National Toxicology Program
U.S. Department of Health and Human Services

Handbook for Preparing Report on Carcinogens Monographs

July 20, 2015

Office of the Report on Carcinogens
Division of the National Toxicology Program
National Institute of Environmental Health Sciences
U.S. Department of Health and Human Services

Foreword

The National Toxicology Program (NTP) is an interagency program within the Public Health Service of the Department of Health and Human Services and is headquartered at the National Institute of Environmental Health Sciences of the National Institutes of Health (NIEHS/NIH). Three agencies contribute resources to the program: NIEHS/NIH, the National Institute for Occupational Safety and Health of the Centers for Disease Control and Prevention, and the National Center for Toxicological Research of the Food and Drug Administration. Established in 1978, the NTP is charged with coordinating toxicological testing activities, strengthening the science base in toxicology, developing and validating improved testing methods, and providing information about potentially toxic substances to health regulatory and research agencies, scientific and medical communities, and the public.

The NTP prepares the Report on Carcinogens (RoC), a science-based public heath document, for the Secretary, Department of Health and Human Services. The Office of the RoC is responsible for carrying out this activity within the NTP and prepares a monograph for each substance evaluated for listing in the RoC (i.e., a candidate substance). The monograph is a literature-based review document that captures the cancer hazard evaluation.

This handbook provides instructions for preparing the RoC monographs. It is based largely on approaches outlined in protocols (i.e., methods) used to prepare RoC monographs starting in 2013, although the methods have since undergone a series of revisions. These monographs, prepared according to the protocols, have been peer reviewed by panels of experts. It is anticipated that this handbook will be refined as new tools for conducting literature-based systematic reviews are developed and from knowledge learned from conducting cancer hazard evaluations on candidate substances with more diverse databases.

Acknowledgments

Handbook Peer Reviewers

David Eastmond	Environmental Toxicology Graduate Program and Department of Cell Biology & Neuroscience University of California Riverside, California
Neela Guha	International Agency for Research on Cancer Monographs Section Lyon, France
Bill Jameson	CWJ Consulting, LLC Cape Coral, FL
Sheila Zahm	Sheila Zahm Consulting Hermon, ME
Lauren Zeise	Office of Environmental Health Hazard Assessment California Environmental Protection Agency

Support for the Preparation of the Handbook

Integrated Laboratory Systems, Inc.,
Morrisville, NC
NIEHS Contract Number HHSN273201100004C

Table of Contents

Introduction

Objectives

This handbook describes the methods and considerations for conducting a literature-based review (i.e., cancer hazard evaluation) of an agent, substance, mixture, or exposure circumstance (collectively referred to as "substance") selected for evaluation for listing in the Report on Carcinogens (RoC). The cancer hazard evaluation is captured in a RoC monograph, and this handbook serves as a resource for those preparing the monographs, including Office of the RoC (ORoC) staff, contractor support staff, and technical advisors. The approach to conducting the cancer hazard evaluation incorporates principles of systematic review, with the goal of increasing transparency (to the public and others) on how the conclusions are reached and strengthening consistency across evaluations of different substances. For each substance under review, a protocol is developed that adapts these methods for scientific issues specific to the substance.

Background Information on the RoC

The RoC is a congressionally mandated (see below) science-based document that identifies potential cancer hazards for people living in the United States. Substances are listed in two categories: *known to be a human carcinogen* and *reasonably anticipated to be a human carcinogen*. The National Toxicology Program (NTP) prepares the report for the Secretary, Department of Health and Human Services (HHS) using a four-part process (http://ntp.niehs.nih.gov/go/rocprocess) and established criteria (summarized in Part G and available at http://ntp.niehs.nih.gov/go/15209). For each listed substance, the RoC includes a substance profile with information from cancer studies that supports the listing, as well as information about use and production, potential sources of exposure, and current federal regulations to limit exposure. Each edition of the RoC is cumulative and consists of substances newly reviewed in addition to those listed in previous editions. The latest edition of the report, the 13th RoC, contains 243 substance profiles (http://ntp.niehs.nih.gov/go/roc13).

Congressional mandate

Section 301(b)(4) of the Public Health Service Act, as amended, requires that the Secretary, HHS, publish an annual report that contains a list of all substances

- which either are *known to be human carcinogens* or may *reasonably be anticipated to be human carcinogens* and
- to which a significant number of persons residing in the United States are exposed.

Process for Preparing the RoC

The process for preparing the RoC has four parts: (1) nomination and selection of candidate substances, (2) scientific evaluation of candidate substances (captured in the draft RoC monographs), (3) public release and peer review of the draft RoC monographs, and (4) HHS approval and release of the latest edition of the RoC. This process is diagrammed in Figure 1.

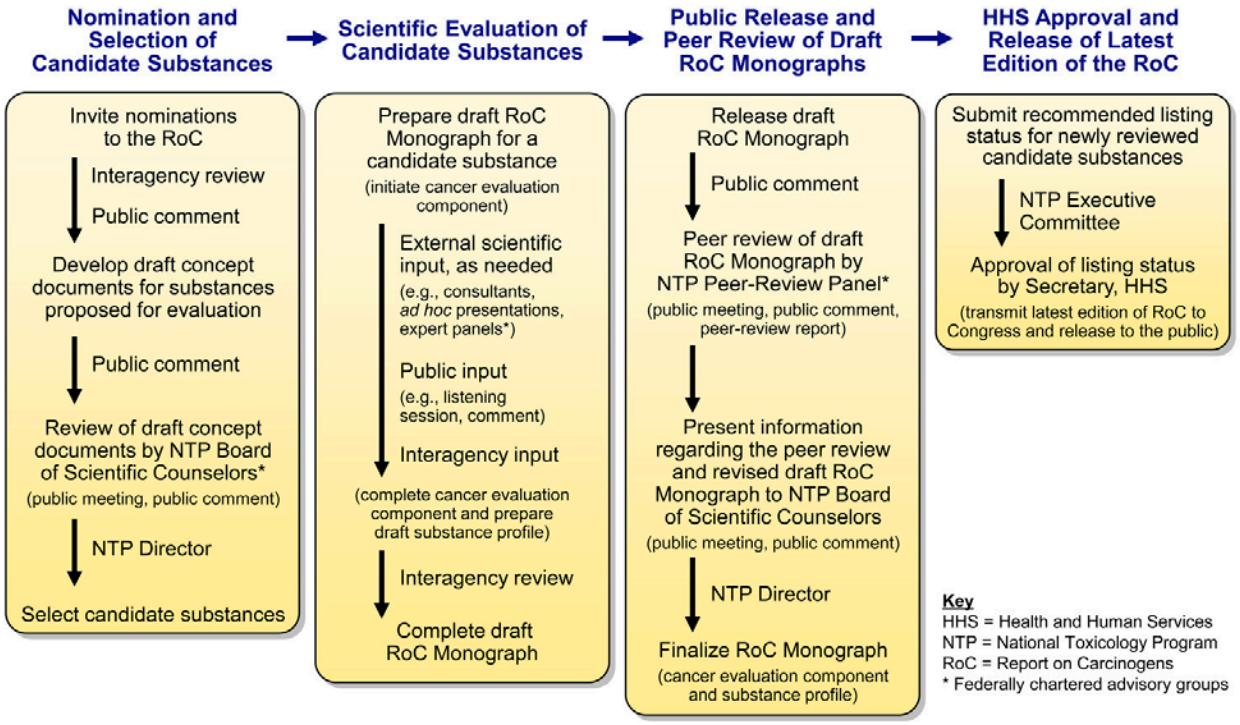


Figure 1. Process for preparation of the Report on Carcinogens

Background Information on RoC Monographs and the Handbook

An RoC monograph has two parts: (1) a cancer hazard evaluation component that reviews all information that may bear on a listing decision, assesses its quality and sufficiency for reaching a listing decision, applies the RoC listing criteria to the relevant scientific information, and recommends an RoC listing status for the candidate substance and (2) a substance profile that contains NTP's preliminary listing recommendation and a summary of the scientific evidence considered key to reaching that recommendation. In general, the cancer evaluation component addresses the following topics, although other topics may be included where relevant to evaluating the carcinogenicity of the candidate substance:

- properties (e.g., chemical, physical, or biological), production, and use
- human exposure
- disposition and toxicokinetics
- cancer studies in humans
- cancer studies in experimental animals
- mechanisms of cancer induction and other related effects (such as genotoxicity)

Information on exposure and properties of the candidate substance must come from publicly available sources, and all scientific information used to evaluate the potential carcinogenicity of a candidate substance must come from peer-reviewed, publicly available sources.

The cancer hazard evaluation component of the RoC monograph (1) presents the literature search strategy and the literature inclusion/exclusion criteria, (2) identifies and describes the studies

relevant for the RoC evaluation, (3) assesses the quality of individual studies and discusses their usefulness for informing the evaluation of carcinogenicity, (4) assesses the level of evidence from human studies or experimental animal studies in applying the RoC listing criteria, and (5) integrates the overall body of evidence (human, animal, and mechanistic) and reaches a preliminary RoC listing recommendation for the substance.

As depicted in Figure 1, the draft RoC monograph is peer reviewed by a NTP panel of experts at a public meeting. Based upon the peer-review comments, ORoC prepares a revised draft RoC Monograph. At a public meeting, the NTP provides the BSC with information regarding the peer review. Following the meeting, ORoC, in concert with the NTP Director, finalizes the RoC Monograph on the candidate substance, including the cancer evaluation component and substance profile, and posts the final monograph on the RoC website.

In addition to providing instructions for preparing each section of the draft monograph (Part 2 of the RoC process), this handbook also briefly discusses steps related to the selection of a candidate substance (Part 1 of the RoC process). The handbook has the following parts:

- Part A: Selection of a candidate substance, planning, and protocol development
- Part B: Identification and selection of studies
- Part C: Evaluation of human exposure data
- Part D: Evaluation of cancer studies in humans
- Part E: Evaluation of cancer studies in experimental animals
- Part F: Evaluation of mechanistic and other relevant data
- Part G: Evidence Integration to Reach a Preliminary Listing Recommendation

Part A: Selection of a Candidate Substance, Planning, and Protocol Development

Planning and research are important throughout much of the cancer hazard evaluation process, including during the initial scoping of the project, development of the concept document and protocol, and assessment of the quality and utility of the individual studies (see Figure A-1). The process is iterative and relies on considerable scientific input.

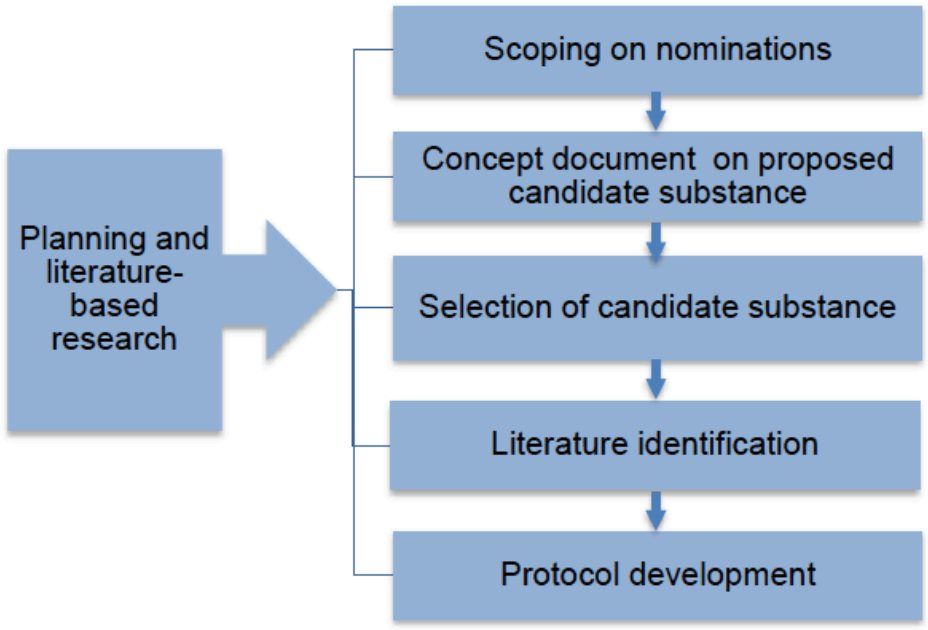


Figure A-1. Planning and literature-based research steps in the cancer hazard evaluation

Planning

Conducting a cancer hazard assessment for a candidate substance begins with a significant phase of planning and literature-based research. The planning process is necessarily iterative and begins with identifying a candidate substance – e.g., scoping the literature on the nominations to determine which ones to propose for review, formulating the rationale and approach for the proposed nominations (concept document) and the process continues after a candidate substance has been selected– e.g., developing the protocol for the cancer hazard evaluation and carrying out the steps outlined in the concept for obtaining relevant scientific and public input to inform the evaluation. During the planning stage, ORoC consults with relevant scientific experts, conducts preliminary literature searches and solicits public comments about specific nominations or candidate substances. All public comments received during the evaluation become part of the public record, are posted on the RoC website, and are considered by NTP and any external advisors during subsequent steps in the evaluation process. For more information on issues related to planning for specific sections of the monograph (such as human cancer studies), see the appropriate parts of this handbook.

Identifying candidate substances

Scoping

Initially, a scoping review is conducted to determine whether there is sufficient information on exposure and carcinogenicity to justify a formal assessment of the substance. ORoC identifies relevant information from various sources, such as authoritative evaluations (e.g., International Agency for Research on Cancer monographs), reviews in the peer-reviewed literature, and preliminary literature searches), other NTP scientists, and subject-specific technical advisors (both government and non-government. When the number of authoritative reviews is limited, more extensive literature searches (scoping reports) may be conducted to determine the available database on a substance. Interagency and public comments are solicited through a *Federal Register* notice to identify information about ongoing studies, recent publications, current production, use patterns, sources of exposure, names of scientific experts with relevant knowledge, and scientific issues important for assessing the carcinogenicity of the substance. Public comments received on the nominations are posted on the RoC website. The interagency and public comments are considered, and NTP selects nominated substances for which to develop a concept document (see below). During this period, additional literature searches may be undertaken that will be used in developing a draft concept document.

Concept Document

The concept document is a brief document that outlines the rationale for the nomination of the substance and the approach to conducting its review. The document includes an overview of information on exposure and the extent and nature of the scientific information. It also identifies key scientific issues, the scope and focus of the monograph and the approach to obtaining public and scientific input regarding these issues. The nature, extent, and complexity of the scientific information on a candidate substance guides the details of the approach used by NTP to evaluate the carcinogenicity of the substance. The approach is tailored to enable ORoC to obtain external advice and address scientific issues in assessing the carcinogenicity of a given candidate substance at various points throughout the process, and in the way that is most appropriate for each substance (e.g., through expert panels, *ad hoc* presentations, individual technical advisors or consultants, public input via listening sessions or comments, and/or interagency input). ORoC revises the concept document based on internal (NTP and interagency) input and shares it with its interagency partners (the National Institute for Occupational Safety and Health and the National Center for Toxicological Research), solicits public comments, and presents it to the NTP Board of Scientific Counselors (BSC) for review at a public meeting.

Selection of the Candidate Substance

Considering comments from the NTP BSC and the public, the NTP Director makes the final determination whether to add the substance to the list of candidate substances for RoC evaluation. Next, ORoC identifies technical advisors and forms a monograph planning team, which includes ORoC staff, contractor staff, NTP/NIEHS staff, and other government scientists. ORoC may also identify non-government scientists to serve as technical advisors.

Literature Search Strategy

With input from the scoping and concept-document processes, information specialists, the monograph planning team, and technical advisors, the ORoC develops a search strategy and

inclusion/exclusion terms, which are reviewed and implemented to identify relevant peer-reviewed studies in several databases. The search strategy is discussed more generally in Part B and more specifically for the various evidence streams (e.g., human, animal or mechanistic) in the relevant parts of this handbook. An appendix to this handbook that contains standard search strings used for monograph topics is available at http://ntp.niehs.nih.gov/go/rochandbook [Note: the literature search appendix will be posted on the ORoC website by November 2015]

Protocol Development

Once a substance is selected for formal review, ORoC, with input from the monograph planning team and technical advisors, develops a protocol that outlines the methods for preparing the draft monograph. It includes, but is not limited to, the literature search strategy, key issues, the focus of the document and the strategy for drafting it, methods for evaluating the quality of studies, and considerations for integrating the evidence to reach level-of-evidence conclusions. This handbook serves as the basis for developing the protocol; however, the protocol is adapted to address issues specific to the candidate substance. Protocol development for specific types of evidence streams (such as human cancer studies) is discussed in the relevant parts of this handbook.

Part B: Identification and Selection of Studies

Introduction and Objective

The objective of the literature search is to identify the literature that is relevant for evaluating the potential carcinogenicity of the candidate substance. In general, this includes literature on the following topics:

- properties (e.g., identification of the substance) and human exposure (focusing on the U.S. population, see Introduction, congressional mandate)
- disposition (absorption, distribution, metabolism, and excretion) and toxicokinetics in experimental animals and humans
- human cancer studies
- studies of cancer in experimental animals
- mechanistic data and other relevant effects
    - genotoxicity and related effects
    - mechanistic considerations

As discussed in Part A, the literature search and selection process is informed by the scoping, by development of the concept document, and by input from information specialists, the monograph planning team, and technical advisors. For some topics (such as human exposure), the RoC monograph may rely on authoritative reviews supplemented by key primary literature, whereas for others (such as cancer studies in experimental animals and humans), the monograph will rely on primary literature. The approach for using authoritative reviews will have been outlined in the concept document that was reviewed by the NTP Board of Scientific Counselors and released for public comments.

The methods for identifying the relevant literature, including the literature search strategy (Section 1) and the review of citations using web-based systematic review software (Section 2), are illustrated in Figure B-1 and discussed below. For each candidate substance, the specific procedures for conducting a literature search strategy and selecting literature will be provided in the protocol and the results of the literature search for each will be summarized in an appendix in the draft RoC monograph.

