evidence is essential to decision making. The objective of embedding systematic methods in the OHAT evaluation processes is to enhance transparency, promote participation by the public and stakeholders in the evaluation process, ensure consistency across evaluations, facilitate updates, and support more general acceptance of evaluations to other agencies.

This document is intended to serve as a handbook, or standard operating procedures (SOP), for the development of systematic review for conducting OHAT evaluations. The SOPs are based on (1) lessons learned from developing protocols for two case studies for implementing systematic review, (2)

consideration of public comments received on systematic review during the past two years, and (3) discussions with experts at other organizations and agencies working on applying methods of systematic review to environmental health and toxicology. It provides an overview of the general OHAT evaluation process, including systematic review methodology, and procedures used to integrate evidence and to support conclusions.

Many aspects of existing methods for systematic review have informed the development of this document, and OHAT has consulted with experts in the [Cochrane Collaboration](#), [Navigation Guide](#), [GRADE Working Group](#), [CAMARADES](#), [SYRCLE](#)¹, and others, to draw upon the experience of experts in the field. New methods are needed for evidence-based evaluation of nonhuman toxicological studies, including mechanistic studies. As these procedures are developed and tested, they will be integrated into the OHAT process for NTP evaluations. The methods proposed in this document will need to be evaluated for their relevance and usefulness in reaching the goals of transparency, consistency, and identification of preventable sources of bias in studies and statistical methods applied to observational epidemiology and non-human toxicology. It is expected that these methods will evolve in response to improvements in toxicity testing, statistical methods and other elements relevant to the goals of OHAT and NTP. The handbook is a living document and will be updated as methodological practices are refined and evaluated and strategies are identified that improve the reliability, ease, and efficiency of conducting systematic reviews. (see “Handbook Peer Review & Updates”).

OHAT EVALUATION PROCESS, SYSTEMATIC REVIEW, AND EVIDENCE INTEGRATION

OHAT Evaluation Process

The OHAT evaluation process includes multiple opportunities for external scientific, public, and interagency inputs and external peer review. These are not limited or changed by the adoption of systematic review methods.

The process for conducting a systematic review and integrating evidence refers to the methods used to conduct the evaluation, which is one component of the overall evaluation process by which OHAT initiates, conducts, and ensures peer review of its evaluations. [Figure 1](#) shows the overall process for evaluations that lead to the development of a formal NTP opinion published in OHAT monographs. OHAT

¹ [GRADE Working Group](#) - Grading of Recommendations Assessment, Development and Evaluation (short GRADE) Working Group began in the year 2000 as an informal collaboration of people with an interest in addressing the shortcomings of present grading systems in health care.

[CAMARADES](#) (Collaborative Approach to Meta-Analysis and Review of Animal Data from Experimental Studies) provides a supporting framework for groups involved in the systematic review and meta-analysis of data from experimental animal studies. As of December 2014, CAMARADES has five global national co-ordinating centres: University of Edinburgh, Florey Institute of Neuroscience & Mental Health, Radboud University Nijmegen Medical Centre, University of California San Francisco and Ottawa Hospital Research Institute.

[SYRCLE](#) (SYstematic Review Centre for Laboratory animal Experimentation) was officially founded in 2012. SYRCLE focuses on the execution of SRs of animal studies towards more evidence-based translational medicine.

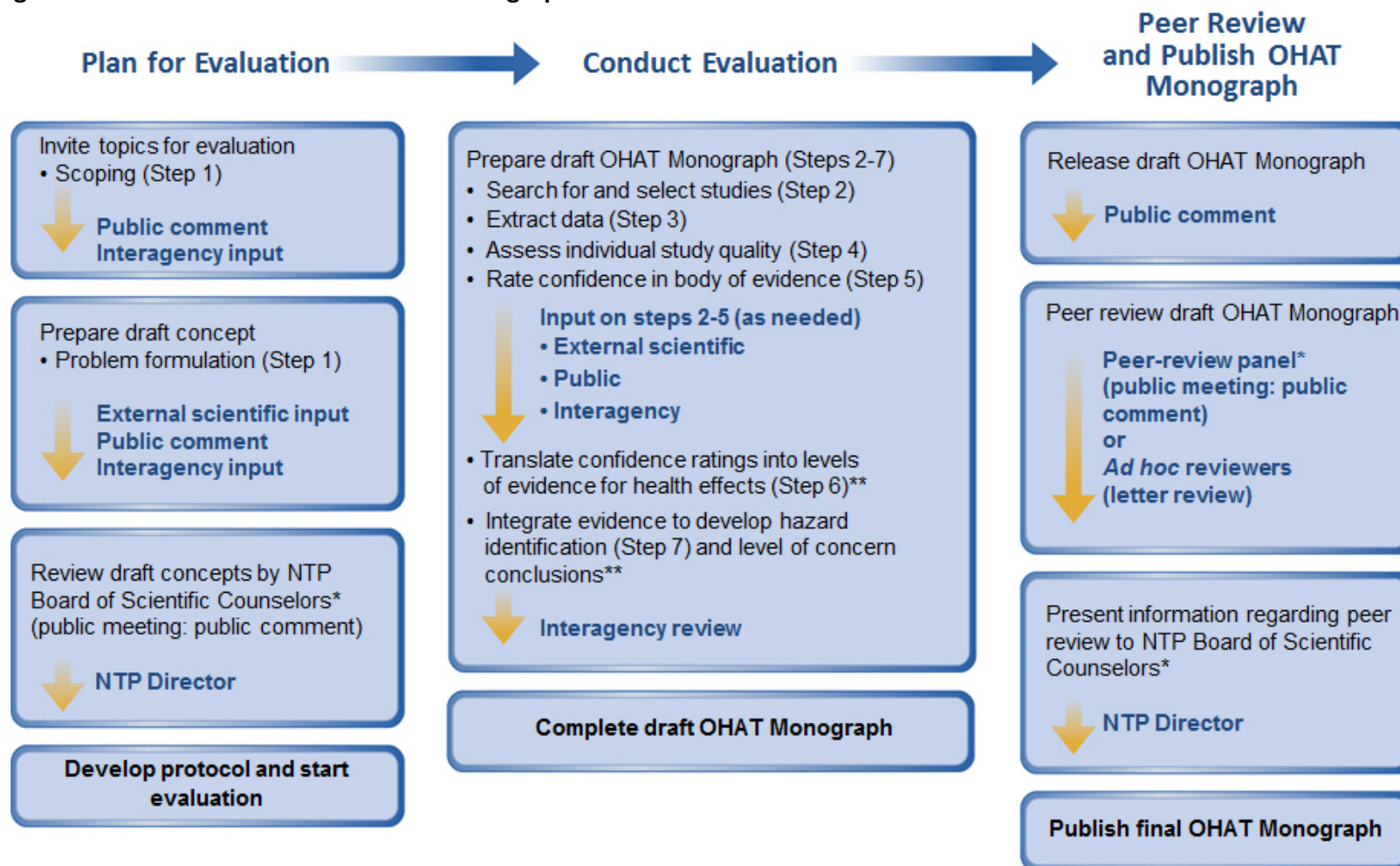
develops formal NTP opinions for *hazard identification* (for non-cancer health outcomes) and *level of concern* conclusions:

- **Hazard Identification Conclusions** – Conclusions on evidence linking an exposure to a non-cancer health outcome based on considering findings from human, animal, and mechanistic² studies: (1) **known** to be a hazard to humans, (2) **presumed** to be a hazard to humans, (3) **suspected** to be a hazard to humans, (4) **not classifiable** as a hazard to humans, or (5) **not identified** as a hazard to humans. **Note: Hazard identification labels are typically expressed by health outcome category (e.g., reproductive toxicant), with support for the label provided by evidence on specific health effects (e.g., infertility).*
- **Level of Concern (LoC) Conclusions** – For LoC conclusions OHAT integrates two categories of evidence: (1) health-outcome data from human, animal, and mechanistic studies to reach hazard identification conclusions and (2) information on the extent of exposure and pharmacokinetics. LoC conclusions are narrative (i.e., non-quantitative) conclusions that use a 5-point scale ranging from “negligible” to “serious” concern for exposure. As part of implementing systematic reviews the NTP will update its LoC framework to ensure integrated consideration of relevant and reliable evidence and to enhance transparency in describing how these conclusions are reached. These strategies will improve the LoC framework as a risk communication tool (expected completion in 2016-2017). The updated LoC framework will be included in a future version of the OHAT handbook.

The evaluation process outlined in [Figure 1](#) applies to formal NTP opinions and is similar for research projects or other literature-review evaluations that do not result in formal NTP opinions, such as state-of-the-science reviews or expert panel workshop reports, which can be published as OHAT monographs or peer-reviewed journal articles.

²Mechanistic data come from a wide variety of studies and are generally not intended to identify a disease phenotype. This source of experimental data includes *in vitro* and *in vivo* laboratory studies directed at identifying the cellular, biochemical, and molecular mechanisms that are related to chemicals that produces particular adverse effects. These studies increasingly take advantage of new “-omics” tools, such as proteomics and metabolomics, to identify early biomarkers of effect. Another broad class of mechanistic data relates to the toxicokinetics of a chemical (NRC 2014a).

Figure 1. Evaluation Process for OHAT Monographs



The use of systematic methods is in the evaluation planning and conduct phases and consists of Steps 1–7 (Rooney *et al.* 2014)

* federally chartered advisory group

** not included in state-of-science evaluation or expert panel workshop report

OHAT Systematic Review and Evidence Integration

In 2012, OHAT began using a 7-step framework for systematic review and evidence integration (Rooney *et al.* 2014), [Figure 2](#). This framework is implemented during the planning and conduct of an evaluation in [Figure 1](#). OHAT's systematic review methodology is conducted in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) Statement criteria (Moher *et al.* 2009).

In brief, systematic review methods for hazard identification use a pre-specified approach to both identify evidence, including selection and collection of studies relevant to the research question, and evaluate evidence from the studies included in the review. Each of these elements is conducted in a transparent and documented manner such that others can follow and replicate the review process from the definition of the topic through the evaluation of the evidence. Pre-specifying, or setting criteria prior to undertaking a systematic review is critical as it ensures the objectivity of the evaluation and that criteria are not developed to support a particular outcome. Pre-specifying criteria also facilitates use of consistent criteria across reviews. As shown in [Figures 2 and 3](#), the process of evidence integration occurs after conducting the systematic review and is used by OHAT to reach one of five possible hazard identification categories: (1) Known, (2) Presumed, (3) Suspected, (4) Not classifiable, or (5) Not identified to be a hazard to humans. After this point, the hazard identification conclusion is considered by OHAT in the context of additional information on human exposure and pharmacokinetics to reach a Level of Concern conclusion ([Figure 3](#)), consistent with current NTP practice.

Figure 2. OHAT Framework for Systematic Review and Evidence Integration

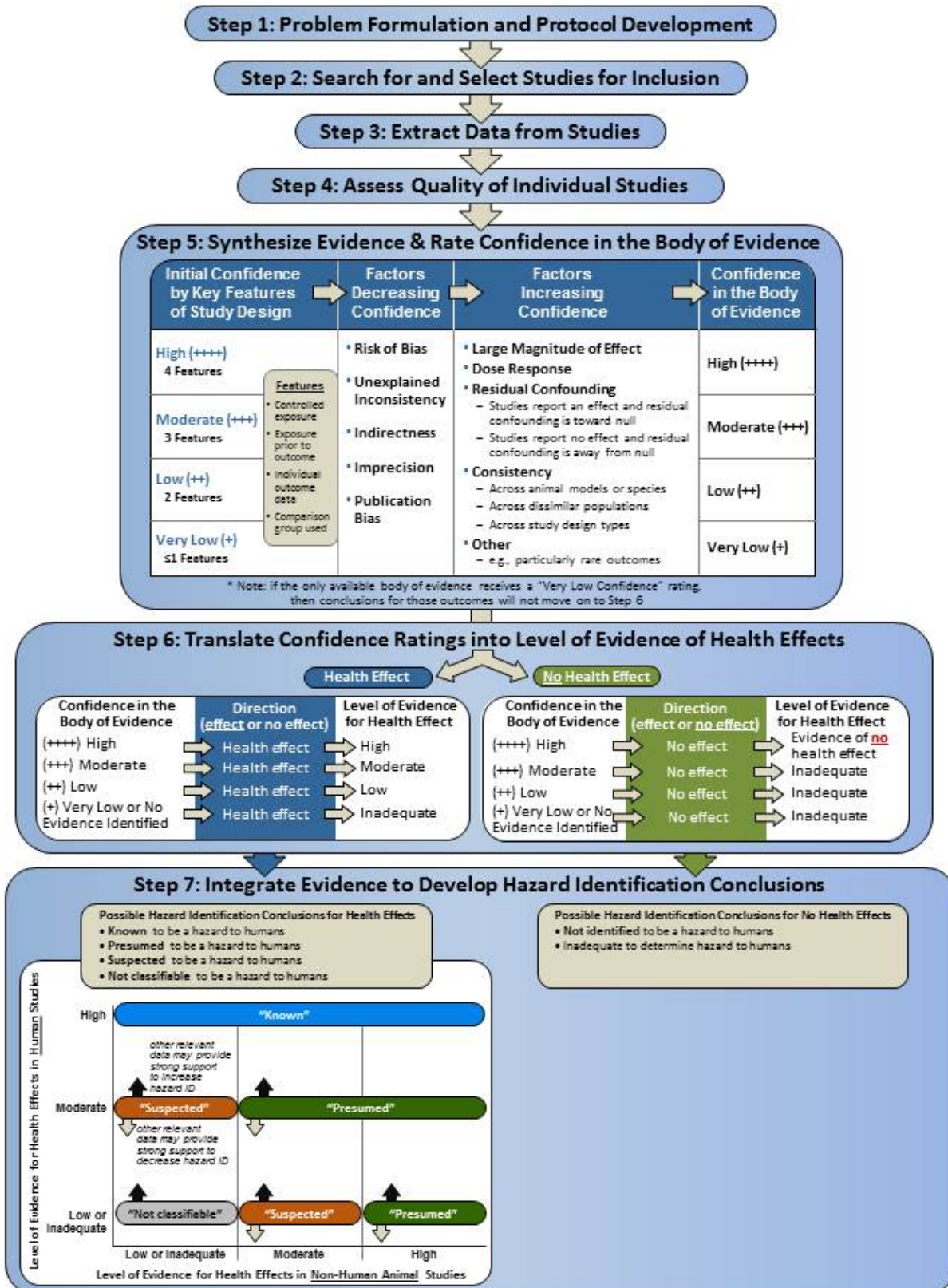
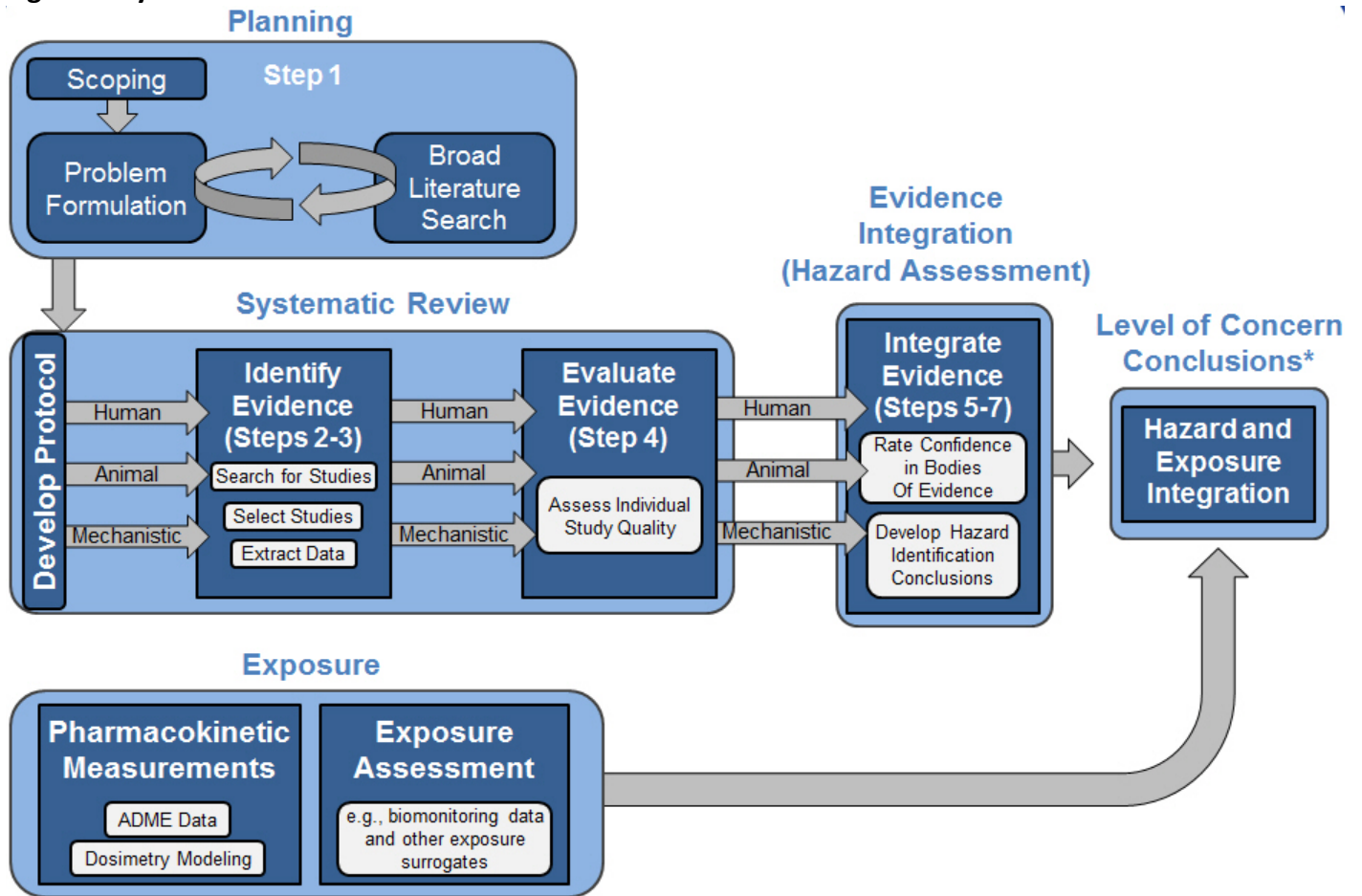


Figure 3. Systematic Review in the Context of an OHAT Hazard Identification or Level of Concern Conclusion



ADME = absorption, distribution, metabolism, excretion

*NTP is currently updating the NTP approach for reaching level of concern conclusions (expected 2016/2017)

STEP 1: FORMULATE PROBLEM AND DEVELOP PROTOCOL

OHAT Process for Identifying Topics and Formulating the Study Question

Nominations

The NTP has developed and maintains an open nomination process for identifying substances or topics to consider for an OHAT evaluation (<http://ntp.niehs.nih.gov/go/27911>). Nominations can come from the public, environmental health researchers, federal or state government agencies, international health organizations, industry, policy makers, labor unions, health care professionals, and others. Nominations to OHAT must be accompanied by the reason for the nomination and, whenever possible, appropriate background information, data, or literature citations. Factors considered in whether to pursue a nomination include concern as a possible public health hazard based on the extent of human exposure and/or suspicion of toxicity, the extent to which the topic has undergone evaluation by other organizations, and whether an OHAT evaluation can contribute to identifying and prioritizing research needs.

Scoping, Problem Formulation, and Development of Draft PECO Statement

This section describes the steps taken to obtain input on nominations and, if selected for OHAT evaluation, to refine the topic based on scientific review and public comment. The goal of this phase is to define the overall objective and formulate a study question that is addressable. The overall objective gives the scope of the evaluation, and the study question is defined through the “PECO” statement (Populations, Exposures, Comparators, and Outcomes). The PECO statement guides the entire review process, including the literature search strategy, inclusion/exclusion criteria, the type of data extracted from studies, and the strategy for synthesis and reporting of results. Proposed topic(s) should be feasible, of high priority, not duplicative, and of high potential impact. Key questions should reflect areas of uncertainty.

Definitions:

- **Scoping** refers to the process of seeking input from federal agencies, the public, and other stakeholders to understand the extent of interest in a nomination, assess the potential impact of conducting an evaluation, and identify related activities that may be underway. This information is used to begin to define the realm of the evaluation and focus the question to ensure that each assessment is as informative and useful as possible for the various groups that will use the evaluation (EPA 1998, NRC 2014a).
- **Problem formulation** (also called topic refinement) refers to the first step in the systematic-review process in which an explicit definition or statement is reached on what is to be evaluated in the assessment and how it is to be evaluated (EPA 1998, NRC 2014a, Rooney *et al.* 2014). Problem formulation is necessary to define the overall objective and PECO statement.

Concepts of scoping and problem formulation are also utilized by the Agency for Healthcare Research (AHRQ 2014) and EPA's Integrated Risk Information System (IRIS) (US EPA)2013) as described in the National Research Committee (NRC) Review of EPA's IRIS Process (NRC 2014a).

Scoping

1. NTP informs the NTP Executive Committee³ about the nomination, solicits input on their interest in the evaluation and its relevance to their agency, and solicits names of agency technical staff that should be involved in the evaluation. Initial decisions on whether to pursue a nomination further are considered based on expected use, impact, potential duplication of effort, and feasibility.
2. NTP solicits public input on the nomination of a substance or topic via a request for information (RFI) that appears in the *NIH Guide for Grants and Contracts*, the *Federal Register*, and/or NTP listserv (<http://ntp.niehs.nih.gov/help/contactus/listserv/>). Requested information typically includes (1) general comments on the nomination; (2) potential areas of focus and key issues; (3) unpublished, ongoing, or planned research; and (4) names of scientists with knowledge relevant to the topic. At this point a webpage for each evaluation is posted on the OHAT website, which is updated as the evaluation progresses.
3. In parallel, the OHAT staff person managing the project (the project lead) organizes an evaluation team (federal staff and contractor staff) who are involved in the entire systematic review process. As needed, OHAT will also engage non-federal technical advisors, who are screened for potential conflicts of interest. Contractor staff members are also screened for potential conflicts of interest. Federal staff members should do a self-evaluation for conflicts of interest. The NTP provides information about the potentially affected companies.

Problem Formulation and Creating PECO Statement

4. The project lead and staff formulate the problem to be reviewed in consultation with an information specialist⁴ (or specialists; it may be necessary to include more than one expert given the range of information sources relevant to OHAT evaluations). With this consultation the evaluation team and technical advisors design strategies for supporting a search of the literature to identify possible health outcomes of interest for the topic under investigation. This process is exploratory in terms of optimizing search strategies that be used to access and collect sources of information. Results from this literature search may be reviewed to inventory or survey the body of literature, and studies may be broadly characterized by evidence stream (human, animal, mechanistic), type of health outcome or endpoint, and type of exposure or exposure assessment. At this step no results are extracted or summarized. Text mining tools such as SWIFT (Sciome

³The NTP Executive Committee provides programmatic and policy oversight to the NTP Director and meets once or twice a year in closed forum. Members of this committee include the heads (or their designees) from the following federal agencies: Consumer Product Safety Commission (CPSC), Department of Defense (DoD), Environmental Protection Agency (EPA), Food and Drug Administration (FDA), National Cancer Institute (NCI), National Center for Environmental Health/Agency for Toxic Substances and Disease Registry (NCEH/ATSDR), National Institute of Environmental Health Sciences (NIEHS), National Institute for Occupational Safety and Health (NIOSH), Occupational Safety and Health Administration (OSHA).

⁴ A person with expertise in information science and systematic review methods as well as subject-specific knowledge, who interacts with the evaluation team and provides advice on the literature search strategy. (NRC 2014a).

Workbench for Interactive, Computer-Facilitated Text-mining) (Howard *et al.* 2014) may be used to inventory/survey studies.

This step provides information for assessing the feasibility of the project and developing the specific study question(s) to be addressed by the systematic review. The preliminary searches of the literature will assist in identifying the breadth and depth of the available literature, which will aid the NTP in determining whether to proceed with a nomination. This step also supports the development of a draft PECO statement (AHRQ 2014).

5. After this preliminary step, in consultation with the evaluation team, the project lead prepares a draft concept document that determines the feasibility of the nomination and, for those nominations determined to be feasible in terms of the availability of relevant information, outlines the proposed approach for conducting the evaluation. Concept documents are used to facilitate review of nominations by the NTP Board of Scientific Counselors (BSC) and the public. The concept document briefly outlines the nomination and rationale, steps taken in problem formulation and, objectives for the evaluation, draft PECO statement, key scientific issues to consider, proposed format of the evaluation (if known, e.g., state-of-science evaluation, formal NTP opinion on hazard identification or level of concern), and significance of the evaluation.

The concept document typically has the following format:

Overall Objective

Background

Nomination History

Overview of human exposure data and health outcome data

Draft PECO Statement

Specific Aims (if known)

Significance

Significance/intended use

Proposed format (if known)

Summary of Problem Formulation Activities

Results of scoping reports

Consideration of public and scientific input

Consideration of potential duplication of effort with recent or ongoing evaluations by others

Consideration of key scientific issues and areas of complexity

6. The project lead presents the draft concept to the NIEHS/NTP Project Review Committee for internal review and revises the draft concept document in response to comments as necessary. The NTP shares draft concepts with its partner agencies, invites their review, and revises the concept document as needed.

7. The concept document is posted on the NTP website for public comments (written and oral) and reviewed by the NTP Board of Scientific Counselors (BSC) during a public meeting (Figure 1). The BSC is asked to consider questions similar to the following when reviewing the concept:
 - Comment on the merit of the proposed evaluation relative to the mission and goals of the NTP. The NTP's stated goals are to provide information on potentially hazardous substances to all stakeholders, develop and validate improved testing methods, strengthen the science base in toxicology, and coordinate toxicology testing programs across the US Department of Health and Human Services (HHS) (<http://ntp.niehs.nih.gov/about/index.html>)
 - Comment on the clarity and validity of the rationale for the proposed evaluation as articulated in the NTP evaluation concept document. Has the scope of the problem been adequately defined? Have the relevant scientific issues been identified and clearly articulated? Are you aware of other scientific issues that need to be considered?
 - Comment on the proposed approach for further developing and refining the evaluation.
 - Rate the overall significance and public health impact of this evaluation as low, moderate, or high.
 - Provide any other comments you feel NTP staff should consider in developing this evaluation.

Develop Protocol

A protocol is a detailed plan or set of steps to be followed in a systematic review and should describe the rationale, objectives for the review, and problem formulation activities (Step 1), describe methods that will be used to locate and select relevant evidence (Step 2), data extraction of included studies (Step 3), critically appraise studies for risk of bias (Step 4), synthesize results from the included studies (Step 5), and reach hazard identification conclusions based on integrating levels of evidence across human, animal, and considering support provided by mechanistic data (Steps 6 and 7) (Higgins and Green 2011, Rooney *et al.* 2014). The concept document forms the basis for Step 1 of the protocol.

Definitions of outcomes especially for non-human toxicology and mechanistic studies are critical at this stage and need to be established to guide the search for information and plans for analysis. Since human clinical health conditions may not exist or are differently defined in non-human species, these endpoints and outcomes may need to be defined in terms that can be measured in the domain being searched (such as non-human toxicology studies). It is likely that more than one outcome will be defined in toxicology as relevant to a human disease. Whenever possible, the same definitions should be used across OHAT systematic reviews on the same outcome. Amendments and updates of the outcome definitions over time should be documented and explained by the authors of the systematic review and the most updated version of the outcome definitions will be preferred.

The protocol is developed based on feedback on the concept document from the NTP BSC, the public, and discussions with the evaluation team/technical advisors. The protocol is posted on OHAT's website. The website is updated when key milestones in the overall systematic review are reached, such as results of the literature search. The availability of these documents/materials is announced via the NTP listserv. Protocols will be submitted to relevant protocol repositories maintained by systematic review organizations.

Protocols may also describe contextual topics, defined as topics that provide important information to support the rationale or conduct of the systematic review but are not study questions addressed in the systematic review (USPSTF 2011). Contextual topics may include a variety of different types of information, such as the current levels of exposure to a chemical or substance; or prevalence, risk factors, and natural history of the health effect in question. These types of topics are generally not the study question addressed through systematic review. Instead, information to address contextual topics may be retrieved via different mechanisms: (1) targeted literature searches, (2) secondary reviews, (3) expert input, or (4) reports identified during the comprehensive literature screening for records relevant to the PECO statement. Contextual topics are not listed as separate questions in the methods section of the final report and are not reported in the results section.

The guidance developed for a protocol in advance of initiating the evaluation is meant to be comprehensive, although it is expected that during the course of an evaluation, relevant topics may be identified that were not anticipated ahead of time. When this occurs, decisions will be made on how to address the issue. When this occurs, decisions will be made on how to address the issue. Any revisions to the protocol will be documented as an update in the protocol. All versions of the protocol will remain available upon request, although the evaluation will usually proceed according to the most updated version of the protocol.

Protocol Format for Step 1

NOTE: The protocol is meant to build on the format of the concept document described earlier, thus there will be duplication in content.

Nomination History (if applicable)

This section describes the history of the nomination (if applicable) and steps the NTP has taken to solicit feedback on the topic under consideration, including *Federal Register* notices, requests for information in the *NIH Guide for Grants and Contracts*, outreach to federal agencies on the NTP Executive Committee, or outreach to other divisions within NIEHS. Provide a brief summary of any comments received during the comment periods.

Background and Significance

This section is a short overview of the topic (approximately 1-3 paragraphs) and describes the rationale/significance for conducting the review (approximately 1-2 paragraphs), emphasizing how it expands on previous reviews, addresses issues that have not been previously evaluated, or otherwise addresses knowledge gaps.

Overall Objective, Specific Aims, and PECO Statement

This section states the overall objective and specific aims of the systematic review, together with the PECO statement used to formulate an answerable question(s) and to provide more specific information about the scope of the review including specific definitions with respect to non-human toxicology and mechanistic studies. This step will guide protocol development in terms of literature search, study eligibility criteria, data extraction, and data analysis and integration (AHRQ 2014).