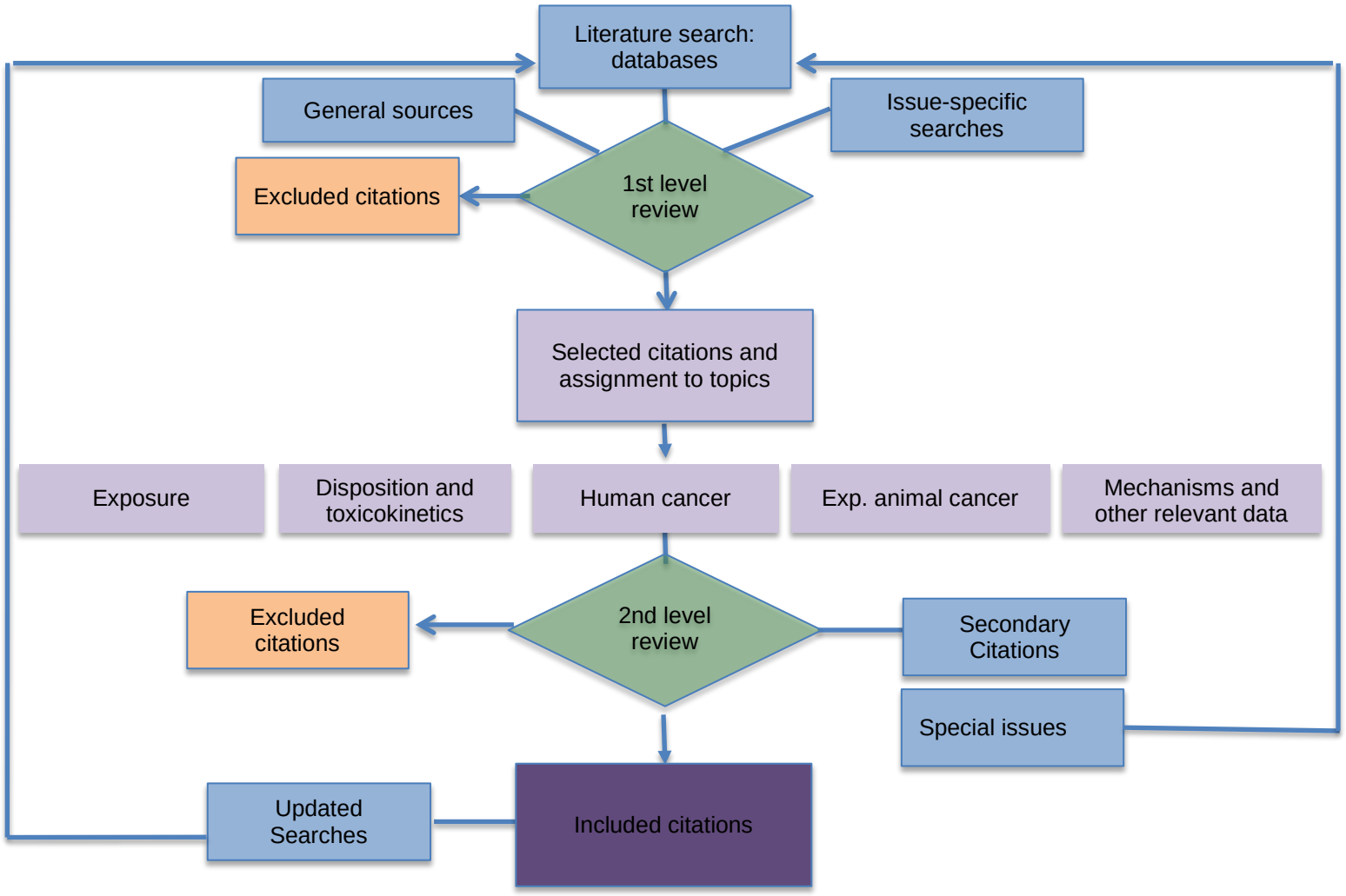


Figure B-1. Literature search strategy

1 Literature Search Strategy

Several approaches are used to identify relevant literature, including the following (shown as blue boxes in Figure B-1):

- ***Database searches:*** The major source for identifying relevant papers on the relevant topics (see below for more details).
- ***General sources:*** Examples include authoritative reviews, government reports, and web-based databases (see Section 3) and exposure-related data searches (see Part C, Human Exposure).
- ***Focused searches for specific scientific issues:*** Typically, issues that are identified at the beginning and during the literature-based review.
- ***Secondary citations:*** Citations identified from literature cited in authoritative reviews or in primary references located by the literature search.
- ***Updated searches:*** Literature searches are updated by either saving the search strategies and rerunning them or by creating monthly alerts in the appropriate databases (e.g., PubMed, Scopus, or Web of Science).

A fundamental step in developing a literature search strategy involves ongoing consultation with an information specialist to develop search terms (e.g., for the substance, end point, and type of evidence or topic) and to select the databases to be searched. Searches conducted in PubMed and at least one other database (such as Web of Science, Scopus, or Embase) are considered on a case-by-case basis, depending on the nature of the candidate substance and the topic. The publication dates searched are also considered on a case-by-case and topic-specific basis; for example, if the review of a specific topic is using information from an authoritative review, the literature search may be restricted to studies on that particular topic published since the review.

Literature searches of the databases generally use search terms for the substance combined with search terms for cancer and/or the types of evidence streams or topics (such as human exposure or epidemiological studies). For chemical substances, search terms usually include the candidate substance, its synonyms, trade name(s) when relevant, the metabolites of the substance, and the chemical class to which the substance belongs. Exposure scenarios or settings are also used for many types of substances or agents (including chemical substances) where exposure could occur in a specific occupational setting or through use of a specific consumer product. Titles, abstracts, and key words are searched. In addition, full-text searches of a custom-made library of specific types of studies (see Part D, Human Cancer Studies, for its current use to identify case-control studies of cancer and specific occupational exposures). These libraries were created in QUOSA (a literature management software) using search terms for the specific types of studies.

Before screening the literature, it is important to conduct "validation" analysis of the literature search; e.g., compare references obtained via the search strategy with "seed studies" (references obtained from authoritative reviews and other data sources, such as those described in Section 3) and to conduct additional literature searches using new search terms, as needed.

2 Screening and Selection of Literature

Citations retrieved from literature searches (and other sources) are uploaded to an EndNote library, and any duplicates are removed. Next, the EndNote library is uploaded to web-based systematic review software such as DistillerSR from Evidence Partners or Health Assessment Workspace Collaborative (HAWC) for multi-level screening using inclusion/exclusion criteria.

In Level 1 screening, the citations are screened based on the title and abstract (where available) by two screeners (members of the monograph planning team), to eliminate studies or articles that do not contain information on the candidate substance or on any of the key topics or questions covered by a monograph (exposure, cancer studies in humans and animals, toxicokinetics, genotoxicity, toxicity, or mechanisms of action). The initial screen is designated as "liberal": it is intended to retrieve a PDF if there is any reasonable possibility that it contains information that could be useful for the review process, and a positive response by only one of the reviewers is sufficient to pass a publication on to the next review level. The initial reviewers assign (or tag) the citation to one of the topic(s) or sections covered by a monograph (see Figure B-1).

In Level 2 screening, the PDFs (i.e., articles) obtained for all citations not excluded at Level 1 are screened by two topic-specific experts, typically the writer and scientific reviewer of the monograph section, using inclusion/exclusion criteria. These criteria are generally similar to the criteria used in Level 1 screening (e.g., information on the candidate substance and topic); however, Level 2 screeners can make more informed judgments about the relevance of the citations than the Level 1 screeners because they have the full texts in addition to titles and abstracts, and thus can sort and tag studies to the relevant topics. Citations at Level 2 may also be redistributed to other topics not identified at Level 1 screening. Depending on the topic, more specific inclusion/exclusion criteria (at either Level 2 or Level 3) may be developed, which are delineated in the protocol. Level 3 reviews, which are also screened by two reviewers) are generally limited to the human cancer and animal tumor studies; for example, in the human cancer section, criteria may be developed to exclude case reports or studies without exposure to the candidate substance or without clear exposure to the substance.

3 Data Sources

The following is a list of the major data sources that are usually searched for information on a specific candidate substance. The list includes authoritative reviews or study reports and web-based resources and/or databases. Sources that are specific for exposure information are in Part C of this handbook.

Biomedical literature databases

- PubMed (always)
- Web of Science
- Scopus
- Embase

Authoritative reviews and reports

- Agency for Toxic Substances and Disease Registry (ATSDR) Toxicological Profiles (http://www.atsdr.cdc.gov/toxprofiles/index.asp)
- California Environmental Protection Agency Proposition 65 hazard identification documents (http://www.oehha.ca.gov/prop65/hazard_ident/hazard_id.html)
- U.S. Environmental Protection Agency (EPA) Integrated Risk Information System (IRIS) (http://cfpub.epa.gov/ncea/iris/index.cfm?fuseaction=iris.showSubstanceList)
- European Chemicals Agency Risk Assessments (http://echa.europa.eu)
- Health Canada Environmental Health Assessments (http://www.hc-sc.gc.ca/index-eng.php)
- International Agency for Research on Cancer (IARC) Monographs (http://monographs.iarc.fr/ENG/Monographs/PDFs/index.php)
- New York State Department of Health — Health Topics A to Z (http://www.health.ny.gov/healthaz/)
- National Academy of Sciences reports and publications (http://www.nationalacademies.org/publications/)
- NTP publications, including, but not limited to, technical reports, nominations for toxicological evaluation documents, RoC, RoC background documents or monographs, and NTP Office of Health Assessment (OHAT) (formerly CERHR) monographs (http://ntp.niehs.nih.gov; search NTP)
- World Health Organization (WHO)/United Nations Environment Programme (UNEP) International Programme on Chemical Safety (IPCS) INCHEM-related documents (http://www.inchem.org/)

Databases or web resources

- Carcinogenic Potency Database (http://toxnet.nlm.nih.gov/cpdb/)
- European Chemicals Agency (http://echa.europa.eu/)
- European Food Safety Authority (http://www.efsa.europa.eu/en/publications.htm)
- International Labour Organization (http://www.ilo.org/global/publications/lang--en/index.htm)
- International Uniform Chemical Information Database (http://iuclid.eu/)
- National Institute for Occupational Safety and Health (NIOSH) Publications (http://www2.cdc.gov/nioshtic-2/)
- United Nations Environment Programme (www.unep.org)
- U.S. National Library of Medicine (NLM) TOXNET (http://toxnet.nlm.nih.gov)

Part C: Evaluation of Human Exposure Data

Introduction and Objective

The objective of the human exposure section of a monograph for a specific substance is to provide information to determine whether a significant number of persons residing in the United States are exposed to a substance, as required by the congressional mandate (see Introduction). The NTP also considers past exposure as fulfilling this criterion because, in part, of the long latency for many types of cancer. The congressional mandate does not provide guidance to interpret "significant" and information on numbers of exposed individuals is rarely available. However, the potential for exposure can be inferred from other types of information, such as that on use and production; occurrence in the environment, workplace, food, or consumer or medical products; and exposure levels in people. The monograph also discusses how people are exposed to the substance and, where relevant, presents information on exposure scenarios that may inform the evaluation of the human cancer studies. The monograph does not conduct a formal exposure assessment or make conclusions about hazards for levels of exposure.

Key questions

Primary question

- Is there exposure to a significant number of persons living in the United States to the candidate substance and, if so, what is the evidence to support this conclusion?

Secondary questions

- How should the candidate substance be defined so it is represents the substance that humans are exposed to?
- What are the properties (such as chemical, physical, or biological) of the substance?
- How are or were people exposed to the candidate substance (sources, settings, levels, frequency, trends)?
- What federal regulations and guidelines limit (or potentially limit) exposure?

This handbook provides instructions on the type of information to include and the organization to follow in drafting the section on human exposure. The approach may vary somewhat depending on the nature and complexity of the substance.

1 Planning and Literature Search Strategy

As discussed in the Introduction, the first steps in writing the monograph are to conduct research to identify the key issues and to identify technical advisors, as needed. This approach will help inform the development of the literature search strategy and the methods for preparing the document. The RoC policy is that exposure information must be publicly available but need not be peer reviewed.

Part of the search strategy involves searches of the following and any other relevant online sources:

- American Conference of Governmental Industrial Hygienists (ACGIH) Threshhold Limit Value/Biological Exposure Indices (TLV/BEI) documentation (available for purchase) (https://www.acgih.org/store/BrowseProducts.cfm?type=cat&id=16)
- ATSDR Toxicological Profiles (http://www.atsdr.cdc.gov/toxprofiles/index.asp)
- Chem Sources Suppliers (http://db.chemsources.com/login.php)
- EPA AP-42, Compilation of Air Pollutant Emission Factors (http://www.epa.gov/ttnchie1/ap42/)
- EPA Chemical Data Reporting (http://www.epa.gov/oppt/cdr/index.html)
- EPA Enforcement and Compliance History Online (ECHO) database (http://www.epa-echo.gov/echo/)
- EPA EJView Database (http://epamap14.epa.gov/ejmap/entry.html)
- EPA High Production Volume (HPV) Challenge Program Chemical List (http://www.epa.gov/hpv/pubs/update/hpvchmlt.htm#download)
- EPA Chemical Data Reporting (CDR) (http://www.epa.gov/oppt/cdr/index.html)
- EPA Locating and Estimating (L&E) Documents — Locating and Estimating Air Toxic Emissions from Sources of (source category or substance) (http://www.epa.gov/ttnchie1/le/)
- EPA/Office of Pesticide Programs (OPP) National Pesticide Information Retrieval System (http://npirspublic.ceris.purdue.edu/ppis/)
- EPA Toxics Release Inventory (http://www.epa.gov/triexplorer)
- U.S. Food and Drug Administration (FDA) Orange Book: Approved Drug Products with Therapeutic Equivalence Evaluations (http://www.fda.gov/cder/ob/default.htm)
- FDA Pesticide Program Residue Monitoring (http://www.fda.gov/Food/FoodborneIllnessContaminants/Pesticides/UCM2006797.htm)
- FDA Total Diet Study (http://www.fda.gov/Food/FoodScienceResearch/TotalDietStudy/ucm184293.htm)
- IHS CyberRegs (http://www.cyberregs.com/)
- Kirk-Othmer Encyclopedia of Chemical Technology (online access through the NIEHS Library)
- Material Safety Data Sheets (MSDS) (http://www.msdsxchange.com/english/xchange_search.cfm)
- National Health and Nutrition Examination Survey (NHANES) (http://www.cdc.gov/nchs/nhanes.htm)
- NIOSH Health Hazard Evaluations (http://www2a.cdc.gov/hhe/search.asp)
- NLM TOXNET: ChemIDplus, Hazardous Substances Data Bank (HSDB), Haz-Map, Household Products Database, TOXMAP (http://toxnet.nlm.nih.gov)

- National Occupational Exposure Survey (NOES) (1981 to 1983) (http://www.cdc.gov/noes/noes4/agtindx3.html)
- Ullmann's Encyclopedia of Industrial Chemistry (http://onlinelibrary.wiley.com/mrw/advanced/search?doi=10.1002/14356007)
- U.S. Air Force Defense Meteorological Satellite Program https://catalog.data.gov/dataset/defense-meteorological-satellite-program-dmsp
- U.S. Coast Guard National Response Center (http://www.nrc.uscg.mil/Default.aspx)
- U.S. Department of Agriculture Pesticide Recordkeeping Program http://www.ams.usda.gov/AMSv1.0/ams.fetchTemplateData.do?template=TemplateQ&navID=PesticideRecordkeepingProgram&rightNav1=PesticideRecordkeepingProgram&topNav=&leftNav=ScienceandLaboratories&page=PesticideRecordkeepingProgram&resultType=
- U.S. Department of Labor, Bureau of Labor Statistics (BLS) (http://www.bls.gov/)
- U.S. Geological Survey Minerals Yearbook (http://minerals.usgs.gov/minerals/pubs/myb.html) and Commodity Sheet Summaries (http://minerals.usgs.gov/minerals/pubs/mcs/)
- U.S. International Trade Commission (USITC) Interactive Tariff and Trade DataWeb (import/export data) (http://dataweb.usitc.gov/scripts/user_set.asp); Schedule B Codes for USITC Database Query (http://www.census.gov/foreign-trade/schedules/b/index.html)
- U.S. Patent and Trademark Office Patent Search (http://www.uspto.gov/patents-application-process/search-patents); Trademark Electronic Search System (TESS) (http://tmsearch.uspto.gov/bin/gate.exe?f=tess&state=4804:91j259.1.1)
- WHO/UNEP IPCS INCHEM-related documents (http://www.inchem.org/)

Information from these sources is supplemented by literature reviews or exposure studies in the primary literature identified from these sources' citations lists and through literature searches of databases (PubMed and typically Scopus or Web of Science). Searches typically use search terms for the candidate substance combined with search terms related to exposure information (see Table C-1 for examples of search terms, including text words and Medical Subject Headings (MeSH) terms). Search terms for the candidate substance may be chemical synonyms or exposure scenarios associated with exposure to the specific substance. The former are usually identified from NLM databases (e.g., ChemIDplus, HSDB), and exposure scenarios are identified from secondary sources. North American Industry Classification System (NAICS) codes reported by industry to TRI and EPA Industrial Sector (IS) codes reported by industry to CDR are used to identify uses that may be potential exposure scenarios for the candidate substance.

Table C-1. Examples of concepts for searches for exposure information

PubMed, Scopus, and Web of Science	MeSH terms used in PubMed
exposure, occurrence	environmental pollutants
oral, dermal, inhalation	environmental pollution
air, water, food, soil	occupational exposure
environmental pollution	
environmental exposure/monitoring	
occupational exposure/monitoring	

Note that these are examples of search terms and not the detailed or fully developed search strings used in actual literature searches.

Citations retrieved from literature searches are uploaded to web-based systematic review software and screened by two reviewers using predefined inclusion/exclusion criteria. Exposure information should be specific for the substance. Studies are initially included in the review if they meet the following inclusion criteria:

- provide information on use and production
- provide information for interpreting biomonitoring data where relevant
- provide occurrence and exposure data (such as levels in the environment or workplace, food, consumer, or medical products; toxics release data; or biomonitoring data)

2 Section Contents and Approach to Drafting

Exposure information usually comes from secondary sources supplemented with primary studies that provide key exposure information. A comprehensive and formal exposure assessment is beyond the scope of this review, and the objective of the section is to succinctly summarize the relevant exposure information. The exposure section usually consists of subsections on the topics listed below, although the organization may change depending on the available database for each candidate substance. Preferably, each subsection should clearly state the conclusion on the topic (such as occupational exposure) and provide a concise summary of the data and information that support the conclusion. As appropriate, data should be visualized in figures and graphs or presented in tables (see Section 3).

Substance identification and properties

- Defines the substance and provides information on chemical and physical or biological properties (see Section 3.1 for examples of table templates).

Use- and production-related data

- Provides information on present and past uses and identifies which are the most widespread or important.
- Provides information on present and past production, export, import, or consumption (see Section 3.1 for an example of a table template).
- Provides information on trends in use or production over time (see Section 3.2 for an example of a graphic to visualize data).

Exposure levels and biological indices of exposure

- Provides information related to interpreting biological indices used in exposure studies.
- Provides data on levels of the substance (or metabolite when relevant) in human tissues or samples measured in studies such as NHANES.
- Provides information on modeled intake levels from various environmental, occupational, or other sources.

Occupational exposure

- Provides information on types of industries, exposure levels, and exposure trends The data may be extracted into a database or a web application (such as Table Builder, which is a custom-made database) that can be used to generate Word tables. (See Section 3.2 for an example of a graphic to visualize data.)
- Provides information on protection measures to limit exposure.

Non-occupational sources of exposure

- Provides information on present or past exposure from the environment, such as releases to the environment and levels in air, water, and soil.
- Provides information on present or past exposure from and/or levels in food, tobacco smoking, and consumer and medical products.

Synthesis of information and conclusions

- Summarizes the data that support a conclusion on whether a significant number of people residing in the United States are (or were) exposed to the candidate substance.
- Summarizes the major sources of exposure to the candidate substance.
- Discusses whether exposure sources, routes, levels, or patterns have changed over time.
- Discusses whether there are changes in the exposure agent due to the environment; for example, the agent in the work setting may undergo biotransformation or degradation in the general environment. Thus, not only will there be the generally understood quantitative differences across scenarios, but the nature of the exposures to workers and the general public may be qualitatively different, as well.

Regulations and guidelines

- Lists regulations from the U.S. regulatory agencies, such as the Consumer Products Safety Commission, FDA, Department of Agriculture, Department of Transportation, EPA, or Occupational Safety and Health Administration.
- Lists occupational guidelines (if relevant), such as those published by ACGIH and NIOSH.

Regulations and guidelines for each candidate substance are identified from the following searches:

- Website searches of 30 U.S. government agencies, health agencies, and other authoritative sources identified from substances currently listed in the RoC. Examples of

websites that report on multiple regulations or guidelines for their agencies include the Consolidated List of Lists for EPA regulations (http://www2.epa.gov/epcra/consolidated-list-lists) and the NIOSH Pocket Guide to Chemical Hazards (http://www.cdc.gov/niosh/npg/) for NIOSH-recommended occupational exposure limits.