Examples of objectives:

- The overall objective of this evaluation is to develop hazard identification conclusions (“known,” “presumed,” “suspected,” or “not classifiable”) about whether a substance is associated with a health effect(s)⁵ by integrating levels of evidence from human, animal, and considering support provided from mechanistic studies.
- The overall objective of this evaluation is to conduct a state-of-the-science evaluation on topic Z based on evidence from human, animal, and mechanistic studies.

Examples of specific aims:

- Identify literature reporting the effects of [substance X] exposure on [health outcome Y], including human, animal, and mechanistic studies. Exclude studies based on preset criteria [e.g., repetitive publications; reviews; incomplete information on exposure or outcome]
- Data extract relevant studies.
- Assess the risk of bias of individual studies using pre-defined criteria.
- Synthesize the evidence using a narrative approach or meta-analysis (if appropriate) considering limitations on data integrating such as heterogeneity, sample size, etc.

OR for state-of-the-science evaluation:

Synthesize evidence focusing on identifying areas of consistency, uncertainty, and data gaps/research needs.

- Rate confidence in the body of evidence for human and animal studies separately according to one of four statements: 1. High, 2. Moderate, 3. Low, or 4. Very Low/No Evidence Available.
- Translate confidence ratings into level of evidence of health effects for human and animal studies separately according to one of four statements: 1. High, 2. Moderate, 3. Low, or 4. Inadequate.
- Combine the level of evidence ratings for human and animal data and consider the degree of support from mechanistic data to reach one of five possible hazard identification categories: (1) Known, (2) Presumed, (3) Suspected, (4) Not classifiable, or (5) Not identified to be a hazard to humans.

Example of PECO statement(s):

Table 1. PECO Statement for an Evaluation of Immunotoxicity Associated with Exposure to Perfluorooctanoic Acid (PFOA) and Perfluorooctane Sulfonate (PFOS)	
PECO	Human
Participants	Humans without restriction based on sex or on life stage at exposure or outcome assessment

⁵The phrases *health outcome* or *health effect* refer to a disease phenotype—for example, various cancer types, asthma, or diabetes—or specific tissue or organ system damage or dysfunction, such as liver damage, kidney damage, perturbed neurologic function, or altered reproductive function (NRC 2014a).

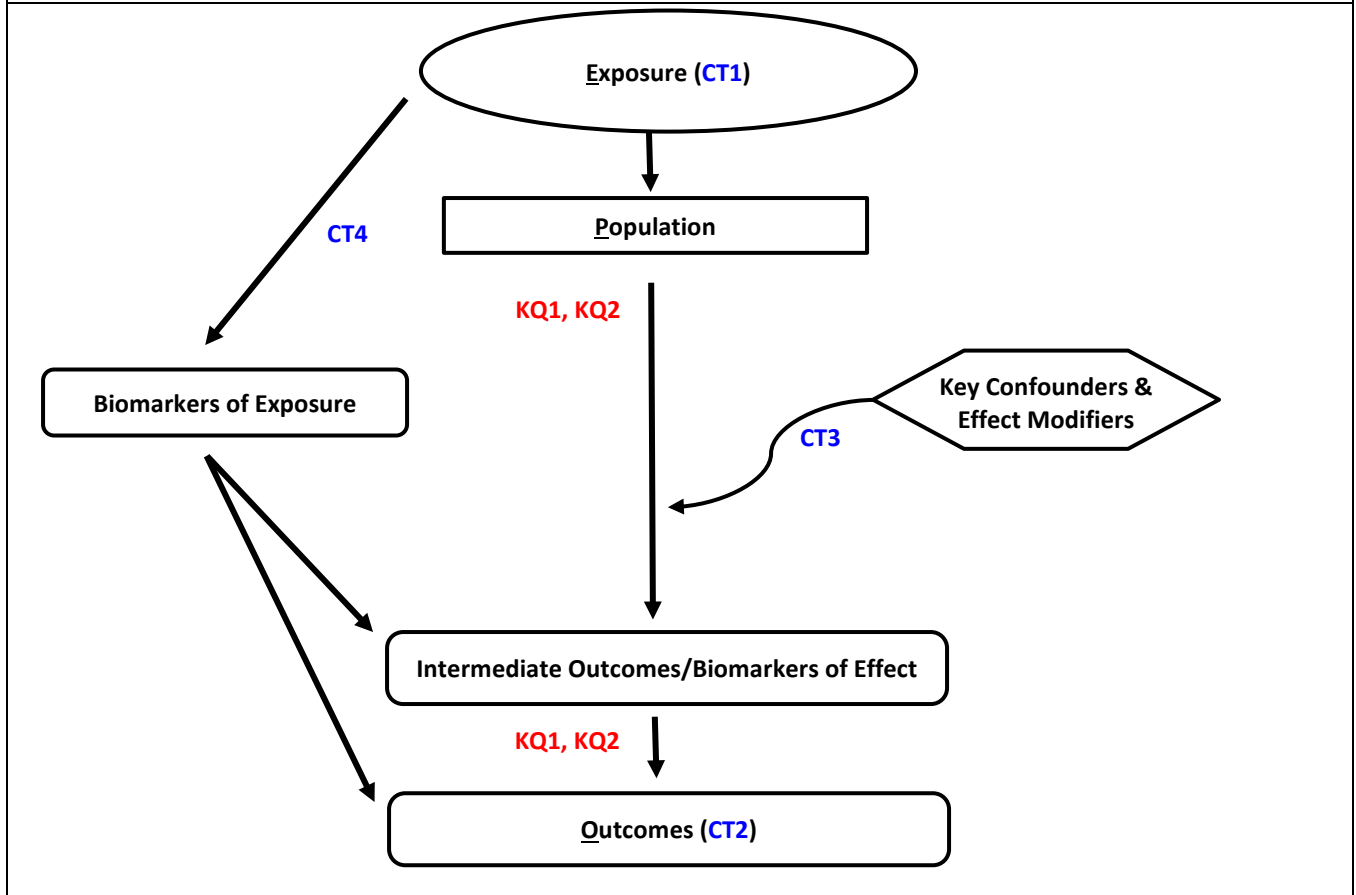
OHAT Handbook (January 9, 2015 REVISION: March 4, 2019)

<u>Exposure</u>	Exposure to PFOA (CAS# 335-67-1) or PFOS (CAS# 1763-23-1) or their salts based on administered dose or concentration, biomonitoring data (e.g., urine, blood, or other specimens), environmental measures (e.g., air, water levels), or indirect measures such as job title
<u>Comparator</u>	Humans exposed to lower levels of PFOA or PFOS
<u>Outcomes</u>	<p>Primary outcomes: Immune-related diseases and measures of immune function: <i>immunosuppression</i> (e.g., otitis, infections, or decreased vaccine antibody response); <i>sensitization and allergic response</i> (e.g., atopic dermatitis or asthma); <i>autoimmunity</i> (e.g., thyroiditis or systemic lupus erythematosus)</p> <p>Secondary outcomes: <i>Immunostimulation</i> (e.g., unintended stimulation of humoral immune function); <i>observational immune endpoints</i> (e.g., lymphocyte counts, lymphocyte proliferation, cytokine levels, serum antibody levels, or serum autoantibody levels)</p>

Key Questions and Analytical Framework

The overall objective can be represented in an analytical framework to provide a schematic that illustrates the key questions considered and types of evidence included in the evaluation (AHRQ 2014) (Figure 4). Contextual topics may also be indicated in the key question table and analytical framework to facilitate transparency in how the evidence is being collected.

Figure 4. Example of Analytical Framework Elements



PECO Statement Key Questions (KQ): Assessed by Systematic Review

KQ1	What is the hazard identification category for an association between exposure to [substance X and [health outcome Y] based on integrating levels of evidence from human and experimental animal studies: 1) Known, (2) Presumed, (3) Suspected, (4) Not classifiable, or (5) Not identified to be a hazard to humans?
KQ2	How does the evidence from other relevant studies (e.g., mechanistic studies) support or refute the biological plausibility of the association between exposure to [substance X] and [health effect Y]?

Examples of Contextual Topics (CT): Not Addressed by Systematic Review**

CT1	What are the use, production, and/or description of current levels of exposure to the chemical or substance in question?
CT2	What are the prevalence, risk factors, and natural history of the health effect in question?
CT3	What are the main potential confounders or effect modifiers that should be considered when assessing internal validity or potential bias of individual studies?
CT4	Are there data that link biomarkers of exposure to intermediate or health outcomes?

*The alternate health outcome question would be, "What is the hazard identification conclusion that environmental substance X is a Y toxicant (e.g., reproductive toxicant) in humans?"

** Contextual topics defined as topics that provide important information to support the rationale or conduct of the systematic review but are not study questions addressed in the systematic review (USPSTF 2011)

Problem Formulation Activities

The section of the protocol should describe and document major decisions made during scoping and problem formulation. It should also describe how key scientific issues will be addressed in the evaluation. Problem formulation activities include discussions of the evaluation design team, preparation of scoping reports⁶ and any external activities, such as concept review by the NTP Board of Scientific Counselors, public comment, or webinars, listening sessions, or workshops undertaken to solicit input on specific scientific or technical issues. Note that any revisions made to the protocol during the course of the systematic review are to be explained and documented in the protocol under “Protocol History and Revisions.”

Common problem formulation activity discussion points:

- Results of scoping reports
- Consideration of public and scientific input
- Consideration of potential duplication of effort with recent or ongoing evaluations by others
- Consideration of key scientific issues and areas of complexity

STEP 2: SEARCH FOR AND SELECT STUDIES FOR INCLUSION

OHAT will take reasonable steps to identify the relevant literature during the search and screening process; however, there are circumstances—especially for projects with large literature bases (e.g., ≥ 10,000 references) that cover several decades—where resource allocation needs must be considered when developing practical approaches. Thus, the specific strategy used may vary across projects with consideration of factors such as the objectives of the evaluation, size and timespan of the literature, heterogeneity of studies, and degree of scientific complexity of the topic.

To complement the literature search strategies described below, OHAT includes opportunities for the public, researchers, and other stakeholders to identify relevant studies that may have been missed by the literature search. OHAT also provides an opportunity for public review of the literature considered for an evaluation. The list of included and excluded studies will be posted on the project’s website (<http://ntp.niehs.nih.gov/go/evals>) once screening is completed and before release of the report, i.e., the draft OHAT monograph, literature publication, or workshop material(s). A second opportunity to identify any missing studies occurs when a draft OHAT monograph is disseminated for public comment prior to peer review (Figure 1).

⁶ A “scoping reports” or “scoping reviews” is a type of review has been defined as “...a form of knowledge synthesis that addresses an exploratory research question aimed at mapping key concepts, types of evidence, and gaps in research related to a defined area or field by systematically searching, selecting, and synthesizing existing knowledge (Colquhoun *et al.* 2014). Methodology guidance has not yet been developed for scoping reviews but OHAT is exploring the option to publish our scoping/problem formulation analyses as scoping reports.

Evidence Selection Criteria

Inclusion and exclusion criteria for study selection are based on the PECO statement. When major limitations (e.g., unreliable methods to assess exposure or health outcome, unknown or very limited external validity of non-human animal models or mechanistic endpoints) for addressing the key questions are known prior to evaluating individual studies, these factors may be used as a basis for excluding those studies during screening. Examples of inclusion and exclusion criteria used to screen articles for relevance and eligibility at both the title-and-abstract and full-text screening stages are detailed in Table 2. The main reason for exclusion at the full-text-review stage is annotated and reported in the study flow diagram (discussed in more detail under “Full-Text Review”).

Table 2. Examples of Inclusion and Exclusion Criteria to Determine Study Eligibility		
	Inclusion Criteria	Exclusion Criteria (may be blank if no specific criteria identified)
Population (Human Studies or Experimental Model Systems)		
human	<ul style="list-style-type: none"> Specify details on lifestage at exposure, geographic setting, clinical sub-population, sex, etc. (e.g., subjects ≤ 18 years of age) Example - PFOA/PFOS immunotoxicity: No restrictions on sex, age, or lifestage at exposure or outcome assessment 	
animal	<ul style="list-style-type: none"> Specify details on lifestage at exposure, species, strain, or sex Example - PFOA/PFOS immunotoxicity: No restrictions on sex, age, species, or lifestage at exposure or outcome assessment 	<ul style="list-style-type: none"> e.g., non-mammalian
mechanistic	<ul style="list-style-type: none"> Specify details on cellular target, cell type, or tissue type Example - PFOA/PFOS immunotoxicity: The principal form of mechanistic studies involves an <i>in vitro</i> exposure system and includes immune measures directed at cellular, biochemical, and molecular mechanisms that explain how exposure to PFOA or PFOS produces immune effects. 	<ul style="list-style-type: none"> Example - PFOA/PFOS immunotoxicity: Studies in non-animal organisms (plants, fungi, protists, archaea, bacteria)
Exposure		
human	<ul style="list-style-type: none"> Specify details on exposure measures, such as biomonitoring data (e.g., urine, blood, or other specimens), environmental measurements (e.g., air, water levels), indirect measures such as job exposure matrix (JEM)(title, or the intervention) Example - PFOA/PFOS immunotoxicity: Exposure to PFOA (CAS# 335-67-1) or PFOS (CAS# 1763-23-1) or their salts based on administered dose or concentration, biomonitoring data (e.g., urine, blood, or other specimens), environmental measures (e.g., air, water levels), or indirect measures such as job title 	
animal	<ul style="list-style-type: none"> Specify details on treatment with substance of interest, dose level, route of administration 	<ul style="list-style-type: none"> e.g., models systems know to have limited relevance to human health

Table 2. Examples of Inclusion and Exclusion Criteria to Determine Study Eligibility		
	Inclusion Criteria	Exclusion Criteria (may be blank if no specific criteria identified)
	<ul style="list-style-type: none"> Example - PFOA/PFOS immunotoxicity: Exposure to PFOA or PFOS or their salts based on administered dose or concentration, bio-monitoring data (e.g., urine, blood, or other specimens), or environmental measures (e.g., air, water levels) 	<ul style="list-style-type: none"> e.g., chemical mixture studies
mechanistic	<ul style="list-style-type: none"> Specify details on treatment with substance of interest and concentration level Example - PFOA/PFOS immunotoxicity: Exposure to PFOA or PFOS or their salts based on administered dose or concentration 	<ul style="list-style-type: none"> e.g., chemical mixture studies
Comparators		
human	<ul style="list-style-type: none"> Unexposed or lowest-exposure group as the referent group (e.g., NHANES-type analyses). Note: in some projects, studies relevant to an evaluation will not have a comparison group, e.g., pharmacokinetic studies. Example - PFOA/PFOS immunotoxicity: Humans exposed to lower levels of PFOA or PFOS 	
animal	<ul style="list-style-type: none"> Vehicle control or lowest-exposure group for observational (wildlife) animal studies Example - PFOA/PFOS immunotoxicity: <i>For experimental studies:</i> animals receiving lower doses of PFOA, PFOS, or vehicle-only treatment <i>For wildlife or observational studies:</i> animals exposed to lower levels of PFOA or PFOS 	
mechanistic	<ul style="list-style-type: none"> Vehicle control Example - PFOA/PFOS immunotoxicity: Cells or tissues receiving lower doses of PFOA, PFOS, or vehicle-only treatment 	
Outcomes		
human	<p>Primary outcomes:</p> <ul style="list-style-type: none"> Most clinically relevant or accepted measures (including established surrogate measures) of health outcome, e.g., functional immune assay such as natural killer (NK) cell activity Example - PFOA/PFOS immunotoxicity: Immune-related diseases and measures of immune function: <i>Immunosuppression</i> (e.g., otitis, infections, or decreased vaccine antibody response) <i>Sensitization and allergic response</i> (e.g., atopic dermatitis or asthma) <i>Autoimmunity</i> (e.g., thyroiditis or systemic lupus erythematosus) <p>Secondary outcomes:</p>	<ul style="list-style-type: none"> Example - PFOA/PFOS immunotoxicity: Immune tissue levels of PFOA or PFOS are not by themselves immune outcomes.

Table 2. Examples of Inclusion and Exclusion Criteria to Determine Study Eligibility		
	Inclusion Criteria	Exclusion Criteria (may be blank if no specific criteria identified)
	<ul style="list-style-type: none"> Less direct, surrogate, or upstream measures of health outcome, e.g., peripheral blood cell counts of NK cells such as CD56 Example - PFOA/PFOS immunotoxicity: <i>Immunostimulation:</i> (e.g., unintended stimulation of humoral immune function) <i>Observational immune endpoints</i> (e.g., lymphocyte counts, lymphocyte proliferation, cytokine levels, serum antibody levels, or serum autoantibody levels) 	
animal	<p>Primary outcomes:</p> <ul style="list-style-type: none"> Most accepted measures (and established surrogate measures) of health outcome, e.g., functional immune assay such as natural killer (NK) cell activity Example - PFOA/PFOS immunotoxicity: Disease resistance assay or measures of immune function: <i>Disease resistance assays</i> (e.g., host resistance to influenza A or trichinella, changes in incidence or progression in animal models of autoimmune disease) <i>Immune function assays following in vivo exposure to PFOA or PFOS</i> (e.g., antibody response, natural killer cell activity, delayed-type hypersensitivity response, phago-cytosis by monocytes, local lymph-node assay) <p>Secondary outcomes:</p> <ul style="list-style-type: none"> Less direct measures, biomarkers of effect, or upstream measures of health outcome, e.g. peripheral blood cell counts of NK cells such as CD335 Example - PFOA/PFOS immunotoxicity: <i>Observational immune endpoints</i> (e.g., lymphoid organ weight, lymphocyte counts or subpopulations, lymphocyte proliferation, cytokine production, serum antibody levels, serum or tissue autoantibody levels, or histo-pathological changes in immune organs) 	<ul style="list-style-type: none"> Example - PFOA/PFOS immunotoxicity: Immune tissue levels of PFOA or PFOS are not by themselves immune outcomes.
mechanistic	<ul style="list-style-type: none"> Outcomes could include key molecular initiating events (MIEs), phenotypic or “apical” outcomes from <i>in vitro</i> studies, results from alternative models such as zebrafish or <i>C. elegans</i>, or <i>ex vivo</i>*** studies. Examples of primary outcomes could include key molecular initiating events, functional assays, or phenotypic endpoints. Example - PFOA/PFOS immunotoxicity: 	

Table 2. Examples of Inclusion and Exclusion Criteria to Determine Study Eligibility		
	Inclusion Criteria	Exclusion Criteria (may be blank if no specific criteria identified)
	<p>Primary outcomes: <i>Immune function assays following <u>in vitro</u> exposure to the test substance</i> (e.g., natural killer cell activity, phagocytosis or bacterial killing by monocytes, proliferation following anti-CD3 antibody stimulation of spleen cells or lymphocytes)</p> <p>Secondary outcomes: <i>Observational immune endpoints following <u>in vitro</u> exposure to the test substance</i> (e.g., general mitogen-stimulated lymphocyte proliferation, cytokine production)</p>	
Publication Type (e.g., specify any language restrictions, use of conference abstracts, etc.)		
	<ul style="list-style-type: none"> • Report must contain original data 	<ul style="list-style-type: none"> • Articles with no original data, e.g., editorials, reviews** • Specify any language restrictions • Studies published in abstract form only, conference presentations or posters
<p>* Ecological studies refer to population surveys with aggregate data on participants. **Relevant reviews are used as background and for reference scanning. ***<i>Ex vivo</i> studies for some endpoints may be considered primary outcomes, e.g., NK cell activity.</p>		

Database Searches

Literature Search Strategy

Strategies for the initial literature search used in problem formulation and any subsequent revisions are developed and refined in consultation with an information specialist, the evaluation team, and any additional subject matter experts as needed.

Development of the search strategy to address the PECO statement begins by identifying relevant search terms through (1) reviewing PubMed's Medical Subject Headings (MeSH) for relevant and appropriate terms, (2) extracting key terminology from relevant reviews and a set of previously identified primary data studies that are known to be relevant to the topic ("test set"), and (3) reviewing search strategies presented in other reviews.

Relevant subject headings and text words are crafted into a search strategy that is designed to maximize the sensitivity and specificity of the search results. Because each database has its own search architecture, the resulting search strategy is tailored to account for each database's unique search functionality. The search strategy is run and the results are assessed to ensure that 100% of the previously identified relevant primary studies were retrieved. The terminology used in the problem-formulation-phase search strategy may need to be revised based on feedback received during the BSC review of the concept or on the posted protocol. Searches for information on mechanisms of toxicity might include studies of other substances that act through related mechanisms.

The search strategy, date of search, and publication dates included in the search are documented in enough detail that the search could be reproduced (Appendix 1), although retrieval of the exact search results may not necessarily occur as databases are updated and changed. The literature search is updated during the evaluation to capture literature published during the course of the review. For OHAT monographs, the last search will occur shortly (e.g., typically around 6 weeks) before the planned release of the draft document for public comment and peer review. Specific stop dates for literature searching are identified at the individual protocol level.

Databases

The following databases will typically be searched:

- Embase
- PubMed
- Scopus
- Toxline
- Web of Science

Specialized literature and data sources, such as those below, are only searched when they contribute to a specific information need (e.g., chemical CAS number search) and/or when the search topic is not complex. With respect to the latter, some of these databases either (1) have word character limits for the search field that preclude searching on very long search strings, (2) do not support running complex Boolean logic strategies, and/or (3) are unable to export results. Note: mechanistic data from NTP's Tox21 and EPA's ToxCast high throughput screening platforms are available via PubChem and EPA ACToR, respectively.

<p>Chemical / Toxicology/Environmental Health</p>	<p>Agricola California EPA Toxicity Criteria Database CHE Toxicant and Disease Database EPA ACToR (Aggregated Computational Toxicology Resource) EPA Chemical Data Access Tool EPA Health & Environmental Research Online (HERO) EPA Integrated Risk Information System (IRIS) EPA Toxicity Reference Database (ToxRefDB) ExPub (includes RTECS) – subscription National Toxicology Program Study Status and Results PAN Pesticide Database PubChem Toxnet (includes CCRIS, DART, Genetox, HSDB, IRIS, ITER) SciFinder – subscription TSCATS</p>
<p>Clinical</p>	<p>CenterWatch Clinical Trials ClinicalTrials.gov Cochrane Central Register of Controlled Trials – subscription Current Controlled Trials (ISRCTN registry) EU Clinical Trials Register</p>

	WHO International Clinical Trials Registry
Grey Literature	DART-Europe (E-Theses) Grey Literature Report OAlster Open Access Theses and Dissertations OpenDOAR Registry of Open Access Repositories Virtual Health Library
Occupational Health	International Labour Organization CISDOC National Institute for Occupational Safety and Health (NIOSH) NIOSH TIC2 European Agency for Safety and Health at Work Labor Occupational Health Program Library (available through LibraryWorld) Occupational Safety and Health Administration (OSHA)
Regional Biomedical Databases	African Index Medicus Latin American and Caribbean Health Science Information (LILACS) Western Pacific Region Index Medicus (WPRIM)
Systematic Reviews	Cochrane Library Database of Promoting Health Effectiveness Reviews (DoPHER) Prospero

Reviews, Letters, Commentaries, or Other Non-Research Articles

The primary goal of the database literature search is to identify original data relevant to addressing the PECO statement and key questions. Thus, relevant reviews, letter, or commentaries without original data will not be part of the included literature but may be used as a source for identifying potentially relevant studies. References identified from reviews, letters, and commentaries will be noted as from “other sources.” These publications are considered for PDF retrieval only if they appear directly relevant. They will be excluded if the title and/or abstract are too general or vague to make a relevance determination. For example, for an evaluation of lead that includes cardiovascular health outcome, “The perils of metals” (no abstract) would be excluded and “Lead and cardiovascular disease?” (no abstract) would be considered for inclusion. Commentaries or letters on specific studies are reviewed during data extraction and risk of bias assessment of the referenced publication to aid in interpretation. Retracted articles are excluded.

Treatment of Special Content Types

OHAT may consider other types of publications in the literature search, including non-English studies, conference abstracts, and theses or dissertations; however, searching for these types of literature can be very resource demanding in terms of time and costs for retrieval, they may require translation (e.g., non-English publications), and obtaining the information required for data extraction may be challenging, especially for abstract-level reports. Decisions to include these types of reports are made on an individual-project level and often determined primarily by the size of the literature base. OHAT recognizes that decisions to potentially exclude the content types described below need to be balanced against concern for introducing bias in the review by excluding categories of studies; for example, the hardest studies to find tend to be those with negative or null results.

Non-English Studies

Decisions on whether to include non-English studies are made on a project-specific basis. For example, non-English studies may be excluded for projects with a large English literature base. For projects where non-English studies are considered for inclusion, they will only advance to full-text review if the title and/or abstract are available in English and sufficiently detailed to make an eligibility determination, and if review of the available information suggests that the article contains original data that are directly relevant. They will be excluded if the title and/or abstract are very general or too vague to make an eligibility determination.

Unpublished Data

NTP only includes publicly accessible, peer-reviewed information in its evaluations. Study sponsors and researchers are invited to submit unpublished data on a project during scoping of the nomination, such as in response to the initial request for information. Additional opportunities for submission of unpublished data occur when the results of the literature search or other project updates are posted on the OHAT website.

If the literature search identifies a study that may be critical to the evaluation and is not peer reviewed, the NTP's practice is to obtain external peer review if the owners of the data are willing to have the study details and results made publicly accessible. The peer review would include an evaluation of the study similar to that for peer review of a journal publication. The NTP would identify and select two to three scientists knowledgeable in scientific disciplines relevant to the topic as potential peer reviewers. Persons invited to serve as peer reviewers would be screened for conflict of interest (COI) prior to confirming their service. In most instances, the peer review would be conducted by letter review. The study authors would be informed of the outcome of the peer review and given an opportunity to clarify issues or provide missing details. OHAT would consider the peer review comments regarding the scientific and technical evaluation of the unpublished study in determining whether to include the study in its evaluation. The study and its related information, if used in the OHAT evaluation, would be included in the systematic review and publicly available. OHAT would acknowledge via a note for the report that the document underwent external peer review managed by the NTP, and the names of the peer reviewers would be identified.

Unpublished data from personal author communication can supplement a peer-reviewed study, as long as the information is made publicly available.

Database Content

Increasingly, relevant evidence for an evaluation may be available in publicly accessible databases and not necessarily in the peer-reviewed literature, e.g., data from NTP's Tox21 and EPA's ToxCast high throughput screening platforms. When peer review is considered appropriate, OHAT anticipates that the validity of assays used in a high throughput screening approach could be peer reviewed and then results from those assays included in current and future systematic reviews, rather than the validity of assays determined every time a different chemical is being assessed.

Conference Abstracts, Grant Awards, and Theses/Dissertations

Decisions on whether to include conference abstracts, presentations, posters, and theses/dissertations are made on a project-specific basis. These records may be tracked during the screening process for use in determining potential publication bias. Findings from these sources that do not eventually appear in the peer-reviewed literature within a reasonable time frame can be an indication of publication bias. Records of these types identified during screening are included when the list of included and excluded studies is posted for public review, so that authors have an opportunity to provide the accompanying published report (if it does not already appear in the list of included studies) or unpublished data. Any unpublished data received from theses and dissertations relevant to the PECO statement(s) would be handled as described under “Unpublished Data.”

Identifying Evidence from Other Sources

In addition to database searches, studies may be identified from other sources, such as reference lists of included literature, the “grey literature” (non-conventional publications, described below), and technical advisors and the public. These resources are screened using the same inclusion/exclusion criteria as for the literature search.

References and Citations of Included Studies

Once the list of included studies is determined, those studies may be the source of additional relevant references. The informationist can use Web of Science and Scopus to capture the references cited in the included studies as well as the publications that cite them. The additional references may be compared against the original search result set and any duplicates removed. The remaining cited references would be evaluated using the same inclusion and exclusion criteria. These studies would be marked as “provided from other sources” in the study selection flow diagram (Figure 5).

Grey Literature

To ensure retrieval of the relevant literature, OHAT may try to identify relevant “grey literature,” which refers to publications that are not commercially published or are not readily available to the public. These publications may include or summarize unpublished data, and their contents and bibliographies are scanned to identify references that were not retrieved from database searches. Examples of grey literature include technical reports from government agencies or scientific research groups, working papers from research groups or committees, and white papers. Any unpublished data identified in these documents relevant to the PECO statement would be handled as described under “Unpublished Data.” Government or public health organizations that routinely produce health assessments include the US Environmental Protection Agency (EPA), Food and Drug Administration (FDA), Agency for Toxic Substances and Disease Registry (ATSDR), National Institute for Occupational Safety and Health (NIOSH), US state environmental agencies (e.g., California Environmental Protection Agency), World Health Organization, European Union, Health Canada, and other international bodies. When numerous risk or hazard evaluations exist, OHAT will preferentially focus on the most recent evaluations. Members of the evaluation team, the public, and technical advisors may identify relevant grey literature. These studies will be marked as “provided from other sources” in the study selection flow diagram (Figure 5).

Public Input

OHAT may attempt to identify relevant literature and information for ongoing studies from scientific and other stakeholder communities through discussions with the evaluation team and a public request for information (RFI) that appears in the *NIH Guide for Grants and Contracts*, the *Federal Register*, and/or NTP listserv (<http://ntp.niehs.nih.gov/go/getnews>), as described above under “Scoping and Problem Formulation.” In addition, the results of the literature screening are posted on the OHAT website for review, with notification of their availability through the NTP listserv as an additional mechanism to identify any relevant studies. References provided by technical advisors, the evaluation team, or members of the public will be noted as “provided from other sources” in the study selection flow diagram (Figure 5).

Screening Process

A web-based, systematic review software program with structured forms and procedures will be used to screen articles for relevance and eligibility to ensure standardization of process, e.g., DistillerSR®, DRAGON (Dose Response Analytical Generator and Organizational Network), or Health Assessment Workspace Collaborative (HAWC).⁷ Initially, results of the literature search are assembled in EndNote software and exact article duplicates removed prior to uploading the references into the systematic review software program. During the screening process, studies are broadly categorized by evidence stream (human, animal, mechanistic), type of health outcome, and type of exposure, as appropriate. This categorization occurs during the title/abstract and/or full-text levels of review, depending on the nature of the specific project.

Title/Abstract Review

In general, two reviewers independently screen all studies at the title and abstract level. If a contractor is used for this step, OHAT prefers the other reviewer to be an NTP staff member. Other approaches, such as machine-learning/text mining in conjunction with an OHAT staff screener will be incorporated as those approaches develop and are validated.