- IHS CyberRegs (http://www.cyberregs.com/), a subscription-based software package providing simultaneous access to the content of all 50 titles of the Code of Federal Regulations

The regulations and guidelines identified for each candidate substance are also added to a database of regulations (over 7,000 entries) for all substances listed in the 13th RoC.

3 Examples of Table Templates and Figures

The following are examples of table templates and figures that have been used in past RoC monographs (NTP 2013, 2014a,b, 2015) or have been newly created. Current plans are to use database or web applications (such as Table Builder) to generate tables for some types of exposure information, such as occupational exposure.

3.1 Example table templates for property and exposure information

The following table templates are used to present information relative to chemical identification, physical and chemical properties, and production, import, and export data that is typically reported for candidate substances that are chemicals. Presentation of information on substances that are not chemicals (e.g., radiation, biological agents, exposure scenarios (such as shift work), mixtures) is decided on a case-by-case basis. Tables for other types of exposure data, such as ambient levels, biomonitoring studies, and toxic releases, are prepared on a case-by-case basis, because these types of data are more heterogeneous in nature.

Chemical identification

Characteristic	Information
Chemical Abstracts index name	
CAS Registry number	
Molecular formula	
Synonyms	

Physical and chemical properties

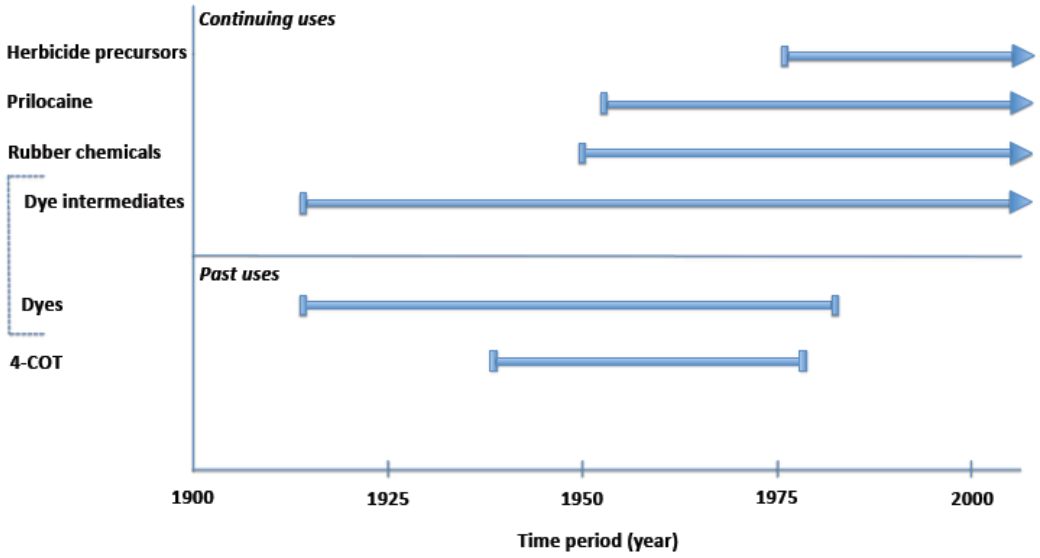
Property	Information
Molecular weight	
Density	
Melting point	
Boiling point	
Log K_{ow}	
Water solubility	
Vapor pressure	
Vapor density relative to air	
Physical state	

Production, export, and import data

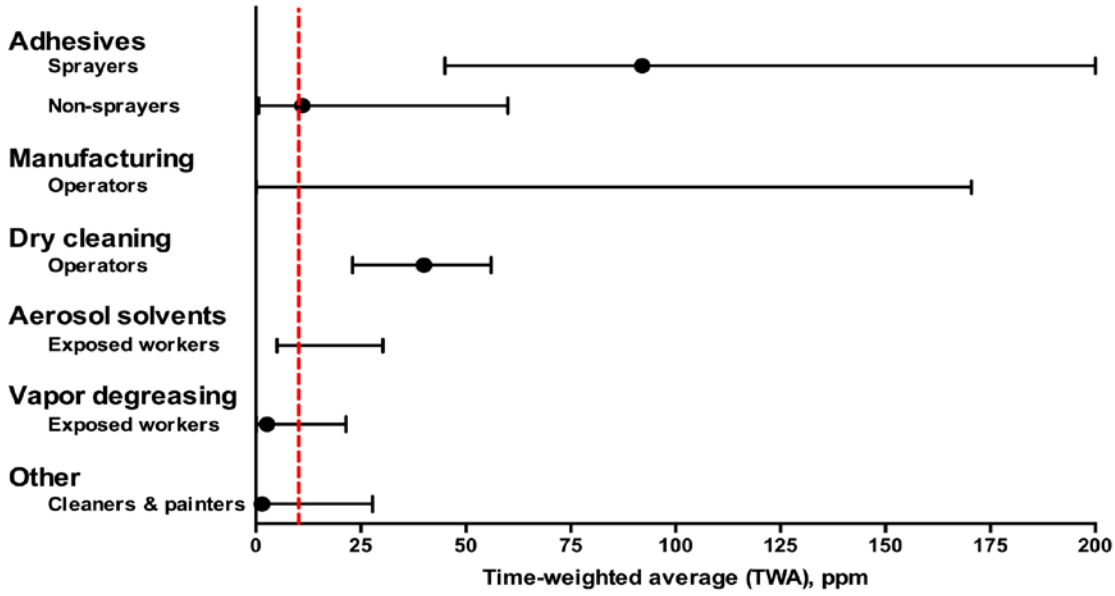
Category	Years covered	Quantity in pounds
U.S. production		
U.S. imports (recent)		
U.S. imports (historical)		
U.S. exports (recent)		
U.S. exports (historical)		

3.2 Examples of figures and graphs for visualizing exposure data

As appropriate, graphs and figures should be used to visualize exposure data in addition to or as an alternative to tables and text. The following figures from past monographs provide examples of how to visualize (1) changes in uses of *ortho*-toluidine over time (Figure 1-2 in NTP 2014b), (2) ambient air monitoring data for 1-bromopropane across different worker populations (Figure 1-2 in NTP 2013), and (3) biomonitoring data related to exposure to pentachlorophenol (Figure 1-2 in NTP 2014a).

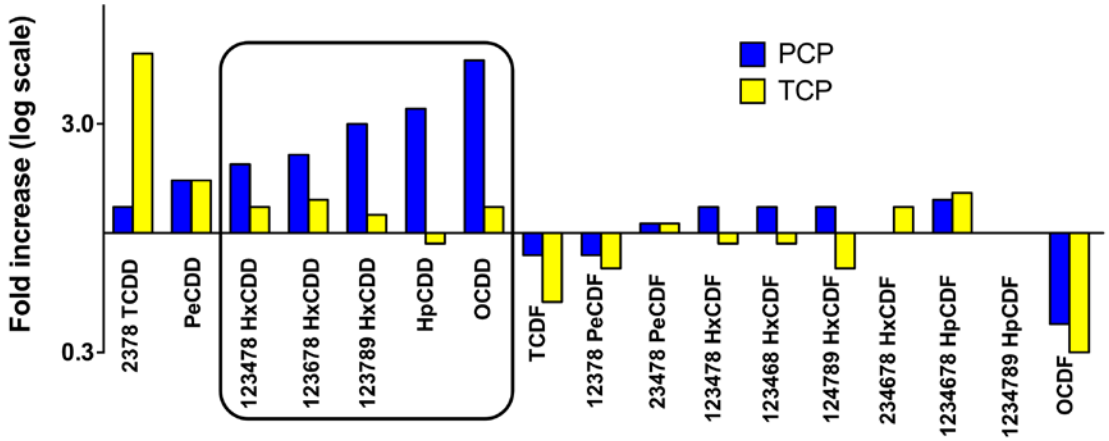


Timeline for continuing (current) and past uses of *ortho*-toluidine in the United States



TWA 1-bromopropane air concentrations across industry sectors

Time-weighted-average 1-bromopropane exposure levels as geometric means (adhesives, other, and vapor degreasing); arithmetic mean (dry cleaning); or not reported (manufacturing and aerosol solvents). The dashed vertical line represents the ACGIH threshold limit value – time-weighted average (TLV-TWA) of 10 ppm.



Dioxin congener patterns for PCP and TCP workers

Relative increase (or decrease) in serum levels of dioxin congeners compared with the reference population for pentachlorophenol (PCP) and trichlorophenol (TCP) workers. Values shown are for PCP-only workers and TCP-only workers. Samples were collected 26 to 62 years after occupational exposure.

Serum levels for dioxin congeners from workers exposed to PCP and TCP were divided by the values from the reference group of unexposed individuals (workers in the same plant who had no known exposure to chlorophenols). The horizontal line at "1" indicates equivalence with the reference group. Bars that extend below the line indicate a lower value for the exposed group than for the reference group. The rectangle drawn around the HxCDD, HpCDD, and OCDD congeners identifies the congener pattern used to distinguish workers exposed to PCP and TCP.

Part D: Evaluation of Human Cancer Studies

Introduction and Objective

This part of the handbook describes the methods and considerations for conducting a systematic cancer hazard evaluation of the evidence from human (epidemiologic) studies for review of a candidate substance for the RoC. As per the introduction, substance refers to agent, substance, mixture, or exposure circumstance. This evaluation includes identifying and reviewing the relevant studies, assessing their utility for informing the hazard evaluation, interpreting their results, applying the RoC listing criteria (below) to the evidence from the studies, and reaching a conclusion about the level of evidence (sufficient, limited, or inadequate) for the carcinogenicity of a candidate substance from studies in humans. The key scientific questions and major steps in the cancer hazard evaluation are described below.

Detailed methods for conducting the evaluation follow this introduction, and examples of tables and figures are provided in Section 6. The approach to the cancer hazard evaluation of human studies is based primarily on the protocols used to prepare RoC monographs on *ortho*-toluidine, pentachlorophenol and by-products of its synthesis, and trichloroethylene. Other resources include a recent Cochrane risk of bias tool developed for non-randomized studies of interventions (e.g., observational) (ACROBAT-NRSI, as reported by Sterne *et al.* 2014), publications of epidemiological methods (cited in the text), the Preamble to the IARC Monographs (IARC 2006), and input from technical advisors (epidemiologists) and other scientists developing systematic review procedures.

RoC listing criteria for evaluating carcinogenicity from studies in humans

- ***Sufficient evidence of carcinogenicity from studies in humans:*** indicates a causal relationship between exposure to the agent, substance, or mixture and human cancer.
- ***Limited evidence of carcinogenicity from studies in humans*****:** a causal interpretation is credible, but alternative explanations, such as chance, bias, or confounding factors, could not adequately be excluded.

Key questions

Primary questions

- Is there a credible association between exposure to the substance and cancer (site specific or all cancer(s) combined)?
- If so, can the association between exposure to the substance and cancer endpoints be explained by chance, bias, or confounding?

Secondary questions

- Which epidemiologic studies should be included in the review?
- What are the potential confounders for cancer risk for the tumor sites of interest in these studies?
- What are key issues for evaluation of the studies?
- What are the methodological strengths and limitations of these studies?

Components of the literature-based cancer hazard assessment

The components of the literature-based assessment are illustrated in Figure D-1, and procedures and considerations for each component are described in Sections 1 through 5.

Conducting a cancer hazard assessment for a candidate substance begins with a significant phase of planning and literature-based research (see the Introduction). The planning consists of identifying technical advisors with relevant expertise, conducting background literature searches to identify scientific issues, consultation with an informational specialist to develop the literature search strategy, and obtaining scientific and public input via webinar, information group or other mechanisms relevant to these issues. The planning process is necessarily iterative, is important in development of the protocol, and overlaps with the literature search and assessment of study utility. "Utility evaluation" refers to the evaluation of study quality (potential for biases) and study sensitivity (see Section 4); it is used to identify which studies are the most informative and facilitates the interpretation of the studies' findings (e.g., confidence in the effect estimates). The cancer hazard assessment consists of interpretation of the individual studies and integration of the evidence across studies to reach a preliminary level-of-evidence conclusion.

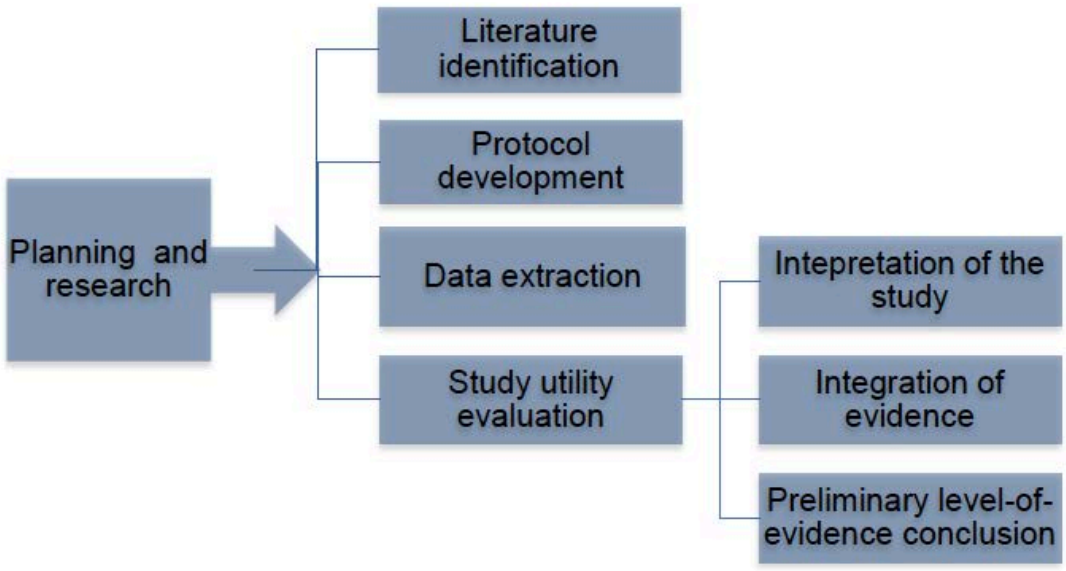


Figure D-1. Components of the literature-based cancer hazard assessment of the human studies

1 Identification and Selection of the Literature

The cancer evaluation component of the draft monograph evaluates all the relevant epidemiologic studies that have assessed exposure to a specific candidate substance and a cancer outcome. As per the RoC process, the studies must be peer reviewed and publicly available (both English and non-English papers are considered). The first step in the process is to develop a literature search strategy, in consultation with an information specialist, and associated

inclusion/exclusion criteria to identify the relevant literature. The second step is to select the primary epidemiologic studies from this database.

Part B of this handbook discusses general procedures used to identify and select relevant literature for preparing the RoC monograph. This section discusses the general literature search strategy and inclusion/exclusion criteria specific for identifying studies for the human cancer section, including (1) primary epidemiologic studies, which form the basis for the cancer evaluation, and (2) supporting literature (e.g., exposure assessment studies) that may be relevant for interpretation of the epidemiologic studies. Recent meta-analyses are also included in the evaluation.

Searches are conducted in PubMed and at least one other database (such as Scopus or Web of Science), depending on the candidate substance, using search terms for the candidate substance and exposure scenarios related to the candidate substance combined (using the Boolean operator "AND") with search terms for epidemiologic studies and with search terms for the outcome (i.e., cancer). Table D-1 lists some general search terms used in most evaluations to identify epidemiologic and cancer studies.

Table D-1. Examples of concepts used in searches for human cancer studies

PubMed, Scopus, and Web of Science		MeSH terms used in PubMed	
Epidemiology terms	**Cancer terms**	**Epidemiology terms**	**Cancer terms**
case-control	cancer	epidemiological studies	Neoplasms
cohort	leukemia	epidemiological methods	
case-referent	lymphoma	occupational exposure/ adverse effects[a]	
case-report	"lymphohematopoietic cancer"[a]	epidemiology[subheading]	
case-series	"multiple myeloma"	etiology[subheading]	
epidemiology	neoplasm		
meta-analysis [publication type]	tumor		
workers			
workmen			
ecological study			

Note that these are examples of search terms and not the detailed or fully developed search string used in the actual literature search.

[a]More specific search terms for lymphohematopoietic cancer may be developed for specific candidate substances.

Relevant literature may also be identified from sources such as authoritative reviews and citations from identified publications, and searches may also be conducted on specific topics. In addition, full-text searches of a custom library may be conducted (such as, the QUOSA scientific literature management software, http://www.elsevier.com/online-tools/quosa) of PDFs of occupational case-control studies, created from past searches of three databases (PubMed, Scopus, and Web of Science) using search terms for occupational exposure and case-control studies.

Citations retrieved from literature searches are uploaded to a web-based systematic review software application and screened by two reviewers using pre-defined inclusion/exclusion criteria. Relevant literature includes primary studies, meta-analyses, and publications with

supporting information, such as those describing methods for exposure assessment. In general, primary studies may be excluded if they (1) do not adequately evaluate exposure specifically to the candidate substance or (2) do not evaluate health effects related to carcinogenicity. Inclusion of studies such as case reports, case-series, or ecological studies is decided on a substance-by-substance basis and will be delineated in the protocol.

2 Initial Literature Review and Protocol Development

At this stage, a brief review of the literature is warranted — noting, for example, the types of studies and exposure assessments — as the basis for deciding which issues or questions need to be addressed in the review. Searches at this stage are typically open-ended regarding cancer end points, except in the case of substances for which there are authoritative reviews (e.g., recent IARC Monographs or National Research Council Reports) that specify the end points of interest for the substance. If only one or two studies exist for a particular cancer site, a decision may be made to exclude that particular cancer end point. In addition, studies or groups of studies that clearly have little utility for the cancer hazard evaluation may be excluded, such as case reports or studies that are not specific for the exposure of interest (for example, drycleaner studies were excluded in the review of trichloroethylene and cancer). In some cases (such as for rare diseases), case reports or case-series may be informative.

Developing the protocol requires understanding what types of studies will be available to inform the hazard assessment. The protocol is written to provide detailed considerations for evaluating study exposure and outcome metrics, co-exposures, the methodologic quality of the study, and potential biases that may be important in evaluating the findings for the hazard evaluation. Protocol development will require background research on the substance, the cancer, and co-exposures and their measurement, taking into consideration input from subject-matter and methodologic experts.

2.1 Identify potential covariates or co-exposures

A key question in the evaluation of the level of evidence conclusion from observational studies is whether any association between the exposure and the potential carcinogen can be explained by confounding. Potential confounders include risk factors that could be associated with *both* exposure to the substance under review *and* the disease outcome(s) of interest and that are not part of the disease pathway. A factor that is not related to the outcome of interest is not considered to be a confounder.

Potential confounders or co-exposures may be quantified or noted by the study authors or may be known from authoritative sources or literature reviewed during the planning phase. Information on occupational co-exposures may be less of a concern in population-based or hospital-based case-control studies, because of the typically low percentage of occupational exposures and broad diversity of jobs found among these study participants. Whether a given co-exposure should be considered as a potential confounder depends on whether there is evidence that the co-exposure is potentially associated with a specific cancer(s) of concern. Potential confounders are also likely to be identified through the expert knowledge of members of the review group or in initial reviews of the literature. Directed acyclic graphs may be used in some cases to help identify potential confounders and also identify whether confounders were controlled for correctly in the analyses. If covariates were included in a multivariable model but should *not*

have been controlled for in an analysis; this may lead to bias in the results (Greenland *et al*. 1999).

2.2 Conduct background research on exposure and outcome metrics

Exposure metrics

A detailed understanding of metrics used to characterize exposure and disease is necessary to assess the quality and utility of studies that contribute to the cancer hazard evaluation.