Reviewers from the evaluation team will be trained using project-specific written instructions in an initial pilot-testing phase that is undertaken on a small subset of the references retrieved. This pilot testing is performed with all team members involved in screening the literature such that everyone reviews the same set of references. Conflicts are examined for opportunities to update and improve the clarity of the inclusion and exclusion criteria, to reduce future conflicts, to limit the number of “unclear” references that move to full-text screening, and to improve accuracy and consistency among screeners. Conflicts are tracked in many systematic review software programs, such as DistillerSR®, which also includes analysis

⁷DistillerSR® (<http://systematic-review.net/>) is a proprietary project management tool for tracking studies through the screening process and storing data extracted from these studies using user-customized forms. ICF International. 2014. From Systematic Review to Assessment Development: Managing Big (and Small) Datasets with DRAGON. <http://www.icfi.com/insights/products-and-tools/dragon-dose-response>. Health Assessment Workspace Collaborative (HAWC): A Modular Web-based Interface to Facilitate Development of Human Health Assessments of Chemicals. <https://hawcproject.org/portal/>.

tools to look at concordance between screeners. Changes to the inclusion and exclusion criteria are documented in the protocol along with a date and an explanation for the modification.

Studies are not considered further when it is clear from the title or abstract that the study does not meet the inclusion criteria. In this respect, title and abstract screening is typically used to exclude studies, and final decisions for inclusion are made at the full-text level. Screening instructions for vague scenarios (e.g., title is general and no abstract is available) are made on a project-specific basis. Typically, for citations where the database contains no abstract, articles will be screened based on titles and PubMed MeSH headings. In case of screening conflicts, screeners will independently review their screening results to confirm the inclusion/exclusion decision and, if needed, discuss discrepancies with the other screener(s). If a true disagreement exists between screeners, the study passes to the full-text review. At that level, true disagreements are resolved by discussion involving another member(s) of the team or, if necessary, through consultation with technical advisors. This approach typically is sufficient to resolve disagreements, although if agreement is not reached, then the study would be included. To ensure quality control, the project lead will perform screening of a minimum of five percent or 5 papers, whichever is greater, of search results eligible for full text review.

Full-Text Review

After completion of the title/abstract screen, full-text articles are retrieved⁸ for those studies that either clearly meet the inclusion criteria or where eligibility to meet the inclusion criteria is unclear. Depending on the size and complexity of the project, full-text review will be either (1) independently conducted by two members of the review team or (2) conducted by one member of the review team, with a second member of the team confirming the exclusion determination of the first reviewer.

The list of included and excluded studies is posted on the project's website (<http://ntp.niehs.nih.gov/go/evals>) once screening has been completed and prior to completion of the report, i.e., the draft OHAT monograph, literature publication, or workshop material(s), to provide an opportunity for public review of the literature considered for an evaluation.

Multiple Publications of Same Data

Multiple publications with overlapping data for the same study (e.g., publications reporting subgroups, additional outcomes or exposures outside the scope of an evaluation, or longer follow-up) are identified by examining author affiliations, study designs, cohort name, enrollment criteria, and enrollment dates. If necessary, study authors will be contacted to clarify any uncertainty about the independence of two or more articles. OHAT will include all publications on the study, select one study to use as the primary, and consider the others as secondary publications with annotation as being related to the primary record

⁸OHAT will initially attempt to retrieve a full-text copy of the study using an automated program, such as QUOSA, when possible, and NIH library services (NIH subscriptions and interlibrary loans). For publications not available through NIH, OHAT will search the Internet and/or may attempt to contact the corresponding author. Studies not retrieved through these mechanisms are excluded and notated as "not available."

during data extraction. The primary study will generally be the publication with the longest follow-up, or for studies with equivalent follow-up periods, the study with the largest number of cases or the most recent publication date. OHAT will include relevant data from all publications of the study, although if the same outcome is reported in more than one report, OHAT will exclude the duplicate data.

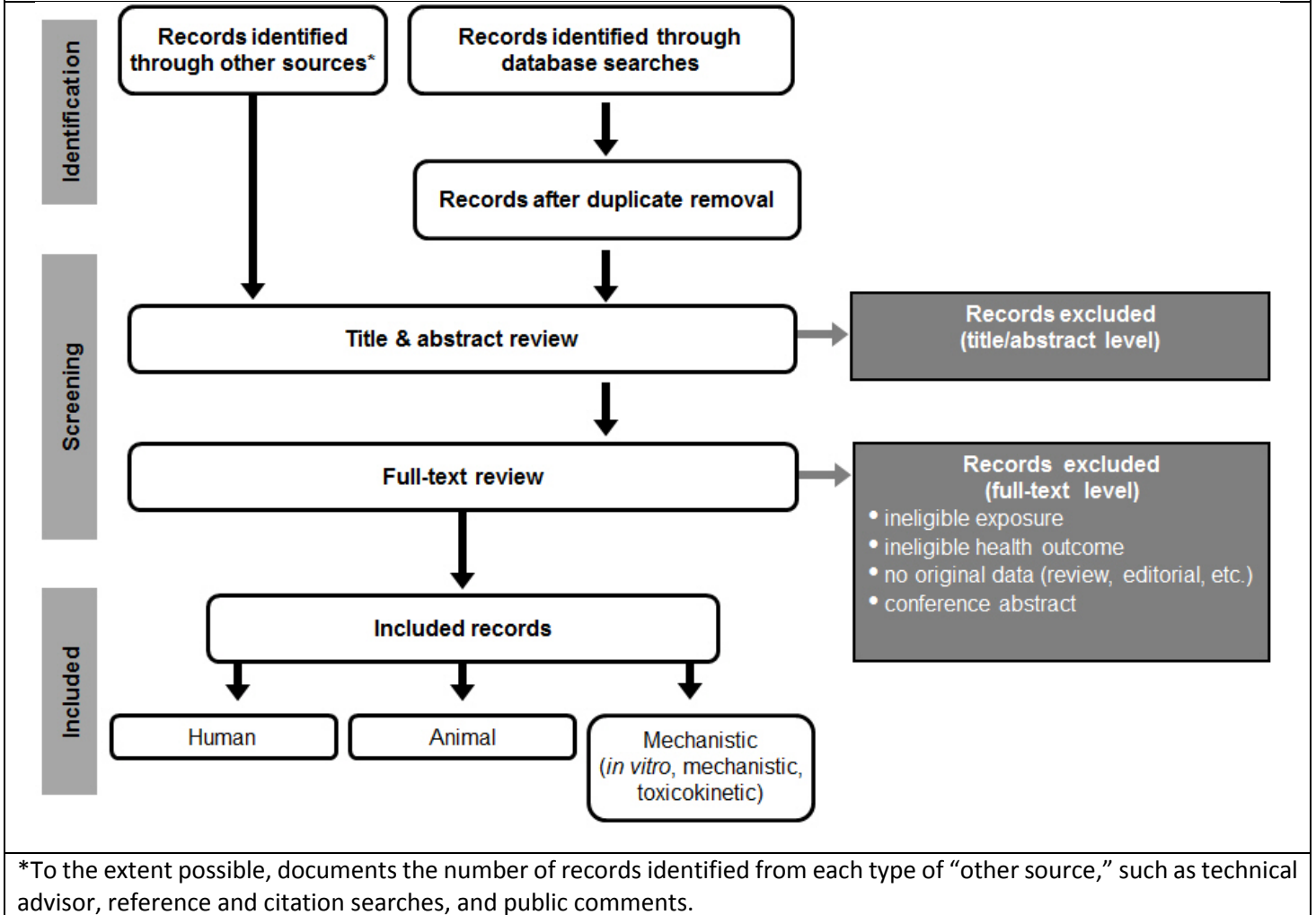
Tracking Study Eligibility and Reporting the Flow of Information

The main reason for exclusion at the full-text-review stage is annotated and reported in the study flow diagram (Figure 5). Commonly used categories for exclusion include the following: (1) is a review, commentary, or letter with no original data; (2) lacks relevant exposure information; (3) lacks relevant health outcome information; or (4) is a conference abstract (and the criteria for including unpublished data, described above, are not met). As appropriate for the evaluation topic, additional reasons for exclusion may be tracked, such as “non-English,” “ineligible study design,” “ineligible human population or experimental model system,” “thesis/dissertation,” or “multiple publication of duplicate data.” Reasons for exclusions identified during data extraction, e.g., multiple publications of same data, are annotated at the full-text review level.

Study Flow Diagram

The study flow diagram is a required element of a systematic review that is used to depict the flow of information through the different phases of the evaluation (Figure 5). It maps out the number of included and excluded records identified, and the reasons for exclusions (Moher *et al.* 2009). If OHAT conducts an updated evaluation, the study flow diagram would have a similar format but distinguish between new and previously included studies (Stovold *et al.* 2014).

Figure 5. Example of a Study Selection Flow Diagram



STEP 3: EXTRACT DATA FROM STUDIES

Data Extraction Process and Data Warehousing

Data extraction is managed with structured forms and study information stored in a database format using specialized software applications, such as ICF International’s [DRAGON](#) (Dose Response Analytical Generator and Organizational Network), [HAWC](#) (Health Assessment Workspace Collaborative), or a similar program. The application used depends on the scope and complexity of the project.

Study information collected during data extraction will be made publicly available when a draft OHAT monograph is released for public comment as an additional quality control strategy. Study information will be transferred to the NTP [Chemical Effects in Biological Systems \(CEBS\) database](#) when an evaluation is considered complete following peer review. The CEBS database serves as a public data repository to facilitate data sharing and analysis of evidence across OHAT evaluations.

At a minimum, two reviewers work independently to extract quantitative and other key data from each study related to the outcome measures under review.. One reviewer enters the data from included articles, and another member of the review team checks the extracted study information against the accompanying article(s) for completeness and accuracy as a quality control measure. Data extractors from the evaluation team will be trained using project-specific written instructions (data dictionary) in an initial pilot phase using a subset of studies. This pilot testing should be performed with all team members who will be involved in data extraction such that everyone extracts data from the same reference or set of references. This phase is undertaken to improve the clarity of the data extraction instructions and to improve accuracy and consistency among extractors. In most cases data extraction precedes assessment of an individual study’s internal validity/risk of bias (Step 4), although it may occur following Step 4 in projects where risk of bias assessment is used to exclude studies as a strategy to potentially reduce the number of studies that require full data extraction, which is costly and time intensive. Studies excluded for this reason would be indicated on the study flow diagram.

Discrepancies during data extraction are initially discussed by extractors and may involve another team member(s) or, if necessary, consultation with technical advisors to resolve disagreements. Information that is inferred, converted, or estimated during data extraction will be marked by brackets, e.g., [n=10]. Mistakes identified during data entry prior to quality control will not be annotated. Corrections made after quality control will be annotated with a rationale. An additional opportunity to identify any errors in data extraction occurs when a draft OHAT monograph is disseminated for public comment prior to peer-review (<http://ntp.niehs.nih.gov/go/38138>).

Missing Data

OHAT will attempt to contact authors of included studies to obtain missing data considered important for evaluating key study findings (e.g., level of data required to conduct a meta-analysis). The evaluation report will note that an attempt to contact study authors was unsuccessful if study researchers do not respond to an email or phone request within one month of the attempt to contact. In addition, draft OHAT monographs are posted for public comment prior to peer review, which provides another opportunity for investigators to comment on the summary of study information and other aspects of the evaluation.

Data Extraction Elements

The data extraction elements listed in [Table 3](#) are typically recorded for studies. These elements are the minimal amount of information for data extraction, and specific projects may include additional data extraction items. The extracted data will be used to help summarize study designs and findings, facilitate assessment of internal validity/risk of bias and/or conduct statistical analyses. See Appendices 3 and 4 for sample formats of how data extraction and risk of bias assessment are presented in reports for individual studies. Elements marked with an asterisk (*) are examples of items that can be used to assess risk of bias in Step 4.

Table 3. Key Data Extraction Elements to Summarize Study Design, Experimental Model, Methodology, and Results	
HUMAN	
Funding	Funding source(s)
	Reporting of conflict of interest (COI) by authors (*reporting bias)
Subjects	Study population name/description

Table 3. Key Data Extraction Elements to Summarize Study Design, Experimental Model, Methodology, and Results	
	Dates of study and sampling time frame
	Geography (country, region, state, etc.)
	Demographics (sex, race/ethnicity, age or lifestage at exposure and at outcome assessment)
	Number of subjects (target, enrolled, n per group in analysis, and participation/follow-up rates) (*missing data bias)
	Inclusion/exclusion criteria/recruitment strategy (*selection bias)
	Description of reference group (*selection bias)
Methods	Study design (e.g., prospective or retrospective cohort, nested case-control study, cross-sectional, population-based case-control study, intervention, case report, etc.)
	Length of follow-up (*information bias)
	Health outcome category, e.g., cardiovascular
	Health outcome, e.g., blood pressure (*reporting bias)
	Diagnostic or methods used to measure health outcome (*information bias)
	Confounders or modifying factors and how considered in analysis (e.g., included in final model, considered for inclusion but determined not needed (*confounding bias)
	Substance name and CAS number
	Exposure assessment (e.g., blood, urine, hair, air, drinking water, job classification, residence, administered treatment in controlled study, etc.) (*information bias)
	Methodological details for exposure assessment (e.g., HPLC-MS/MS, limit of detection) (*information bias)
	Statistical methods (*information bias)
Results	Exposure levels (e.g., mean, median, measures of variance as presented in paper, such as SD, SEM, 75th/90th/95th percentile, minimum/maximum); range of exposure levels, number of exposed cases
	Statistical findings (e.g., adjusted β , standardized mean difference, adjusted odds ratio, standardized mortality ratio, relative risk, etc.) or description of qualitative results. When possible, OHAT will convert measures of effect to a common metric with associated 95% confidence intervals (CI). Most often, measures of effect for continuous data are expressed as mean difference, standardized mean difference, and percent control response. Categorical data are typically expressed as odds ratio, relative risk (RR, also called risk ratio), or β values, depending on what metric is most commonly reported in the included studies and on OHAT's ability to obtain information for effect conversions from the study or through author query.
	If not presented in the study, statistical power can be assessed during data extraction using an approach that can detect a 10% to 20% change from response by control or referent group for continuous data, or a relative risk or odds ratio of 1.5 to 2 for categorical data, using the prevalence of exposure or prevalence of outcome in the control or referent group to determine sample size. For categorical data where the sample sizes of exposed and control or referent groups differ, the sample size of the exposed group will be used to determine the relative power category. Recommended sample sizes to achieve 80% power for a given effect size, i.e., 10% or 20% change from control, will be compared to sample sizes used in the study to categorize statistical power. Studies will be considered adequately powered when sample size for 80% power is met.
	Observations on dose response (e.g., trend analysis, description of whether dose-response shape appears to be monotonic, non-monotonic)

Table 3. Key Data Extraction Elements to Summarize Study Design, Experimental Model, Methodology, and Results	
Other	Documentation of author queries, use of digital rulers to estimate data values from figures, exposure unit, and statistical result conversions, etc.
ANIMAL	
Funding	Funding source(s)
	Reporting of COI by authors (*reporting bias)
Animal Model	Sex
	Species
	Strain
	Source of animals
	Age or lifestage at start of dosing and at health outcome assessment
	Diet and husbandry information (e.g., diet name/source)
Treatment	Chemical name and CAS number
	Source of chemical
	Purity of chemical (*information bias)
	Dose levels or concentration (as presented and converted to mg/kg bw/d when possible)
	Other dose-related details, such as whether administered dose level was verified by measurement, information on internal dosimetry (*information bias)
	Vehicle used for exposed animals
	Route of administration (e.g., oral, inhalation, dermal, injection)
	Duration and frequency of dosing (e.g., hours, days, weeks when administration was ended, days per week)
Methods	Study design (e.g., single treatment, acute, subchronic (e.g., 90 days in a rodent), chronic, multigenerational, developmental, other)
	Guideline compliance (i.e., use of EPA, OECD, NTP or another guideline for study design, conducted under GLP guideline conditions, non-GLP but consistent with guideline study, non-guideline peer-reviewed publication)
	Number of animals per group (and dams per group in developmental studies) (*missing data bias)
	Randomization procedure, allocation concealment, blinding during outcome assessment (*selection bias)
	Method to control for litter effects in developmental studies (*information bias)
	Use of negative controls and whether controls were untreated, vehicle-treated, or both
	Report on data from positive controls – was expected response observed? (*information bias)
	Endpoint health category (e.g., reproductive)
	Endpoint (e.g., infertility)
	Diagnostic or method to measure endpoint (*information bias)
Statistical methods (*information bias)	
Results	Measures of effect at each dose or concentration level (e.g., mean, median, frequency, and measures of precision or variance) or description of qualitative results. When possible, OHAT will convert measures of effect to a common metric with associated 95% confidence intervals (CI). Most often, measures of effect for continuous data will be expressed as mean difference, standardized mean difference, and percent control response. Categorical data will be expressed as relative risk (RR, also called risk ratio).

Table 3. Key Data Extraction Elements to Summarize Study Design, Experimental Model, Methodology, and Results	
	No Observed Effect Level (NOEL), Lowest Observed Effect Level (LOEL), benchmark dose (BMD) analysis, statistical significance of other dose levels, or other estimates of effect presented in paper. Note: The NOEL and LOEL are highly influenced by study design, do not give any quantitative information about the relationship between dose and response, and can be subject to author’s interpretation (e.g., a statistically significant effect may not be considered biologically important). Also, a NOEL does not necessarily mean zero response. Ideally, the response rate at specific dose levels is used as the primary measure to characterize the response.
	If not presented in the study, statistical power can be assessed during data extraction using an approach that assesses the ability to detect a 10% to 20% change from control group’s response for continuous data, or a relative risk or odds ratio of 1.5 to 2 for categorical data, using the outcome frequency in the control group to determine sample size. Recommended sample sizes to achieve 80% power for a given effect size, i.e., 10% or 20% change from control, will be compared to sample sizes used in the study to categorize statistical power. Studies will be considered adequately powered when sample size for 80% power is met.
	Observations on dose response (e.g., trend analysis, description of whether dose-response shape appears to be monotonic, non-monotonic)
	Data on internal concentration, toxicokinetics, or toxicodynamics (when reported)
Other	Documentation of author queries, use of digital rulers to estimate data values from figures, exposure unit, and statistical result conversions, etc.
IN VITRO	
Funding	Funding source(s)
	Reporting of COI by authors (*reporting bias)
Cell/Tissue Model	Cell line, cell type, or tissue
	Source of cells/tissue (and validation of identity)
	Sex of human/animal of origin
	Species
	Strain
Treatment	Chemical name and CAS number
	Concentration levels (as presented and converted to μM when possible)
	Source of chemical
	Purity of chemical (*information bias)
	Vehicle used for experimental/control conditions
	Duration and frequency of dosing (e.g., hours, days, weeks when administration was ended, times per day or week)
Methods	Guideline compliance (i.e., use of EPA, OECD, NTP or another guideline for study design, conducted under GLP guideline conditions, non-GLP but consistent with guideline study, non-guideline peer-reviewed publication)
	Randomization procedure, allocation concealment, blinding during outcome assessment (*selection bias)
	Number of replicates per group (*information bias)
	Percent serum/plasma in medium
	Use of negative controls and whether controls were untreated, vehicle-treated, or both
	Report on data from positive controls – was expected response observed? (*information bias)
	Endpoint health category (e.g., endocrine)

Table 3. Key Data Extraction Elements to Summarize Study Design, Experimental Model, Methodology, and Results

	Endpoint or assay target (e.g., estrogen receptor binding or activation)
	Name and source of assay kit
	Diagnostic or method to measure endpoint (e.g., reporter gene) (*information bias)
	Statistical methods (*information bias)
Results	No Observed Effect Concentration (NOEC), Lowest Observed Effect Concentration (LOEC), statistical significance of other concentration levels, AC50, or other estimates of effect presented in paper. Note: The NOEC and LOEC are highly influenced by study design, do not give any quantitative information about the relationship between dose and response, and can be subject to author's interpretation (e.g., a statistically significant effect may not be considered biologically important). Also, a NOEC does not necessarily mean zero response.
	Observations on dose response (e.g., trend analysis, description of whether dose-response shape appears to be monotonic, non-monotonic)
Other	Documentation of author queries, use of digital rulers to estimate data values from figures, exposure unit, and statistical result conversions, etc.
Elements marked with an asterisk (*) are examples of items that can be used to assess internal validity/risk of bias in Step 4.	

STEP 4: ASSESS INTERNAL VALIDITY OF INDIVIDUAL STUDIES

Internal Validity (“Risk of Bias”)

Individual human, animal, and *in vitro* studies will be assessed for internal validity (commonly referred to as “risk of bias” (RoB) in systematic review) by considering aspects relevant for specific study designs. Assessment of risk of bias is related to but distinguished from the broader concept of assessment of methodological quality (Higgins and Green 2011).

- **Bias** is a systematic error, or deviation from the truth, in results or inferences. Biases can operate in either direction: different biases can lead to underestimation or overestimation of the true effect. Biases can vary in magnitude: some are small (and trivial compared with the observed effect), and some are substantial (so that an apparent finding may be entirely due to bias). Even a particular source of bias may vary in direction: bias due to a particular design flaw (e.g., lack of allocation concealment) may lead to underestimation of an effect in one study but overestimation in another study. It is usually impossible to know to what extent biases have affected the results of a particular study, although there is good empirical evidence that particular flaws in the design, conduct, and analysis of randomized studies lead to bias. Because the results of a study may in fact be unbiased despite a methodological flaw, it is more appropriate to consider **risk of bias** (Higgins and Green 2011).
- **Quality** refers to the critical appraisal of included studies to evaluate the extent to which study authors conducted their research to the highest possible standards (Higgins and Green 2011).

Assessment of methodological quality is distinguished from assessment of risk of bias by Cochrane for several reasons, including the following: (1) risk of bias more directly addresses the extent to which results of included studies should be relied on; (2) a study may be performed to the highest possible standards yet still have an important risk of bias (e.g., blinding of subjects or study personnel may not have been

conducted or be impossible to achieve); (3) some markers of quality in research, such as obtaining ethical approval, performing a sample-size calculation, and reporting adequately, are unlikely to have direct implications for risk of bias; and (4) an emphasis on risk of bias overcomes ambiguity between the quality of reporting and the quality of the underlying research (Higgins and Green 2011).

Table 4 presents an overview of the types of biases considered for experimental (human or animal) and observational studies and explains how the types of biases are addressed in specific RoB assessment tools. OHAT’s current RoB tool (Table 5) is consistent with methods used by other groups or recent guidance recommendations (Higgins and Green 2011, Viswanathan *et al.* 2012, Krauth *et al.* 2013, Hooijmans *et al.* 2014, Johnson *et al.* 2014b, Koustas *et al.* 2014, NRC (National Research Council) 2014a, Sterne *et al.* 2014). The development, assessment, and validation of assessment tools that address the types of evidence typically considered in environmental health–observational human, experimental animal, and *in vitro* studies is currently an active area of methods development and lacks validation. Thus, refinements to the OHAT tool may occur to facilitate harmonization with other organizations conducting systematic reviews in environmental health.

Table 4. Types of Study Biases		
Types of Bias	Description	Risk of Bias Questions/Domains
Experimental Studies (Human or Animal)		
Selection ^{1,2,5}	Systematic differences between exposed and control groups in baseline characteristics that result from how subjects are assigned to groups. ¹ Selection bias has also been used to refer to associations of study group assignments with demographic, clinical, or social characteristics (i.e., confounding bias). ⁵	<ul style="list-style-type: none"> • Random/adequate sequence generation^{2,4,5,7,8} • Allocation concealment^{2,4,5,7,8} • Participants analyzed within groups to which they were originally assigned⁵ • Similar baseline characteristics⁸ • Design or analysis accounted for confounding^{4*,5,8} or modifying^{4*,5}
Performance ^{1,2,5}	Systematic differences between groups with regard to how the groups are handled, or in exposure to factors other than the exposure/intervention of interest. ^{1,2} Examples include deviations from the study protocol, contamination of the control group with the exposure, and inadequate blinding of providers and participants. ⁵	<ul style="list-style-type: none"> • Blinding of participants and/or personnel^{2,4,7,8} • Adherence to study protocol^{4,5} • Consideration of other exposures that might bias results^{4*,5} • Random housing within the room (animal studies)⁸
Detection/ Measurement/ Information ^{1,2,5}	Systematic differences between exposed and control groups with regard to how outcomes are assessed. Detection bias includes measurement errors (or measurement limitations) related to exposure or outcomes that occur during the course of the study. ¹	<ul style="list-style-type: none"> • Blinding of outcome assessment^{2,4,5,7,8} • Exposure assessment/ intervention^{4,5} • Measurement of outcomes^{4,5} • Measurement of confounding factors⁴ • Bias in inferential statistics⁵ • Similar length of follow-up in prospective studies⁵ • Random presentation at outcome assessment (animal studies)⁸
Missing Data/Attrition/ Exclusion ^{1,2,5}	Systematic differences between exposed and control groups in withdrawal from the study	<ul style="list-style-type: none"> • Incomplete or missing outcome data^{2,4,5,7,8}

Table 4. Types of Study Biases		
Types of Bias	Description	Risk of Bias Questions/Domains
	or exclusion from analysis. ¹ This issue is usually referred to as selection bias in observational studies. ³	
Reporting ^{1,2,5}	Selective reporting of entire studies, outcomes, or analyses. ¹ Systematic differences between reported and unreported findings. ^{2,5} Can include potential for bias in reporting through source of funding. ⁵	<ul style="list-style-type: none"> • Selective reporting^{2,3,4,5,6,7,8} • Conflict of interest^{5,6,7}
Other ²	Bias due to problems not covered elsewhere. ²	<ul style="list-style-type: none"> • Other sources of bias^{2,4,6,7,8}
Observational Studies		
Selection ^{1,3,5}	Differences in the distribution of risk factors between exposed and non-exposed groups can occur at baseline or during follow-up. ¹ In observational studies, selection bias has often been used as a synonym for confounding, but recent efforts encourage consideration of selection bias and confounding as distinct. ³	<ul style="list-style-type: none"> • Selection of participants into the study,^{3,4,5,6} e.g., similar baseline characteristics, application of inclusion/exclusion criteria, recruitment strategy
Confounding ^{1,3,5}	Occurs when one or more factors that predict the outcome of interest are also associated with exposure status. ³	<ul style="list-style-type: none"> • Design or analysis accounted for confounding^{3,4,5,6} or modifying^{4,5} • Consideration of other exposures that might bias results^{3*,4,5*} • Time-varying confounding³
Performance ⁵	Systematic differences between groups with regard to how the groups are handled, exposure to factors other than the exposure/intervention of interest. ¹ Performance bias can also be referred to as departure from intended interventions in non-randomized studies of interventions. ³	<ul style="list-style-type: none"> • Departure from intended exposure/intervention³ • Consideration of other exposures that might bias results^{4*,5} • Adherence to study protocol^{4,5}
Detection/ Measurement/ Information ^{1,3,5}	Systematic differences between exposed and control groups with regard to how outcomes are assessed. Detection bias includes measurement errors (or measurement limitations) related to exposure or outcomes that occur during the course of the study. ^{1,3}	<ul style="list-style-type: none"> • Blinding of outcome assessment^{3,4,5,6} • Exposure assessment^{3,4,5,6} • Measurement of outcomes^{3,4,5,6} • Measurement of confounding factors^{4,5} • Bias in inferential statistics⁵ • Similar length of follow-up in prospective studies, time between exposure and outcome assessment in cases and control⁵
Missing Data/Attrition/ Exclusion ^{3,5}	Systematic differences between exposed and control groups in withdrawal from the study or exclusion from analysis that can occur when the analysis does not include all	<ul style="list-style-type: none"> • Incomplete or missing outcome data^{3,4,5,6}





Types of Bias	Description	Risk of Bias Questions/Domains
	participants (e.g., differential loss during follow-up, non-response). ^{1,3} This issue is often referred to as selection bias in observational studies, but recent efforts encourage consideration of selection and missing data as distinct for observational studies. ^{3,5}	
Reporting ^{1,3,5}	Selective reporting of entire studies, outcomes, or analyses. ¹ Systematic differences between reported and unreported findings. ^{2,5} Can include potential for bias in reporting through source of funding. ⁵	<ul style="list-style-type: none"> • Selective reporting^{2,3,4,5,6,7,8} • Conflict of interest^{5,6,7}
Other	Bias due to problems not covered elsewhere. ²	<ul style="list-style-type: none"> • Other sources of bias^{2,4,6,7,8}

Sources: Based on ¹NAS 2014, Table 5-1 (NRC 2014a); ²Higgins and Green 2011, Table 8.4.a and 8.5.a (Higgins and Green 2011), in development; ³Cochrane Risk Of Bias Assessment Tool for Non-Randomized Studies of Interventions (ACROBAT-NRSI), Version 1.0.0, 24 September 2014 (Sterne *et al.* 2014); ⁴OHAT RoB tool (Rooney *et al.* 2014); ⁵AHRQ guidance (Viswanathan *et al.* 2012); ⁶Navigation Guide RoB tool for human studies (Johnson *et al.* 2014b); ⁷Navigation Guide RoB tool for animal studies (Kousta *et al.* 2014); and ⁸SYRCLE's RoB tool for animal studies (Hooijmans *et al.* 2014).

*Tool includes item, but it appears under a different RoB type than presented in this table

The OHAT RoB tool that takes a parallel approach to evaluating risk of bias from human and non-human animal studies (Table 5) to facilitate consideration of RoB across evidence streams with common terms and categories. Risk of bias domains and questions for experimental animal were based on established guidance for experimental human studies (randomized clinical trials). Instructions for response are provided in a guidance document tailored to the specific evidence stream and type of human study design in the detailed guide for using the tool (<http://ntp.niehs.nih.gov/go/38673>).