Interpretation of study findings may be altered by the type of exposure assessment, such as questionnaire data, monitoring data (e.g., ambient or personal air levels or biological monitoring), or for occupational studies, job or job-task exposure matrices (JEM or JTEM), or expert assessments that link monitoring data or job processes to the subject's occupational history (e.g., job or department titles, task descriptions, duration of employment, or calendar years worked).

Searches for information such as environmental scenarios related to exposure, consumer products and uses, production methods, anticipated levels of exposure to the substance, and interpretation of various exposure metrics, such as intensity, duration, or calendar years employed, should be conducted in the context of the particular type of study. For example, for studies assessing exposure by biological markers, it is important to know the specificity of the biomarker for the exposure of interest and for timing of exposure. Researching information on relevant time windows of exposure relevant to the disease endpoint is also important.

The various methods of exposure assessment (e.g., JEM, in-person data collection, or proxy data collection) may have implications for the interpretation of the findings; thus, it is important during this phase to obtain knowledge about various methods used in the studies of interest, which can be used in protocol development.

Outcome metrics

Prior to the evaluation, it is important to understand the methods used to obtain vital status or cancer incidence, the expected rates of cancer mortality or incidence for the end points of interest, and the implications of long or short survival rates for interpreting the use of mortality or incidence rates in a study. Similarly, changes in diagnostic methods and criteria and coding systems for cancer over time may have implications for various subtypes of cancer (especially some of the lymphohematopoietic cancers), and any such changes should be understood prior to the assessment. Also, the latency period between exposure and the diagnosis of cancer, which can differ among various types of cancers, should be researched prior to the evaluation, to provide a basis for understanding the sensitivity of the study to detect the occurrence of cancer.

3 Systematic Extraction of Data from the Epidemiologic Studies

Two independent reviewers extract data (such as methods and results) from the individual studies into a database or web application (such as Table Builder) in a systematic manner using standardized instructions and questions. The database contains fields that are specific for the various types of extracted information (such as study population characteristics, exposure and disease assessment, analytical methods, and results). The instructions for data extraction (questions and considerations) describe the specific type of information that should be

summarized or entered into each field. The fields from the database are used to populate tables for the monograph. (See Section 6 for examples of tables for extracted data on population characteristics and methodologies.)

For studies in which multiple updates or re-analyses have been published, the reviewer usually extracts data from the most recently published follow-up or update for each cancer end point included in the study. If there is overlap between the study populations, the publication with the most complete or relevant follow-up of the study population is usually reported. Information (such as exposure data or reanalyses) from relevant publications may also be included in the review if it is needed to assess the study.

Quality assurance of data extraction and database entry are accomplished by (1) review of each data entry by an independent reviewer and (2) resolution of any discrepancies by mutual discussion with reference to the original data source.

4 Assessment of the Utility of the Individual Epidemiologic Studies

This section describes the assessment of the utility of the individual studies, including an overview of the approach (Section 4.1), considerations in assessing each type of bias or other factors related to study utility (Section 4.2), and considerations in reaching an overall judgment on the utility of each study to inform the cancer hazard evaluation. This step is completed prior to the cancer hazard evaluation. (See Section 6 for examples of tables and figures for reporting on study utility.)

4.1 Overview of the approach for assessing study utility

For the purpose of this documents study utility (i.e., informativeness) is defined as the ability to inform the cancer hazard evaluation. Biases in observational studies are often classified into three major categories: (1) selection bias, (2) information bias, and (3) confounding (Rothman *et al.* 2008). In addition, studies should have adequate reporting methods (von Elm *et al.* 2007) and apply appropriate analytical methods for calculating effect estimates. Finally, studies with greater sensitivity to detect an effect (e.g., having adequate numbers of exposed cases, exposure levels, durations, ranges, windows of exposure, and lengths of follow-up) are also considered to be more informative for the evaluation, although studies with lesser sensitivity may not suffer from bias *per se.*

4.1.1 Domains for evaluation of study quality and sensitivity

Each primary study is systematically evaluated for its ability to inform the cancer hazard evaluation by two independent reviewers using five domains related to study quality and one domain related to study sensitivity (diagrammed in Figure D-2). The evaluation of the potential for bias in each domain is captured by a core question. Domains are similar to those used in previous evaluations of human studies for the RoC (such as *ortho*-toluidine, pentachlorophenol and byproducts of its synthesis, and trichloroethylene). Core questions are largely similar (with some exceptions) to those being developed for the U.S. EPA IRIS Toxicological Reviews. A series of signaling and follow-up questions are used to address specific issues related to the core question. These questions are concerns that epidemiologists usually consider for each type of bias and are not meant to be a checklist. Some of these concerns (such as the healthy worker

effect) could be considered in more than one domain, but are to be evaluated in only one domain for the cancer hazard evaluation.

The overall evaluation of study utility is derived from integrating the domain-level judgments. To determine if there is potential bias operating within a study, each characteristic of the actual study is compared with that of an “ideal” study for a specific end point and exposure (see Section 4.2). However, the potential for a given bias in a study does not necessarily mean that the findings of the study should be disregarded. When there is adequate information, a judgment is made on the direction of the potential bias (over- or under-estimate of the effect estimate, or unknown) and the potential magnitude of the distortion of the bias on the effect estimate although information related to the latter is rarely available.

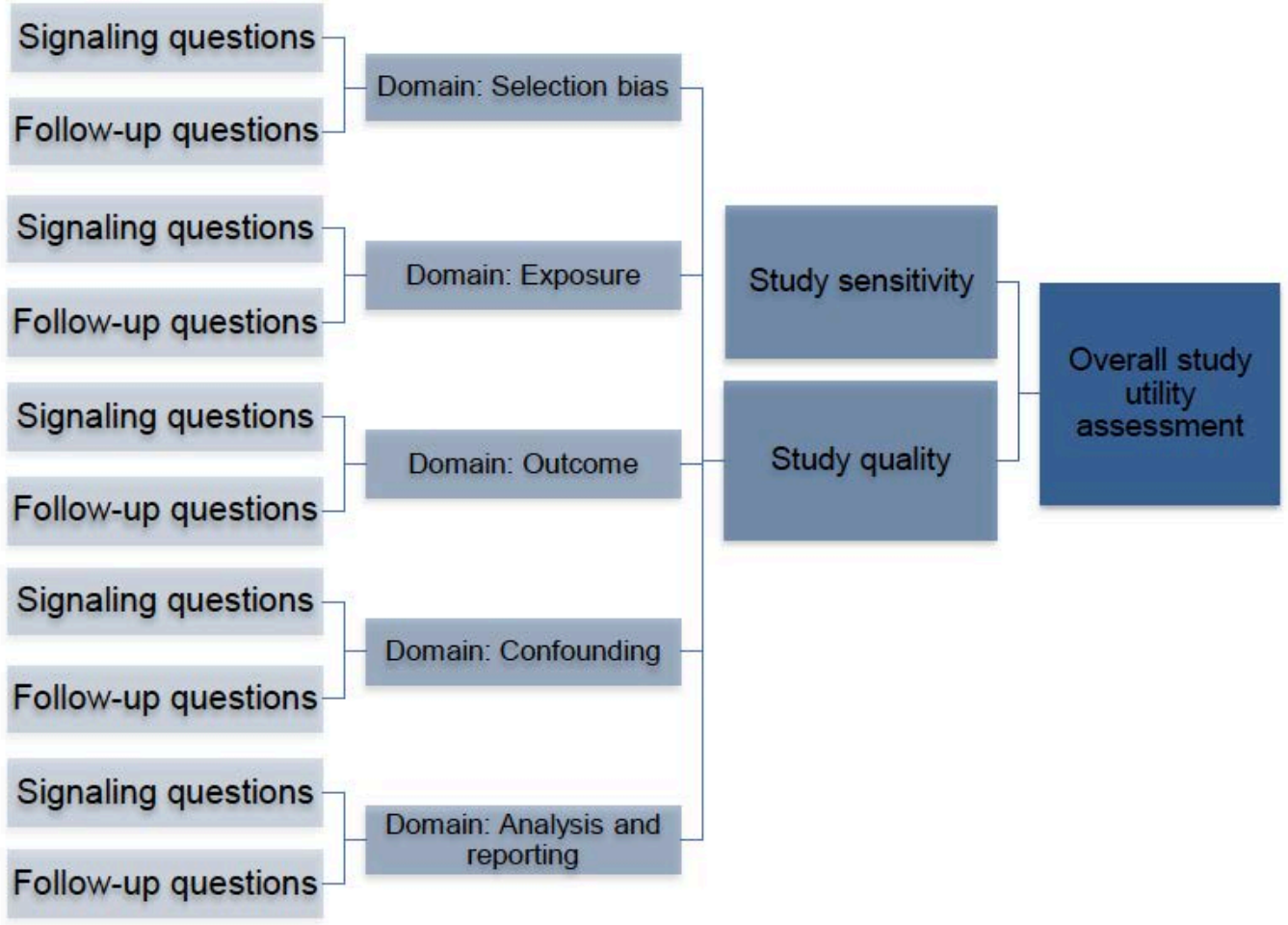


Figure D-2. Schematic of the approach to systematic review of study utility

A domain-level judgment is made for each domain, which is captured by a core question. The signaling and follow-up questions represent specific issues related to answering the core question.

This step is completed prior to interpretation of the individual study's findings and assessment of the evidence across studies. Differences are resolved by mutual discussion with reference to the original data source. A small subset of studies will be used in a "pilot" phase to discuss and resolve any ambiguity before proceeding with evaluation of the full set of studies. Study authors may be contacted if there is inadequate information to evaluate a signaling question. The approach of using signaling and follow-up questions for evaluating the potential for different types of bias and reaching conclusions about the quality of the studies, as well as some (but not all) of the study domains for quality, are somewhat similar to those used by other systematic review methodologies (e.g., Acrobat-NRSI, Sterne *et al.* 2014). Terms used in the evaluation are defined below, and the evaluation of the specific domains follows the scheme shown in Figure D-2. The overall evaluation of the utility of the study (including the judgment terms) is discussed in Section 4.3.

4.1.2 Domain-level judgment: Responses to core questions

The signaling and follow-up questions are used to provide transparency in answering the core question (e.g., domain-level judgment), rather than responding separately for each signaling or follow-up question. In some cases, a rating may not be possible due to the complexity of the issues and the discussion will be captured by narrative text. An example of this was the evaluation of the exposure assessment in the RoC Monograph on Trichloroethylene (NTP 2015). When adequate information is available, a judgment is made on the direction and distortion of each bias. The responses and general considerations are outlined below, and more specific considerations for each domain are discussed in Section 4.2.

- ***Low/minimal concerns:*** Information on the study design and methodologies indicates that they are close to the ideal study characteristics and that the potential for bias is low or minimal, recognizing the general limitations of observational studies. (+++, high quality)
- ***Some concerns:*** The study design or methodologies are less than ideal, indicating possible bias. (++, medium quality)
- ***Major concerns:*** The information on the study design or methodologies suggests that the potential for a specific type of bias is high. However, depending on the direction and distortion of the potential bias, the study may have some limited utility. (+, low quality)
- ***Critical concern:*** Distortion of bias would make the study findings unreliable for cancer hazard identification. ("0" rating)
- ***No information:*** The information in the study is inadequate to evaluate the level of concern for the domain.
- ***Direction of bias:*** ↑(away from the null or overestimate), ↓(towards the null or underestimate), not known.

4.2 Considerations in evaluating the potential for biases and confounding

4.2.1 Selection and attrition bias

Selection bias arises when study participants are not selected from the same underlying source population. This happens when the relationship between the exposure and disease is different for those who participated and for those who should have been eligible for the study, including those

who did not participate (Rothman *et al.* 2008). Selection bias can be a concern in any type of epidemiological study, but is most often a factor in case-control studies.

Case-control studies are at risk for selection bias because differing probabilities of selection for cases and controls are inherent in the study design (Pearce *et al.* 2007). To reduce bias, cases and controls should be selected from the same underlying population (or cohort, in the case of nested case-control studies) and should be representative of the population from which they are selected.

Apart from the use of inappropriate control groups, other types of selection bias in a case-control study include self-selection bias and Berkson's bias. Self-selection bias occurs when participants self-refer into a study, as the reasons for the self-referral may be associated with the outcome under study. Berkson's bias is often associated with hospital-based case-control studies and occurs when the controls are hospitalized for an exposure that is related to the disease of interest (Rothman *et al.* 2008).

Selection bias is usually less of a concern in cohort studies with complete recruitment and follow-up, as the cohort itself acts as the source population (Pearce *et al.* 2007). One type of selection bias that can occur in a cohort study is attrition bias. This can occur when follow-up of participants is incomplete and when loss to follow-up is related to both exposure and disease status. The quality of the case-ascertainment methods and the percentage of loss to follow-up are considered in the assessment. In cancer studies, the evaluation of completeness of follow-up often overlaps with outcome, because similar methods (such as the use of cancer registries) may be used for both.

In general, the evaluation of attrition bias considers methods of obtaining vital status and number of cases/deaths, but not the methods of diagnosis. Ascertainment of vital status usually relies on data such as death-certificate data, medical records, and/or cancer registry data, with medical records being the least preferred. In the United States (and other industrialized countries), death-certificate data, either in the form of Social Security or National Death Index files in the United States, are considered to be mostly complete. The completeness of cancer registry incidence data can vary by, for example, collection methods, region, and calendar period. The United States has no central national cancer registry, which makes it more difficult, especially for individuals who migrate to other states, to obtain complete follow-up information of a cohort.

Incomplete follow-up that is not related to both exposure and disease (nondifferential) can reduce the statistical power of the study.

An additional concern in occupational studies is the healthy worker effect (which can be considered as both a type of selection bias and confounding, but is addressed under selection bias). The healthy worker effect (HWE) includes both the selection of healthy workers into the workplace (healthy-worker hire effect [HWHE]) and the selection of unhealthy workers out of the workplace (healthy worker survival effect [HWSE]). The HWHE effect occurs when workers must meet minimal health criteria to begin working and thus are healthier than the general population. HWHE biases the findings towards the null and can be partially controlled for by conducting an internal analysis that compares the exposed workers with the unexposed workers instead of with the general population. HWSE may occur when healthier workers continue to work, whereas less healthy workers may transfer to jobs with lower exposures, take time off, or

leave work prior to disease or death. As a result, the unhealthy workers have the shortest employment duration, which may underestimate any exposure-response relationships. Controlling for (time-related) employment status may help reduce biases from HWSE (Pearce *et al.* 2007).

Cohorts that consist entirely of workers identified at one point in time (i.e., they include both prevalent and incident hires) have been found to overrepresent long-term healthy workers and underestimate disease prevalence. Left truncation, a concept that overlaps with HWSE, occurs when prevalent workers who are at risk for disease do not remain observable at the start of follow-up. Prevalent workers may be healthier and not representative of all workers hired before the start of the study. This bias can be corrected somewhat by restricting the study (or analysis) to incident or recent prevalent hires. In some cases, the bias can be analytically corrected with sufficient data (e.g., G-estimation, inverse-probability-of-treatment methods, and use of censoring weights). However, the variables needed to correct the bias are typically unmeasured or unavailable in most occupational studies. Although the direction of the bias from HWE is typically towards the null, its magnitude can be estimated given sufficient information on the proportions of prevalent workers and the length of the follow-up period. (For more information, see Pearce *et al.* 2007, Applebaum *et al.* 2011, and Picciotto *et al*. 2013).

Table D-2 describes the core, signaling, and follow-up questions and general considerations for assessing the potential for selection and attrition bias. More specific and complete considerations (e.g., for all rating categories) may be developed in the protocol for each candidate substance.

Table D-2. Selection and attrition bias: Questions and responses

Core question Is there concern that selection into the study (or out of the study) was related to both exposure and to outcome?		
Signaling questions	**Follow-up questions**	**Responses[a]**
Case-control studies Is there concern that cases and controls may not have been selected from the same underlying population during a similar time period? Are there concerns that eligibility criteria (inclusion/exclusion), recruitment strategies, or participation of cases and controls may have been related to exposure or disease status?	**Case-control studies** If there is concern about the potential for bias, what is the predicted direction or distortion of the effect estimate (if there is enough information)?	***Low/minimal concerns (+++) rating*** Cases and controls were selected from the same population by similar methods and criteria. There is no evidence that selection of the subjects was related to both exposure and disease. The cohort is clearly defined (e.g., includes the relevant exposed, non-exposed, or referent group for a specific time period/location), with no evidence that follow-up differed between exposed and non-exposed subjects. There is no evidence of HWE, or appropriate methods were used to address the potential bias. ***Critical concerns (0) rating*** There is strong evidence that selection or attrition of subjects was clearly related to both exposure and disease.
Cohort studies Is there concern about HWHE or that non-exposed subjects may not have been selected from the same underlying population during a similar time period?[b] Are there concerns about HWSE, prevalent hires, or left truncation and/or that follow-up time and start of exposure did not coincide (for diseases with short latencies)? Is there a concern that follow-up was incomplete?[c]	**Cohort studies** If there is concern about the potential for selection or attrition bias, what is the predicted direction or distortion of the effect estimate (if there is enough information)? If so, were appropriate analyses performed to address the potential bias?[c] If so, is there concern that completeness of follow-up is related to both exposure and disease?	

[a]Considerations for responses for other rating categories (e.g., "some" or "major") may be defined in the protocol for a specific candidate substance.

[b]HWHE and HWSE can also be considered as confounders, since the potential bias may be attenuated by using appropriate statistical analysis. However, since the evaluation involves some issues related to selection, these issues are usually considered as selection bias and are not evaluated in both domains.

[c]This evaluation includes consideration of the methods for case ascertainment. For cancer end points, follow-up methods often overlap with outcome assessment (e.g., mortality databases and death certificates are used for tracking both vital status and cause of death). Length of follow-up is considered in the evaluation of study sensitivity.

4.2.2 Exposure misclassification

One of the most important aspects of a study is the ability to correctly classify the study subjects (at the individual level) with respect to their exposure status. This involves an evaluation of the quality of the exposure assessment methods and information on the exposure setting.

Quantitative estimates of each individual's exposure to the substance of interest that use multiple metrics (such as cumulative, peak, and average intensity of exposure) are ideal. In addition, the exposure assessment methods should measure or ascertain exposure (or a surrogate metric correlated with exposure) that occurs prior to disease outcome and during a relevant window of exposure for the end point of interest.

In occupational studies, exposure assessment is often based on job- or job-task exposure matrices or expert assessments that link the subject's occupational history (e.g., job or department titles; task descriptions, including frequency; duration of employment; or calendar years worked) with plant or workplace exposure data (e.g., monitoring data, production methods or applications, or protection procedures) that are plant- and calendar-year-specific. Detailed information on job tasks, exposure setting, and any use of personal protective equipment improves the exposure assessment.

The assessment of environmental exposures in geographical or ecological studies would ideally rely on the likely sources of individual exposure levels (such as ambient levels of airborne or water pollutants, household dust, or residence in or near sources of environmental contamination) and/or biological monitoring data; in some cases, a range of relevant surrogate measures may improve the assessment. This assessment is often supplemented by the use of questionnaires to establish individual patterns of exposure (e.g., consumption of drinking water or duration of residence near a pollution source). Typically, quantitative or qualitative estimates of exposure based on aggregate measures of exposure are subject to considerable error for individuals, for both ever-exposure and the exposure categories used (e.g., in evaluating exposure-response relationships).

Assessment of other (i.e., non-environmental or occupational) exposures, such as biological agents, pharmaceuticals, and chemotherapy agents, ionizing or ultraviolet radiation, dietary contaminants and supplements, or lifestyle factors such as smoking, typically rely on a combination of one or more of medical and clinical data or records, biological monitoring (e.g., cotinine in urine), or participant questionnaires.