The response options for each RoB question are:

	Definitely Low risk of bias: There is direct evidence of low risk of bias practices (May include specific examples of relevant low risk of bias practices)
	Probably Low risk of bias: There is indirect evidence of low risk of bias practices OR it is deemed that deviations from low risk of bias practices for these criteria during the study would not appreciably bias results, <u>including consideration of direction and magnitude of bias</u>
	Probably High risk of bias: There is indirect evidence of high risk of bias practices OR there is insufficient information (e.g., not reported or “NR”) provided about relevant risk of bias practices
	Definitely High risk of bias: There is direct evidence of high risk of bias practices (May include specific examples of relevant high risk of bias practices)

Bias Domains and Questions	Experimental Animal¹	Human Controlled Trials²	Cohort	Case-control³	Cross-sectional	Case Series
Selection Bias						
1. Was administered dose or exposure level adequately randomized?	X	X				
2. Was allocation to study groups adequately concealed?	X	X				
3. Did selection of study participants result in appropriate comparison groups?			X	X	X	
Confounding Bias						
4. Did the study design or analysis account for important confounding and modifying variables?			X	X	X	X
Performance Bias						
5. Were experimental conditions identical across study groups?	X					
6. Were the research personnel and human subjects blinded to the study group during the study?	X	X				
Attrition/Exclusion Bias						
7. Were outcome data complete without attrition or exclusion from analysis?	X	X	X	X	X	
Detection Bias						
8. Can we be confident in the exposure characterization?	X	X	X	X	X	X
9. Can we be confident in the outcome assessment?	X	X	X	X	X	X
Selective Reporting Bias						
10. Were all measured outcomes reported?	X	X	X	X	X	X
Other Sources of Bias						
11. Were there no other potential threats to internal validity (e.g., statistical methods were appropriate and researchers adhered to the study protocol)?	X	X	X	X	X	X
¹ Experimental animal studies are controlled exposure studies. Non-human animal observational studies could be evaluated using the design features of observational human studies such as cross-sectional study design. ² Human Controlled Trials (HCTs): studies in humans with a controlled exposure, including Randomized Controlled Trials (RCTs) and non-randomized experimental studies ³ Cross-sectional studies include population surveys with individual data (e.g., NHANES) and population surveys with aggregate data (i.e., ecological studies).						

Excluding or Analyzing Studies Based on Aspects of Study Quality

Decisions on whether to exclude studies based on study quality are made on a project-specific basis and may be influenced by the goal of the evaluation (i.e., OHAT monograph presenting a formal NTP opinion versus state-of-the-science evaluation) and by consideration of the available evidence identified during problem formulation. For example, data from case studies are often excluded from projects, especially those with a large evidence base, although for some topics case studies may be the primary human evidence, in which case it would be inappropriate to exclude them.

More than one strategy may be used to exclude studies based on consideration of study quality. Ideally, key factors are identified as exclusion criteria in the PECO framework (e.g., exclusion of case reports, or use of high risk of bias exposure or health outcome assessment methods). OHAT may also utilize a 3-tier system (Table 6), in which studies in Tier 3 are excluded during assessment of individual risk of bias because there is high concern for bias on key element(s). The key elements would be determined on a project-specific basis, and for observational human studies they would typically include exposure assessment, outcome assessment, and confounding/selection. OHAT has received a variety of opinions on its proposed tiering strategy, ranging from support to concern that it resembles a scoring or scaling system of the type explicitly discouraged in the Cochrane handbook (see chapter 8.3) (Higgins and Green 2011). We do not consider our tiering approach to represent scaling, which is described in the Cochrane handbook as follows: “scores for multiple items are added up to produce a total.” Our tiering approach is conceptually consistent with an approach outlined in the Cochrane handbook for reaching summary assessments of risk of bias (see chapter 8.7, Table 8.7.a) and with methods used in certain AHRQ protocols (Higgins and Green 2011, AHRQ 2012b). Similarly, the recently developed **A Risk Of Bias Assessment Tool for Non-Randomized Studies of Interventions (ACROBAT-NRSI)** includes a framework for using responses to individual risk of bias domains to reach conclusions on overall risk of bias for a study, which includes a provision that a study may have critical risk of bias and be considered “too problematic to provide any useful evidence and should not be included in any synthesis” (Sterne *et al.* 2014). The tiering approach outlined by OHAT favors inclusion of studies unless they are problematic in multiple key aspects of study quality, an approach that offsets concerns about potentially excluding studies based on a single measure, which could seriously limit the evidence base available for an evaluation, given the type of studies available in environmental health.

OHAT uses strategies recommended by Cochrane for synthesizing study findings when risks of bias vary across studies: (1) restrict primary analysis to studies with lower risk of bias and perform a sensitivity analysis to show how conclusions might be affected if studies at high risk of bias were included; (2) present multiple (stratified) analysis; or (3) present all studies and provide a narrative discussion of risk of bias, ideally through a structured approach like GRADE (Higgins and Green 2011). It is also possible that risk of bias assessment and a tiering approach for assessing study quality might occur prior to data extraction as a strategy to potentially reduce the number of studies that require full data extraction, which is costly and time intensive. Studies excluded for this reason would be indicated on the study flow diagram. This strategy has appeal in terms of efficiency of conducting a systematic review, especially for topics with a large literature base, but it would preclude being able to conduct a sensitivity analysis on excluded studies.

Table 6. Example of Approach for Determining Tiers of Study Quality for Individual Observational Studies		Risk of Bias Domains and Ratings										
		Key Criteria			Other RoB Criterion							
		Can we be confident in the exposure characterization?	Can we be confident in the outcome assessment?	Did the study design or analysis account for	Other RoB criteria	Other RoB criteria	Other RoB criteria	Other RoB criteria	Other RoB criteria	Other RoB criteria	Other RoB criteria	
<ul style="list-style-type: none"> Tier 1: A study must be rated as “definitely low” or “probably low” risk of bias for key elements AND have most other applicable items answered “definitely low” or “probably low” risk of bias. <p>Example of key risk of bias elements for observational human studies:</p> <ul style="list-style-type: none"> ○ Can we be confident in the exposure characterization? ○ Can we be confident in the outcome assessment? ○ Does the study design or analysis account for important confounding variables? <ul style="list-style-type: none"> Tier 2: Study meets neither the criteria for 1st or 3rd tiers. Tier 3: A study must be rated as “definitely high” or “probably high” risk of bias for key elements AND have most other applicable items answered “definitely high” or “probably high” risk of bias. 												
Category	Guidance											
1 st tier	- “definitely low” or “probably low” risk of bias for key items AND - “definitely low” or “probably low” risk of bias for most other applicable criteria	+	++	+	-	+	+	+	+	+	-	
2 nd tier												
	example 1	-	+	++	++	--	+	-	+	+	+	+
	study does not meet criteria for “low” or “high”											
	example 2	+	++	+	+	-	-	-	--	+	-	+
	example 3	--	-	--	++	-	+	+	+	+	+	+
3 rd tier	- “definitely high” or “probably high” risk of bias for key items AND - “definitely high” or “probably high” risk of bias for most other applicable criteria											
		--	-	-	--	+	-	-	+	--	+	--
Risk of bias response options for individual items												
	++	Definitely low risk of bias	--	Definitely high risk of bias								
	+	Probably low risk of bias	-	Probably high risk of bias								

Studies are evaluated on all applicable risk of bias questions based on study design. The rating or answer to each risk of bias question is selected on an outcome basis prior to determining the tier from 4 options: definitely low risk of bias (++), probably low risk of bias (+), probably high risk of bias (-), or definitely high risk of bias (--).

Consideration of Funding Source and Disclosure of Conflict of Interest

Financial conflicts of interest (COI) related to funding source may raise the risk of bias in design, analysis, and reporting (Viswanathan *et al.* 2012), but there is debate on whether COI should be considered a risk of bias element (Lundh *et al.* 2012, Viswanathan *et al.* 2012, Bero 2013, Krauth *et al.* 2013). Currently, Cochrane recommends collecting and evaluating COI information but it is not considered a specific item in the Cochrane risk of bias tool or ACROBAT-NRSI (Higgins and Green 2011, Sterne *et al.* 2014) while the Navigation Guide includes COI as a risk of bias element (Johnson *et al.* 2014b, Koustas *et al.* 2014). OHAT's practice is not to exclude studies based on funding source and not to consider financial COI as a specific risk of bias element. However, OHAT collects information about funding source during data extraction and considers it at multiple points in the evaluation. Funding source is recommended as a factor to consider when evaluating risk of bias of individual studies for selective reporting, and then again for evaluating the body of evidence for publication bias (Viswanathan *et al.* 2012). Funding source should be considered as a potential factor to explain apparent inconsistency within a body of evidence. Also, since many journals now require a COI statement regarding funding, it should be recognized that newer studies may appear to be at greater risk than older studies because of changes in journal reporting standards (Viswanathan *et al.* 2012).

Consideration of Timing and Duration of Exposure and Route of Administration

The issue of timing and duration of exposure as well as route of administration in most cases relates to directness or applicability and not risk of bias: "Did the study design address the topic of the evaluation?" However, there may be instances where these factors are best considered as part of risk of bias – for example, if there are differences in the duration of follow-up across study groups or differences in the time point for assessing exposure across study groups.

In other cases, a limited duration of exposure or duration of follow-up may be problematic based on the health outcome being evaluated; for example, a short duration of time between exposure and health outcome assessment would be inappropriate for evaluating the association with a chronic disease. Ideally, factors such as this should be considered in the PECO statement for study eligibility. If not considered in the PECO statement, a case could be made for addressing these issues as part of risk of bias, or else later in the evaluation during assessment of directness or applicability; both approaches have been proposed (Koustas *et al.* 2014, Rooney *et al.* 2014). OHAT will consider these issues as part of directness/applicability unless attempts to harmonize methods with other organizations conducting systematic reviews indicate preference for a different strategy.

Risk of Bias Assessment Process

Subject matter experts (technical advisors) may be retained to review guidance for assessing risk of bias. Guidance on exposure assessment, health outcome assessment, selection, and confounding will change across evaluations; however, other risks of bias items are less likely to need project-specific customization. For observational human studies, the guidance on assessing confounding will be based on feedback from

the experts, assessment of potential impact of confounding variables from other studies, initial review of the literature, causal diagrams (directed acyclic graphs), and/or resources such as the PhenX Toolkit.⁹

Two members of the evaluation design team will independently make risk of bias determinations for each study across all bias domains/question and then compare their results to identify discrepancies and attempt to resolve them. Any remaining discrepancies will be assessed by the project lead and, if needed, other members of the evaluation design team and/or technical advisors. If, upon further discussion, the evaluation team cannot reach agreement on a risk of bias determination for a particular domain, the more conservative judgment will be selected (e.g. if one reviewer makes a judgment of ‘yes’ and the other makes a judgment of ‘probably yes’, the ‘probably yes’ judgment will be used). In addition, draft OHAT monographs are posted for public comment prior to peer review, which provides an opportunity for investigators to comment on risk of bias assessment of included studies.

Missing Information for Risk of Bias Assessment

OHAT will attempt to contact authors of included studies to obtain missing information considered important for evaluating risk of bias. The product of the evaluation (e.g., monograph, report, or publication) will note that an attempt to contact study authors was unsuccessful if study researchers do not respond to an email or phone request within one month of the attempt to contact. If additional data or information are acquired from study authors, risk of bias judgments will be modified to reflect the updated study information.

Exposure Assessment

Evaluation of exposure assessment is included in OHAT’s RoB tool and includes consideration of methodological quality, sensitivity and validation of the methods used, and degree of variation in participants (described below). We recognize that the factors we consider when assessing the quality of exposure may not necessarily be systematic sources of bias as the concept is described by Cochrane, and we consider this a topic for future method/terminology refinement (see discussion on bias and quality above and “Handbook Peer Review and Updates”).

Experimental studies (and studies assessing internal dosimetry):

- Purity of test compound – Ideally, the purity of the test material is stated and confirmed (and not considered unacceptably low, or unless studying an environmental mixture or commercial compound on purpose) (NTP 2013a).
- Stability and homogeneity of stock material and formulation – Ideally, these factors have also been verified and fall within acceptable ranges. Studies should also provide information about

⁹The PhenX Toolkit (www.phenx.org) is a publicly available free resource that identifies scientifically accepted and standard measures related to assessment of complex diseases, phenotypic traits, and environmental exposures. PhenX measures are selected by working groups of domain experts using a consensus process that includes input from the scientific community. Use of PhenX measures facilitates combining data from a variety of studies and makes it easier for researchers to expand a study design beyond the primary research focus.

consumption through measurement of the dosing medium and dose intake quantity, e.g., feed or water consumption (NTP 2013a).

Observational studies (and studies assessing internal dosimetry):

- Specificity of the biomarker of exposure – Is the biomarker derived from one parent chemical or multiple parent chemicals? (LaKind *et al.* 2014)
- Method sensitivity (detection/quantification limits) – Limits of detection and quantification are low enough to detect chemicals in a sufficient percentage of the samples to address the research question (NTP 2013a, LaKind *et al.* 2014).
- Methods requirements – Minimal concern when instrumentation provides unambiguous or a high degree of confidence to identify and quantitate the biomarker at the required sensitivity (NTP 2013a, LaKind *et al.* 2014).
- Exposure variability and misclassification – Includes factors such as consideration of adequacy of a single measurement to capture long-term exposures. Non-persistent chemicals may have a high degree of individual variability when samples are collected at different time points (LaKind *et al.* 2014).
- Considerations of whether there is sufficient variation in exposure levels across groups to potentially identify associations with health outcomes.
- Adequacy of indirect measures (like drinking water or air levels), self-reported measures, questionnaires, or job exposure matrices to characterize exposure
- Availability of information to determine whether peak or average exposure levels are most important for the health outcome of interest
- Biomarker stability after collection – Ideally, samples have a known history and documentation of stability, and no loss is identified. If losses occurred, concerns for exposure assessment may not be severe if differences between low and high exposure can be qualitatively expressed (LaKind *et al.* 2014).
- Sample contamination – Samples are contamination-free from time of collection to time of measurement (e.g., by use of certified analyte-free collection supplies and reference materials, and by appropriate use of blanks both in the field and lab), and research includes documentation of the steps taken to provide the necessary assurance that the study data are reliable (NTP 2013a, LaKind *et al.* 2014).
- Matrix adjustment – Ideally, results are provided for both adjusted and non-adjusted concentrations (when adjustment is needed). There may be more concern for quality of the exposure assessment if recommended adjustments were not conducted and/or there is no established method for adjustment (LaKind *et al.* 2014).

Risk of bias assessors will be trained using project-specific instructions in an initial pilot-testing phase that is undertaken on a small subset of the included studies (see Appendix 2 for example quick project specific reference guide). This pilot testing should be performed with all team members who will be involved in the RoB assessment so that everyone assesses the same set of studies. Assessors should note potential ambiguities that do not clearly distinguish the criteria for assigning the RoB rating for any question. These ambiguities and any rating conflicts across the team should be examined for opportunities to update and

improve the clarity of the protocol guidance for any of the RoB questions or study types. Revisions to the guidance will reduce future conflicts and improve consistency among assessors. It is also expected that information about confounding and other important issues may be identified during or after data extraction, which can lead to further refinement of the RoB instructions (Sterne *et al.* 2014). Major changes to the RoB guidance (e.g., those that result in revision of response) should be documented in the protocol along with a date and an explanation for the modification.

Risk of bias is independently assessed using structured forms by two assessors for each study, and conflicts are resolved by consensus, arbitration by a third member of the review team, and/or consultation with technical advisors, as needed. If a contractor is used for this step, one of the reviewers should be an OHAT staff member. Space is provided in the form for free-text response to justify each answer or provide context. Brief direct quotations from the text of the study should be used when appropriate.

The RoB tool used should include an option to judge the direction of putative bias for each question or domain. For some questions or domains, the bias is most easily thought of as directional towards or away from the null, and for others (in particular confounding, selection bias, and forms of measurement bias such as differential misclassification), the bias is thought of as an increase or decrease in the effect estimate independent of the null. In some cases, it could be difficult to reach a conclusion, as the bias may go in either direction. A clear rationale with scientific support for judging the likely direction of the bias should be provided, and reviewers should not attempt to guess it (Sterne *et al.* 2014).

STEP 5: SYNTHESIZE EVIDENCE AND RATE CONFIDENCE IN BODY OF EVIDENCE

Considering and Conducting a Meta-Analysis

Heterogeneity within the available evidence will determine the type of evidence synthesis that is appropriate. We anticipate that in many cases, the types of environmental health studies will have disparate exposure and outcome assessments that will not lend themselves to formal statistical meta-analysis. In those cases a narrative synthesis of the available evidence is the most appropriate approach.

OHAT's process for considering whether and how to conduct a meta-analysis is very similar to Navigation Guide methodology (Johnson *et al.* 2013, Koustas *et al.* 2013, Johnson *et al.* 2014a). Summaries of main characteristics for each included study will be compiled and reviewed by two reviewers in pairs to determine comparability between studies, identify data transformations necessary to ensure comparability, and determine whether biological heterogeneity is a concern. The main characteristics evaluated across all eligible studies include the following:

Human Studies

- Study design (e.g., cross-sectional, cohort)
- Details on how participants were classified into exposure groups, if any (e.g., quartiles of exposure concentration)
- Details on source of exposure data (e.g., questionnaire, area monitoring, biomonitoring)
- Concentrations of the chemical(s) for each exposure group
- Health outcome(s) reported
- Conditioning variables in the analysis (e.g., variables considered confounders)

- Type of data (e.g., continuous or dichotomous), statistics presented in paper, ability to access raw data
- Variation in degree of risk of bias at individual study level

Animal Studies

- Experimental design (randomized or not, acute or chronic, multigenerational, etc.)
- Animal model used (species, strain, sex, and genetic background)
- Age of animals (at start of treatment, mating, and/or pregnancy status)
- Developmental stage of animals at treatment and outcome assessment
- Dose levels, frequency of treatment, timing, duration, and exposure route
- Health outcome(s) reported
- Type of data (e.g., continuous or dichotomous), statistics presented in paper, ability to access raw data
- Variation in degree of risk of bias at individual study level

We expect to require input from topic-specific experts to help assess whether studies are too heterogeneous for meta-analysis to be appropriate. Situations where it may not be appropriate to include a study are (1) data on exposure or outcome are too different to be combined, (2) there are concerns about high risk of bias, or (3) other circumstances may indicate that averaging study results would not produce meaningful results. Considerations for conducting a meta-analysis on animal data may differ from those for human data (Vesterinen *et al.* 2014). For example, a greater degree of heterogeneity across studies may be expected (species, strain, route of administration) and effects may be more correlated (or dependent) as compared to human studies from the use of shared control or treatment groups (multi-armed studies), multiple comparisons from one study, group housing, source of animals, etc. When it is inappropriate or not feasible to quantitatively combine results, OHAT will narratively describe or visually present findings.

To assess the impact of existing-study heterogeneity on the meta-analysis, the I^2 statistic will be calculated. The I^2 index is not dependent on the number of studies and can be used to quantify the amount of heterogeneity and provide a measure of the degree of inconsistency in the studies' results ($I^2 = [(Q-df)/Q] \times 100\%$). The I^2 statistic will be evaluated by considering the magnitude/direction of the effect, the extent of evidence of heterogeneity, and Cochrane's guide to interpretation as follows:

- 0% to 40%: might not be important
- 30% to 60%: may represent moderate heterogeneity
- 50% to 90%: may represent substantial heterogeneity
- 75% to 100%: considerable heterogeneity

If determining whether a meta-analysis can be conducted, we will consult with a statistician to identify appropriate statistical methods for analyzing the data and to determine whether further modifications of effect size are required prior to performing a meta-analysis. In general, random-effect models are used to account for potential heterogeneity across studies. Consultation with a statistician will guide identification of the statistical approach that is most appropriate for the study data available. Statistical analyses may be conducted using Comprehensive Meta-Analysis, SAS, or R statistical package.

If the type or source of exposure data differs among studies (e.g., biomonitoring data, estimates from dietary intake or dust concentrations), the data will be normalized when possible to the same metric of concentration or intake. Similarly, we will transform the data on health outcomes, when possible, to convert to common metrics. For example, OHAT may attempt to convert binary outcomes to odds ratio (OR) or relative risk (RR) as the effect measure. For continuous outcomes, effects measures such as absolute mean difference, standardized mean difference, or normalized mean difference (e.g., percent control response) can be used (Vesterinen *et al.* 2014). The scale of the available data is primarily used to determine the choice of effect measure (Fu *et al.* 2011, Vesterinen *et al.* 2014). Absolute mean differences can be used if findings are reported with the same or similar scale, and standardized mean difference (SMD) is typically used when the outcome is measured using different scales. Percent control response can be helpful to assess dissimilar but related outcomes measured with different scales, e.g., fat mass and percent fat mass. If there is a mixture of outcome measurements such that some data are expressed as an empirical or percent change in outcome measurement while other data are expressed as a prevalence of the outcome, then the possibility of including both types of data into one analysis will be explored. The results from subgroup, combined, and any sensitivity analyses will be compared.

Sensitivity Analysis and Meta-Regression

Sensitivity analyses will be performed by examining the effects of including excluded studies with particularly heterogeneous results as well as by performing subgroup analyses based on excluding subsets of studies with shared characteristics that prompted exclusion that might be influential.

If possible, i.e., if there are enough studies; we will assess potential publication bias by developing funnels and performing Egger regression on the estimates of effect size. In addition, if these methods suggest that publication bias is present, we will use trim and fill methods to predict the impact of the hypothetical “missing” studies (Vesterinen *et al.* 2014).

If there is significant study-level heterogeneity, then OHAT may conduct stratified analyses or multivariate meta-regression in an attempt to determine how much heterogeneity can be explained by taking into account both within- and between-study variance (Vesterinen *et al.* 2014). Multivariate meta-regression approaches are especially useful for assessing the significance of associations between study design characteristics. These approaches are considered most suitable if there are at least six to ten studies for a continuous variable and at least four studies for a categorical variable (Fu *et al.* 2011).

Confidence Rating: Assessment of Body of Evidence

The confidence rating for a given health outcome is developed by considering the strengths and weaknesses in a collection of human and animal studies that constitute the body of evidence. The confidence rating reflects confidence that the study findings accurately reflect the true association between exposure to a substance and an effect. The confidence rating approach described below ((Rooney *et al.* 2014); [Figure 6](#)) is based primarily on guidance from the Grading of Recommendations Assessment, Development and Evaluation (GRADE) Working Group (Balshem *et al.* 2011, Guyatt *et al.* 2011a). The GRADE framework is applied most often to evaluate the quality of evidence and the strength of recommendations for health care interventions based on human studies (typically randomized clinical trials). The appeal of the GRADE framework is that (1) it is widely used (Guyatt *et al.* 2011f), (2) it is conceptually similar to the approach used by the Agency for Healthcare Research and Quality for grading the strength of a body of evidence of human studies (AHRQ 2012a), (3) the Cochrane Collaboration has

adopted the principles of the GRADE system for evaluating the quality of evidence for outcomes reported in systematic reviews (Higgins and Green 2011), and (4) the GRADE Working Group is committed to method development/validation and has recently established subgroups to focus on application of GRADE to environmental health and animal studies. Embedded within the GRADE approach is consideration of principles that are consistent with causation as discussed by Sir Austin Bradford Hill (Hill 1965, Schünemann *et al.* 2011). Aspects of this handbook that address Hill considerations for causality are discussed further in Step 6.

None of the previous systematic review frameworks (GRADE, AHRQ, and the Cochrane Collaboration) address approaches for considering animal studies, *ex vivo*, or *in vitro* studies—defined here as other than whole-animal studies and including cell systems, computational toxicology, and *in silico* methods. In addition, the guidance provided by GRADE, AHRQ, and the Cochrane Collaboration is less developed for observational human studies compared to randomized clinical trials. For these reasons OHAT uses a framework that includes a number of refinements to GRADE that were necessary to accommodate the need to integrate data from multiple evidence streams (human, animal, *in vitro*) and focus on observational human studies rather than the randomized clinical trials. This is important because ethical considerations virtually preclude use of human controlled intervention studies to test the hazards of substances in order to address environmental health questions. Controlled exposure studies sometimes appear in the environmental health literature, e.g., air pollution studies in asthmatics (Vagaggini *et al.* 2010), although they are not designed to assess the potential for serious, irreversible, or long-term health effects. Occasionally “natural experimental” studies may occur where individuals are exposed to substances or where exposures is interrupted by nature or factors outside of the control of the investigator, e.g., effects of ionizing radiation in people living near Hiroshima (Shore 2014), impact of smoking bans on health (Sargent *et al.* 2004); reduction in air pollution during the Atlanta and Beijing Olympic games. However, these studies may lack adequate exposure information. Typically the human studies available for environmental health assessments are observational studies of cross-sectional, case-control, cohort, or case reports/series design. Thus, the most widely available data for addressing environmental health questions are human observational epidemiology and experimental animal studies and these data need consideration with clear appreciation for their inherent strengths and limitations (Oxman *et al.* 2006, Silbergeld and Scherer 2013).

The Navigation Guide also applies a modified version of GRADE to environmental health topics (Johnson *et al.* 2014b, Koustas *et al.* 2014). However, the experience with GRADE in the environmental health context is as yet limited and empirical evaluations of using GRADE in this context are also limited. Future collaborations with the GRADE Working Group, the Navigation Guide, and others will aim to evaluate the use of GRADE for addressing environmental health topics. Thus, methodological changes may occur, and OHAT will preferentially utilize a framework that facilitates harmonization with other organizations that conduct systematic reviews in environmental health.

To date, the framework described below has only been applied to human and animal studies and should be applicable to other evidence streams such as mechanistic data, which include outcomes from *in vitro*, mechanistic, cellular, or genomic studies. As a future research effort, OHAT is developing a framework for mechanistic data that is conceptually similar to the approach for human and animal studies.

Four descriptors are used to indicate the level of confidence in the body of evidence for human and animal studies:

- **High Confidence (++++)** in the association between exposure to the substance and the outcome. The true effect is highly likely to be reflected in the apparent relationship.
- **Moderate Confidence (+++)** in the association between exposure to the substance and the outcome. The true effect may be reflected in the apparent relationship.
- **Low Confidence (++)** in the association between exposure to the substance and the outcome. The true effect may be different from the apparent relationship.
- **Very Low Confidence (+)** in the association between exposure to the substance and the outcome. The true effect is highly likely to be different from the apparent relationship.

In the context of identifying research needs, a conclusion of “High Confidence” indicates that further research is very unlikely to change confidence in the apparent relationship between exposure to the substance and the outcome. Conversely, a conclusion of “Very Low Confidence” suggests that further research is very likely to have an impact on confidence in the apparent relationship. It is possible that a single well-conducted study may provide sufficient evidence of toxicity or health effect. This is consistent with the US EPA’s minimum evidence necessary to determine if a potential hazard exists: data demonstrating an adverse reproductive (or developmental) effect in a single appropriate, well-executed study in a single test species (EPA (US Environmental Protection Agency) 1991, 1996).

Available studies on a particular outcome are initially grouped by key study design features, and each grouping of studies is given an initial confidence rating by those features (Figure 6). This initial rating (column 1) is downgraded for factors that decrease confidence in the results (risk of bias, unexplained inconsistency, indirectness or lack of applicability, imprecision, and publication bias) and upgraded for factors that increase confidence in the results (large magnitude of effect, dose response, consistency across study designs/populations/animal models or species, and consideration of residual confounding or other factors that increase our confidence in the association or effect). The reasons for downgrading (or upgrading) confidence may not be due to a single domain of the body of evidence. If a decision to downgrade is borderline for two domains, the body of evidence is downgraded once in a single domain to account for both partial concerns based on considering the key drivers of the strengths or weaknesses. Similarly, the body of evidence is not downgraded twice for what is essentially the same limitation (or upgraded twice for the same asset) that could be considered applicable to more than one domain of the body of evidence. Consideration of consistency across study designs, human populations, or animal species is not included in the GRADE guidance (Guyatt *et al.* 2011a); however, it is considered in the modified version of GRADE used by OHAT (Rooney *et al.* 2014).

Confidence ratings are independently assessed by federal staff on the evaluation review team, and discrepancies are resolved by consensus and consultation with technical advisors as needed. Confidence ratings are summarized in evidence profile tables (see Table 7 for format and Appendix 3 for an example).

The confidence ratings are then used to develop conclusions related to (1) evidence of health effect and research needs in a state of the science evaluation or (2) evidence of health effect, research needs, and hazard identification category for a level of concern evaluation.

Figure 6. Assessing Confidence in the Body of Evidence

Initial Confidence by Key Features of Study Design	Factors Decreasing Confidence	Factors Increasing Confidence	Confidence in the Body of Evidence
High (++++) 4 Features	<ul style="list-style-type: none"> • Risk of Bias • Unexplained Inconsistency • Indirectness • Imprecision • Publication Bias 	<ul style="list-style-type: none"> • Large Magnitude of Effect • Dose Response • Residual Confounding <ul style="list-style-type: none"> – Studies report an effect and residual confounding is toward null – Studies report no effect and residual confounding is away from null • Consistency <ul style="list-style-type: none"> – Across animal models or species – Across dissimilar populations – Across study design types • Other <ul style="list-style-type: none"> – e.g., particularly rare outcomes 	High (++++)
Moderate (+++) 3 Features			Moderate (+++)
Low (++) 2 Features			Low (++)
Very Low (+) ≤1 Features			Very Low (+)

- Features**
- Controlled exposure
 - Exposure prior to outcome
 - Individual outcome data
 - Comparison group used

Note: if the only available body of evidence for an outcome receives a “Very Low” confidence rating, then the conclusion for that outcome will not move forward to Step 6. From Figure 1 in Rooney et al. (2014).