Toxicological, mechanistic, and other types of information related to the optimal time window of exposure for a specific type of cancer (taking into account the latency period) may also inform the evaluation of the exposure assessment. This evaluation overlaps somewhat with the assessments of the optimal length of follow-up (i.e., follow-up would begin with the appropriate window of exposure), statistical analyses, and study design.

In-person interviews are typically preferred over mailed or phone interviews, and information obtained from the subject is preferred over information from proxy respondents. However, for some sensitive variables (e.g., histories of illicit drugs, abortions, sexual behavior), mail or phone interviews have been more accurate and had less gender discrepancies than in-person interviews. Ideally, exposure-assessment investigators and interviewers should be blinded to the disease status of the study participants. Of these, the blinding of the investigators conducting exposure assessments is considered the most important; blinding of in-person interviewers may not be feasible, depending on, for example, the health of the subject with cancer. For studies using biomarkers, knowledge of the sensitivity and specificity of the method for measuring the biomarker in various biological media (e.g., urine, plasma, fat, or soft tissue) is important, together with the limit of detection. Some markers are non-selective and can also be markers for

other compounds. Knowledge of the time period over which the biomarker reflects dose is also important for determining the exposure period over which the biomarker is a valid indicator.

In general, exposure is better characterized in most occupational cohort studies than in geographical or ecological cohort studies or population- or hospital-based case-control studies. Misclassification of exposure in cohort studies is almost always nondifferential and usually results in a bias towards the null (i.e., an underestimate of the true risk). When there are more than two exposure categories, the direction of the bias is not always clear, but it may result in attenuation of the exposure-response relationship.

Recall bias is less likely to be a concern in case-control studies in which occupational exposure is assigned based on job titles, occupations, work history, or other types of occupational data than in studies using self-assessment of chemical-specific exposures or other types of exposures (e.g., use of questionnaires with exposure checklists). For self-reported exposure, recall bias is often differential and biases towards an overestimate of the effect; however, it can also be nondifferential. “Reverse causality” may a concern for case-control studies that measure exposure after disease diagnosis.

Table D-3 describes the core, signaling, and follow-up questions and general considerations for assessing the potential for exposure misclassification. More specific and complete considerations (e.g., for all rating categories) may be developed in the protocol for each candidate substance.

Table D-3. Exposure misclassification: Questions and responses

Core question Is there concern that the exposure assessment methods did not distinguish between exposed and non-exposed people or among exposure categories at a relevant time window of exposure?		
Signaling questions	**Follow-up questions**	**Responses**[a]
Is there concern that the subjects were misclassified with respect to ever-exposure? Is there concern that the exposure classification did not capture the variability of exposure? Is there concern that the exposure assessment did not capture the relevant time window or metric of exposure?[b] Is there concern that knowledge (e.g., observation or recall bias) or presence of the outcome (e.g., reverse causality) for exposure may potentially bias the exposure assessment? Is there concern that missing exposure data (including methods used to input data) may have resulted in exposure misclassification?	Did any misclassification vary by exposure category? If there is concern that there is exposure misclassification, is it differential or nondifferential, and what is the predicted direction or distortion of the effect estimate (if there is adequate information)?	***Low/minimal concerns*** ***(+++) rating*** The exposure assessment methods have good sensitivity and specificity, leading to reliable classification (or discrimination) with respect to ever-exposure, exposure level, timing, or other relevant metrics. Alternatively, the exposure assessment methods may be less than ideal, but detailed information on exposure setting allows for discrimination between exposed and non-exposed and among exposure categories. ***Critical concerns*** ***(0) rating*** Exposure assessment is not at the individual level or is not likely to reflect individual exposure. The study has poor sensitivity and specificity, resulting in poor discrimination between exposed and non-exposed and among exposure categories.

[a]Considerations for responses for other rating categories (e.g., "some" or "major") may be defined in the protocol for a specific candidate substance.

[b]Potential overlap with study sensitivity; any overlap is addressed in the protocol for a specific candidate substance.

4.2.3 Outcome misclassification

Diagnosis of the type of cancer typically relies upon data such as death certificate data, medical records, or cancer registry data. Incidence data from population-based cancer registry sources, medical records, or hospital pathology data are generally more detailed and accurate than death certificate data. Ideally, cases of cancer should be histologically confirmed and/or undergo independent pathology review (e.g., on a subset of the cases) by the study investigator; this is more likely to be conducted in case-control studies than in cohort studies.

Particular cancers, such as non-Hodgkin lymphoma, may have subtypes that can pose difficulties for classification, especially over time. The classification of subtypes of some cancers has changed over the course of several editions of the *International Classification of Diseases* and may present challenges if histological data are unavailable to confirm subtypes. The potential for misclassification of such cancers may be greater at some time points than at others, and

especially when classification is based on death certificate mortality data. In addition, for cancers with heterogeneous subtypes (e.g., leukemia), some diagnoses may combine subtypes, thereby diluting the effect of the exposure on any particular subtype (e.g., myeloid leukemia).

Cancer incidence data may be considerably more informative than mortality data (depending on ascertainment, reporting, and diagnostic accuracy) for cancers with longer survival times and good treatment prognoses, such as non-Hodgkin lymphoma and urinary-bladder cancer. For cancers with lower survival, both incidence and mortality data may be of similar utility, assuming an adequate length of follow-up. (The evaluation of length of follow-up is usually considered in evaluation of study sensitivity.)

Nondifferential misclassification of cancer (not related to exposure status) would most likely result in the loss of statistical power and an underestimation of the risk estimate.

Table D-4 describes the core, signaling, and follow-up questions and general considerations for assessing the potential for outcome misclassification. More specific and complete considerations (e.g., for all rating categories) may be developed in the protocol for each candidate substance.

Table D-4. Outcome misclassification: Questions and responses

Core question Is there concern that the outcome measure does not reliably distinguish between the presence or absence (or degree of severity) of the outcome?[b]		
Signaling questions	**Follow-up questions**	**Responses[a]**
Is there concern that the diagnosis of disease is incomplete? If mortality data are used, do they adequately reflect incidence? Is there concern that the disease was not accurately diagnosed? Does misclassification vary across exposure groups? Is there concern that the non-diseased group may have disease? Is there concern about observation bias?	Is there concern that any outcome misclassification is either differential or nondifferential? What is the predicted direction or distortion of the effect estimate (if there is adequate information)?	***Low/minimal concerns (+++) rating*** Outcome methods clearly distinguish between diseased and non-diseased subjects. Follow-up and diagnoses are conducted independent of exposure status. ***Critical concerns (0) rating*** There is strong evidence that the methods do not discriminate between diseased and non-diseased subjects and/or that follow-up and diagnoses are likely related to exposure status.

[a]Considerations for responses for other rating categories (e.g., "some" or "major") may be defined in the protocol for a specific candidate substance.

[b]For cancer end points, follow-up methods often overlap with outcome assessment (e.g., mortality databases and death certificates are used for tracking both vital status and cause of death).

4.2.4 Potential for confounding

Confounding occurs when the comparison groups under study (the exposed versus the unexposed groups in a cohort and the case versus control groups in a case-control study) have different background risks of disease (Pearce *et al.* 2007), in effect mixing the association of interest with the effects of other factors. Potential confounders include any exposures or risk factors that could be associated with both exposure and causally with the disease outcome(s) of interest and that are not part of the disease pathway. The potential for confounding in a study can be controlled in the design phase or in the analysis phase. One option in the design phase is restriction — limiting the study to only those subjects for whom potential confounders fall within a narrow range of values (for example, enrolling only males into a study). Another method of confounder control in the design phase is through matching of cases and controls. In the analysis phase of a study, confounding can be controlled for through statistical techniques such as stratification and multivariable methods.

The ability to control for any confounding factor is predicated on that factor being accurately measured and quantified in the study. Assessment of the quality of measurement of exposure to the confounding factor is similar to that for measurement of the exposure of interest (see Section 4.2.2, Exposure misclassification). If information is not available on the risk factor, it may also be possible to conduct sensitivity analyses (indirect adjustment) to evaluate the direction and extent of the potential confounding. This usually requires that the magnitude of the effect estimate for the confounder and disease be known and that information is available to estimate the prevalence of the confounder among the exposed and comparison groups (Pearce *et al.* 2007).

The healthy worker effect is both a special type of confounding and a type of selection bias, and is described in detail in Section 4.2.1, Selection and attrition bias.

Table D-5 describes the core, signaling, and follow-up questions and general considerations for assessing the adequacy of the methods and other information to address potential confounding. More specific and complete considerations (e.g., for all rating categories) may be developed in the protocol for each candidate substance.

Table D-5. Methods for evaluating potential confounding: Questions and responses

Core question Is there concern that either the methods are inadequate or there is inadequate information to evaluate potential confounding?[b]		
Signaling questions	**Follow-up questions**	**Responses[a]**
Is there concern about the measurement of co-exposures or lifestyle risk factors measured in the study?	If no data are provided about confounders, are surrogate data on potential confounders available?	***Low/minimal concerns*** ***(+++) rating*** The study measured all relevant potential confounders and/or used appropriate analyses or designs to address them.
Is there concern that the design or analysis may not adequately address important confounding through matching, stratification, multivariable analysis, or other approaches?	Is there additional information available to evaluate potential confounding or conduct sensitivity analyses (indirect adjustment)?	***Critical concerns*** ***(0) rating*** There is strong evidence that the effects of the exposure cannot be distinguished from the effects of potential confounders.

[a]Considerations for responses for other rating categories (e.g., "some" or "major") may be defined in the protocol for a specific candidate substance.

[b]The evaluation of potential confounding is considered in the interpretation of the study; this assessment is limited to the adequacy of the study methods or useful information.

4.2.5 Selective reporting

When selective or partial reporting is based on the direction, magnitude, or statistical significance of exposure effect estimates, then reporting bias can occur. Selective outcome reporting occurs when the effect estimate for an outcome measurement was selected from among analyses with several outcome measurement instruments and reflected the most favorable result or subcategories (Sterne *et al.* 2014).

Selective analysis reporting occurs when results were selected from exposure effects estimated in several ways, but were reported only for one (or a subset) of the outcomes. Evidence for selective analysis reporting can come from the selection of analyses of a subgroup from a larger cohort: the cohort for analysis may have been selected from a larger cohort for which data were available on the basis of a more interesting finding. Subgroups defined in unusual ways (e.g., an unusual classification of subgroups by dose or dose frequency) may provide evidence of such selective reporting (Sterne *et al.* 2014).

Table D-6 describes the core, signaling, and follow-up questions and general considerations for assessing the potential for bias from selective reporting.

Table D-6. Selective reporting: Questions and responses

Core question Is there concern that the study does not provide results for all relevant measures and participants, biasing its interpretation?		
Signaling questions	**Follow-up questions**	**Responses**[a]
Is there concern that while several types of data were collected, only a subset were reported, or that data were reported for only some subgroups?		***Low/minimal concerns (+++) rating*** There is no evidence that reporting of the data or analyses were limited to only a subset of the data that were collected. ***Critical concerns (0) rating*** There is strong evidence that selective reporting of data or analyses compromised the interpretation of the study.
Is there concern that while several analyses of the exposure-disease relationship may have been performed, only one or a subset of analyses was reported?	Are the analyses needed for the evaluation available (e.g., from the study authors)?	

[a]Considerations for responses for other rating categories (e.g., "some" or "major") may be defined in the protocol for a specific candidate substance.

4.2.6 Analysis

Where adequate data are available, studies should evaluate exposure-response relationships and latency or conduct subgroup analyses (especially for subgroups exposed at higher levels or for longer durations). (This overlaps somewhat with study sensitivity.) Analysis bias may also arise from inappropriate data assumptions, models, or statistical methods used to evaluate the overall findings, exposure-response relationships, latency, or confounding. Bias can also result from controlling for variables in the pathway between exposure and response or for variables unrelated to both the exposure and outcome. In some studies, such as case-control studies evaluating exposure to numerous substances without clear hypotheses, appropriate methods should be conducted to account for multiple comparisons.

Table D-7 describes the core and signaling questions and general considerations for assessing the potential for bias from analysis. More specific and complete considerations (e.g., for all rating categories) may be developed in the protocol for each candidate substance. (No follow-up questions are identified.)

Table D-7. Analysis: Questions and responses

Core question Is there concern that the data assumptions and analysis were not adequate or that the study did not conduct relevant analysis of the available data?	
Signaling questions	**Responses**[a]
Is there concern about whether the data assumptions used in the statistical analysis were adequate (e.g., were the data appropriately log transformed)? If the study data were adequate, did the study evaluate exposure-response and latency or conduct subgroup analyses (especially for subgroups exposed at higher levels or for longer durations)?[b] Is there concern about the adequacy of the models used to evaluate the overall findings, exposure-response relationships, latency, or confounding? Is there evidence of over-controlling for confounding?	***Low/minimal concerns*** ***(+++) rating*** The study used relevant data and appropriate assumptions and methods of analysis. ***Critical concerns*** ***(0) rating*** There is strong evidence that the study analytical methods were so limited that the findings were uninterpretable or distorted.

[a]Considerations for responses for other rating categories (e.g., "some" or "major") may be defined in the protocol for a specific candidate substance.
[b]Overlaps somewhat with study sensitivity.

4.2.7 Study sensitivity

Factors that increase the ability of a study to detect an effect (if present) include moderate to large numbers of exposed and non-exposed participants or cases and controls; evidence of substantial exposure (e.g., level, duration, frequency, or probability) during an adequate time window; an adequate range in exposure levels or duration, allowing for evaluation of exposure-response relationships; and an adequate length of follow-up in cohort studies. When both exposure and disease are rare, statistical power is largely determined by the number of exposed cases (Thomas 2009, as cited in NRC 2014).

Assessment of study sensitivity requires integration of the various factors; for example, a study evaluating effects from low levels of exposure most likely will need larger numbers of exposed subjects than studies of subjects exposed at higher levels. Some of these factors may overlap with exposure assessment, outcome assessment, and analysis; however, it is possible to have a well-designed study that may not be informative for cancer evaluation because of low sensitivity (such as a cohort study evaluating rare cancers). Poor study sensitivity may make it harder to detect an effect (if present) and may also help explain heterogeneity across studies (see Section 5.2).

Table D-8 describes the core and signaling questions and general considerations for evaluating study sensitivity. More specific and complete considerations (e.g., for all rating categories) may be developed in the protocol for each candidate substance. No follow-up questions are identified.

Table D-8. Study sensitivity: Questions and responses

Core question Does the study have adequate sensitivity to detect an effect from exposure (if present)?	
Signaling questions	**Responses**[a]
Are the numbers of exposed cases adequate for detection of an effect in the exposed population and/or subgroups of the exposed population? Are the levels, duration, time window, or range of exposure of the population at risk in cohort and case-control studies sufficient or adequate for detection of an effect of exposure? Is the follow-up period adequate to allow for a cancer induction period?	***Low/minimal concerns*** ***(+++) rating*** The study has an adequate number of exposed subjects, with substantial exposure (level, duration, or range) and with adequate duration of follow-up for latency. ***Critical concerns*** ***(0) rating*** A modest or small study with few exposed subjects and/or exposure is minimal.

[a]Considerations for responses for other rating categories (e.g., "some" or "major") may be defined in the protocol for a specific candidate substance.

4.3 Overall assessment of study utility

The overall utility of a study is based on consideration of both the potential for bias (i.e., study quality) and study sensitivity. Serious concerns about study quality will result in a lower utility ranking. However, a well-designed study with low sensitivity (e.g., having few exposed or expected cases for a specific end point) could be given a low utility ranking. Where adequate information is available for a study, a judgment is made on the direction and distortion of its overall biases or whether it has low sensitivity to detect an effect.

Studies with a critical concern about bias in at least one domain are usually considered to have inadequate utility and are not brought forward to the cancer hazard evaluation. Studies with major concerns in all domains may also be excluded from the evaluation, depending on the direction and distortion of the biases but will be evaluated on a case-by-case basis. The overall judgment of study utility is not meant to be an algorithm that sums up the ratings across domains. Different domains may be given greater weight depending on issues important for the specific candidate substance. This evaluation occurs prior to the cancer hazard evaluation (i.e., interpretation of the study's findings).

Study utility-level judgment

- ***High*** (low/minimal concerns about most potential biases, high or moderate sensitivity rating)
- ***Moderate*** (low/minimal or some concerns about most potential biases, high or moderate sensitivity rating)
- ***Moderate/low*** (some or major concerns about several potential biases, sensitivity rating varies)
- ***Low*** (major concerns about several potential biases, sensitivity rating varies)
- ***Inadequate*** (critical concerns about any bias, sensitivity rating varies)

5 Cancer Hazard Evaluation

This section outlines the approaches to reaching a conclusion on the level of evidence (sufficient, limited, or inadequate) for the carcinogenicity of the substance from studies in humans. The conclusions regarding the assessment of study utility are carried forward to the cancer hazard evaluation, which consists of two phases: the evaluation of the evidence from the individual studies (Section 5.1) and integration of the evidence across studies to reach a preliminary level-of-evidence conclusion (Section 5.2). Studies with the highest utility (i.e., lowest risk of bias and greatest sensitivity to detect an effect) are given the most weight in the assessment. The identification of the potential for specific types of uncontrolled bias or confounding and the assessment of study sensitivity are also used to interpret the findings from studies and to help explain heterogeneity across studies.

Application of the RoC listing criteria to the body of studies on a specific substance involves evaluating (1) whether there is credible evidence for an association between exposure to the substance and cancer and (2) whether such an observed association can be explained by chance, bias, or confounding. Several considerations — strength of the association, consistency across studies, evidence of an exposure-response gradient, and temporality of exposure (Hill 1965) — are used to help guide the evaluation of these questions. However, it should be noted that that these are not criteria; with the exception of temporality, each and every element is not required in order to demonstrate causality (Rothman and Greenland 2005). Figure D-4 shows a schematic of the cancer assessment approach.

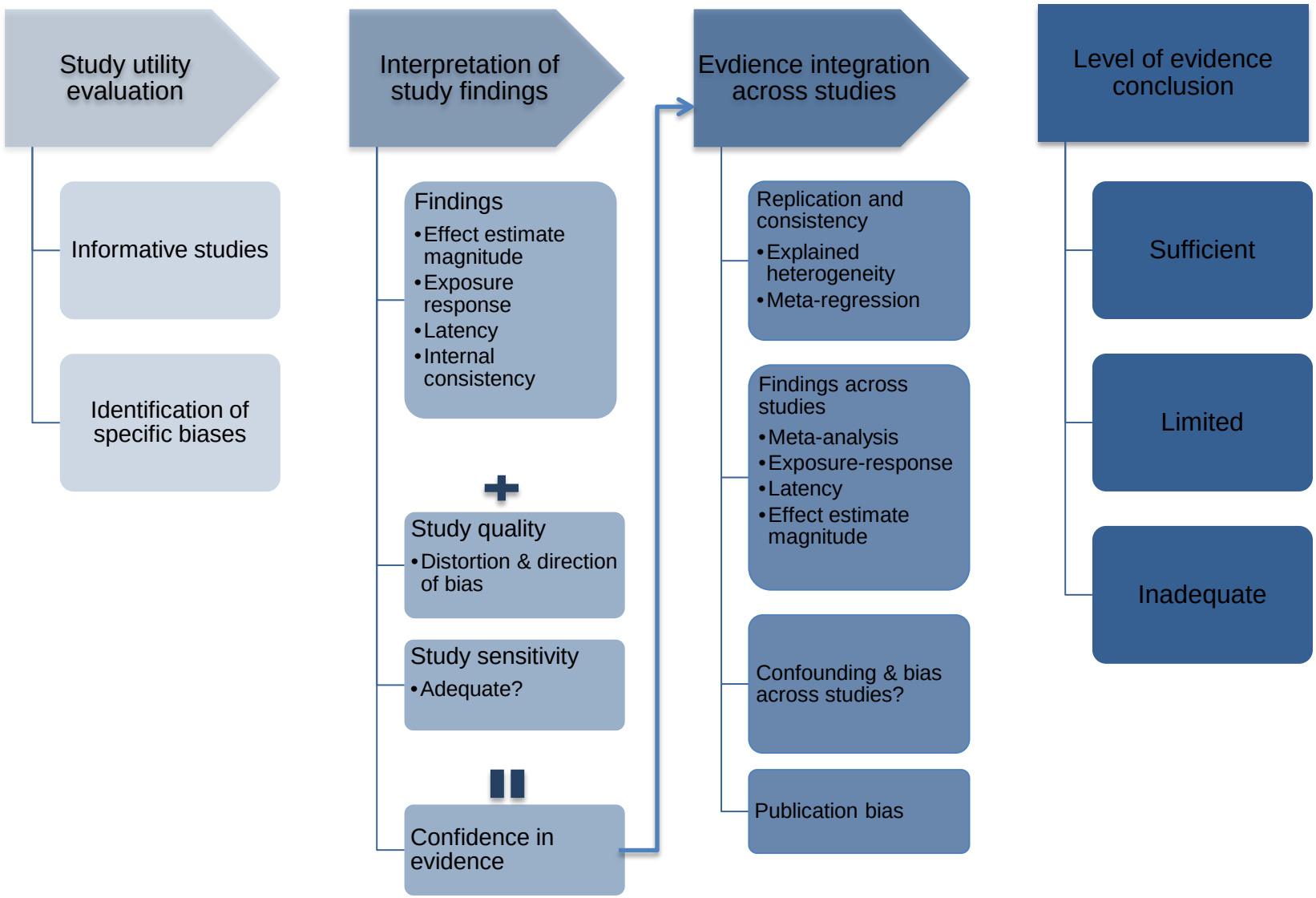


Figure D-3. Approach for evaluating evidence from human cancer studies

5.1 Evaluation of the evidence from the individual studies

The presence of potential bias (such as selection bias or information bias from misclassification of exposure or outcome) or confounding in a study does not necessarily mean that the study should be excluded from the assessment. Conclusions about the evidence from each study should consider the strengths and weaknesses of the study, the direction and distortion of the biases, and the strength of the association between exposure to the substance and the cancer end point.

This section discusses the evaluation of potential confounding and the interpretation of the study's findings, given its strengths and weaknesses, to reach conclusions regarding confidence in the evidence, Confounding is considered in several steps in the cancer hazard evaluation and therefore merits a cohesive discussion.

5.1.1 Evaluation of potential confounding

A key question in the evaluation of the level of evidence from human studies is whether an association (if any) between exposure to a substance and cancer can be explained by confounding. As discussed in Section 4.2.4, confounding occurs when the comparison groups under study have different background risks of disease. Potential confounders include any exposures or other factors that are associated with both exposure and the disease outcome(s) of interest and that are not part of the disease pathway. In occupational studies, co-exposures that highly correlate with exposure to the substance of interest are of potential concern, especially if there are few published data regarding the relationship between the exposure and the outcome.

The evaluation of potential confounding takes into account the following factors:

- Identification of the potential confounders (see Section 2, Protocol Development).
- Assessment of the adequacy of the study design and the analytical or statistical methods used to control for potential confounders (see Section 4, Assessment of the Utility of the Individual Epidemiologic Studies).
- Assessment of whether there is sufficient information on the potential confounders to allow evaluation of the potential for confounding (see Section 4).
- Examination of the magnitude of the risk estimate for exposure to the substance of interest or the strength of exposure-response relationships for specific cancer end points. (This step is also important for ruling out potential unknown confounders).

Tables are created to facilitate the systematic evaluation of the potential confounders for each study; the tables summarize the rationale for whether a suspected confounder can be ruled out in a study (such as statistical adjustment). This information is then used to create a table to evaluate specific potential confounders across studies (see Section 6.3). This approach does not address potential confounding from unmeasured factors.

5.1.2 Evaluating confidence in the study findings

Confidence in a study's findings (e.g., evidence for or against an association) involves considering the strength of the association, the potential for specific biases or confounding, the direction and distortion of those biases or confounding, and the sensitivity of the study to detect an effect. This is especially important for studies where there is a major concern about a potential

bias. For example, if a study finds an association between exposure and disease despite concern about bias towards the null, the findings could be considered as supporting evidence. However, if the direction of the bias is unknown or away from the null, that study would probably not be considered in the integration of the evidence across studies.

Several factors (discussed below) are considered in reaching conclusions concerning confidence in a study's findings (i.e., whether the study provides evidence of an association). These are not meant to be algorithms or requirements; decisions should be based on scientific judgment.

Strength of the observed association between exposure to the substance and cancer

The strength of the association can be important in evaluating whether specific confounders or biases can explain the observed association between exposure and the cancer. When the magnitude is large, the effects of potential confounding (known, residual, or unknown) or bias are typically minor. Biases or confounding may have a greater effect when the effect estimates are small. However, the magnitude of the risk estimate should be judged with consideration to the direction and distortion of the bias or confounding. For example, there may be data (such as sampling) to suggest that potential confounding from smoking could only explain 10% of an increase; therefore, one could have confidence in a study reporting an effect estimate of relatively low magnitude.

Evidence for an exposure-response gradient

As with the magnitude of an association, a positive exposure-response relationship can help rule out bias, confounding, and chance, and can provide convincing evidence of a credible association between exposure and disease. This is important for both identified confounders and unknown confounders. Dose-response curves for established carcinogens include direct monotonic, inverse monotonic, J- or U- shaped, or plateau-shaped relationships. Radiation has a dose-response curve that plateaus, due to cell killing at high doses. Many occupational exposures have attenuation of risks at high doses for a variety of reasons (Stayner *et al.* 2003). There may be biological or methodological reasons for not observing a gradient, and the absence of evidence for an exposure-response relationship is not strong evidence *per se* for the absence of a causal association.

Evidence for associations with appropriate latency

The strength of the association between exposure and cancer risk may be stronger in analyses using lagged models that are consistent with knowledge of the latency of a specific type of cancer or other experimental data.

Internal consistency

Examples of internal consistency include findings that are similar in both external and internal analyses or for different metrics of exposure. However, inconsistency may be attributed to design features (e.g., such as HWE) or biological reasons (e.g., a specific metric may be related to the mode of action of a specific substance).

The following terms and considerations are used to describe the confidence in the reported effect estimate of individual studies. This evaluation requires scientific judgment; these considerations are *not* strict criteria or checklists.

- ***Evidence of an association (increased or decreased):*** The presence of a statistically significant risk, evidence of an exposure-response relationship, or patterns showing internal consistency from a well-designed study. These studies have a limited potential for (or small distortion from) bias, or any bias that may be operating tends mainly towards the null hypothesis producing an underestimate of the risk estimate (for a positive association). Methods used to assess confounding or information available on potential confounders indicate that potential confounding is unlikely to account for all of the excess or reduced risk.
- ***Some evidence of an association:*** Evidence of an association, but the strength of the association is not likely to account for potential confounding or bias.
- ***Null:*** Effect estimates are close to 1.0, but most potential bias is towards the null, or the study has low sensitivity to detect an effect.
- ***Inconclusive:*** Study findings vary, but it is unclear whether all the excess or decreased risk can be explained by potential bias or confounding and/or the direction of bias is unknown.

5.2 Integration of the scientific evidence across human cancer studies

The final step in the assessment is to integrate the evidence across studies, giving greater weight to the most informative studies, in order to reach a preliminary listing recommendation. In some cases, quantitative assessments — meta-analyses (either published by others or conducted for the review) — will be considered. Meta-analyses contribute to the qualitative assessment but are not by themselves the basis for a level-of-evidence conclusion. Many of the following Hill considerations, similar to those elements mentioned above for evaluating confidence in the evidence of the individual studies, are used in the overall assessment discussed below. However, as mentioned previously, it should be noted that these are not criteria; with the exception of temporality, each and every element is not required in order to demonstrate causality (Rothman and Greenland 2005). In addition, these elements may overlap, and are best considered in an integrative manner in the cancer hazard assessment.

Temporality

Exposure must occur before the disease outcome.

Replication, chance, and consistency of findings across studies

It seems reasonable that a positive (or negative) association needs to be replicated in more than one study in order to rule out chance and reach a level-of-evidence conclusion. However, it is difficult to establish more precise considerations, because the degree of replication may depend on the nature of the studies and the strength of the association observed in the studies. For example, findings from multi-center or multi-cohort studies of different populations would have greater weight than findings from a single factory or small case-control studies. In addition, weak associations may need to be replicated in more studies than strong associations (provided that the strong associations are relatively precise and not driven by small numbers of exposed cases).

Consistency needs to be evaluated in the context of study quality (e.g., variations in outcome definitions or exposure assessment methodologies), study sensitivity (e.g., levels or duration of exposure of the population, exposure windows, or length and completeness of follow-up), or

other differences in population characteristics or study methodologies. Consistency can be evaluated through the use of meta-regression methods.

Strength of observed associations between exposure to the substance and cancer across studies

The strength of the association, as measured by the magnitude of the effect estimate, may be difficult to evaluate across studies (in the absence of a meta-analysis), since effect estimates are likely to vary across studies for a number of reasons (e.g., differences in exposure conditions, outcome measurements, and populations). Although a higher magnitude may provide greater confidence that an association is not due to chance, bias, or confounding, this is not required in order to demonstrate causality. There are many examples of weak associations between exposure to a substance and an end point that are nevertheless considered to be causal (e.g., environmental tobacco smoke and lung cancer).

Evidence for an exposure-response gradient

If adequate information on exposure levels (or duration) is available, exposure-response relationships can be evaluated across studies, in addition to within individual studies.

Evidence for associations with appropriate latency

Latency may also be evaluated across studies and may help to explain heterogeneity of results across studies.

Alternative explanations of chance, bias, or confounding

Chance, bias, and confounding can be evaluated across studies, in addition to being considered within individual studies. The finding of consistent positive associations that are replicated across studies in different populations, with different study designs, and in different occupational settings reduces the likelihood that specific biases or potential confounders in individual studies explain the positive associations.

Publication bias

Publication bias occurs when the findings of published studies differ from those of unpublished studies; in particular, null findings may be more likely to be unpublished, while published studies may be more likely to report an effect.

Most of the available methods for evaluating publication bias (such as funnel plots and “trim and fit”) are used in meta-analysis and also in evaluating small-study-size effects; but these methods may be subject to error (Macaskill *et al.* 2001). Publication bias may be less of a concern for qualitative evaluations relying on more informative studies.

6 Examples of Table Templates and Figures

The following table templates and figures are taken or modified from the RoC Monograph on Trichloroethylene (NTP 2015) or have been created for illustration for this handbook. Current plans are to use a database or web-based application (such as Table Builder) to generate the study description, study quality, and evidence-based tables.

6.1 Study description tables

A table is created for each study containing information (generated from database fields) describing the study population and methodologies (but not the findings). These tables contain information used in the study utility assessment and are usually provided in an appendix to the monograph. Separate templates are provided below for cohort and case-control studies. For each endpoint, tables are created to summarize the quality and sensitivity of individual studies and the overall evaluation across studies. These include the evaluation of each study according to the five quality domains and one sensitivity domain. The rating for each domain and its rationale is given (see Templates Figure). These tables would appear in the appendix.

Study descriptions and methodologies: Cohort studies

Field	Description
Reference	Reference Related references
Location	
Enrollment dates	
Population characteristics	Population description Eligibility criteria Population size (exposed and unexposed) Loss to follow-up Referent group
Exposure assessment	Type Details
Outcome assessment	
Exposure information	Exposure levels, duration, range, setting, and other exposure information
Coexposures	Occupational or environmental co-exposures (not lifestyle factors)
Analysis methods and control for confounding	Study type Analytical methods Covariates Other information (such as confounder considered in analysis or design).
All-cause and all-cancer mortality/incidence	

Study descriptions and methodologies: Case-control studies

Field	Description		
Reference	Reference Related references		
Location			
Enrollment dates			
Population		Cases	Controls
	Population size Eligibility criteria Participation rate		
	Matching criteria		
Exposure assessment	Type Details		
Outcome assessment			
Exposure information	Exposure levels, duration, range, setting, and other exposure information		
Co-exposures	Occupation or environmental co-exposures (not lifestyle factors)		
Analysis methods and control for confounding	Study type Analytical methods Covariates Other information (such as confounder considered in analysis or design).		
All-cause and all-cancer mortality/incidence			

6.2 Study utility tables and figures

Tables are created to summarize the quality and sensitivity of individual studies and the overall evaluation across studies. For each study, the rating and the rationale are provided for the five domains of study quality and the one domain for study sensitivity. A table is also created for the overall evaluation, which consists of the rating for each domain, the overall rating, and the rationale for that rating. Templates are provided below for (1) selection bias and potential confounding, (2) information bias (exposure and outcome misclassification), (3) selective reporting and analysis, (4) information related to study sensitivity and its rating, and (5) the overall rating of the study. In addition, tables showing utility rankings across all of the studies for a specific endpoint may be prepared, similar to the table used in the evaluation of trichloroethylene to complement and serve as a reference for forest plots, which stratify studies by their utility rankings. Examples of such tables are shown below and would appear in the cancer hazard evaluation section of the monograph, not in an appendix.

Summary of study quality: Selection bias and potential confounding

Study	Selection bias	Methods to evaluate potential confounding
Reference	*Rating:* 0, +, ++, +++ Direction ↑, ↓, not known *Rationale*	*Rating:* 0, +, ++, +++ Direction ↑, ↓, not known *Rationale*

Summary of study quality: Information bias (exposure and outcome)

Study	Exposure misclassification	Outcome misclassification
Reference	*Rating:* 0, +, ++, +++ Direction ↑, ↓, not known *Rationale*	*Rating:* 0, +, ++, +++ Direction ↑, ↓, not known *Rationale*

Summary of study quality: Selective reporting and analysis

Study	Selective reporting	Analysis
Reference	*Rating:* 0, +, ++, +++ *Rationale*	*Rating:* 0, +, ++, +++ *Rationale*

Study sensitivity: Information and rating template

Study, Rating	Exposed cases	Reported or estimated exposure levels or duration	Follow-up and other information
Reference *Rating:* 0, +, ++, +++ Rationale			

Overall study utility

Study	Selection	Exposure	Outcome	Methods to evaluate potential confounding	Analysis/selective reporting	Sensitivity	Overall rating
Reference	0, +, ++, +++ Direction ↑, ↓, not known	0, +, ++, +++ Direction ↑, ↓, not known	0, +, ++, +++ Direction ↑, ↓, not known	0, +, ++, +++ Direction ↑, ↓, not known	0, +, ++, +++	0, +, ++, +++	High; moderate; moderate-low; low; inadequate Direction ↑, ↓, not known

Although a separate column has been created for each domain, this does not imply that all the domains contribute equally to the overall study rating, and the overall rating may also be based on an integration of these factors.

Utility ranking across studies for (specific end points or cohort studies)

High Low to minimal concern about selection and information biases Moderate (++) to high (+++) study sensitivity rating Adequate consideration of confounders	Study 1 Study 5
Moderate Low to minimal or some concern about selection bias Some concern about nondifferential exposure or outcome misclassification Study sensitivity rating ++ or +++ Low to some concerns about methods to evaluate potential confounding	Study 2 ++ Study 4 not known
Moderate to low Some concern about selection bias Some to major concern about exposure misclassification or outcome Study sensitivity rating varies Some concerns about methods to evaluate potential confounding	Study 3 ↓ and/or ++ Study 7 not known
Low Some to major concern about selection bias Major concern about nondifferential exposure or outcome misclassification Study sensitivity rating varies Some to major concerns about methods to evaluate potential confounding	Study 6 ↓ ↓ and/or + Study 8 ↑

Studies for a specific end point are broadly grouped into study utility categories ranging from high (top) to low (bottom studies). The right hand column also provides information on the direction of bias or sensitivity of the study for the moderate to low studies: tan (or ↑) = bias away from the null or overestimate of the effect estimate; blue (or ↓) = biases towards the null or underestimate of the effect estimate; * = indicates low sensitivity. The degree of shading (darkest most severe) or number or arrows or * indicate the severity of the bias.

6.3 Potential confounding evaluation tables

Tables (i.e., matrices) may be created to facilitate the systematic evaluation of potential confounders for each study and across studies, recognizing that the evaluation is complex and relies on scientific judgment. It is anticipated that a database will be used to enter answers to key questions about each potential confounder. For each study, a tabular report will be created from the database that consists of the potential confounders (variables) and a series of questions and answers that provide information to evaluate concerns about confounding in that study. The list of variables is limited to those substances for which there is reasonable concern about potential confounding: either (1) they are risk factors for the outcome and could be related to exposure status or (2) they are occupational co-exposures correlated with exposure to the substance of interest and linked to the outcome of interest, or their relationship with the outcome has not been widely studied. It is important to note that the table allows for other information to be used in the

evaluation of confounding to allow for flexibility and consideration of different issues that may be study and candidate substance specific.

The final conclusion about whether each potential confounder is likely to confound the exposure-disease relationship in each study is included in a table summarizing the potential for confounding across studies.

Matrix for evaluating confounding in individual studies

Questions	Variables of concern		
	1	2	3
Is the variable addressed appropriately in the statistical analysis?[a]	yes, no, somewhat	yes, no, somewhat	yes, no, somewhat
Is the variable distributed similarly across case-control status?[b]	yes, no, somewhat, NR	yes, no, somewhat, NR	yes, no, somewhat, NR
Is the variable distributed similarly across exposure status or not strongly correlated with exposure? (Provide correlation coefficient and levels of co-exposures compared with exposure, if available.)	yes, no, somewhat, NR	yes, no, somewhat, NR	yes, no, somewhat, NR
Is the variable associated with the outcome of the study? If so, what is the EE and/or evidence of an E-R?	EE/E-R	EE/E-R	EE/E-R
Is an occupational co-exposure associated with the outcome in other studies? If so, what is the level of evidence conclusion for the association?	e.g., sufficient, limited	e.g., sufficient, limited	e.g., sufficient, limited
Do other types of information suggest that confounding is not likely?[c]	yes/no info	yes/no info	yes/no info
Is it reasonable to rule out confounding; e.g., is the variable unlikely to explain all the excess (or decreased) risk associated with the exposure?[d]	yes/no rationale	yes/no rationale	yes/no rationale

NR = not reported; EE = effect estimate; E-R = exposure-response relationship; info = information.
[a]Adjusted for or considered in the analysis; "somewhat" may be used if there are concerns about residual confounding.
[b]For example, the variable is a matching factor, or data suggest it occurs at a similar frequency or level in cases and controls.
[c]For example, there is no exposure-related increase for end points (such as emphysema) that are linked to the potential confounder (such as smoking).
[d]Where possible, indicate how much of the excess risk can be explained by the confounder. Examples of the rationales are generally answers to the preceding questions, such as whether there was adequate control for confounding in statistical analysis.