Table 7. Evidence Profile Table Format

Body of Evidence	Risk of Bias	Unexplained Inconsistency	Indirectness	Imprecision	Publication Bias	Magnitude	Dose Response	Residual Confounding	Consistency Across Species/ Model	Final Rating
Example of the type of information that should be in an evidence profile										
Evidence stream (human or animal)	Serious or not serious	Serious or not serious	Serious or not serious	Serious or not serious	Detected or undetected	Large or not large	Yes or no	Yes or no	Yes or no	Final Rating
(# Studies) Initial Rating	<ul style="list-style-type: none"> Describe trend Describe key questions Describe issues 	<ul style="list-style-type: none"> Describe results in terms of consistency Explain apparent inconsistency (if it can be explained) 	<ul style="list-style-type: none"> Discuss use of upstream indicators or populations with less relevance 	<ul style="list-style-type: none"> Discuss ability to distinguish treatment from control Describe confidence intervals 	<ul style="list-style-type: none"> Discuss factors that might indicate publication bias (e.g., funding, lag) 	<ul style="list-style-type: none"> Describe magnitude of response 	<ul style="list-style-type: none"> Outline evidence for or against dose response 	<ul style="list-style-type: none"> Address whether there is evidence that confounding would bias toward null 	<ul style="list-style-type: none"> Describe cross-species, model, or population consistency 	High, Moderate, or Low

Initial Confidence Based on Study Design

An initial confidence rating for the body of evidence for a specific outcome is determined by the ability of the study design to ensure that exposure preceded and was associated with the outcome (Figure 6, column 1). This ability is reflected in the presence or absence of four key study design features used to delineate the studies for initial confidence ratings: (1) the exposure to the substance is experimentally controlled, (2) the exposure assessment demonstrates that exposures occurred prior to the development of the outcome (or concurrent with aggravation/amplification of an existing condition), (3) the outcome is assessed on the individual level (i.e., not through population aggregate data), and (4) an appropriate comparison group is included in the study. The first key feature, “controlled exposure,” reflects the ability of experimental studies in humans and animals to largely eliminate confounding by randomizing allocation of exposure. Therefore, these studies usually have all four features and receive an initial rating of “High Confidence.” Observational studies do not have controlled exposure and are differentiated by the presence or absence of the three remaining study design features. For example, prospective cohort studies usually have all three remaining features and receive an initial rating of “Moderate Confidence” (Table 8).

Study Design	Controlled Exposure	Exposure Prior to Outcome	Individual Outcome Data	Comparison Group Used	Initial Confidence Rating
Human controlled trial ^a	likely	likely	likely	likely	high
Experimental animal	likely	likely	likely	likely	high
Cohort	unlikely	may or may not	likely	likely	low to moderate
Case-control	unlikely	may or may not	likely	likely	low to moderate
Cross-sectional ^b	unlikely	unlikely	likely	likely	low
Ecologic ^b	unlikely	may or may not	may or may not	likely	very low to moderate
Case series/report	unlikely	may or may not	likely	unlikely	very low to low

^aHuman controlled trial study design as used here refers to studies in humans with a controlled exposure, including randomized controlled trials and non-randomized experimental studies.

^bCross-sectional study design as used here refers to population surveys with individual data (e.g., NHANES), as distinct from population surveys with aggregate data on participants (i.e., ecologic studies).

These study design features are distinct from the risk of bias assessment, as they consider only the presence or absence of a factor (e.g., was a comparison group used?) and not its relative quality captured in risk of bias (e.g., was the comparison group appropriate?). Observational animal studies (“wildlife studies”) are considered using the same study design features. The initial ratings are the starting points based on the four study design features, and then studies are evaluated for factors that would downgrade or upgrade confidence in the evidence for a given outcome.

Domains That Can Reduce Confidence

On an outcome-by-outcome basis, five properties for a body of evidence (risk of bias across studies, unexplained inconsistency, indirectness, imprecision, and publication bias) are used to determine if the initial confidence rating based upon the four study design features should be downgraded (Figure 6, column 2).

Risk of Bias Across Studies

In this step, risk of bias for a given outcome is considered across studies.

Summary of Risk of Bias Ratings for Each Outcome

A visual summary of the risk of bias ratings for each outcome is prepared for the outcome of interest by evidence stream, e.g., one for human studies and one for animal studies (see [Table 9](#) for an example of a summary of risk of bias for a set of animal studies). This summary provides an overview of the general strengths and weaknesses for all studies included in the analysis. In addition, it highlights particular risk of bias items that could be explored when evaluating inconsistency within the evidence base.

This analysis can also be useful when considering risk of bias in the context of direction of bias and magnitude of effect. For example, if most human studies are high risk of bias due to non-differential misclassification of exposure, it will generally bias results towards the null; however, differential misclassification can bias towards or away from the null, and consideration of the source, direction, and magnitude of potential biases in the body of evidence is required to interpret findings (Szklo and Nieto 2007).

Table 9. Example of a Visual Summary of Risk of Bias Ratings for Animal Studies

Risk of Bias Question	Study 1	Study 2	Study 3	Study 4	Study 5	Study 6	Study 7	Study 8	Study 9	Study 10	Study 11	Study 12	Study 13	Study 14	Study 15	Study 16	Study 17	Study 18	Study 19
Randomization	+	-	++	++	-	++	+	+	++	-	-	-	+	+	+	-	-	+	++
Allocation concealment	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-
Confounding (design/analysis)	++	+	++	++	++	+	++	++	++	++	+	++	++	+	-	-	-	-	++
Unintended exposure	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Identical experimental conditions	++	++	+	+	++	++	++	++	++	+	++	+	++	++	++	++	++	++	++
Adhere to protocol	+	+	+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Blinding of researchers during study	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-
Missing outcome data	-	+	++	++	--	-	+	-	-	+	--	-	-	+	++	+	++	+	++
Assessment of confounding variables	+	+	++	++	++	-	+	+	++	++	+	+	+	++	++	-	+	+	++
Exposure characterization	++	-	+	+	-	-	+	+	-	-	-	+	+	+	+	+	+	-	+
Outcome assessment	+	+	+	+	+	+	++	+	+	-	++	+	+	+	+	+	+	+	+
Blinding of outcome assessors	+	+	+	+	++	+	+	+	+	+	+	+	--	+	++	+	+	+	+
Outcome reporting	+	+	+	++	--	+	+	+	+	-	+	+	--	+	+	+	++	-	+

Key:

Definitely low risk of bias

++

Probably low risk of bias

+

Probably high risk of bias

-

Definitely high risk of bias

--

Studies are evaluated on all applicable risk of bias questions based on study design. The rating or answer to each risk of bias question is selected on an outcome basis prior to determining the tier from 4 options: definitely low risk of bias (++), probably low risk of bias (+), probably high risk of bias (-), or definitely high risk of bias (--).

Consideration of Whether to Downgrade Confidence Based on Risk of Bias

The strategy for assessing risk of bias differs depending on whether confidence ratings will be primarily used to identify research needs for a state-of-science evaluation or to reach formal NTP conclusions on hazard identification. Downgrading for risk of bias should reflect the entire body of studies; therefore, the decision to downgrade should be applied conservatively. The decision to downgrade should be reserved for cases for which there is substantial risk of bias across most of the studies composing the body of evidence.

Confidence Ratings to Identify Research Needs

All studies providing data on a given health outcome, regardless of the risk of bias tier for each individual study, are considered when developing confidence ratings. OHAT will use the approach described earlier in Step 4 for categorizing individual studies as “Tier 1,” “Tier 2,” or “Tier 3” risk of bias together with the guidance presented in Table 10 when considering the extent to which confidence should be downgraded based on risk of bias across studies.

Table 10. Guidance on When to Downgrade for Risk of Bias Across Studies		
Downgrade	Interpretation	Guidance
“Not likely”	Plausible bias unlikely to seriously alter the results	Most information is from Tier 1 studies (low risk of bias for all key domains).
“Serious”	Plausible bias that raises some doubt about the results	Most information is from Tier 1 and 2 studies.
“Very serious”	Plausible bias that seriously weakens confidence in the results	The proportion of information from Tier 3 studies at high risk of bias for all key domains is sufficient to affect the interpretation of results.

If Tier 3 risk of bias studies are omitted from the confidence-rating phase, OHAT may conduct analyses to assess the extent to which inclusion of the Tier 3 risk of bias studies altered conclusions, e.g., by comparing consistency of findings from studies in the Tier 3 risk of bias with findings from studies in the Tiers 1 and 2 risk of bias.

Unexplained Inconsistency

Inconsistency, or large variability in the direction or magnitude of individual study effect estimates for comparable measures of association that cannot be explained, reduces confidence in the body of evidence (Guyatt *et al.* 2011d, AHRQ 2012a). Reasons for variation in such measures may relate to study design, model misspecification and to factors such as differences between studies in lengths of follow-up or age structures. Inconsistency that can be explained, such as variability in study populations, would not be eligible for a downgrade. Potential sources of inconsistency across studies are explored, including consideration of population or animal model (e.g., cohort, species, strain, sex, lifestage at exposure and assessment); exposure or treatment duration, level, or timing relative to outcome; study methodology (e.g., route of administration, methodology used to measure health outcome); conflict of interest, and statistical power and risk of bias. Generally, there is no downgrade when identified sources of inconsistency can be attributed to study design features such as differences in species, timing of exposure, or health outcome assessment. There is no downgrade for inconsistency in cases where the evidence base

consists of a single study. In this case, consistency is unknown and is documented as such in the summary of findings table.

Risk of bias of individual studies in the body of evidence will also be considered when there is inconsistency of findings across studies. If differences in risk of bias explain the heterogeneity of findings, then OHAT will reconsider the decision on whether or not to downgrade for risk of bias in developing the confidence rating.

The statistical power of studies will also be considered if OHAT detects an inconsistency of findings across studies. OHAT may omit underpowered studies from consideration when determining confidence ratings, especially in cases where a meta-analysis is not feasible for pooling results across studies. If underpowered studies are omitted from the confidence-rating phase, OHAT may conduct analyses to assess the extent to which inclusion of these studies would alter conclusions, e.g., by comparing consistency of findings. **Note:** Consideration of the statistical power of studies remaining in the confidence ratings is formally part of the evaluation of imprecision (see below).

No single measure of consistency is ideal, and the following factors are considered when determining whether to downgrade for inconsistency: (1) similarity of point estimates, (2) extent of overlap between confidence intervals, and (3) results of statistical tests of heterogeneity, e.g., Cochran's Q (chi-square, χ^2), I^2 , or τ^2 (tau square). Tests for statistical heterogeneity are less reliable when there are only a few studies. See [Table 11](#) for examples and additional details on guidance.

Cochran's Q: A statistical test for heterogeneity distributed as a chi-square (χ^2) statistic, which tests the null hypothesis that all studies have the same underlying magnitude of effect; a low p-value ($p < 0.1$) indicates significant heterogeneity (Higgins and Green 2011). The level of significance for χ^2 is often set at 0.1 because of the low power of the test to detect heterogeneity. A rule of thumb is if χ^2 is larger than the degrees of freedom (df, number of studies minus 1), then heterogeneity is present. The χ^2 statistic has low power to detect heterogeneity when there are few studies, or, conversely, it may detect heterogeneity of minimal biological or clinical importance when the number of studies is large.

I^2 : Preferred index that is not dependent on the number of studies and can be used to quantify the amount of heterogeneity and provide a measure of the degree of inconsistency in the studies' results ($I^2 = [(Q - df)/Q] \times 100\%$). I^2 represents the percentage of the total variation across studies due to heterogeneity rather than sampling error or chance, with values ranging from 0% (no observed heterogeneity) to 100%.

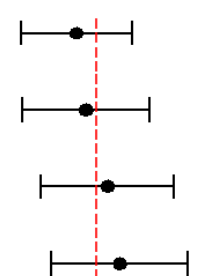
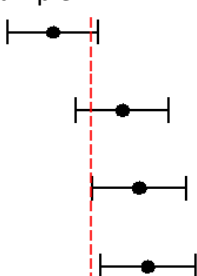
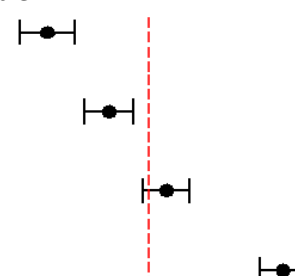
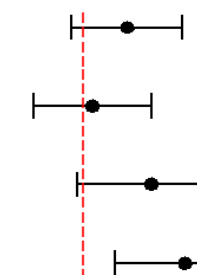
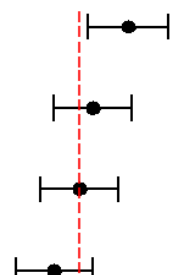
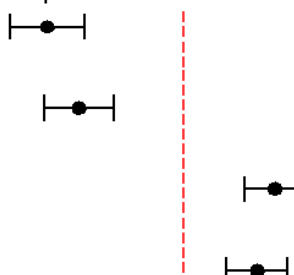
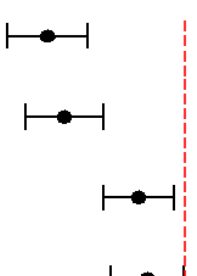
Thresholds for the interpretation of I^2 can be misleading, since the importance of the observed value of I^2 depends on (1) the magnitude and direction of effects and (2) the strength of evidence for heterogeneity (e.g., p-value from the chi-square test, or a confidence interval for I^2). A rough guide for interpretation of I^2 is as follows (Higgins and Green 2011):

- 0% to 40%: might not be important
- 30% to 60%: may represent moderate heterogeneity
- 50% to 90%: may represent substantial heterogeneity
- 75% to 100%: considerable heterogeneity

Tau square (T^2 , τ^2 , τ^2): An estimate of the between-study variance in a random-effects meta-analysis. A τ^2 close to 0 would be strict homogeneity, and > 1 suggests the presence of substantial statistical heterogeneity.

Table 11. Factors to Consider in Addressing Consistency of Results When Variation Cannot Be Explained by Methodological Factors

----- = null hypothesis

“Not serious”	“Serious”	“Very serious”
<ul style="list-style-type: none"> • Point estimates similar • Confidence intervals overlap • Statistical heterogeneity is non-significant ($p \geq 0.1$) • I^2 of $\leq 50\%$ 	<ul style="list-style-type: none"> • Point estimates vary • Confidence intervals show minimal overlap • Statistical heterogeneity has low p-value ($p \leq 0.1$) • I^2 of $> 50\%$ to 75% 	<ul style="list-style-type: none"> • Point estimates vary widely • Confidence intervals show minimal or no overlap • Statistical heterogeneity has low p-value ($p \leq 0.1$) • I^2 of $> 75\%$
<p>Example A</p>  <p>χ^2 p-level = 0.767; $I^2 = \ll 1\%$; $\tau^2 = \ll 1$</p>	<p>Example A</p>  <p>χ^2 p-level = 0.017; $I^2 = 71\%$; $\tau^2 = 0.044$</p>	<p>Example A</p>  <p>χ^2 p-level = < 0.001; $I^2 = 98\%$; $\tau^2 = 1.022$</p>
<p>Example B</p>  <p>χ^2 p-level = 0.241; $I^2 = 29\%$; $\tau^2 = 0.046$</p>	<p>Example B</p>  <p>χ^2 p-level = 0.068; $I^2 = 58\%$; $\tau^2 = 0.025$</p>	<p>Example B</p>  <p>χ^2 p-level = < 0.001; $I^2 = 98\%$; $\tau^2 = 0.774$</p>
<p>Example C</p>  <p>χ^2 p-level = < 0.001; $I^2 = 86\%$; $\tau^2 = 0.111$ * there is less concern for numerical estimates of heterogeneity because point estimates are in the same direction</p>		

Directness and Applicability

Directness refers to the applicability, external validity, generalizability, and relevance of the studies in the evidence base in addressing the objectives of the evaluation (AHRQ Guyatt *et al.* 2011c, 2012a). Directness addresses the question, “Did the study design address the topic of the evaluation?”

To determine whether to downgrade confidence based on indirectness, OHAT considers factors related to (1) relevance of the animal model to outcome of concern, (2) directness of the endpoints to the primary health outcome(s), (3) nature of the exposure in human studies and route of administration in animal studies, and (4) duration of treatment in animal studies and length of time between exposure and outcome assessment in animal and prospective human studies. The appropriateness of the window of exposure given the health outcome measured is generally considered as part of the evaluation for directness and applicability (i.e., “Are the results of the study credible?” versus “Did the study design address the topic of the evaluation?”). However, there may be cases where time between exposure and health outcome assessment is considered a risk of bias. For example, if there were differences in the duration of follow-up across study groups, this would be a source of bias considered under detection bias. Duration of follow-up is also relevant to the indirectness or applicability of a study if the duration of follow-up was not sufficient for developing the outcome of interest (e.g., a 6-week study of cancer endpoints). In this case, an otherwise well-designed and well-conducted study may suffer from indirectness despite having low risk of bias (Viswanathan *et al.* 2012).

Relevance of the Animal Model to Human Health

- Rats, mice, and other mammalian model systems: Studies conducted in mammalian model systems are assumed relevant for humans (i.e., not downgraded) unless compelling evidence to the contrary is identified during the course of the evaluation. The applicability of specific health outcomes or biological processes in non-human animal models is outlined in the PECO-based inclusion and exclusion criteria, with the most accepted relevant/interpretable outcomes considered “primary” and less direct measures, biomarkers of effect, or upstream measures of health outcome considered “secondary.” OHAT recognizes that interpreting the relevance for humans of specific outcomes or events in non-human animals is often very challenging and lacking in empirical support.
- Genetically modified rodent models; bird, reptile amphibian, fish, and other non-mammalian vertebrate model systems: the validity of these model systems to address human health is not as well established as the use of unmodified mammalian model systems. For this reason, studies conducted in these model systems are generally downgraded for directness unless data suggest otherwise. Evidence that supports phylogenetic similarity and/or the concordance of findings in these model systems with findings from traditional toxicological species should be considered when determining whether or not to downgrade.
- Invertebrate model systems: Validity of these model systems to address many outcomes relevant to human health is not well established. For this reason, studies conducted in non-mammalian vertebrates are generally downgraded for directness. Evidence that supports phylogenetic conservation or mechanism or response similarity and/or the concordance of findings in these model systems with findings from traditional toxicological species should be considered when determining the extent to which to downgrade.

Exposure

- *Human studies:* Human studies are not downgraded for directness regardless of the exposure level or setting (e.g., general population, occupational settings, etc.). In OHAT's process, the applicability of a given exposure scenario for reaching a "level of concern" for a certain subpopulation is considered after hazard identification. For that subpopulation the health effect is interpreted in the context of what is known regarding the extent and nature of human exposure (Twombly 1998, Medlin 2003, Jahnke *et al.* 2005, Shelby 2005).
- *Dose levels used in animal studies:* There is no downgrading for dose level used in experimental animal studies because it is not considered as a factor under directness for the purposes of reaching confidence ratings for evidence of health effects. OHAT recognizes that the level of dose or exposure is an important factor when considering the relevance of study findings. In OHAT's process, consideration of dose occurs after hazard identification as part of reaching a "level of concern" conclusion when the health effect is interpreted in the context of what is known regarding the extent and nature of human exposure (Twombly 1998, Medlin 2003, Jahnke *et al.* 2005, Shelby 2005).
- *Route of administration in animal studies:* External dose comparisons used to reach level of concern conclusions need to consider internal dosimetry in animal models, which can vary based on route of administration, species, age, diet, and other cofactors. The most commonly used routes of administration (i.e., oral, dermal, inhalation, subcutaneous) are generally considered direct for the purposes of establishing confidence ratings. Pharmacokinetic data are also considered. Other routes of administration are more likely to be considered indirect (e.g., intraperitoneal, water for aquatic species, or culture media for culture media for cells, *ex vivo* preparations, or invertebrates).

Duration of Treatment and Window of Time Between Exposure and Outcome Assessment

Studies that assess health outcomes following longer periods of exposure and follow-up are generally anticipated to be more informative than studies of shorter duration, e.g., acute toxicity studies lasting from hours to several days. When possible, studies of too short a duration of exposure or follow-up should be excluded as part of the PECO criteria. However, in many cases, defining "too short" is difficult to support based on empirical data, and duration of exposure/follow-up may need to be considered as part of directness and applicability. Duration of treatment and window of time between exposure and outcome are factors considered when evaluating consistency of results across studies.

Imprecision

Precision is the degree of certainty surrounding an effect estimate with respect to a given outcome (AHRQ 2012a). A precise estimate enables the evaluator to determine whether there is an effect (i.e., it is different from the comparison group). OHAT uses 95% confidence intervals as the primary method to assess imprecision (Guyatt *et al.* 2011b). OHAT also considers whether the studies are adequately powered when assessing precision, an issue that is especially important when interpreting findings that do not provide support for an association. Approaches such as "optimal information size" (OIS) can be used to assess precision for dichotomous and continuous outcomes (Guyatt *et al.* 2011b). This analysis calculates the sample size required for an adequately powered individual study, referred to as the OIS threshold or criterion (OIS calculator available at <http://www.stat.ubc.ca/~rollin/stats/ssize/>). In a meta-analysis, the threshold for precision is met when the total sample size for the meta-estimate is as great as, or greater than, the OIS threshold.

As noted earlier, OHAT may omit statistically underpowered studies from consideration when determining confidence ratings, especially in cases where a meta-analysis is not feasible for pooling results across studies. If underpowered studies are omitted from the confidence-rating phase, OHAT may conduct analyses to assess the extent to which inclusion of these studies would alter conclusions, e.g., by comparing consistency of findings.

When a meta-analysis is inappropriate or not feasible, precision is primarily based on the range of effect size estimates in the evidence base (AHRQ 2012a). Data are generally considered imprecise for ratio measures (e.g., OR) when the ratio of the upper to lower 95% CI for most studies is ≥ 10 , and for absolute measures (e.g., percent control response) when the absolute difference between the upper and lower 95% CI for most studies is ≥ 100 . If a meta-analysis is conducted, the same 95% confidence interval assessment is made based on the meta-estimate of the association. See [Table 12](#) for a tabular summary of the guidance OHAT will use to assess imprecision.

Often it is difficult to distinguish between wide confidence intervals due to inconsistency and those due to imprecision, which leads to the question of whether to downgrade once or twice. In most cases, a single downgrade for one of these domains is sufficient (AHRQ 2012a). Thus, in most cases where the body of evidence is downgraded for inconsistency in the direction of effect, OHAT will not further downgrade for imprecision. However, it is considered appropriate to downgrade twice if studies are both very inconsistent (e.g., [Table 11](#), see “very serious” example B) and imprecise.

Table 12. Factors to Consider When Evaluating Imprecision of Results	
Not serious	<ul style="list-style-type: none"> No or minimal indications of large standard deviations (i.e., $SD > \text{mean}$) For ratio measures (e.g., odds ratio, OR) the ratio of the upper to lower 95% CI for most studies (or meta-estimate) is < 10; or for absolute measures (e.g., percent control response) the absolute difference between the upper and lower 95% CI for most studies (or meta-estimate) is < 100.
Serious	Does not clearly meet guidance for “not serious” or “very serious”
Very serious	<ul style="list-style-type: none"> Large standard deviations (i.e., $SD > \text{mean}$) For ratio measures (e.g., OR) the ratio of the upper to lower 95% CI for most studies (or meta-estimate) is ≥ 10; or for absolute measures (e.g., percent control response) the absolute difference between the upper and lower 95% CI for most studies (or meta-estimate) is ≥ 100.

Publication Bias

OHAT characterizes publication bias as “undetected” (no downgrade) or “strongly suspected” as recommended by GRADE (Guyatt *et al.* 2011e). In general, studies with statistically significant results are more likely to be published than studies without statistically significant results (“negative studies”) (Guyatt *et al.* 2011e). Thus, some degree of publication bias is likely on any topic; however, downgrading is reserved for cases where the concern is serious enough to significantly reduce confidence in the body of evidence. Below are some issues OHAT will consider when determining whether to downgrade for publication bias:

- Early positive studies, particularly if small in size, are suspect. Reviews performed early, when only few initial studies are available, tend to overestimate effects (reviewed in Guyatt *et al.* 2011e)]. There may be publication lag time for “negative” studies, and it may take time for

other authors to replicate the early studies. It may be helpful to compare study findings by publication year to determine if this appears to be an issue. In meta-analyses, statistical approaches can be used to calculate meta-estimates at the end of each year to note changes in the summary effect.

- Publication bias should be suspected when studies are uniformly small, particularly when sponsored by industries, non-government organizations (NGOs), or authors with conflicts of interest (reviewed in Guyatt *et al.* 2011e). When possible, OHAT will evaluate findings by funding source or by whether the author(s) reported a conflict of interest.
- Funnel plots, Egger’s regression, and trim and fill techniques can be used to visualize asymmetrical or symmetrical patterns of study results to help assess publication bias when adequate data for a specific outcome are available. Funnel plots and other approaches are less reliable when there are only a few studies.
- The identification of abstracts or other types of grey literature that do not appear as full-length articles within a reasonable time frame (around 3 to 4 years) can be another indication of publication bias (AHRQ 2012a).

Domains That Can Increase Confidence

Four properties for a body of evidence (large magnitude of effect, dose response, plausible confounding that would have an impact on the observed association, and consistency across study designs and experimental model systems) are used to determine if the initial confidence rating should be upgraded (Figure 6, column 3). Large magnitude of effect, dose response, and residual confounding (or “all plausible confounding”) are considered in the GRADE and AHRQ guidelines (AHRQ Guyatt *et al.* 2011g, 2012a). OHAT has added an additional factor to address consistency across human study designs and animal species or animal model systems.

Large Magnitude of Association or Effect

GRADE has guidance for determining when effects might be considered “large” in human studies based primarily on modeling studies that suggest confounding alone is unlikely to explain associations with a relative risk (RR)¹⁰ greater than 2 (or less than 0.5) and very unlikely to explain associations with an RR greater than 5 (or less than 0.2) (Guyatt *et al.* 2011g). Hence, the GRADE Working Group has previously suggested guidelines for rating quality of evidence up by one category (typically from low to moderate) for associations greater than 2, and up by two categories for associations greater than 5 (Guyatt *et al.* 2011g). The rapidity of the response compared with natural progression of the condition can also be considered when determining whether there is a large magnitude of association or effect. However, there is concern about applying the numerical RR guidance from GRADE in environmental health because relatively “small” effects of the type most often observed (such as increases in blood pressure or decreases in IQ associated with lead) can have major public health impacts on a population basis when considering the tails of the normal distribution, and most of the effect is associated with those tails.

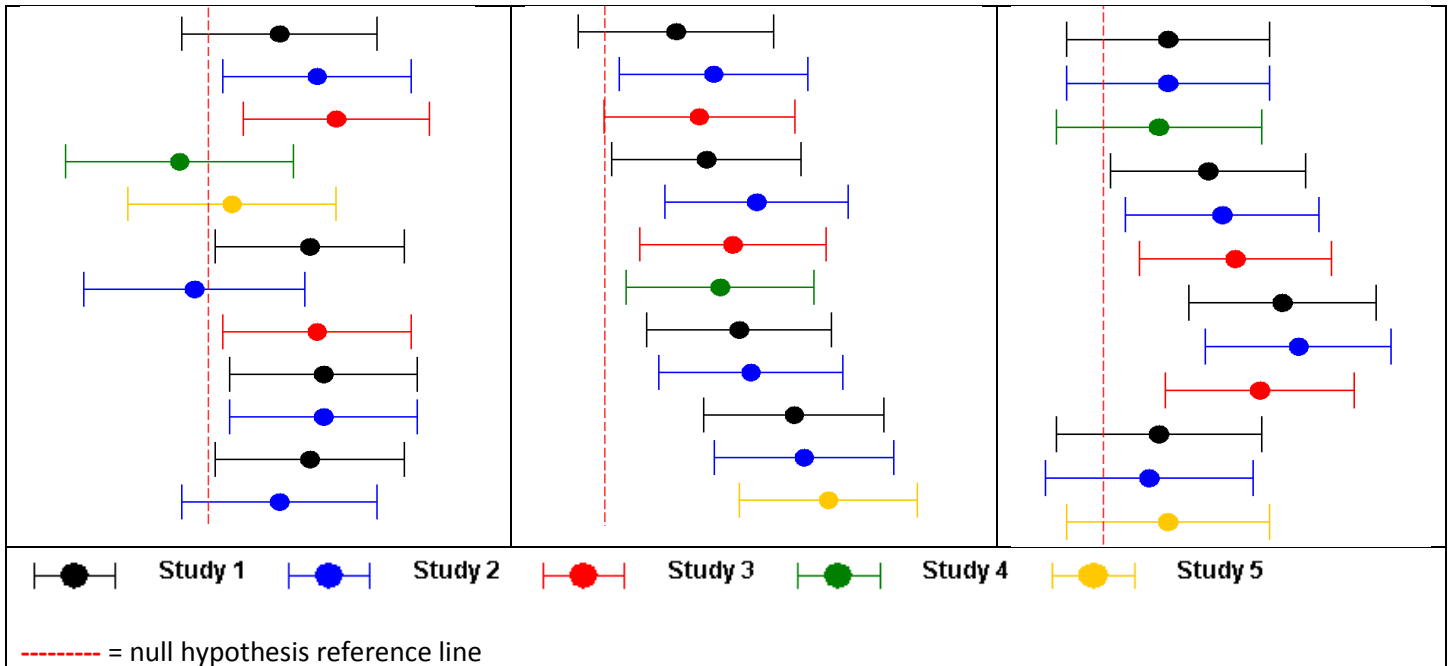
¹⁰When the baseline risk is low (< 20%), the RR and odds ratio (OR) are similar. When the baseline risk is high (> 40%), then the ORs can be much larger in magnitude than RRs, and a higher threshold for ORs to be considered large might be appropriate.

Thus, considerations for identifying a large magnitude of effect, also sometimes referred to as strength of association or strength of response, are made on a project-specific basis based on discussion by the evaluation team and consultation with technical advisors as needed. Determining whether the magnitude of the effect is large includes consideration of the effect being measured and the background prevalence or rate for that effect, the species and dose range utilized in experimental studies, exposure pattern in human studies including peaks, magnitude and duration.

Dose Response

OHAT will upgrade for evidence of a monotonic dose-response gradient (Guyatt *et al.* 2011g) and for evidence of a non-monotonic dose response when data fit the expected pattern, i.e., prior knowledge leads to expectation for non-monotonic dose response, and/or non-monotonic dose response is consistently observed in the evidence base. Patterns of dose response are evaluated within and across studies when considering whether to upgrade (Table 13). Effect size data may be visually sorted (1) by study in order to assess dose response within studies and consistency of dose response across studies of similar dose or exposure levels, and (2) by dose or exposure level to assess dose response across the entire evidence base.

Table 13. Conceptual Examples of Upgrade Decisions for Evidence of Dose-Response Gradient		
No Upgrade	Evidence of Gradient (Monotonic)	Evidence of Gradient (Non-Monotonic)
Example A. Findings sorted by study and then by dose or exposure level (low to high)	Example B. Findings sorted by study and then by dose or exposure level (low to high)	Example C. Findings sorted by study and then by dose or exposure level (low to high)
Example A. Findings across studies sorted by exposure or dose level (low to high)	Example B. Findings across studies sorted by exposure or dose level (low to high)	Example C. Findings across studies sorted by exposure or dose level (low to high)



Residual Confounding or Other Related Factors That Would Increase Confidence in the Estimated Effect

This element primarily applies to observational studies. Residual confounding (also referred to as “all plausible confounding” or “residual biases”) refers to consideration of unmeasured determinants of an outcome unaccounted for in an adjusted analysis that are likely to be distributed unequally across groups (Guyatt *et al.* 2011g). If a study reports an effect or association despite the presence of residual confounding, confidence in the association is increased. Since this confounding can push in either direction, confidence in the results is increased when the body of evidence is potentially biased by factors counter to the observed effect. Upgrading should be considered when there are indications that residual confounding or bias would underestimate an apparent association or treatment effect (i.e., bias towards the null), or suggest a spurious effect when results suggest no effect.

Examples of residual bias towards the null that would strengthen confidence in finding an effect: The “healthy worker” effect is one example that was observed initially in studies of occupational diseases; workers usually exhibit lower overall death rates than the general population because workers may leave employment due to perceived or actual health effects and in many industries severely ill and disabled people are excluded from employment. Another example of residual bias towards the null is outlined in the GRADE guidance (Guyatt *et al.* 2011g) of a systematic review of HIV infection and condom use. The effect estimate from five studies was statistically significant with condom use showing a protective effect compared with no condom use. In two of the studies, the number of sexual partners was also considered (Detels *et al.* 1989, Difranceisco *et al.* 1996). These studies found that condom users were more likely to have more sexual partners, yet the studies did not adjust for number of partners in their final analyses. Had the number of partners been considered in the meta-analysis, it might have strengthened the effect estimate in favor of condom use.

Example of residual bias pushing toward a spurious positive effect that would strengthen confidence in finding no association: An example, also taken from the GRADE guidance (Guyatt *et al.* 2011g), considers two observational studies (Taylor *et al.* 1999, Elliman and Bedford 2001) that failed to confirm a well-publicized association between vaccination and autism, which was widely discredited and eventually retracted (Wakefield *et al.* 1998). After the widespread initial publicity, it was empirically confirmed that parents of autistic children were more likely to remember their vaccine experience than parents of children diagnosed before the publicity (Andrews *et al.* 2002). Parents of non-autistic children were presumed to also be less likely to remember their children's vaccinations. Thus, the negative findings of the observational studies, despite the demonstrated recall bias, increase the confidence that there is no association and could be the basis of an upgrade to the confidence rating.

Cross-Species/Population/Study Consistency

Three types of consistency in the body of evidence can be used to support an increase in confidence in the results:

- across animal studies—consistent results reported in multiple experimental animal models or species
- across dissimilar populations—consistent results reported across populations (human or wildlife) that differ in factors such as time, location, and/or exposure
- across study types—consistent results reported from studies with different design features, e.g., between prospective cohort and case-control human studies or between chronic and multigenerational animal studies

Other

Additional factors specific to the topic being evaluated may be considered in rating confidence in the body of evidence, such as specificity of the association in cases where the effect is rare or unlikely to have multiple causes. For example, the observation of cases of clear cell adenocarcinoma, a rare kind of vaginal and cervical cancer, in a group of women in their teens and early twenties was highly unusual, and subsequent investigation determined that it resulted from *in utero* exposure to diethylstilbestrol (DES) (<http://www.cdc.gov/des/consumers/daughters/index.html>). This particularly rare outcome in an unusual population increases confidence in the association despite being based on small observational human studies. OHAT does not anticipate routinely using the “other” category for upgrading confidence across the body of studies for the majority of evaluations. However, if during the course of an evaluation an important additional factor for upgrading confidence becomes evident, OHAT would consult experts on use of the additional factor, and a change in the categories for rating confidence in the body of evidence would be noted as a revision to the protocol.

Combine Confidence Conclusions for All Study Types and Multiple Outcomes

Conclusions are primarily based on the evidence with the highest confidence when considering evidence across study types and multiple outcomes. Confidence ratings are initially set based on key design features of the available studies for a given outcome (e.g., for experimental studies separately from observational studies). The studies with the highest confidence rating form the basis for the confidence conclusion for each evidence stream. As outlined previously, consistent results across studies with different design

features increase confidence in the combined body of evidence and can result in an upgraded confidence rating moving forward to Step 6.

After confidence conclusions are developed for a given outcome, conclusions for multiple outcomes can be developed. When outcomes are biologically related, they may inform confidence on the overall health outcome, and confidence conclusions can be developed in two steps. Each outcome is first considered separately. Then, the related outcomes are considered together and re-evaluated for properties that relate to downgrading and upgrading the body of evidence. This approach is especially helpful in circumstances where conclusions can be informed by evidence for which there is lower confidence. For instance, a less confident body of evidence may support the higher confidence body of evidence and thereby contribute to the conclusion.

REVISION: Consideration Across Multiple Exposures

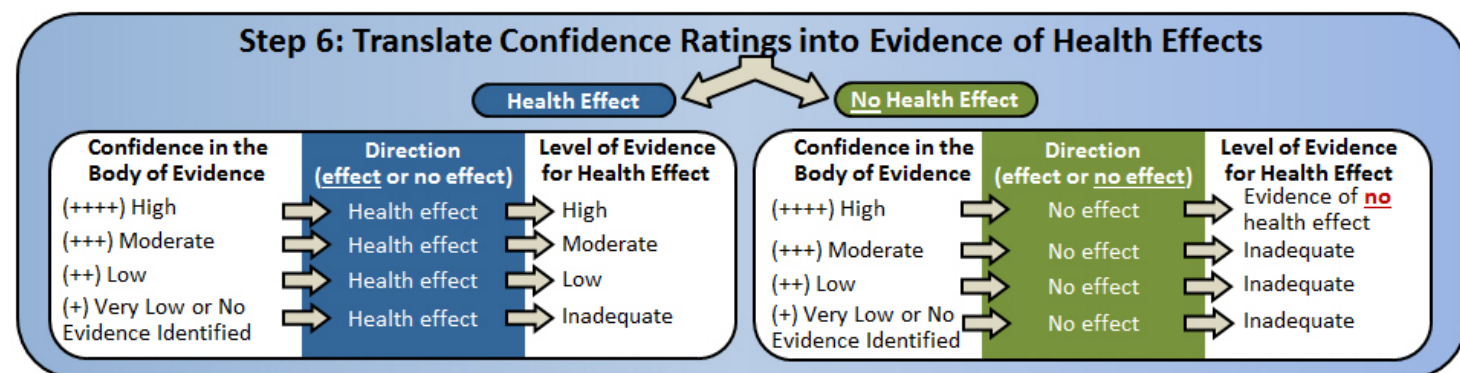
REVISION: When individual chemical or physical agent exposures are components of a broader relevant exposure derived from a common source, collectively they may inform overall confidence in the association of that broader exposure with the health effect. Confidence conclusions can be developed in three steps. Each individual exposure is first considered separately and a confidence rating in the body of evidence is reached. Then mechanistic data or other relevant considerations should be used to determine: 1) if the individual exposures could independently affect the health outcome and 2) if there is evidence of an exposure-dependent relationship between the exposure and the health effect. 3) If both scenarios are true for the exposure, then evidence from the individual exposures is considered together and re-evaluated for properties that relate to downgrading or upgrading confidence in the body of evidence.

STEP 6: TRANSLATE CONFIDENCE RATINGS INTO LEVEL OF EVIDENCE FOR HEALTH EFFECT

The level of evidence in Step 6 of OHAT's framework is assessed separately for human and experimental animal data. A similar approach for mechanistic data is under development.

The conclusions for the level of evidence for health effects reflect the overall confidence in the association between exposure to the substance reached in Step 5 ("high," "moderate," "low," or "very low") and the nature of the effect ("health effect" or "no health effect"). Five descriptors are used to categorize the level of evidence: "high," "moderate," "low," "evidence of no health effect," and "inadequate evidence" (Figure 7). Three descriptors ("high," "moderate," and "low" level of evidence) directly translate from the ratings of confidence in the evidence reached in Step 5 that exposure to the substance is associated with a health effect. If the Step 5 conclusion is "very low" or no evidence is identified, then the Step 6 level-of-evidence conclusion is characterized as "inadequate evidence." The descriptor "evidence of no health effect" is used to indicate confidence that the substance is not associated with a health effect. Because of the inherent difficulty in proving a negative, the conclusion "evidence of no health effect" is only reached when there is high confidence in the body of evidence.

Figure 7. Translate Confidence Ratings into Evidence of Health Effect Conclusions



Evidence Descriptors	Definition
High Level of Evidence	There is high confidence in the body of evidence for an association between exposure to the substance and the health outcome(s).
Moderate Level of Evidence	There is moderate confidence in the body of evidence for an association between exposure to the substance and the health outcome(s).
Low Level of Evidence	There is low confidence in the body of evidence for an association between exposure to the substance and the health outcome(s), or no data are available.
Evidence of No Health Effect	There is high confidence in the body of evidence that exposure to the substance is not associated with the health outcome(s).
Inadequate Evidence	There is insufficient evidence available to assess if the exposure to the substance is associated with the health outcome(s).

Although the conclusions describe associations, a causal relationship is implied. Table 14 outlines how the Hill considerations on causality (Hill 1965) are related to the process for evaluating confidence in the body of evidence and then integrating the evidence (similar to GRADE approach as described in Schünemann *et al.* 2011).

Hill Consideration	Relationship to the OHAT Approach
Strength	Considered in upgrading the confidence rating for the body of evidence for large magnitude of effect and in downgrading the confidence rating for imprecision .
Consistency	Considered in upgrading the confidence rating for the body of evidence for consistency across study types, across dissimilar populations, or across animal species ; and in integrating the body of evidence among human, animal, and other relevant data; also in downgrading the confidence rating for the body of evidence for unexplained inconsistency .
Temporality	Considered in initial confidence ratings by key features of study design; for example, experimental studies have an initial rating of “High Confidence” because of the increased confidence that the controlled exposure preceded outcome.
Biological gradient	Considered in upgrading the confidence rating for the body of evidence for evidence of a dose-response relationship.
Biological plausibility	Considered in examining dose-response relationships and developing confidence-rating conclusions across biologically related outcomes, particularly outcomes along a pathway to disease. Other relevant data that

	inform plausibility, such as physiologically based pharmacokinetic and mechanistic studies, are considered in integrating the body of evidence. Also considered in downgrading the confidence rating for the body of evidence for indirectness .
Experimental evidence	Considered in setting initial confidence ratings by key features of study design and downgrading the confidence rating for risk of bias .

STEP 7: INTEGRATE EVIDENCE TO DEVELOP HAZARD IDENTIFICATION CONCLUSIONS

For determining the appropriate hazard identification category, the evidence streams for human studies and animal studies, which have remained separate through the previous steps, are integrated along with other relevant data, such as supporting evidence from *in vitro* studies.

Integration of Human and Animal Evidence

Hazard identification conclusions are initially reached by integrating the highest level-of-evidence conclusion for a health effect(s) from the human and the animal evidence streams. On an outcome basis, this approach applies to whether the data support a health effect conclusion or provide evidence of no health effect. Hazard identification conclusions may be reached on individual outcomes (health effects) or groups of biologically related outcomes, as appropriate, based on the evaluation’s objectives and the available data. The five hazard identification conclusion categories are as follows:

- Known to be a hazard to humans
- Presumed to be a hazard to humans
- Suspected to be a hazard to humans
- Not classifiable as a hazard to humans
- Not identified as a hazard to humans

When the data support a health effect, the level-of-evidence conclusion for human data from Step 6 is considered together with the level of evidence for non-human animal data to reach one of four hazard identification conclusions (Figure 8). If one evidence stream (either human or animal) is characterized as “Inadequate Evidence,” then conclusions are based on the remaining evidence stream alone (which is equivalent to treating the missing evidence stream as “Low” in Step 7).

If the human data provide a high level of evidence of no health effect from Step 6, then that conclusion is considered together with the level-of-evidence conclusion for non-human animal data. If the human conclusion of no health effect is supported by animal evidence of no health effect, the hazard identification conclusion is “not identified.”

OHAT hazard identification labels are similar to those used in the Globally Harmonized System of Classification and Labelling of Chemicals (GHS)¹¹, although they should not be considered equivalent because of differences in definition and strategies used to integrate data. For example, GHS conclusions for reproductive toxicity are based on an unstructured strength-of-evidence approach, whereas conclusions for specific target-organ toxicity can be based on the administered dose level in an animal study where significant and/or severe effects are observed.

REVISION: Reaching Hazard Conclusions from Human Health Data Alone

REVISION: The evidence integration approach (Figure 8) outlines how all four hazard categories could be reached when there is human evidence and “low or inadequate” confidence in the non-human animal evidence. Characteristics of a body of evidence can differ such that moderate confidence in a body of evidence for human data alone may support a hazard conclusion of suspected in some cases and presumed in other cases. The justification for the final hazard conclusion will be based on transparent evaluation criteria appropriate for the body of evidence and scientific judgement.

REVISION: Moderate confidence in the human data (with no animal data/low confidence in available animal data) will result in either a conclusion of “suspected to be hazard to humans” or “presumed to be a hazard to humans” based on scientific judgement as to the robustness of the body of evidence that supports moderate confidence and consideration of the potential impact of additional studies.

REVISION: The hazard rating reflects the likelihood that additional studies could impact the conclusions. For “suspected”, there is a reasonable expectation that the data from new studies would impact the hazard conclusion and result in a change in the hazard rating. For “presumed”, there is a low expectation that new studies would impact the hazard conclusion.

- **REVISION:** For example, bodies of evidence that would lead to a conclusion of suspected to be a hazard include, but are not limited to: 1) a single well-designed and conducted study including multiple populations with small group sizes and/or a small magnitude of effect; 2) a few well-designed and conducted studies with small study populations or group sizes and/or small magnitude of effect; or 3) a larger number of studies with some inconsistencies in outcomes but an overall small magnitude of effect across the body of evidence.
- **REVISION:** For example, bodies of evidence that would lead to a conclusion of presumed to be a hazard include, but are not limited to: 1) a few well-designed and conducted studies with large magnitude of effect; 2) a few well-designed and conducted studies with large study populations or group sizes with a small magnitude of effect; or 3) a larger number of studies showing a consistent pattern of a small magnitude of effect across the body of evidence.

¹¹GHS addresses classification of chemicals by types of hazard and proposes harmonized hazard communication elements, including labels and safety data sheets:
www.unece.org/trans/danger/publi/ghs/ghs_rev05/05files_e.html

Consideration of Mechanistic Data

The NTP does not require mechanistic or mode-of-action data in order to reach hazard identification conclusions, although when available, this and other relevant supporting types of evidence may be used to raise (or lower) the category of the hazard identification conclusion. Mechanistic, or mode of action, data come from a wide variety of studies that are not intended to identify a disease phenotype. This source of experimental data includes *in vitro* and *in vivo* laboratory studies directed at cellular, biochemical, and molecular mechanisms that explain how a chemical produces particular adverse effects. These studies increasingly take advantage of new “-omics” tools, such as proteomics and metabolomics, to identify early biomarkers of effect. Toxicokinetic information is sometimes considered a type of mechanistic data (NRC 2014a).

If mechanistic data provide strong support for biological plausibility of the relationship between exposure and the health effect, the hazard identification conclusion may be upgraded (indicated by black “up” arrows in the Step 7 graphic in [Figure 8](#)) from the one initially derived by considering the human and non-human animal evidence together. It is envisioned that strong evidence for a relevant biological process from mechanistic data could result in a conclusion of “suspected” in the absence of human epidemiology or experimental animal data. It is theoretically possible that mechanistic data could provide strong opposition for biological plausibility of the relationship between exposure and the health effect. If such a case arises, the hazard identification conclusion may be downgraded (indicated by gray “down” arrows in the Step 7 graphic in [Figure 8](#)). OHAT is working on developing a more structured approach for considering mechanistic data and sees similarities to the factors considered in Step 5 for rating confidence in the body of evidence from human and animal studies ([Figure 9](#)). In the meantime, evaluations of the strength of evidence provided by mechanistic data are made on a project-specific basis based on discussion by the evaluation team and consultation with technical advisors as needed.

Figure 8. Hazard Identification Scheme

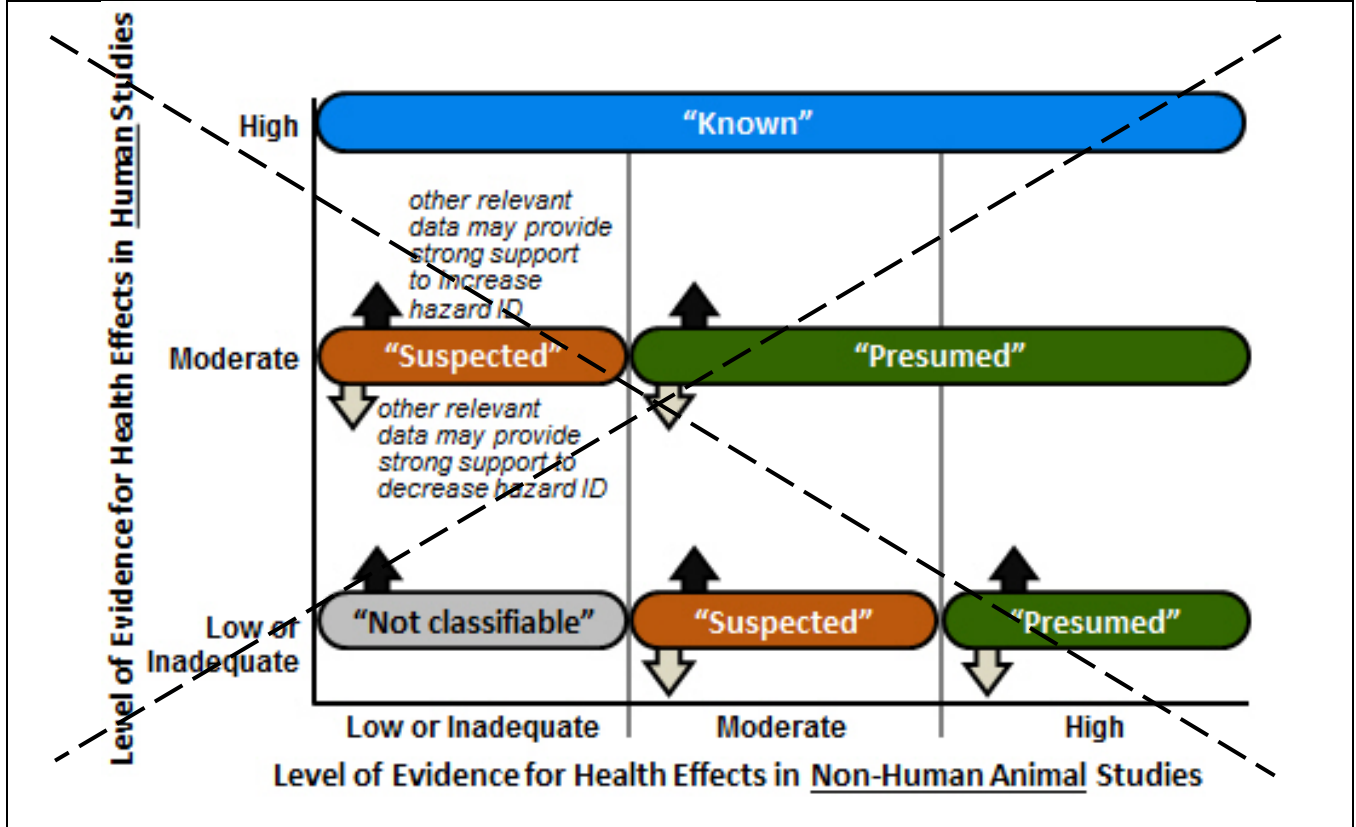


Figure 8. REVISION: Hazard Identification Scheme

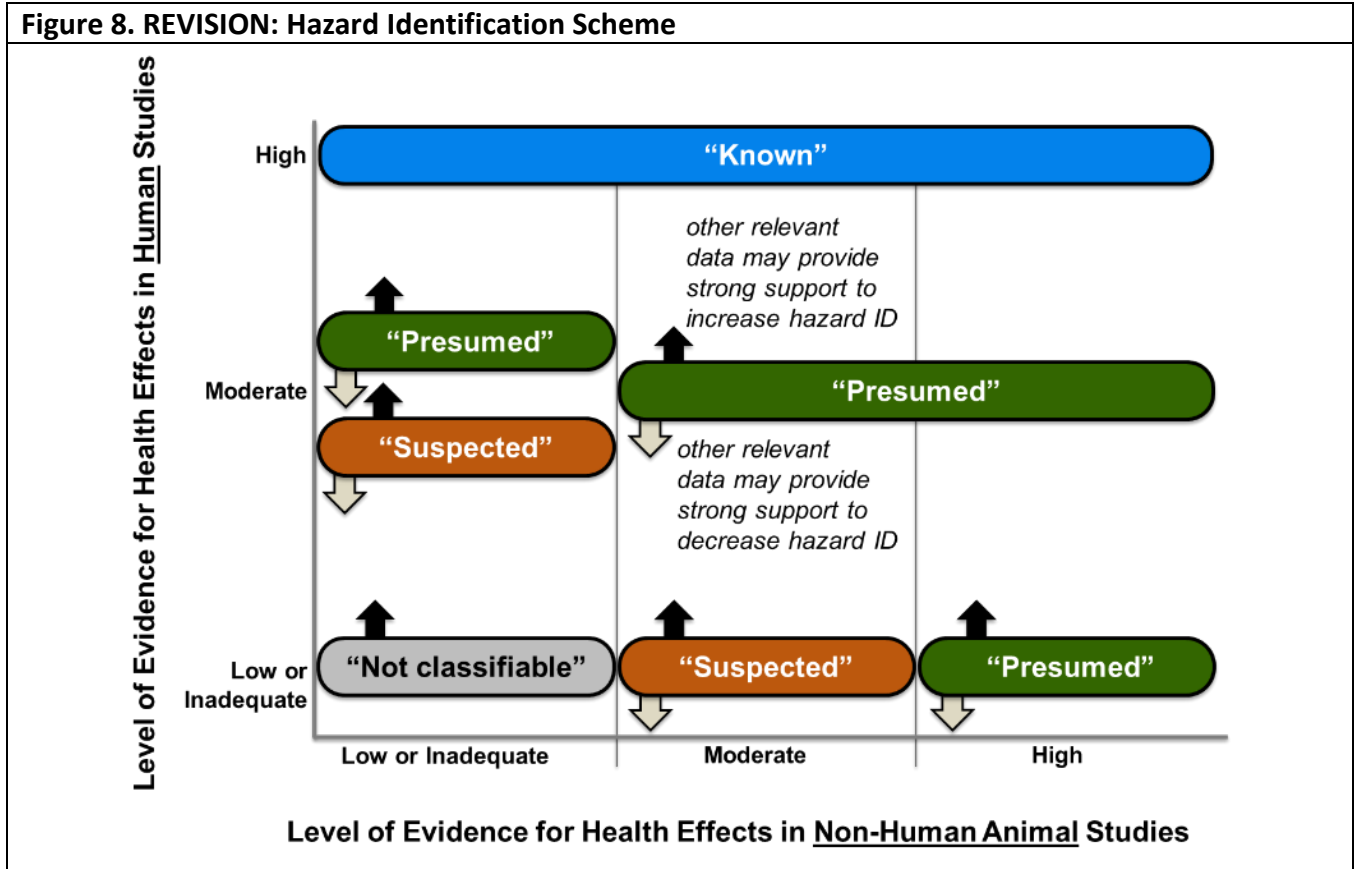


Figure 9. Factors Considered in Evaluating the Support for Biological Plausibility When Mechanistic Data Are Available

Confidence-Rating Factors for Mechanistic Data

- ↑ magnitude of effect ≈ potency
- ↑ dose response
- ↑ consistency
 - across cellular targets on same pathway
- ↓ unexplained inconsistency
 - across studies of same endpoint
- ↓ directness/applicability ≈ relevance
 - relevance of pathway to humans
 - concentration
- ↓ risk of bias/internal validity
- ↓ publication bias

Method Develop Needs

Guidance on similarity profiling – extrapolation from chemicals with more established toxicity

Factors Considered for Human and Animal Evidence

Factors Increasing Confidence

- ↑ magnitude of effect
- ↑ dose response
- ↑ consistency (species/populations)
- ↑ other
 - residual confounding

Factors Decreasing Confidence

- ↓ unexplained inconsistency
- ↓ indirectness/applicability
- ↓ risk of bias/internal validity
- ↓ publication bias
 - imprecision

ABOUT THE PROTOCOL

Contributors

Evaluation Team

Evaluation teams are composed of federal staff and contractor staff. Contractor staff members are screened for potential conflicts of interest. Federal staff members should do a self-evaluation. Epidemiologists and toxicologists on OHAT evaluation teams should have at least three years' experience and/or training in reviewing studies, including summarizing studies and critical review (e.g., assessing study quality and interpreting findings). Experience in evaluating occupational or environmental studies is preferred. Team members should have at least a master's degree or equivalent in epidemiology, toxicology, environmental health sciences, or a related field.

Name	Affiliation
Jane Doe, PhD	NIEHS/NTP, Project Lead
Joe Smith, MD	NIEHS/NIH
<i>Contract support: Assisted in literature screening, data extraction and risk of bias assessment</i>	
Mary Jane, PhD	Company name

Technical Advisors

Technical advisors are outside experts retained on an as-needed basis to provide individual advice to the NTP for a specific topic. Potential technical advisors are screened for conflict of interest prior to their service. Depending upon the situation, the potential conflict of interest is acknowledged, or the person is disqualified from service. Service as a technical advisor does not necessarily indicate that an advisor has read the entire protocol or endorses the final document.

Name	Affiliation
Jane Doe, PhD	East Carolina University, Department of Pharmacology and Toxicology
Joe Smith, MD	NIEHS/NIH

*any conflicts of interest should be stated here

Sources of Support

National Institute of Environmental Health Sciences/Division of the National Toxicology Program

Protocol History and Revisions

Date	Activity or revision
March 26, 2013:	Protocol posted on OHAT website
May 13, 2013:	Risk of bias guidance updated

DATA DISPLAY AND SOFTWARE

Data Display

Tables and graphical displays of study findings are used to reduce text volume and to enhance the clarity and transparency of evidence synthesis. Text in an OHAT monograph represents a concise synthesis of the evidence and does not include long descriptions of individual studies.

Detailed information for individual studies is presented in appendix tables (see Appendix 4 for templates for human, animal, and *in vitro* studies). *Ex vivo*, cellular, genomic, or mechanistic outcomes reported in eligible animal or human studies are included in the animal and human tables and are primarily summarized and interpreted with results from mechanistic studies.

Graphical displays are preferentially included in the main body of the report, ideally based on effect size using a forest plot or exposure-response array format (for human and animal studies) or a concentration-specific response for *in vitro* studies (see Appendix 5 for templates for human, animal, and *in vitro* studies prepared with MetaData Viewer and Inkscape).

Software

OHAT uses a variety of software programs in its evaluations, including (but not limited to) the following:

- *Comprehensive Meta-Analysis* (www.meta-analysis.com): Used to compute effect sizes and to conduct meta-analysis and meta-regression, and to generate statistics for evaluating consistency of data.
- *DistillerSR*® (<http://systematic-review.net/>): Systematic review software primarily used to facilitate tracking of studies through the screening process. Includes capabilities for creating forms to help categorize studies or do a basic level of data extraction.
- *DRAGON, Dose Response Analytical Generator and Organizational Network* (<http://www.icfi.com/insights/products-and-tools/dragon-dose-response>): Software platform that facilitates the conduct of comprehensive human health assessments that require systematic review and synthesis. Includes structured data extraction forms for toxicologic, epidemiologic, and *in vitro* studies. DRAGON has a modular structure and project management capabilities.
- *Endnote* (<http://endnote.com/>): Reference management software.
- *GraphPad Prism*® (www.graphpad.com/scientific-software/prism/): Used to prepare graphs, such as x versus y plots.
- *HAWC, Health Assessment Workspace Collaborative* (<https://hawcproject.org/portal/>): A modular, web-based interface that facilitates development of human health assessments of chemicals. Includes capabilities for screening; categorizing studies; preparing reports; carrying out structured data extraction for toxicologic, epidemiologic, and *in vitro* studies; and enabling interactive, web-based visual displays of data.
- *Inkscape* (<http://inkscape.org/en/>): Open-source, vector graphics editor. It uses [Scalable Vector Graphics](http://www.w3.org/) (SVG), an open XML-based [W3C](http://www.w3.org/) standard as the native format.

- *MetaData Viewer* (ntp.niehs.nih.gov/go/tools_metadataviewer) (Boyles *et al.* 2011): Used to visually display data based on Microsoft Excel file input, mostly based on effect size, which allows for sorting and filtering of data to help assess patterns of findings in complex data sets.
- Microsoft Office Suite
- *OpenEpi* (http://www.openepi.com/Menu/OE_Menu.htm): A free and open-source software for epidemiologic statistics that provides statistics for counts and measurements in descriptive and analytic studies, stratified analysis with exact confidence limits, matched-pair and person-time analysis, sample-size and power calculations, random numbers, sensitivity, specificity and other evaluation statistics, R x C tables, chi-square for dose response, and links to other useful sites.
- *Quosa Information Manager* (<http://www.quosa.com>): Used to manage personal biomedical literature collections, including batch retrieval of PDF copies of studies.
- *SWIFT (Sciome Workbench for Interactive, Computer-Facilitated Text-mining)*: Text-mining/machine-learning tool to help prioritize literature search results based on test set (“seed” studies); identifies overrepresented words, concepts, and phrases; enables categorization of studies based on subtopics (i.e., health outcome, chemical, evidence stream).
- *Universal Desktop Ruler* (www.AVPSoft.com): Used to digitally estimate numerical data from graphs presented in included studies.

TIME AND COST ESTIMATES

For an individual study, the following table estimates the time required for title/abstract review, full-text review, data extraction, and risk of bias assessment. These estimates assume familiarity with the software platforms DistillerSR®, DRAGON, or HAWC.