Matrix for evaluating confounding across studies

Study	Reasonably rule out confounding?		
	Variable 1	Variable 2	Variable 3
Reference	yes/no rationale	yes/no rationale	yes/no rationale

6.4 Visualization of the evidence or findings across studies

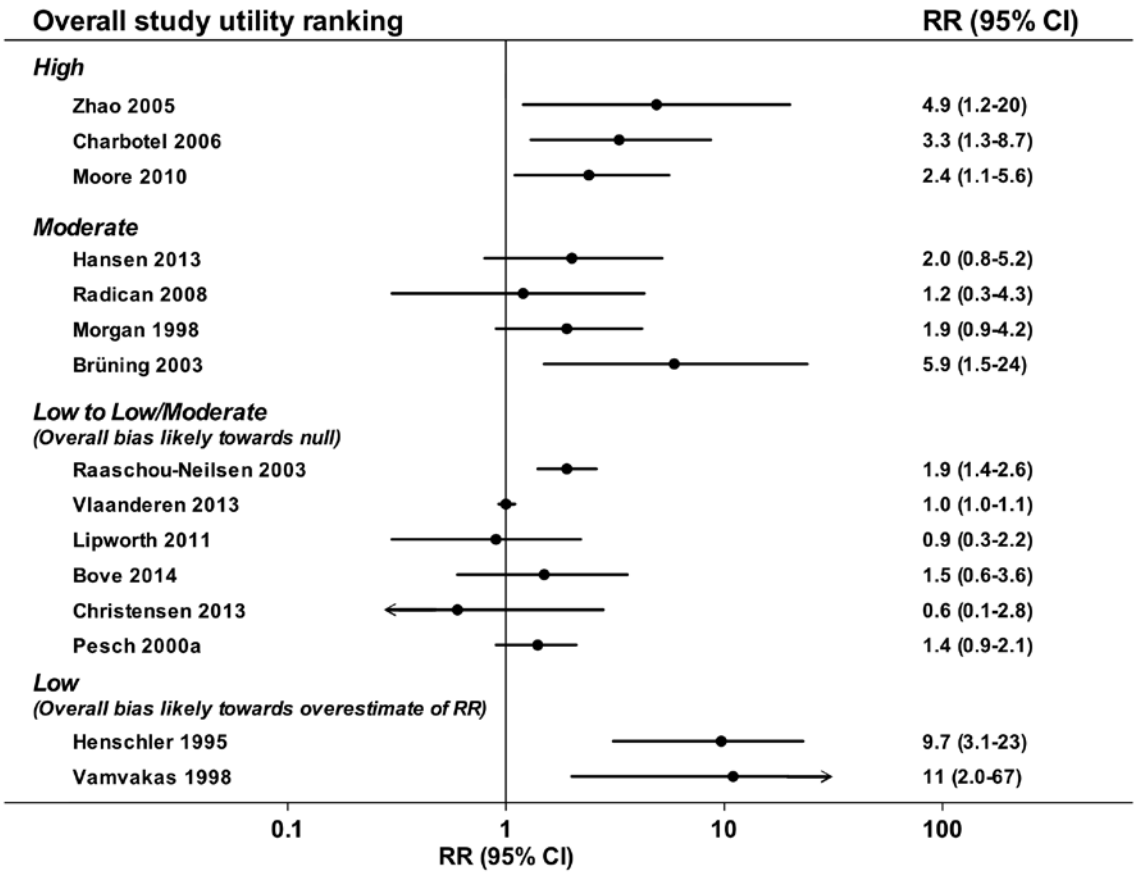
The findings and evidence across studies may be presented in tabular format or in figures such as forest plots. Evidence-based tables are prepared for each outcome (cancer site) of interest. If there are few studies on an outcome, the outcomes may be grouped together in the same study. The tables provide concise information on the population and exposure methods (detailed information is available in the description tables), exposure information (such as level), covariates used in the analysis, the strength and limitations of the study, and the confidence in the study's finding. An example of a template is provided below, and examples of actual tables are available in the RoC Monograph on Trichloroethylene.

Evidence-based tables

Study	Study size (N) Exposure assessment	Exposure groups	RR or OR (95% CI) No. of exposed cases/controls	Interpretation
Reference				Exposure information (e.g., level, duration) Covariates: Strengths: Limitations: Confidence in evidence

Forest plot examples from the RoC Monograph on Trichloroethylene

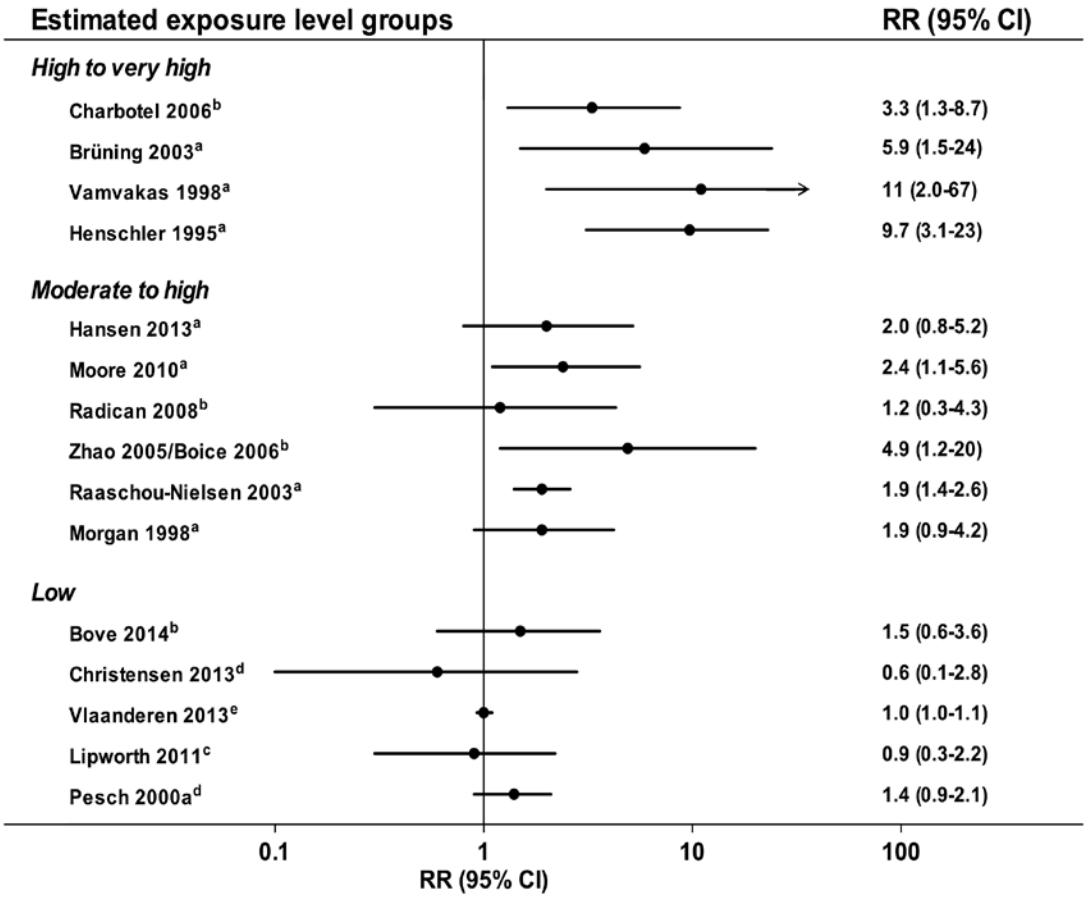
For the cancer hazard evaluation of trichloroethylene, forest plots were constructed using risk estimates for kidney cancer and the highest exposure category. Studies were grouped by study utility ranking (first plot) and by broad groups of estimated exposure (for the highest exposure category). The plots illustrate how heterogeneity among studies can be explained by study quality, study utility or by exposure level. Forest plots could be created or meta-regression analysis could be conducted to look at specific biases, elements of study sensitivity, or other factors that may explain heterogeneity.



Kidney cancer and high exposure to trichloroethylene

Effect estimate and 95% CI for high exposure to trichloroethylene and kidney cancer by study utility category and overall prediction of direction of any bias for low-utility studies.

Figure 4-3 from the RoC Monograph on Trichloroethylene.



Kidney cancer and estimated exposure level for trichloroethylene

Effect estimate and 95% CI for high exposure to trichloroethylene and kidney cancer and estimated exposure level. Different metrics of exposure were graphed and are as follows:
[a]Exposure intensity.
[b]Cumulative exposure.
[c]Exposure duration.
[d]Categories including confidence of probability of exposure with level and/or duration.
[e]Cumulative exposure measures that included exposure prevalence.

Figure 4-4 from the RoC Monograph on Trichloroethylene.

Part E: Evaluation of Cancer Studies in Experimental Animals

Introduction and Objective

This part of the handbook describes the methods and considerations for conducting a systematic cancer hazard evaluation of the evidence from studies in experimental animals for review of a candidate substance for the RoC. This includes identifying and reviewing the relevant studies, assessing study quality and interpreting results, applying the RoC listing criteria (below) to the evidence from the studies, and reaching a conclusion about the level of evidence (sufficient or not sufficient). The key scientific questions and the major steps in the cancer hazard evaluation are listed below. Detailed methods for conducting the evaluation follow this introduction, followed by examples of tables and figures.

Although this handbook describes general methods common to all evaluations, specific protocols will be developed that adapt these methods, identify scientific issues, and develop considerations specific for each candidate substance. This section of this handbook is an adaptation of the protocol used to prepare the RoC Monograph on Pentachlorophenol and By-products of Its Synthesis (NTP 2014a). Where possible, adaptations were made for harmonization with the protocol for human cancer studies and to reflect input from NTP toxicologists.

RoC listing criteria for evaluating carcinogenicity from studies in experimental animals

- ***Sufficient evidence of carcinogenicity from studies in experimental animals:*** An increased incidence of malignant and/or a combination of malignant and benign tumors (1) in multiple species or at multiple tissue sites, or (2) by multiple routes of exposure, or (3) to an unusual degree with regard to incidence, site, or type of tumor, or age at onset.

Key questions

Primary question

- What is the level of evidence (sufficient or not sufficient) for carcinogenicity of the candidate substance from studies in experimental animals?

Secondary questions

- Which experimental animal studies should be included in the review?
- What are key issues for evaluation of the studies?
- What are the methodological strengths and limitations of these studies?
- What are the target tissue sites?

Components of the literature-based cancer hazard assessment

The components of the cancer hazard evaluation of animal studies are the same as those for the human cancer studies (see Figure D-1) and are listed below. The procedures and considerations for each component are described in Sections 1 through 5. Section 6 provides examples of table templates and graphs.

- Planning and research (see the Introduction)
- Literature identification and selection (Section 1, below)

- Protocol development (Section 2, below)
- Data extraction (Section 3, below)
- Study utility evaluation (Section 4, below)
- Cancer hazard evaluation (Section 5, below)

1 Identification and Selection of the Relevant Literature

The cancer evaluation component of the draft monograph evaluates all the relevant cancer studies in experimental animals on exposure to a specific candidate substance. As per the RoC process, studies must be peer reviewed and publicly available. As with human cancer studies, the first step is to develop a literature search strategy and associated inclusion/exclusion criteria to identify the relevant literature (such as reviews, supporting literature, and primary studies), and the second step is to select the primary experimental animal studies from this database. The general approach to identifying and selecting relevant literature is discussed in Part B of this handbook; this section discusses the literature search strategy and inclusion/exclusion criteria specific to studies in experimental animals.

Searches are conducted in PubMed and at least one other database (such as Scopus or Web of Science) using search terms for the candidate substance combined with search terms related to cancer and experimental animal studies (see Table E-1 for examples of search terms). Search terms for the candidate substance may be chemical synonyms, which are usually identified from National Library of Medicine databases (e.g., ChemIDplus or HSDB). Relevant literature may also be identified from sources such as authoritative reviews, IARC monographs, the TOXNET Carcinogenicity Potency Database, PHS 149 (*Survey of Compounds Which Have Been Tested for Carcinogenic Activity*), the TOXNET Chemical Carcinogenesis Research Information System, citations from identified publications, and searches on specific topics. Searches specific for the candidate substance will be developed in the protocol for that substance.

Table E-1. Examples of concepts used in searches for cancer studies in experimental animals

Pubmed, Scopus, and Web of Science		MeSH terms used in PubMed	
Animal terms	**Cancer terms**	**Cancer terms**	**Animal terms**
animal	cancer	neoplasms	models, animal
mouse	neoplasm	carcinogens	animal experimentation
mice	carcinogens		animals, laboratory
rat	malignancy		
hamster	oncogene		
“guinea pig”	tumor		
rabbit			
monkey			
dog			
fish			

Note that these are examples of search terms and not the detailed or fully developed search string used in the actual literature search.

Citations retrieved from literature searches are uploaded to web-based systematic review software and screened by two reviewers using pre-defined inclusion/exclusion criteria.

Studies are initially included in the evaluation if they meet the following inclusion criteria:

- measure neoplastic (benign, malignant) end points
- have non-cancer data that is informative for a cancer assessment, such as reporting preneoplastic lesions
- describe non-neoplastic lesions that are considered part of a morphologic continuum to neoplasia
- provide information on chronic study dose selection (such as a subchronic or short-term toxicity study used for chronic study dose selection)

Studies meeting these criteria typically include studies such as traditional cancer bioassays, initiation-promotion and co-carcinogen studies, and studies in genetically modified animals, which are intended to readily detect carcinogens (such as *Tp53* mouse, *RasH2* mouse, etc.). Studies with no concurrent control group or poor reporting of study design or results may be excluded from further consideration on a case-by-case basis.

2 Protocol Development

Developing the protocol requires an understanding of the types of studies available to inform a hazard assessment, and the protocol is usually written after an initial review of the literature. The protocol provides detailed instructions and considerations for evaluating study exposure conditions and outcome metrics, the methodologic quality of the study, and other issues that may be important for evaluating the findings for the hazard evaluation.

3 Systematic Extraction of Data from the Experimental Animal Studies

Two independent reviewers extract data (such as methods and findings) from the individual studies into a database or web application (such as Table Builder or HAWC) in a systematic manner using standardized instructions and questions. The database contains fields that are specific for the various types of extracted information (such as species, strain, sex, route, dosing regimen, duration, and results). The instructions for data extraction (questions and considerations) describe the specific type of information that should be summarized or entered into each field. The fields are used to populate tables used in the monograph. (See Section 6 for examples of tables for extracted data.)

Study data include neoplasm location and histotype, animal survival, tumor incidence, and statistical significance. If the study authors did not perform statistical analysis, NTP will calculate pairwise analysis of neoplasm incidence relative to control group(s) using Fisher's exact test and analysis for trend across treatment groups using the Cochran-Armitage test, and will note that this was calculated by NTP.

Quality assurance of data extraction and database entry are accomplished by (1) double-checking of each data entry by the two independent reviewers and (2) flagging of any discrepant entries and resolution by mutual discussion with reference to the original data source.

4 Assessment of the Utility of the Individual Studies in Experimental Animals

This section describes the assessment of the utility (i.e., informativeness) of the individual studies, including the steps in the process, responses for each step, signaling questions to

evaluate study utility (internal validity and sensitivity) and external validity, and the overall judgment of the utility of the study to inform the cancer hazard evaluation. This step is completed prior to the cancer hazard evaluation. (See Section 6 for examples of tables for reporting on study quality and utility.)

4.1 Steps in the assessment of study utility

Each primary study is systematically evaluated for its ability to inform the cancer hazard evaluation by two independent reviewers using a series of signaling questions related to the following study performance elements: study design, exposure conditions, outcome assessment, potential confounding, and statistics and reporting (Figure E-1). These questions highlight concerns that toxicologists usually consider when evaluating study utility and are used to increase transparency, but are not meant to be a checklist. The potential for a given bias in a study does not necessarily mean that the findings of the study should be disregarded; and when adequate information is available, the direction of the bias (away or towards the null) should be considered. The rating for each question of whether there is a potential bias or limitation is based on a comparison of the study element with that of the "ideal" study for a specific end point and exposure. In some cases, a rating may not be possible due to the complexity of the issues and the discussion will be captured by narrative text. Each element contains questions related to potential for bias as well as questions related to study sensitivity, which is the ability of the study to detect a "true" risk. This approach differs somewhat from that for the human studies, in which sensitivity questions are part of a separate domain; however, for the assessment of animal studies, there are question level rather than domain level judgments, thus sensitivity ratings are still separated from the risk of bias ratings.

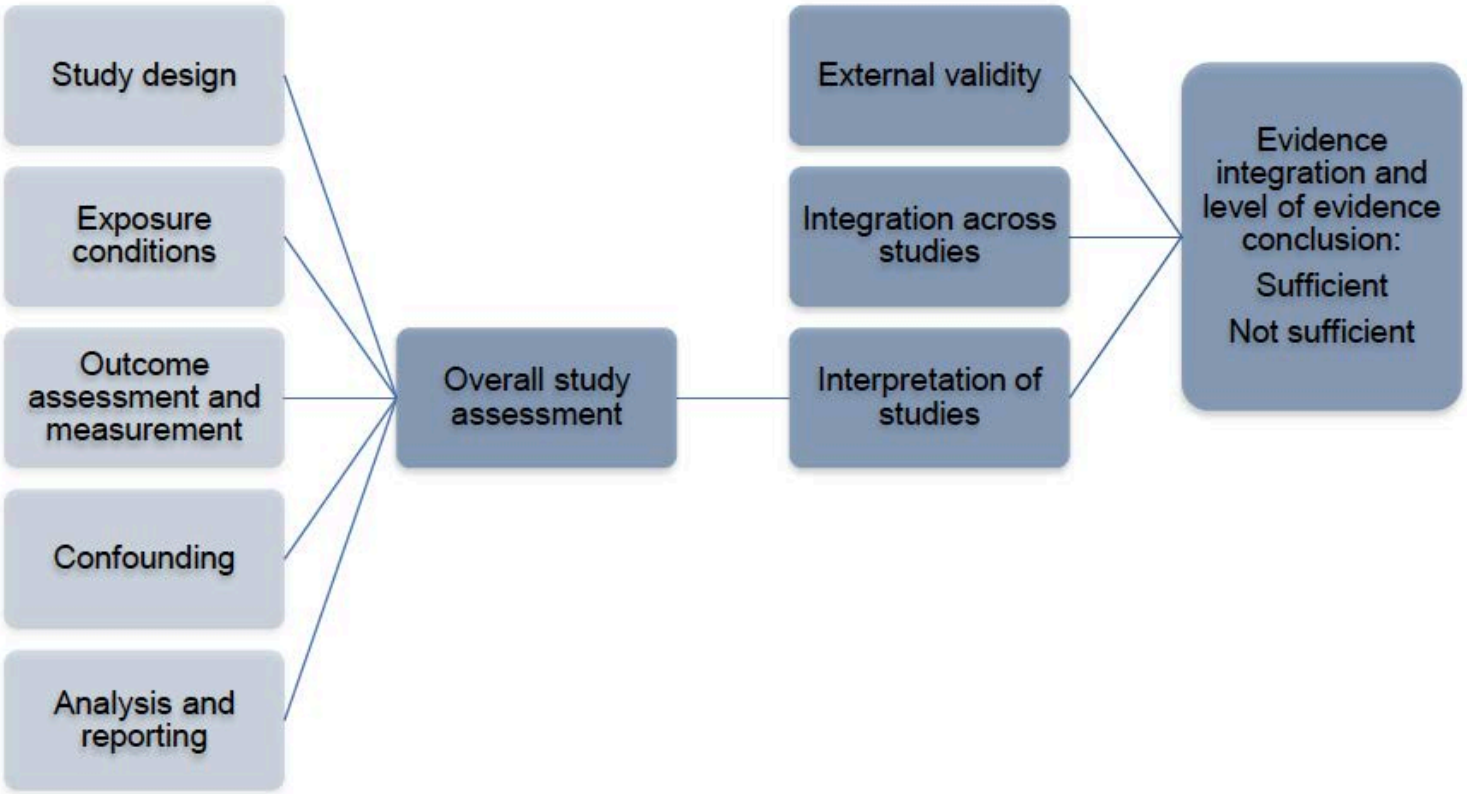


Figure E-1. Approach for evaluating evidence from cancer studies in experimental animal studies

This step is completed prior to interpretation of the individual study's findings and assessment of the level of evidence across studies. Differences are resolved by mutual discussion with reference to the original data source. A small subset of studies is used in a "pilot" phase to discuss and resolve any ambiguity before proceeding with evaluation of the full set of studies. Study authors may be contacted if there is inadequate information to evaluate a signaling question.