Phase	Time Estimate per Study*	Cost Estimate (\$100/hour)
Title and abstract review (per screener)	10-20 seconds (180-360 per hour)	~5.5-11 hours to review 1000 studies (\$550-\$1100)
Title and abstract screening + characterization of relevant studies by evidence stream (human, animal, mechanistic), type of health outcome, and type of exposure (per screener)	30 seconds (120 per hour)	~16.6 hours to review 1000 studies (\$1660)
Full-text screening + characterization of relevant studies by evidence stream (human, animal, mechanistic), type of health outcome, and type of exposure	5-10 minutes (6-12 per hour, depending on number of exposure measures/outcomes)	~80-170 hours to review 1000 (\$8000-\$17 000) ~8-17 hours to review 100 (\$800-\$1700)
Data extraction	1.5-3.5 hours (depending on study complexity)	~150-350 hours for 100 studies (\$15 000-\$35 000)
Risk of bias assessment	0.5-1.5 hours (depending on study complexity)	~50-150 hours for 100 studies (\$5000-\$15 000)

*Time estimates after pilot phase. During the pilot phase, time estimates for each step may double. Pilot-testing study number estimates: title and abstract review (100 studies), full-text review (30 studies), and data extraction (2-5 studies, depending on diversity of study designs).

QC = quality control

HANDBOOK PEER REVIEW AND UPDATES

REVISION: Handbook History and Updates

Date	Release or Update
January 9, 2015	Release of OHAT Handbook
March 4, 2019	Update and Clarification of OHAT Handbook

Peer Reviewers (January 9, 2015 Release)

Name	Affiliation
Daniele Mandrioli, MD	Johns Hopkins Bloomberg School of Public Health, Department of Environmental Health Sciences
Malcolm Macleod, PhD	University of Edinburgh, Centre for Clinical Brain Sciences
David Richardson, PhD	University of North Carolina Gillings School of Global Public Health, Department of Epidemiology
Roberta Scherer, PhD	Johns Hopkins Bloomberg School of Public Health, Department of Epidemiology
Ellen K Silbergeld, PhD	Johns Hopkins Bloomberg School of Public Health, Department of Environmental Health Sciences
Tracey Woodruff, PhD, MPH Patrice Sutton, MPH	University of California San Francisco, Department of Obstetrics and/Gynecology and Philip R. Lee (PRL) Institute for Health Policy Studies

Future Considerations

The handbook will be updated as methodological practices are refined and strategies identified that improve the ease and efficiency of conducting a systematic review. A number of changes suggested during peer review were not incorporated into the current version because (1) the changes relate to method development and are more efficiently addressed through collaborations with other environmental health groups promoting systematic review and structured frameworks for evidence integration, (2) additional OHAT systematic reviews need to be conducted to help assess the feasibility of a proposed practice across a broad range of topics, or (3) a range of opinions was expressed and considered in light of NTP programmatic policies and consistency with other federal agencies.

Areas for further consideration and/or method development:

Format

- Restructure the OHAT Handbook along the lines of the Cochrane Handbook such that each step is its own chapter and each chapter starts with a short summary of “key points” followed by the more lengthy instructions. This would allow better separation of the systematic review concept from OHAT process. NOTE: Expect to add this to future versions through collaboration with the Evidence-Based Toxicology discussion group (Mandrioli *et al.* 2014)
- Add a glossary

General

- Improve clarity on when systematic review methods would be used to identify and assess exposure, mechanistic, and toxicokinetic data
- Harmonize terminology and methods with other groups
- Consider developing scoping reports or scoping reviews. This type of review has been defined as "...a form of knowledge synthesis that addresses an exploratory research question aimed at mapping key concepts, types of evidence, and gaps in research related to a defined area or field by systematically searching, selecting, and synthesizing existing knowledge (Colquhoun *et al.* 2014).
- Develop and validate the methods and tools needed for consideration of non-human and mechanistic studies. Even the step of problem formulation has challenges at present especially in defining outcomes when the clinical endpoint does not occur in a non-human model. Assessing sources of heterogeneity and definition of appropriate statistical models are also underdeveloped in toxicology (Silbergeld and Scherer 2013, Ioannidis 2014).

Step 2

- Routine inclusion of non-English studies, given factors such as resource allocation, feasibility, and potential bias introduced to the evaluation
- Consideration of non-peer-reviewed data, e.g., should assess the impact of excluding
- Consideration of a study's statistical power as an exclusion criterion
- Consider establishing criteria or thresholds for screening agreement during pilot phase

Step 3

- Consider establishing criteria or thresholds for accuracy of data extraction during pilot phase

Step 4

- Reconsider nomenclature for describing process of assessing internal validity of studies – the term "risk of bias" is used in systematic review, but strong preference by some to change terminology to "bias," "sources of bias," or something similar.
- Exposure assessment needs more clarity on how to consider in terms of risk of bias, methodology quality, and statistical power/sensitivity based on variation and degree of exposure in subjects.
- Consider adding financial conflict of interest as an element of risk of bias
- Method work needed to determine empirical support for risk of bias elements for observational and experimental animal studies
- Consideration of confounding needs more thought, e.g., how to consider potential impact of factors, consideration of magnitude of estimate and not just p-value
- Consider establishing criteria or thresholds for agreement during pilot phase

Step 5

- Need method development work on establishing initial confidence in the evidence approach, especially for observational studies
- Need method development work to create a structured framework for considering mechanistic data
- How is directness/applicability at the individual-study level considered? (Currently, directness is considered in Step 5 across a collection of studies).
- Need to assess framework for integrating across diverse sources of "indirect" evidence
- Dose-response gradient, e.g., consideration of non-monotonic dose response, needs additional guidance
- Guidance for upgrading evidence, i.e., does current GRADE guidance adequately address animal studies, which might start high but are downgraded for directness to a greater extent than human studies?

Step 6

- Consider providing more detail on level of evidence descriptors, similar to the format used by the Navigation Guide.

REFERENCES

- AHRQ (Agency for Healthcare Research and Quality). 2012a. Grading the Strength of a Body of Evidence When Assessing Health Care Interventions: An Update (Draft Report). Available at <http://effectivehealthcare.ahrq.gov/search-for-guides-reviews-and-reports/?pageaction=displayproduct&productid=1163> [accessed 30 July 2012].
- AHRQ (Agency for Healthcare Research and Quality). 2012b. Interventions for Adults with Serious Mental Illness Who are Involved with the Criminal Justice System. Available at http://effectivehealthcare.ahrq.gov/ehc/products/406/1259/SMI-in-CJ-System_ResearchProtocol_20120913.pdf [accessed 26 September 2012].
- AHRQ (Agency for Healthcare Research and Quality). 2014. AHRQ Training Modules for the Systematic Reviews Methods Guide. Available at <http://www.effectivehealthcare.ahrq.gov/index.cfm/tools-and-resources/slide-library/> [accessed 11 October 2013].
- Andrews N, Miller E, Taylor B, Lingam R, Simmons A, Stowe J, Waight P. 2002. Recall bias, MMR, and autism. *Archives of disease in childhood* 87(6): 493-494.
- ATSDR (Agency for Toxic Substances and Disease Registry). 2012. The Future of Science at ATSDR: A Symposium, Atlanta, GA, US Department of Health and Human Services (DHHS) Agency for Toxic Substances and Disease Registry (ATSDR).
- Balshem H, Helfand M, Schunemann HJ, Oxman AD, Kunz R, Brozek J, Vist GE, Falck-Ytter Y, Meerpohl J, Norris S, Guyatt GH. 2011. GRADE guidelines: 3. Rating the quality of evidence. *Journal of clinical epidemiology* 64(4): 401-406.
- Bero LA. 2013. Why the Cochrane risk of bias tool should include funding source as a standard item. *The Cochrane database of systematic reviews* 12: ED000075.
- Birnbaum LS, Thayer KA, Bucher JR, Wolfe MS. 2013. Implementing systematic review at the National Toxicology Program: Status and next steps. *Environmental health perspectives* 121(4): A108-109.
- Boyles AL, Harris SF, Rooney AA, Thayer KA. 2011. Forest Plot Viewer: a fast, flexible graphing tool. *Epidemiol* 22(5): 746-747.
- Bucher JR, Thayer K, Birnbaum LS. 2011. The Office of Health Assessment and Translation: A problem-solving resource for the National Toxicology Program. *Environmental health perspectives* 119(5): A196-197.
- Carwile JL, Michels KB. 2011. Urinary bisphenol A and obesity: NHANES 2003-2006. *Environmental research* 111(6): 825-830.
- Colquhoun HL, Levac D, O'Brien KK, Straus S, Tricco AC, Perrier L, Kastner M, Moher D. 2014. Scoping reviews: time for clarity in definition, methods, and reporting. *Journal of Clinical Epidemiology* 67(12): 1291-1294.
- Detels R, English P, Visscher BR, Jacobson L, Kingsley LA, Chmiel JS, Dudley JP, Eldred LJ, Ginzburg HM. 1989. Seroconversion, sexual activity, and condom use among 2915 HIV seronegative men followed for up to 2 years. *Journal of acquired immune deficiency syndromes* 2(1): 77-83.

OHAT Handbook (January 9, 2015 REVISION: March 4, 2019)

- Difranceisco W, Ostrow DG, Chmiel JS. 1996. Sexual adventurousness, high-risk behavior, and human immunodeficiency virus-1 seroconversion among the Chicago MACS-CCS cohort, 1984 to 1992. A case-control study. *Sexually transmitted diseases* 23(6): 453-460.
- EFSA (European Food Safety Authority). 2010. Application of systematic review methodology to food and feed safety assessments to support decision making. Available at: <http://www.efsa.europa.eu/en/efsajournal/pub/1637.htm> [accessed 18 January 2012]. *EFSA Journal* 8(6): 1637 [1690 pp.].
- Elliman DA, Bedford HE. 2001. MMR vaccine--worries are not justified. *Archives of disease in childhood* 85(4): 271-274.
- EPA (US Environmental Protection Agency). 1991. Guidelines for Developmental Toxicity Risk Assessment. U.S. Environmental Protection Agency, Risk Assessment Forum, Washington, DC, EPA/600/FR-91/001, 1991. <http://cfpub.epa.gov/ncea/cfm/recordisplay.cfm?deid=23162#Download> [accessed 3 August 2014]. .
- EPA (US Environmental Protection Agency). 1996. Guidelines for Reproductive Toxicity Risk Assessment. U.S. Environmental Protection Agency, Risk Assessment Forum, Washington, DC, 630/R-96/009, 1996.. <http://www.epa.gov/raf/publications/guidelines-reproductive-tox-risk-assessment.htm> [accessed 3 August 2014]. .
- EPA (US Environmental Protection Agency) (US Environmental Protection Agency). 1998. *Guidelines for Ecological Risk Assessment*. EPA/630/R-95/002F. Washington, DC: Office of Prevention Pesticides and Toxic Substances. Available: <http://www.epa.gov/raf/publications/pdfs/ECOTXTBX.PDF>.
- EPA (US Environmental Protection Agency). 2013. Materials Submitted to the National Research Council Part I: Status of Implementation of Recommendations. Environmental Protection Agency: Integrated Risk Information System Program. http://www.epa.gov/iris/pdfs/IRIS%20Program%20Materials%20to%20NRC_Part%201.pdf [accessed 22 February 2013]. .
- Ferguson SA, Law CD, Jr., Abshire JS. 2011. Developmental treatment with bisphenol a or ethinyl estradiol causes few alterations on early preweaning measures. *Toxicological sciences : an official journal of the Society of Toxicology* 124(1): 149-160.
- Fu R, Gartlehner G, Grant M, Shamliyan T, Sedrakyan A, Wilt TJ, Griffith L, Oremus M, Raina P, Ismaila A, Santaguida P, Lau J, Trikalinos TA. 2011. Conducting quantitative synthesis when comparing medical interventions: AHRQ and the Effective Health Care Program. *J Clin Epidemiol* 64(11): 1187-1197.
- Guyatt GH, Oxman AD, Akl EA, Kunz R, Vist G, Brozek J, Norris S, Falck-Ytter Y, Glasziou P, Debeer H, Jaeschke R, Rind D, Meerpohl J, Dahm P, Schunemann HJ. 2011a. GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. *Journal of Clinical Epidemiology* 64(4): 383-394.
- Guyatt GH, Oxman AD, Kunz R, Brozek J, Alonso-Coello P, Rind D, Devereaux PJ, Montori VM, Freyschuss B, Vist G, Jaeschke R, Williams JW, Jr., Murad MH, Sinclair D, Falck-Ytter Y, Meerpohl J, Whittington C, Thorlund K, Andrews J, Schunemann HJ. 2011b. GRADE guidelines 6. Rating the quality of evidence--imprecision. *Journal of clinical epidemiology* 64(12): 1283-1293.

OHAT Handbook (January 9, 2015 REVISION: March 4, 2019)

- Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, Alonso-Coello P, Falck-Ytter Y, Jaeschke R, Vist G, Akl EA, Post PN, Norris S, Meerpohl J, Shukla VK, Nasser M, Schunemann HJ. 2011c. GRADE guidelines: 8. Rating the quality of evidence--indirectness. *Journal of clinical epidemiology* 64(12): 1303-1310.
- Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, Alonso-Coello P, Glasziou P, Jaeschke R, Akl EA, Norris S, Vist G, Dahm P, Shukla VK, Higgins J, Falck-Ytter Y, Schunemann HJ. 2011d. GRADE guidelines: 7. Rating the quality of evidence--inconsistency. *Journal of clinical epidemiology* 64(12): 1294-1302.
- Guyatt GH, Oxman AD, Montori V, Vist G, Kunz R, Brozek J, Alonso-Coello P, Djulbegovic B, Atkins D, Falck-Ytter Y, Williams JW, Jr., Meerpohl J, Norris SL, Akl EA, Schunemann HJ. 2011e. GRADE guidelines: 5. Rating the quality of evidence--publication bias. *Journal of clinical epidemiology* 64(12): 1277-1282.
- Guyatt GH, Oxman AD, Schunemann HJ, Tugwell P, Knottnerus A. 2011f. GRADE guidelines: A new series of articles in the Journal of Clinical Epidemiology. *Journal of clinical epidemiology* 64(4): 380-382.
- Guyatt GH, Oxman AD, Sultan S, Glasziou P, Akl EA, Alonso-Coello P, Atkins D, Kunz R, Brozek J, Montori V, Jaeschke R, Rind D, Dahm P, Meerpohl J, Vist G, Berliner E, Norris S, Falck-Ytter Y, Murad MH, Schunemann HJ. 2011g. GRADE guidelines: 9. Rating up the quality of evidence. *Journal of clinical epidemiology* 64(12): 1311-1316.
- Higgins J, Green S. 2011. Cochrane Handbook for Systematic Reviews of Interventions. Version 5.1.0 (updated March 2011). <http://handbook.cochrane.org/> [accessed 3 February 2013].
- Hill AB. 1965. The Environment and Disease: Association or Causation? *Proc R Soc Med* 58: 295-300.
- Hooijmans CR, Rovers MM, de Vries RB, Leenaars M, Ritskes-Hoitinga M, Langendam MW. 2014. SYRCLE's risk of bias tool for animal studies. *BMC medical research methodology* 14: 43.
- Howard BE, Shah R, Walker K, Pelch K, Holmgren S, Thayer K. 2014. Use of text-mining and machine learning to prioritize the results of a complex literature search. *Society of Toxicology (SOT)*. 53rd annual meeting. Phoenix, AZ (March 23-27, 2014).
- Hugo ER, Brandebourg TD, Woo JG, Loftus J, Alexander JW, Ben-Jonathan N. 2008. Bisphenol A at environmentally relevant doses inhibits adiponectin release from human adipose tissue explants and adipocytes. *Environmental health perspectives* 116(12): 1642-1647.
- Ioannidis JP. 2014. How to make more published research true. *PLoS medicine* 11(10): e1001747.
- IOM (Institute of Medicine). 2011. Finding What Works in Health Care: Standards for Systematic Reviews. http://www.nap.edu/openbook.php?record_id=13059&page=R1 [accessed 13 January 2013].
- Jahnke GD, Iannucci AR, Scialli AR, Shelby MD. 2005. Center for the evaluation of risks to human reproduction--the first five years. *Birth defects research. Part B, Developmental and reproductive toxicology* 74(1): 1-8.
- Johnson PI, Sutton P, Atchley D, Koustas E, Lam J, Robinson K, Sen S, Axelrad D, Woodruff TJ. 2013. Applying the Navigation Guide: Case Study #1: The impact of developmental exposure to perfluorooctanoic acid (PFOA) on fetal growth (Final protocol) <http://prhe.ucsf.edu/prhe/navigationguide.html> [accessed 29 November, 2014]

OHAT Handbook (~~January 9, 2015~~ REVISION: March 4, 2019)

- Johnson PI, Koustas E, Vesterinen HM, Sutton P, Atchley D, Kim AN, Campbell M, McDonald J, Bero L, Sen S, Axelrad D, Zeise L, Woodruff TJ. 2014a. Applying the Navigation Guide: Case Study #2: Reproductive and developmental effects of exposure to triclosan (Protocol) <http://prhe.ucsf.edu/prhe/navigationguide.html> [accessed 29 November, 2014]
- Johnson PI, Sutton P, Atchley DS, Koustas E, Lam J, Sen S, Robinson KA, Axelrad DA, Woodruff TJ. 2014b. The Navigation Guide-Evidence-Based Medicine Meets Environmental Health: Systematic Review of Human Evidence for PFOA Effects on Fetal Growth. *Environmental health perspectives*.
- Koustas E, Lam J, Sutton P, Johnson PI, Atchley D, Robinson K, Sen S, Axelrad D, Woodruff TJ. 2013. Applying the Navigation Guide: Case Study #1. The impact of developmental exposure to perfluorooctanoic acid (PFOA) on fetal growth. A systematic review of the non-human evidence (Final protocol) <http://prhe.ucsf.edu/prhe/navigationguide.html> [accessed 29 November, 2014].
- Koustas E, Lam J, Sutton P, Johnson PI, Atchley DS, Sen S, Robinson KA, Axelrad DA, Woodruff TJ. 2014. The Navigation Guide-Evidence-Based Medicine Meets Environmental Health: Systematic Review of Nonhuman Evidence for PFOA Effects on Fetal Growth. *Environmental health perspectives*.
- Krauth D, Woodruff TJ, Bero LCINEHPM, A P. 2013. Instruments for assessing risk of bias and other methodological criteria of published animal studies: a systematic review. *Environmental health perspectives* 121(9): 985-992.
- LaKind JS, Sobus JR, Goodman M, Barr DB, Furst P, Albertini RJ, Arbuckle TE, Schoeters G, Tan YM, Teeguarden J, Tornero-Velez R, Weisel CP. 2014. A proposal for assessing study quality: Biomonitoring, Environmental Epidemiology, and Short-lived Chemicals (BEES-C) instrument. *Environment international* 73C: 195-207.
- Lam J, Koustas E, Sutton P, Johnson PI, Atchley DS, Sen S, Robinson KA, Axelrad DA, Woodruff TJ. 2014. The Navigation Guide-Evidence-Based Medicine Meets Environmental Health: Integration of Animal and Human Evidence for PFOA Effects on Fetal Growth. *Environmental health perspectives*.
- Lundh A, Sismondo S, Lexchin J, Busuioac OA, Bero LCINGF, Pmid. 2012. Industry sponsorship and research outcome. *The Cochrane database of systematic reviews* 12: MR000033.
- Mandrioli D, Sillbergeld E, Bero L. 2014. Preparation of Evidence Based Toxicology Handbook. <https://colloquium.cochrane.org/meetings/evidence-based-toxicology-handbook>. Cochrane Colloquium expert meeting. Hyderabad, India (September 26, 2014).
- Medlin J. 2003. New arrival: CERHR monograph series on reproductive toxicants. *Environmental health perspectives* 111(13): A696-698.
- Moher D, Liberati A, Tetzlaff J, Altman DG. 2009. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *Journal of Clinical Epidemiology* 62(10): 1006-1012.
- Murray HE, Thayer KA. 2014. Implementing systematic review in toxicological profiles: ATSDR and NIEHS/NTP collaboration. *Journal of environmental health* 76(8): 34-35.
- NRC (National Research Council). 2014a. Review of EPA's Integrated Risk Information System (IRIS) Process (http://www.nap.edu/catalog.php?record_id=18764) [accessed 1 January 2015].

OHAT Handbook (~~January 9, 2015~~ REVISION: March 4, 2019)

- NRC (National Research Council). 2014b. Review of the Environmental Protection Agency's State-of-the-Science Evaluation of Nonmonotonic Dose-Response Relationships as they Apply to Endocrine Disrupters (http://www.nap.edu/catalog.php?record_id=18608) [accessed 1 January 2015].
- NTP (National Toxicology Program). 2012a. Board of Scientific Counselors December 11, 2012 meeting. Meeting materials available at <http://ntp.niehs.nih.gov/go/9741> [accessed 21 February 2013].
- NTP (National Toxicology Program). 2012b. Board of Scientific Counselors June 21-22, 2012 meeting. Meeting materials available at <http://ntp.niehs.nih.gov/go/9741> [accessed 21 February 2013].
- NTP (National Toxicology Program). 2013a. Draft Protocol for Systematic Review to Evaluate the Evidence for an Association Between Perfluorooctanoic Acid (PFOA) or Perfluorooctane Sulfonate (PFOS) Exposure and Immunotoxicity. Available: http://ntp.niehs.nih.gov/ntp/ohat/evaluationprocess/pfos_pfoa_immuneprotocoldraft.pdf [accessed 1 October 2014].
- NTP (National Toxicology Program). 2013b. Draft OHAT Approach for Systematic Review and Evidence Integration for Literature-based Health Assessments – February 2013. <http://ntp.niehs.nih.gov/go/38138> [accessed 26 January 2013].
- Oxman AD, Schunemann HJ, Fretheim A. 2006. Improving the use of research evidence in guideline development: 7. Deciding what evidence to include. *Health research policy and systems / BioMed Central* 4: 19.
- Rooney AA, Boyles AL, Wolfe MS, Bucher JR, Thayer KA. 2014. Systematic review and evidence integration for literature-based environmental health science assessments. *Environmental health perspectives*.
- Sargent RP, Shepard RM, Glantz SACINBMJJ, author reply P. 2004. Reduced incidence of admissions for myocardial infarction associated with public smoking ban: before and after study. *BMJ (Clinical research ed.)* 328(7446): 977-980.
- Schünemann H, Hill S, Guyatt G, Akl EA, Ahmed F. 2011. The GRADE approach and Bradford Hill's criteria for causation. *J Epidemiol Community Health* 65(5): 392-395.
- Shelby MD. 2005. National Toxicology Program Center for the Evaluation of Risks to Human Reproduction: guidelines for CERHR expert panel members. *Birth defects research. Part B, Developmental and reproductive toxicology* 74(1): 9-16.
- Shore RE. 2014. Radiation impacts on human health: certain, fuzzy, and unknown. *Health physics* 106(2): 196-205.
- Silbergeld E, Scherer RW. 2013. Evidence-based toxicology: Strait is the gate, but the road is worth taking. *Altex* 30(1): 67-73.
- Sterne JAC, Higgins JPT, Reeves BC, on behalf of the development group for ACROBAT-NRSI. 2014. ACROBAT-NRSI: A Cochrane Risk Of Bias Assessment Tool for Non-Randomized Studies of Interventions. <https://sites.google.com/site/riskofbiastool/> [accessed 24 September 2014].
- Stovold E, Beecher D, Foxlee R, Noel-Storr A. 2014. Study flow diagrams in Cochrane systematic review updates: an adapted PRISMA flow diagram. *Systematic reviews* 3: 54.
- Szklo M, Nieto FJ. 2007. *Epidemiology: Beyond the basics (2nd edition)* 2nd ed., Sudbury, MA: Jones and Bartlett Publishers.

OHAT Handbook (January 9, 2015 REVISION: March 4, 2019)

- Taylor B, Miller E, Farrington CP, Petropoulos MC, Favot-Mayaud I, Li J, Waight PA. 1999. Autism and measles, mumps, and rubella vaccine: no epidemiological evidence for a causal association. *Lancet* 353(9169): 2026-2029.
- Twombly R. 1998. New NTP centers meet the need to know. *Environmental health perspectives* 106(10): A480-483.
- USPSTF (U.S. Preventive Services Task Force). 2011. USPSTF Procedural Manual. AHRQ Publication No. 08-05118-EF. August 2011.
<http://www.uspreventiveservicestaskforce.org/uspstf08/methods/procmanual.htm>
[accessed 16 September, 2014].
- Vagaggini B, Bartoli ML, Cianchetti S, Costa F, Bacci E, Dente FL, Di Franco A, Malagrino L, Paggiaro P. 2010. Increase in markers of airway inflammation after ozone exposure can be observed also in stable treated asthmatics with minimal functional response to ozone. *Respiratory research* 11: 5.
- Vesterinen HM, Sena ES, Egan KJ, Hirst TC, Churolov L, Currie GL, Antonic A, Howells DW, Macleod MR. 2014. Meta-analysis of data from animal studies: a practical guide. *Journal of neuroscience methods* 221: 92-102.
- Viswanathan M, Ansari M, Berkman ND, Chang S, Hartling L, McPheeters LM, Santaguida PL, Shamliyan T, Singh K, Tsertsvadze A, Treadwell JR. 2012. Assessing the risk of bias of individual studies when comparing medical interventions (March 8, 2012). Agency for Healthcare Research and Quality Methods Guide for Comparative Effectiveness Reviews. March 2012. AHRQ Publication No. 12-EHC047-EF. Available at: www.effectivehealthcare.ahrq.gov/, or direct link at <http://www.effectivehealthcare.ahrq.gov/index.cfm/search-for-guides-reviews-and-reports/?pageaction=displayproduct&productid=998> [accessed 3 January 2013].
- Wakefield AJ, Murch SH, Anthony A, Linnell J, Casson DM, Malik M, Berelowitz M, Dhillon AP, Thomson MA, Harvey P, Valentine A, Davies SE, Walker-Smith JA. 1998. Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children. *Lancet* 351(9103): 637-641 [RETRACTION: *Lancet*. 2010 Feb 2016;2375(9713):2445].
- Woodruff TJ, Sutton P. 2014. The Navigation Guide Systematic Review Methodology: A Rigorous and Transparent Method for Translating Environmental Health Science into Better Health Outcomes. *Environmental health perspectives*.
- Ye X, Kuklenyik Z, Needham L L, and Calafat A. M. 2005. Automated on-line column-switching HPLC-MS/MS method with peak focusing for the determination of nine environmental phenols in urine. *Analytical Chemistry* 77: 5407-5413.

TYPICAL PROTOCOL APPENDICES

Appendix 1: Database-Specific Search Strategies

* Provide data ranges included in search and the date when search was performed

COCHRANE LIBRARY x results date range: date of search:	The exact search terminology would be listed here
EMBASE x results date range: date of search:	
EPA ACToR x results date range: date of search:	CAS Number
PubChem x results date range: date of search:	CAS Number
PUBMED x results date range: date of search:	
SCOPUS x results date range: date of search:	
Toxline x results date range: date of search:	
WEB OF SCIENCE x results date range: date of search:	

Appendix 2: Example of Quick Reference Instructions for Risk of Bias

Observational (Human or Wildlife) Risk of Bias Quick Answers					
#	Question	Definitely Low Direct evidence(D)	Probably Low Indirect(IN)	Probably High Indirect (IN) or missing	Definitely High Direct (D) evidence
1	Randomization	NA	NA	NA	NA
2	Allocation concealment	NA	NA	NA	NA
3	Comparison group	<ul style="list-style-type: none"> Co/CrSe-D-similar (same pop, criteria, response rate) CaCo-D-similar Ca/Co 	<ul style="list-style-type: none"> Co/CrSe-IN-similar groups OR differences OK CaCo-IN-similar Ca/Co OR differences OK 	<ul style="list-style-type: none"> Co/CrSe-IN-not similar (pop, criteria, rate) CaCo-IN-not similar INSUFFICIENT info. 	<ul style="list-style-type: none"> D-not similar (very dissimilar, response rate, or different time frame)
4	Confounding <ul style="list-style-type: none"> design and analysis AND variables assessed AND other exposures 	<ul style="list-style-type: none"> D-appropriately adjusted 	<ul style="list-style-type: none"> IN adjustments OR JUDGED OK 	<ul style="list-style-type: none"> IN-confounders differed INSUFFICIENT info. OR none considered 	<ul style="list-style-type: none"> D-confounders differed
		<ul style="list-style-type: none"> AND variables assessed well-established methods AND same TIME OR acceptable methods AND same TIME <u>PLUS</u> OTHER (e.g., small cv) 	<ul style="list-style-type: none"> AND IN-acceptable methods AND TIME OR JUDGED OK (age, sex, wt.) 	<ul style="list-style-type: none"> IN-insensitive method IN-TIME differed INSUFFICIENT info. 	<ul style="list-style-type: none"> D-insensitive method D-TIME differed
		<ul style="list-style-type: none"> Not present or adjusted, including assessed with well-established methods 	<ul style="list-style-type: none"> IN-not present/adjusted OR JUDGED OK INSUFFICIENT info. and LOW/gen. pop. exposures 	<ul style="list-style-type: none"> IN-unbalanced other present/not adjusted INSUFFICIENT info. and HIGH exposures /occupational Not reported relevant to endpoint (phytoest. diet) 	<ul style="list-style-type: none"> D-unbalanced other exposure present/not adjusted, or not well measured
5	Experimental conditions	NA	NA	NA	NA
6	Blinding (during study)	NA	NA	NA	NA
7	Complete outcome data	<ul style="list-style-type: none"> D-no loss OR addressed and documented 	<ul style="list-style-type: none"> IN-no loss OR addressed OR JUDGED OK 	<ul style="list-style-type: none"> IN-big loss NOT addressed INSUFFICIENT info. 	<ul style="list-style-type: none"> D-big loss NOT addressed

OHAT Handbook (January 9, 2015 REVISION: March 4, 2019)

Observational (Human or Wildlife) Risk of Bias Quick Answers (continued)					
#	Question	Definitely Low Direct evidence(D)	Probably Low Indirect(IN)	Probably High Indirect (IN) or missing	Definitely High Direct (D) evidence
8	Exposure characterization	<ul style="list-style-type: none"> • LOD reported and not near values 	<ul style="list-style-type: none"> • IN-LOD not near values • OK if LOD not reported 	<ul style="list-style-type: none"> • IN-insensitive methods • IN-TIME differed • INSUFFICIENT info. • IN-LOD near values 	<ul style="list-style-type: none"> • D-insensitive method • D-TIME differed • D-LOD near values
9	Outcome assessment <ul style="list-style-type: none"> • Outcome • Blinding 	<ul style="list-style-type: none"> • D-well-established methods • Co/CaCo-D-AND TIME • Acceptable methods AND TIME <u>PLUS</u> OTHER (e.g. internal control, small cv) 	<ul style="list-style-type: none"> • IN-acceptable methods • Co/CaCo-IN-AND TIME • OR JUDGED OK (age, sex, weight) 	<ul style="list-style-type: none"> • IN-insensitive method • Co/CaCo IN-TIME differed • INSUFFICIENT info. 	<ul style="list-style-type: none"> • D-insensitive method • Co/CaCo D-TIME differed
		<ul style="list-style-type: none"> • D-blinding 	<ul style="list-style-type: none"> • IN-blinding • OR JUDGED OK • OR steps to minimize bias 	<ul style="list-style-type: none"> • IN-no blinding • INSUFFICIENT info. 	<ul style="list-style-type: none"> • D-no blinding
10	Outcome reporting	<ul style="list-style-type: none"> • D-all in detail 	<ul style="list-style-type: none"> • IN-all, e.g. sig. dif. or not • OR analyses planned 	<ul style="list-style-type: none"> • IN-not all reported • INSUFFICIENT info. 	<ul style="list-style-type: none"> • D-not all reported
11	No other threats <ul style="list-style-type: none"> • specified in protocol OR • e.g., statistics • e.g., adhere to protocol 	<ul style="list-style-type: none"> • D-OTHER IN PROTOCOL 	<ul style="list-style-type: none"> • IN-OTHER IN PROTOCOL 	<ul style="list-style-type: none"> • IN-NOT OTHER • INSUFFICIENT info. 	<ul style="list-style-type: none"> • D- NOT OTHER
		<ul style="list-style-type: none"> • Stats–appropriate • Stats–if required, test for homogeneity 	<ul style="list-style-type: none"> • Stats-IN-appropriate 	<ul style="list-style-type: none"> • Stats-IN-inappropriate • Stats-if required, no test for homogeneity • INSUFFICIENT info. 	<ul style="list-style-type: none"> • Stats – D-inappropriate or errors
		<ul style="list-style-type: none"> • D-no protocol deviations 	<ul style="list-style-type: none"> • IN–no deviation • INSUFFICIENT info. • OR JUDGED OK 	<ul style="list-style-type: none"> • IN-large deviations 	<ul style="list-style-type: none"> • D-large deviations

OHAT Handbook (January 9, 2015 REVISION: March 4, 2019)

Animal Risk of Bias Quick Answers					
#	Question	Definitely Low Direct evidence(D)	Probably Low Indirect(IN)	Probably High Indirect (IN) or missing	Definitely High Direct (D) evidence
1	Randomization	<ul style="list-style-type: none"> randomization METHOD blocked design w/method 	<ul style="list-style-type: none"> “random” NO METHOD 	<ul style="list-style-type: none"> IN-non-random INSUFFICIENT info. 	<ul style="list-style-type: none"> D- non-random
2	Allocation concealment	<ul style="list-style-type: none"> allocation concealment 	<ul style="list-style-type: none"> IN-concealment OR JUDGED OK for lack of concealment 	<ul style="list-style-type: none"> IN-lack INSUFFICIENT info. 	<ul style="list-style-type: none"> D-lack
3	Comparison group	NA	NA	NA	NA
4	Confounding <ul style="list-style-type: none"> design and analysis AND variables assessed AND other exposures 	<ul style="list-style-type: none"> adjust weight AND other (e.g. blocked kill design) 	<ul style="list-style-type: none"> adjust weight only IN adjustments OR JUDGED OK 	<ul style="list-style-type: none"> IN-confounders differed IN-no adjust weight INSUFFICIENT info. 	<ul style="list-style-type: none"> D-confounders differed D-no adjust weight
		<ul style="list-style-type: none"> AND variables assessed well-established methods AND same TIME OR acceptable methods AND same TIME PLUS OTHER (e.g., small cv) 	<ul style="list-style-type: none"> AND IN-acceptable methods AND TIME OR JUDGED OK (age, sex, wt.) 	<ul style="list-style-type: none"> IN-insensitive method IN-TIME differed INSUFFICIENT info. 	<ul style="list-style-type: none"> D-insensitive method D-TIME differed
		<ul style="list-style-type: none"> Not present or adjusted, including assessed with well-established methods 	<ul style="list-style-type: none"> IN-not present/adjusted OR JUDGED OK INSUFFICIENT info. 	<ul style="list-style-type: none"> IN-unbalanced other present/not adjusted Not reported relevant to endpoint (phytoest. diet) 	<ul style="list-style-type: none"> D-unbalanced other exposure present/not adjusted, or not well measured
5	Experimental conditions	<ul style="list-style-type: none"> Identical conditions and same vehicle control 	<ul style="list-style-type: none"> No report of differences IN same vehicle control OR JUDGED OK dif. veh. 	<ul style="list-style-type: none"> IN-differences No report vehicle control INSUFFICIENT info. 	<ul style="list-style-type: none"> D-differences D-untreated control D-diff. vehicle control
6	Blinding (during study)	<ul style="list-style-type: none"> D-blinding during study 	<ul style="list-style-type: none"> IN-blinding during study Blinding not possible AND steps to minimize bias 	<ul style="list-style-type: none"> IN-no blinding AND no steps to minimize bias INSUFFICIENT info. 	<ul style="list-style-type: none"> D-no blinding AND no steps to minimize bias
7	Complete outcome data	<ul style="list-style-type: none"> D-no loss OR addressed and documented 	<ul style="list-style-type: none"> IN-no loss OR addressed OR JUDGED OK 	<ul style="list-style-type: none"> IN-big loss NOT addressed INSUFFICIENT info. 	<ul style="list-style-type: none"> D-big loss NOT addressed

OHAT Handbook (January 9, 2015 REVISION: March 4, 2019)

Animal Risk of Bias Quick Answers (continued)					
#	Question	Definitely Low Direct evidence(D)	Probably Low Indirect(IN)	Probably High Indirect (IN) or missing	Definitely High Direct (D) evidence
8	Exposure characterization	<ul style="list-style-type: none"> Independent assess pure stability purity ≥ 99% LOD reported and not near values 	<ul style="list-style-type: none"> IN or supplier assess pure IN-stability IN-purity ≥ 99% OR ≥ 98%, JUDGED 2% OK IN-LOD not near values OK if LOD not reported 	<ul style="list-style-type: none"> IN-insensitive method IN-TIME differed IN-stability import. not tested or controlled INSUFFICIENT info. IN-LOD near values 	<ul style="list-style-type: none"> D-insensitive method D-TIME differed D-stability import. not tested or controlled D-LOD near values
9	Outcome assessment <ul style="list-style-type: none"> Outcome Blinding 	<ul style="list-style-type: none"> D-well-established methods AND same TIME Acceptable methods AND TIME <u>PLUS</u> OTHER (e.g. internal control, small cv) 	<ul style="list-style-type: none"> IN-acceptable methods AND same TIME OR JUDGED OK (age, sex, weight) 	<ul style="list-style-type: none"> IN-insensitive method IN-TIME differed INSUFFICIENT info. 	<ul style="list-style-type: none"> D-insensitive method D-TIME differed
		<ul style="list-style-type: none"> D-blinding 	<ul style="list-style-type: none"> IN-blinding OR JUDGED OK OR steps to minimize bias 	<ul style="list-style-type: none"> IN-no blinding INSUFFICIENT info. 	<ul style="list-style-type: none"> D-no blinding
10	Outcome reporting	<ul style="list-style-type: none"> D-all in detail 	<ul style="list-style-type: none"> IN-all, e.g. sig. dif. or not OR analyses planned 	<ul style="list-style-type: none"> IN-not all reported INSUFFICIENT info. 	<ul style="list-style-type: none"> D-not all reported
11	No other threats <ul style="list-style-type: none"> specified in protocol OR e.g., statistics e.g., adhere to protocol 	<ul style="list-style-type: none"> D-OTHER IN PROTOCOL 	<ul style="list-style-type: none"> IN-OTHER IN PROTOCOL 	<ul style="list-style-type: none"> IN-NOT OTHER INSUFFICIENT info. 	<ul style="list-style-type: none"> D- NOT OTHER
		<ul style="list-style-type: none"> Stats–appropriate Stats–if required, test for homogeneity 	<ul style="list-style-type: none"> Stats-IN-appropriate 	<ul style="list-style-type: none"> Stats-IN-inappropriate Stats-if required, no test for homogeneity INSUFFICIENT info. 	<ul style="list-style-type: none"> Stats – D-inappropriate or errors
		<ul style="list-style-type: none"> D-no protocol deviations 	<ul style="list-style-type: none"> IN–no deviation INSUFFICIENT info. OR JUDGED OK 	<ul style="list-style-type: none"> IN-large deviations 	<ul style="list-style-type: none"> D-large deviations

Appendix 3: Example of an Evidence Profile Table: PFOS/PFOA and Functional Antibody Response

Body of Evidence	Risk of Bias	Unexplained Inconsistency	Indirectness	Imprecision	Publication Bias	Magnitude	Dose Response	Residual Confounding	Consistency Across Species/Model	Final Rating
Endpoint: Functional antibody response (example of a “hypothetical” illustration for PFOS)										
Animal	not serious	not serious	not serious	not serious	undetected	not large	yes (increase)	no	no	HIGH
(8 PFOS Studies) <u>Initial Rating</u> High	<ul style="list-style-type: none"> General low Key questions <ul style="list-style-type: none"> ○ Randomize = mixed low and probably high ○ Outcome = low Probably high for allocation concealment 	<ul style="list-style-type: none"> Consistent suppression Potential inconsistent response, but differed by: <ul style="list-style-type: none"> ○ Species (rat vs mouse), ○ Outcome (IgG vs IgM), ○ Antigen (SRBC vs KLH) 	<ul style="list-style-type: none"> SRBC IgM response by PFC or ELISA are among best measures of antibody response 	<ul style="list-style-type: none"> General small, confidence interval (CI) Non-overlapping CIs between control and exposed 	<ul style="list-style-type: none"> No evidence of lag bias Funding <ul style="list-style-type: none"> ○ Government ○ Universities ○ Industry 	<ul style="list-style-type: none"> Not sufficiently large to overcome potential bias 	<ul style="list-style-type: none"> Dose response observed in multiple studies 	<ul style="list-style-type: none"> No evidence of confounding that would bias toward null 	<ul style="list-style-type: none"> All positive results from mice 	Started high No serious downgrades Upgrade for dose response Final rating would be High

Appendix 4: Template Options for Tabular Data Summary

Human Studies

Template Option 1: Human Study				
Reference, Study Design, & Population	Health Outcome	Exposure	Statistical Analysis	
<p>(Carwile and Michels 2011) Study design: cross-sectional Adults who participated in the 2003/04 and 2005/06 National Health and Nutrition Examination Survey (NHANES) and had a spot urine sample analysed for BPA N: 2747 Location: US, NHANES national survey Sex (% male): ♂♀(49.6%) Sampling time frame: 2003-2006 Age: 18-74 years Exclusions: pregnant women, participants with missing urinary BPA, creatine, BMI, or covariate data Funding source: NIH National Research Service Award (NRSA) Author conflict of interest: not reported</p>	<p>Diagnostic and prevalence in total cohort:</p> <p>Obesity: BMI ≥ 30 (n = 932, 34.3%) Overweight: 25 ≤ BMI < 30 (n = 864, 31.8%) Elevated waist circumference (WC): > 102 cm in ♂ or ≥ 88 cm in ♀ (n = 1330, 50%)</p> <p>*BMI = body mass index (kg/m²)</p>	<p>Exposure assessment: Urine (µg/g creatinine or ng/ml and creatinine as adjustment variable) measured by online SPE-HPLC-MS/MS (Ye 2005)</p> <p>Exposure levels: 2.05 µg/g creatinine (geometric mean), 1.18-3.33 (25-75th percentile) Q1: ≤ 1.1 ng/ml Q2: 1.2-2.3 ng/ml Q3: 2.4-4.6 ng/ml Q4: > 4.7 ng/ml</p>	<p>Obesity & overweight: polytomous regression Elevated WC: logistic regression Adjustment factors: sex, age, race, urinary creatinine, education, smoking Statistical power: Appears to be adequately powered based on ability to detect an OR of 1.5 with 80% power using Q1 prevalence of 40.4% obesity, 44.4% overweight, and 46% elevated WC</p>	adjOR (95% CI)
				Obesity
				Q2 vs Q1: 1.85 (1.22, 2.79)
				Q3 vs Q1: 1.60 (1.05, 2.44)
				Q4 vs Q1: 1.76 (1.06, 2.94)
				Overweight
				Q2 vs Q1: 1.66 (1.21, 2.27)
				Q3 vs Q1: 1.26 (0.85, 1.87)
				Q4 vs Q1: 1.31 (0.80, 2.14)
				Elevated WC
Q2 vs Q1: 1.62 (1.11, 2.36)				
Q3 vs Q1: 1.39 (1.02, 1.90)				
Q4 vs Q1: 1.58 (1.03, 2.42)				
RISK OF BIAS ASSESSMENT				
<i>Risk of bias response options for individual items:</i>				
Bias Domain	Criterion	Response		
Selection	Was administered dose or exposure level adequately randomized?	n/a	Not applicable	
	Was allocation to study groups adequately concealed?	n/a	Not applicable	
	Were the comparison groups appropriate?	++	Yes, based on quartiles of exposure	
Confounding	Does the study design or analysis account for important confounding and modifying variables?	++	Yes (sex, age, race, urinary creatinine, education, smoking), but no adjustment for nutritional quality, e.g., soda consumption	
	Did researchers adjust or control for other exposures that are anticipated to bias results?	+	No, but not considered to present risk of bias in general population studies	
Performance				

OHAT Handbook (January 9, 2015 REVISION: March 4, 2019)

Template Option 1: Human Study			
	Were experimental conditions identical across study groups?	n/a	Not applicable
	Did deviations from the study protocol have an impact on the results?	+	No deviations reported
	Were the research personnel and human subjects blinded to the study group during the study?	n/a	Not applicable
Attrition	Were outcome data incomplete because of attrition or exclusion from analysis?	+	Not considered a risk of bias, excluded observations (≤ 87 for any analysis) based on missing BMI or covariate data
Detection	Were the outcome assessors blinded to study group or exposure level?	++	Yes, BPA levels not known at time of outcome assessment
	Were confounding variables assessed consistently across groups using valid and reliable measures?	++	Yes, used standard NHANES methods
	Can we be confident in the exposure characterization?	++	Yes, NHANES methods are considered “gold standard” for urinary BPA
	Can we be confident in the outcome assessment?	++	Yes, used standard diagnostic criteria
Selective Reporting	Were all measured outcomes reported?	++	Yes, primary outcomes discussed in methods were presented in results section with adequate level of detail for data extraction
Other	Were there any other potential threats to internal validity (e.g., inappropriate statistical methods)?	++	None identified
			1st Tier for risk of bias

RISK OF BIAS

<i>Risk of bias response options for individual items:</i>	
++	definitely low risk of bias
+	probably low risk of bias
-	probably high risk of bias
--	definitely high risk of bias
n/a	not applicable

Animal Studies

Template Option 1: Animal Study								
Reference, Animal Model, and Dosing		Health Outcome	Results					
<p>(Ferguson et al. 2011) Species: rat Strain (source): Sprague-Dawley (NCTR Breeding colony derived from Charles River Crl: COBS CD (SD) BR Rat, Outbred) Sex: ♂♀ Doses: 0.0025 or 0.025 mg/kg/day BPA Purity (source): > 99% (TCI America) Dosing period: GD6-21 (via dam) and PND 1-21 to pup Route: oral gavage Diet: low-phytoestrogen chow (TestDiet 5K96 [irradiated pellets], Verified Casein Diet 10 IF; TestDiet), low levels of daidzein (< 0.34 ppm) and genistein (< 0.58 ppm) measured in three separate samples Controls: naïve and vehicle control of 0.3% (by weight) aqueous solution of carboxymethylcellulose (CMC) sodium salt Funding source: National Center for Toxicological Research/Food and Drug Administration Author conflict of interest: not reported Comments: 0.005 or 0.010 mg/kg/day ethinyl estradiol (EE₂) used as postive control</p>		<p>Endpoints: leptin & ghrelin measured by ELISA Age at assessment: PND 21 N = 10-17 for males; 13-15 for females Statistical analysis: two-way ANOVAs with treatment and sex as factors Control for litter effects: one offspring/sex/litter Statistical power: underpowered (sample size is < 50% required) to detect a change of 10%-25% control</p>		Group	Mean ± SE	% control (95%CI)*	Mean ± SE	% control (95%CI)*
				Leptin	Males		Females	
				Naive	5.0 ± 1.0		5.8 ± 1.1	
				Vehicle	4.7 ± 0.6		5.5 ± 0.8	
				0.0025 BPA	4.2 ± 0.5	-10.6 (-44.6, 23.6)	4.1 ± 0.7	-25.5 (-69.4, 18.5)
				0.025 BPA	4.7 ± 1.7	0 (-75.2, 75.2)	3.3 ± 0.4	-40 (-77.1, -2.9)
				0.005 EE ₂	3.8 ± 0.8	-19.2 (-67.4, 29.1)	4.5 ± 1.2	-18.2 (-77.7, 41.4)
				0.010 EE ₂	3.1 ± 0.4	-34.0 (-69.6, 1.5)	3.2 ± 0.5	-41.8 (-83.7, 0.02)
				Ghrelin				
				Naive	1.913 ± 0.179		2.085 ± 0.357	
				Vehicle	1.688 ± 0.139		1.953 ± 0.250	
				0.0025 BPA	1.567 ± 0.227	-7.2 (-39.8, 25.5)	1.693 ± 0.170	-13.3 (-45.2, 18.6)
				0.025 BPA	1.760 ± 0.193	4.3 (-22.6, 31.2)	1.508 ± 0.140	-22.7 (-53.8, 8.2)
				0.005 EE ₂	1.755 ± 0.210	4.0 (-24.5, 32.4)	1.823 ± 0.183	-6.6 (-38.5, 25.2)
0.010 EE ₂	1.667 ± 0.201	-1.2 (-29.9, 27.4)	1.623 ± 0.184	-16.9 (-50.4, 16.6)				
*Average group size (rounded up when needed) was used to estimate percent control response (14 for males; 14 for females).								
RISK OF BIAS ASSESSMENT								
Risk of bias response options for individual items:								
Bias Domain	Criterion	Response						
Selection	Was administered dose or exposure level adequately randomized?	++	Yes, "randomly assigned to treatment within their body weight stratum"					
	Was allocation to study groups adequately concealed?	+	Not reported, but lack of adequate allocation concealment at study start not expected to appreciably bias results					
	Were the comparison groups appropriate?	n/a	Not applicable					
Confounding	Does the study design or analysis account for important confounding and modifying variables?	+	No, neither litter size or body weight considered as covariates in analysis, but not clear these need to be considered for endpoints reported in study					
	Did researchers adjust or control for other exposures that are anticipated to bias results?	++	Yes, low phytoestrogen diet and polysulfone cages with only trace BPA used; levels of BPA in other housing equipment measured					

OHAT Handbook (January 9, 2015 REVISION: March 4, 2019)

Template Option 1: Animal Study			
Performance	Were experimental conditions identical across study groups?	+	Assumed yes
	Did deviations from the study protocol have an impact on the results?	+	No deviations reported
	Were the research personnel and human subjects blinded to the study group during the study?	+	Not reported, but lack of adequate allocation concealment during conduct of study not feasible and not expected to appreciably bias results for this study
Attrition	Were outcome data incomplete because of attrition or exclusion from analysis?	+	Yes, but dead or missing (assumed cannibalized) offspring documented and were generally evenly distributed across groups
Detection	Were the outcome assessors blinded to study group or exposure level?	+	Not reported, but not considered a risk of bias for these endpoints (hormone levels) because measurement is not subjective
	Were confounding variables assessed consistently across groups using valid and reliable measures?	n/a	Not applicable given that confounding/modifying variables were not included
	Can we be confident in the exposure characterization?	++	Yes, purity > 99% and dosing solutions measured and were very close to target doses
	Can we be confident in the outcome assessment?	++	Yes, used standard kits and inter assay coefficients of variation < 4%
Selective Reporting	Were all measured outcomes reported?	++	Yes, primary outcomes discussed in methods were presented in results section with adequate level of detail for data extraction
Other	Were there any other potential threats to internal validity (e.g., inappropriate statistical methods)?	++	None identified, potential litter effects were controlled for experimentally
			1st Tier for risk of bias

RISK OF BIAS

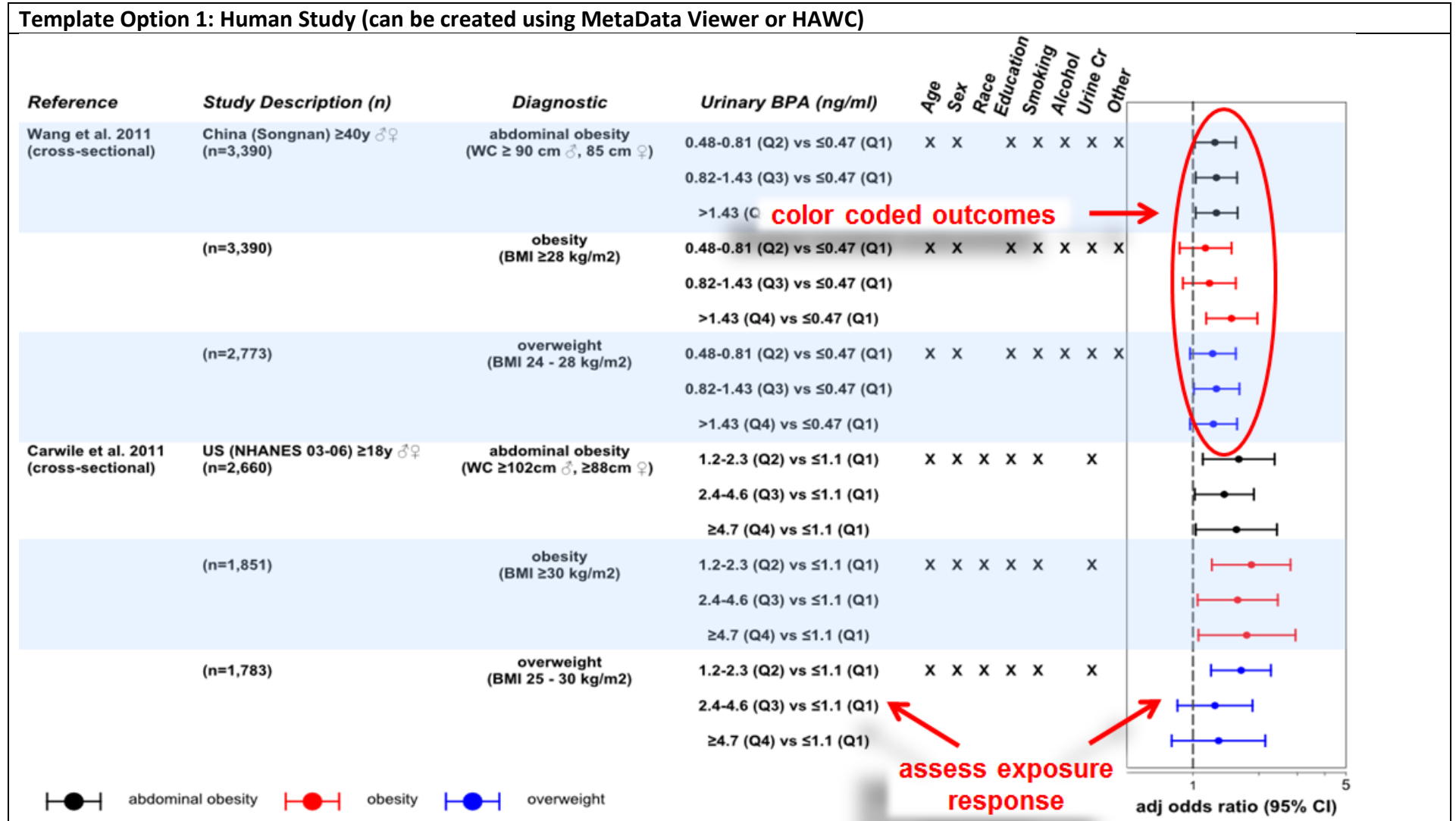
<i>Risk of bias response options for individual items:</i>	
++	definitely low risk of bias
+	probably low risk of bias
-	probably high risk of bias
--	definitely high risk of bias
n/a	not applicable

In Vitro Studies

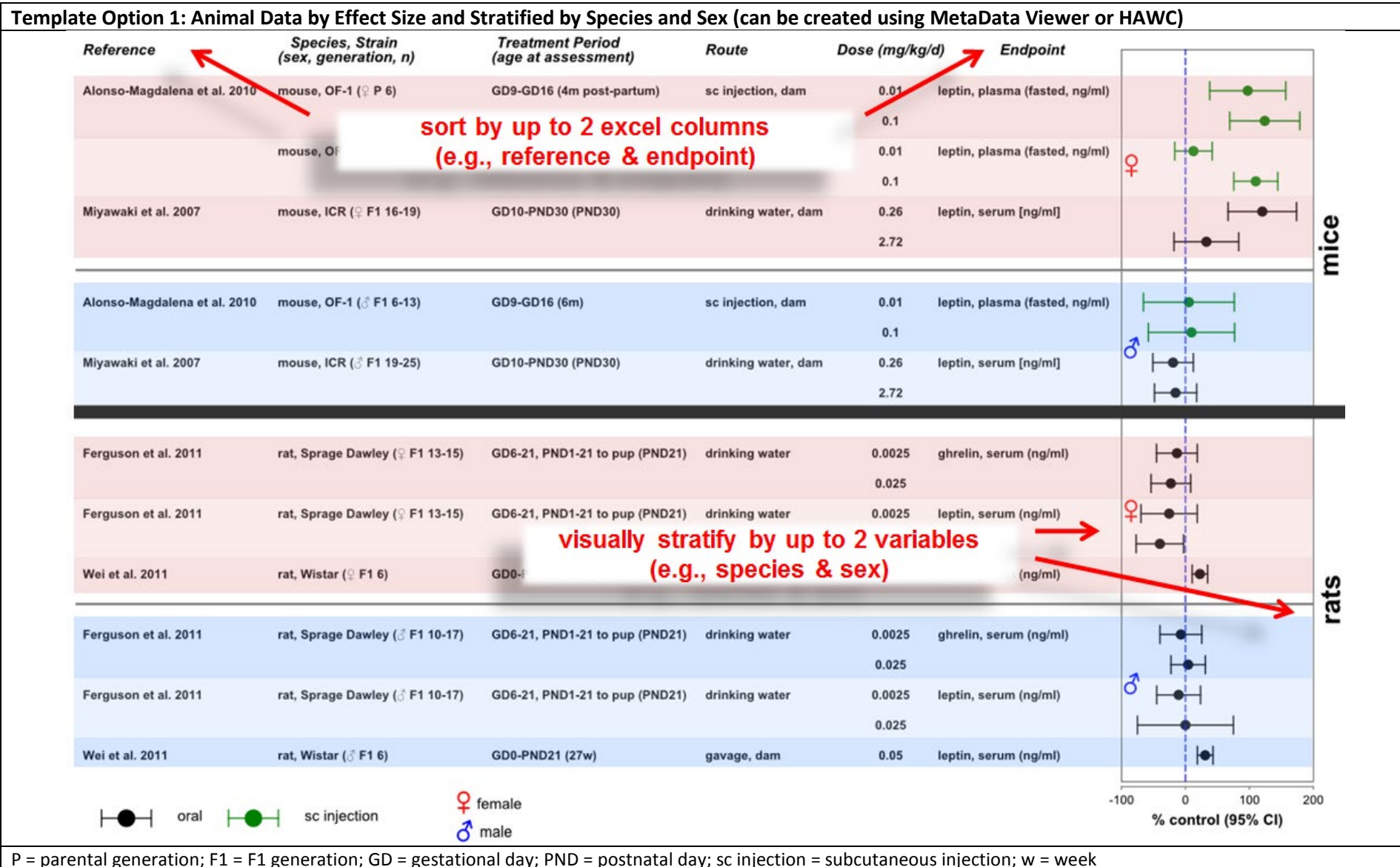
Template Option 1: <i>In Vitro</i> Study		
Reference, Model, and Treatment	Endpoint	Concentration (μM) Specific Findings
(Hugo et al. 2008) Species: human Cell-line/source: explants from breast (8 women undergoing breast reduction surgery) and abdominal subcutaneous adipose (9 women undergoing abdominoplasty) Sex: ♀ Concentrations: 0.0001, 0.001, 0.01, 0.1 μ M BPA Purity (source): > 99% (Sigma-Aldrich) Vehicle: < 0.001% EtOH Treatment period: 6h Replicates: Results based on mean of 6 determinations Funding source: NIH, Department of Defense, Susan G. Komen Breast Cancer Foundation Author conflict of interest: authors declare no competing interest Comments: non-monotonic dose response; response consistent with estradiol positive control	Adiponectin release, breast adipose (ng/100 mg/6h):	0.0001(↓), 0.001(↓), 0.01, 0.1
	Adiponectin release, abdominal adipose (ng/100 mg/6h):	0.0001(↓), 0.001(↓), 0.01, 0.1
↑ = statistically significant increase reported by authors, ↓ = statistically significant decrease reported by authors		

Appendix 5: Template Options for Graphical Data Display

Human Studies



Animal Studies



P = parental generation; F1 = F1 generation; GD = gestational day; PND = postnatal day; sc injection = subcutaneous injection; w = week

In Vitro Studies