Signaling questions: Responses

- ***Low/minimal concerns:*** Information on study design and methodologies indicates that they are close to the ideal study characteristics and that the potential for bias is unlikely or minimal (+++ rating).
- ***Some concerns:*** Study design or methodologies are less than ideal, indicating possible bias (++ rating).
- ***Major concerns:*** Study designs or methodologies suggest that the potential for a specific type of bias is high. However, depending on the nature of the bias, the study may have some limited utility (+ rating).
- ***Critical concerns:*** Study design or methodologies suggest that the bias is critical and would make study findings unreliable for hazard identification (0 rating).
- ***No information on the study:*** The information in the study is inadequate to evaluate the level of concern.

The overall evaluation of the utility of the study is based on an integration of the responses to the signaling questions and the judgment terms are similar to those for the human cancer studies (see Section 4.4). The study utility evaluation is used to identify the most informative studies and informs the interpretation of the study findings (see Section 5). Studies are also evaluated for elements relevant to external validity (interpreting the findings for relevance to humans) (see Section 4.4).

4.2 Study utility evaluation

Signaling questions and considerations for each of the different types of bias and for sensitivity are listed below. Some study elements may overlap between different domains or between study quality and sensitivity and will only be considered in one domain in the evaluation. Study assessments may indicate that a study should not be carried forward to the full evaluation, or they may be used to indicate, across the body of evidence, which findings have more utility in the hazard evaluation than others.

4.2.1 Study design

Study elements that are key for informing the cancer hazard evaluation include randomization of the animals to dose groups and the use of an appropriate comparison group (e.g., ideally, unexposed, sham-treated concurrent controls). The absence of an appropriate control group, by itself, may be sufficient for judging a study inadequate for the cancer hazard evaluation although in some cases historical controls may serve in place of concurrent controls. The experimental design of some studies evaluating co-carcinogens may not include untreated (or vehicle) concurrent controls, but they generally include positive controls, which can result in acceptable study quality. The availability of historical control data from the testing laboratory can be helpful

in assessing the significance of a finding, especially in the case of rare tumors, low-powered studies, or assessment of background tumor incidences, in particular when background rates are high. Historical controls should resemble the concurrent controls with respect to species, sex, strain, diet and other factors influencing tumor response (IARC 2006). However, the concurrent controls are considered to be the most relevant comparison group for evaluating potential exposure-related tumor effects. The treatment of animals in each dose group should be identical except for exposure status (e.g., control, dosed). In addition, the age of the animals at the start of the study should be relevant for the hypothesized mode of action for the specific candidate substance. For cancer studies, animals are usually 6 to 8 weeks old but for some candidate substances, younger age or prenatal exposure may be more relevant.

There are also several study design issues that are related to study sensitivity. The study should use an animal model that is sensitive for detecting tumors and does not have high background rates of the observed tumors or is well characterized with respect to background tumor rates, survival, and growth rates. Studies in both sexes are more informative than those testing only one sex. Adequate statistical power to detect an effect is based on the number of animals used in a study and their survival to study termination, the incidences of tumors in control vs. treated group(s), and the rarity of the tumor. Poor animal husbandry conditions (feed, water, bedding, housing, care, or environmental conditions) or treatment-related survival effects (leading to deaths before there is sufficient time for tumor formation) may lead to high mortality and decrease the statistical power.

Table E-2. Study design: Questions and responses

Signaling questions[a]	Follow-up questions	Responses[a]
Is there concern that the study design did not include randomization of animals to dosed groups? Is there concern that the concurrent control group was not adequate for evaluating effects across treatment groups? Are historical control data reported? (No rating given) Are there concerns about the age of the animals for evaluating potential effects?	Are the historical controls similar to the concurrent controls?	***Low/minimal concerns (+++) rating*** Animals are randomized to control and experimental groups. Controls are as similar as possible to the exposed animals e.g., appropriate vehicle controls. ***Critical concerns (0) rating*** No concurrent or relevant historical control (that could serve as concurrent controls) is available. Clear evidence that the animals were not randomized to treated groups
Sensitivity question Is there concern that the animal model (source, species, strain, sex,) is not sensitive for detecting an effect? Is there concern that there is inadequate statistical power (number of animals per dose and control group) to detect a neoplastic effect, if present?		***Low/minimal concerns (+++) rating*** The study used a sensitive animal model for detecting potential carcinogenic effects and used adequate numbers of animals for most tumor types. Survival in both treated and control groups was adequate and did not reduce statistical power. ***Critical concerns (0) rating*** The study had very small numbers of animals and/or used a resistant animal model.

[a]For animal studies, there is not a core question and ratings are provided for each signaling question. Follow-up questions are meant to add clarification to the signaling questions, and a rating is not provided for these questions.
[b]Considerations for responses for other rating categories (e.g., "some" or "major") may be defined in the protocol for a specific candidate substance. The responses for the two rating categories address issues for all questions; in the study quality evaluation, ratings would be provided for each specific question.

4.2.2 Exposure conditions

Ideally, a study should use a chemical preparation or material that is representative of the candidate substance (in terms of purity and stability), so that any observed effects can be attributed to the candidate substance, and the identity of the substance should have been confirmed. Inhalation studies should also consider the impact of an aerosol generation system on the purity, stability, particle size, and homogeneous distribution of the substance. The animals should be exposed to high enough doses (resulting in tolerable toxicity) for a sufficiently long duration to assess carcinogenicity (usually approaching the lifetime of the animal for non-persistent substances). Ideally, this dose should not limit survival of the animals over the exposure period, except as a result of tumor formation. Treatment-related survival effects may provide information on the adequacy of the dose(s). When relevant, monitoring of food and water consumption and inhalation exposure should be done to estimate dose levels. Some of these questions, such as short exposure duration period or low dose overlap or are related to

sensitivity; for example, a study using low doses (depending on the number of treated animals) for a short duration may limit the ability to see a true effect. Another question related to sensitivity involves the preference (not requirement) for using more than one dose groups so that dose-response relationships can be evaluated.

Table E-3. Exposure conditions: Questions and responses

Signaling questions[a]	Follow-up questions	Responses[a]
Is there concern that the chemical characterization and dose formulations (e.g., conformation, homogeneity, purity, solubility, and stability) and delivery of the chemical (actual vs. desired dose) were not adequate to support attribution of any neoplastic effects to the substance? Is there concern that the dosing regimen (dose selection and dose groups or other factors) was not adequate for detection of a neoplastic effect (if present) or attribution of any neoplastic effects to the substance?	If there is concern about the selection of the dose levels, do the doses appear to be too high or not high enough? What would be the direction of any bias from inadequate dose selection?	***Low/minimal concerns (+++) rating*** The chemical is representative of the candidate substance. Dose selection is based on subchronic or other studies and the high dose is high enough to result in tolerable toxicity and provided for almost the lifetime of the animal. Minimal treatment related survival effects (unless mortality is related to tumors). ***Critical concerns (0) rating*** Chemical is not representative of the candidate substance. Severe toxicity in all treatment groups affecting survival.
Sensitivity Is there concern that the exposure duration period was not adequate for detection of a neoplastic effect, if present? Is the study design adequate to evaluate dose response relationships (e.g., more than one dose)		***Low/minimal concerns (+++) rating*** The study had multiple treatment groups to evaluate exposure response relationships. ***Critical concerns (0) rating*** The study used a very low dose and/or treated animals for a very short period of time.

a For animal studies, there is not a core question and ratings are provided for each signaling question. Follow-up questions are meant to add clarification to the signaling questions, and a rating is not provided for these questions.
[b] Considerations for responses for other rating categories (e.g., "some" or "major") may be defined in the protocol for a specific candidate substance. The responses for the two rating categories address issues for all questions; in the study quality evaluation, ratings would be provided for each specific question.

4.2.3 Outcome (end-point) assessment and measurement

Ideally, each study should include full gross necropsies of all tissues and histopathological examination of the majority of them. Pathology and/or diagnostic procedures and tissues examined should be accurately reported. Studies that examined only tissues of interest may be informative for that tissue or organ, but the evaluation should note limitations for other organs or

tissues including all lesions that can progress to tumor formation. This question overlaps with sensitivity in that studies not evaluating all tissues may miss treatment-related outcomes. All treatment and control groups should have been assessed in the same way. In addition, outcomes should be measured after an appropriate latency period, which for cancer usually means that experimental animals are observed for most of their lifetime.

Table E-4. Outcome assessment and measurement: Questions and responses

Signaling questions[a]	Follow-up questions	Responses[a]
Is there concern that the methods used to assess tumor outcome or the pathology procedures (necropsy, gross pathology, histology, or diagnosis) were not adequate for attribution of the effects to the exposure? Is there concern that not all treatment and control groups were assessed in the same way and in balanced blocks, to avoid bias? For example, was sectioning of organs done in a consistent manner across all treatment and control groups?	If gross pathology or histopathology was not conducted on all tissues, were the tissues with less complete pathology assessment potential target sites (as identified in other animal studies or human studies, or predicted based on mode of action)?	***Low/minimal concerns*** ***(+++) rating*** Complete necropsies and gross pathology reporting for all tissues; histopathology examination on most tissue tissues. ***Major concerns*** ***(+) rating*** Pathology assessment only done on some tissues and not on potential target tissues.
Sensitivity Is the study duration (observation period) adequate to detect a neoplastic effect, if present?		***Low/minimal concerns*** ***(+++) rating*** The study duration is close to the lifetime of the animals. ***Critical concerns*** ***(0) rating*** Study duration is less than one year and no neoplasms are observed.

a For animal studies, there is not a core question and ratings are provided for each signaling question. Follow-up questions are meant to add clarification to the signaling questions, and a rating is not provided for these questions.
[b] Considerations for responses for other rating categories (e.g., "some"") may be defined in the protocol for a specific candidate substance. The responses for the two rating categories address issues for all questions; in the study quality evaluation, ratings would be provided for each specific question.

4.2.4 Potential for confounding

Some sources of potential confounding in animal studies are the use of an impure chemical that contains other potential carcinogens and inadequate animal husbandry conditions and lack of monitoring for pathogens. Potential confounding may arise from carcinogens present in the animal feed, water, or bedding; the presence of disease or parasites; or housing of the animals with experiments for other potential carcinogens. Treatment-related body weight may also be a potential source of confounding. It is also important to use an appropriate vehicle control.

Table E-5. Potential for confounding: Questions and responses

Signaling questions[a]	Follow-up questions	Responses[a]
Is there concern for potential confounding?	If there is concern for potential confounding, is there enough information to determine the relative impact of the confounding?	***Low/minimal concerns*** ***(+++) rating*** The study used a pure testing agent with no/minimal concern for potential contaminant carcinogens and adequate animal husbandry conditions. There were no treatment-related body weight effects ***Critical concerns*** ***(0) rating*** Strong evidence that the presence of contaminant carcinogens in the testing agent or poor animal husbandry conditions would compromise interpretation of the findings

a For animal studies, there is not a core question and ratings are provided for each signaling question. Follow-up questions are meant to add clarification to the signaling questions, and a rating is not provided for these questions.
[b] Considerations for responses for other rating categories (e.g., "some" or "major") may be defined in the protocol for a specific candidate substance. The responses for the two rating categories address issues for all questions; in the study quality evaluation, ratings would be provided for each specific question.

4.2.5 Reporting and analysis

Each study should adequately report incidence data and use appropriate statistical methods. If statistical tests are not reported, the study should at a minimum present incidence data for specific tumors, so that statistical tests can be conducted. If there is evidence of a decreased survival effect, the studies should have used adequate statistical methods (such as the poly-3 test) to control for these effects. Ideally, studies using several dose groups would include trend analysis to evaluate dose-response relationships. Analyses of benign and malignant tumors from the same tissue type should be reported both separately and combined; tumors of the same cellular origin may be combined (McConnell *et al.* 1986).

Table E-6. Reporting and analysis: Questions and responses

Signaling questions[a]	Follow-up questions	Responses[a]
Is there concern that reporting of the data and statistical analysis are inadequate for evaluating the results? Is there concern that different types of tumors were not accurately combined in the analysis?	If statistical analyses were not conducted, is there adequate reporting of the data to conduct statistical testing such as Fisher's exact test for pairwise comparisons?	***Low/minimal concerns*** ***(+++) rating*** The study used and reported relevant data and appropriate methods of analysis. Analyses were adjusted for survival when relevant and tumors were accurately combined in the analysis ***Critical concerns*** ***(0) rating*** There is strong evidence that reporting of data and analytical methods were so limited that the findings were not interpretable.

[a] For animal studies, there is not a core question and ratings are provided for each signaling question. Follow-up questions are meant to add clarification to the signaling questions, and a rating is not provided for these questions.

[b] Considerations for responses for other rating categories (e.g., "some" or "major") may be defined in the protocol for a specific candidate substance. The responses for the two rating categories address issues for all questions; in the study quality evaluation, ratings would be provided for each specific question

4.3 Overall assessment of study utility

The overall utility (ability of the study to inform the cancer hazard evaluation) of a study is based on consideration of both the potential for bias (limitations) and study sensitivity. Studies having elements with major concerns may still be considered in the evaluation or can be considered to provide support to the more informative studies. Studies with critical concerns about important issues will generally be considered to be inadequate to inform the evaluation.

It should also be noted that some concerns about a study element (such as inadequate observation and/or exposure period or statistical power) would decrease the study's sensitivity to detect an effect. If positive findings were described despite these limitations, these studies would inform a cancer assessment.

In some cases, there is inadequate information to answer a specific question. The interpretation of how inadequate information affects the overall study quality evaluation depends on the extent and importance of the missing information and is evaluated on a case-by-case basis. Some studies, such as co-carcinogen studies, have less utility for determining whether a substance is a cancer hazard but may have utility regarding mechanism or other issues; utility would be rated based on the purpose of the study.

Study utility-level judgment

- ***High*** (low concerns about most potential biases and high sensitivity)
- ***Moderate*** (some concerns about many potential biases)
- ***Low*** (major concerns about several biases)
- ***Inadequate*** (critical concerns about some potential biases)

4.4 External validity or interpretation

Some issues relevant to interpretation of the study findings in experimental animals for evaluating potential human carcinogenicity include the route of exposure and mode of action (which would involve other relevant information, such as substance disposition). Studies of exposure by routes that may be less relevant to human exposure are not usually excluded from the cancer hazard assessment; route of exposure is evaluated on a case-by-case basis (see Section 5). Neoplasms observed in experimental animals are considered to be relevant to humans unless there is *compelling* evidence indicating that they occur by a mechanism that does not operate in humans. Other relevant data, such as mechanisms of carcinogenicity, are evaluated in a different monograph section, and the conclusions are brought forward to the overall cancer evaluation section of the monograph, which integrates mechanistic evidence with evidence from human and experimental animal studies to reach a preliminary listing recommendation.

- Is there concern that the route of exposure was not adequate for evaluating the potential for human carcinogenicity?
- Is there concern that tumor formation occurs by a mechanism that would not operate in humans?

5 Cancer Hazard Evaluation

This section outlines the approaches to interpreting the findings of a study, identifying exposure-related tissue sites, integrating the evidence across studies, applying the RoC listing criteria, and reaching a level-of-evidence conclusion (e.g., sufficient, not sufficient) on the carcinogenicity of the substance from studies in experimental animals. The conclusions regarding the assessment of study utility are carried forward to the cancer hazard evaluation, which consists of two phases: the evaluation of the evidence from the individual studies (Section 5.1), and the integration of the evidence across studies to reach a preliminary level-of-evidence conclusion (Section 5.2). Studies with the greatest utility to inform the cancer hazard evaluation (as described in Section 4) are given the most weight in the evaluation.

5.1 Evaluation of the evidence from the individual studies

The findings of each study are interpreted with respect to their limitations and strengths (identified as described in Section 4). For example, positive findings from studies receiving poor ratings for sensitivity (such as low power or short duration study) should not be discounted. The following factors are taken into consideration in determining whether an effect (e.g., increased incidence in a specific tumor type) is treatment related: statistical significance with respect to concurrent controls and dose-related trends, non-neoplastic lesions, lesion progression, decreased latency, tumor multiplicity, tumor incidence, historical control range, animal survival, species, sex, strain, and rarity of tumor. It important to note that the form of the dose-response curves can vary and are not always monotonic; factors such as the absorption, metabolic activation, DNA damage and mechanistic related events, can be factors related to the study, such as survival among the treatment groups (IARC 2006). The evaluation of potential for confounding in an individual study should consider the magnitude of the effect, the adequacy of the controls and whether the potential confounder can modify effects across exposure groups.

5.2 Integration of the scientific evidence across studies

The final step in the evaluation of the evidence from experimental animals is to integrate the evidence (i.e., for treatment-related tumors) across studies, apply the RoC listing criteria (see the Introduction), and reach a listing recommendation. For most databases, heterogeneity in findings is often explained by difference in experimental conditions (e.g., same species, sex, strain, doses, duration, route) and there are few studies conducted using the exact same experimental conditions. As mentioned previously, the most informative studies (highest quality and sensitivity) are given the most weight, and positive findings from these studies are considered to provide evidence of a treatment-related tumor effects. Moderate and low quality studies can also be used in the assessment, especially when it is unlikely that biases in the studies would cause a false positive; replication of findings in several studies also increase the confidence for treatment related effect. In general, two studies (by different exposure routes or in different species) reporting positive findings of malignant or combination of malignant and benign tumors or one study reporting findings at multiple tissue sites fulfill the RoC criteria (see Introduction) for sufficient evidence of carcinogenicity from studies in experimental animals. In addition, positive findings from one robust study can also fulfill the criteria if the tumors are rare, occur at early onset, or at a high incidence. The spectrum of neoplastic response, from pre-neoplastic lesions and benign tumors to malignant neoplasms of a specific tumor type is relevant for the evaluation of whether increases in benign tumors are likely to progress to malignancy.

Mechanistic data are evaluated in a separate section of the monograph and are integrated with the human and animal data to reach a preliminary listing recommendation. The evidence from studies in experimental animals is considered to be relevant to humans unless there is *compelling* evidence to suggest otherwise. Although the relevance of the route of the exposure to humans is considered, findings of tumors at a similar tissue site following exposure by different routes of exposure strengthen the evidence for carcinogenicity.

6 Examples of Table Templates and Figures

Tables typically used in the monograph section on studies in experimental animals include (1) study quality/utility tables and (2) evidence-based tables for each tumor site, which report study methods, findings, major study strengths and limitations, and other relevant comments.

Study quality/utility tables are created to summarize the study quality, sensitivity, and overall evaluation across studies. For each study, the rating and the rationale are provided for the five domains of study quality and one domain of study sensitivity. Table templates are provided below for (1) population, potential confounding, and analysis and reporting, (2) exposure conditions, outcome measurement, and sensitivity, and (3) overall study utility, which summarizes the ratings for the six domains and the overall rating of the study. These tables are usually included in an appendix to the monograph; however, the study utility table may be included in the cancer hazard evaluation section of the monograph.

In general, evidence-based tables are created for each tumor site of interest or for groups of tumor sites (if the database is small), but the organization of the table may vary according to the database. An example of a template is provided that can be generated with the custom web application Table Builder, although this is an area of continual development.

Data may also be visualized in graphs for oral presentation purposes or potentially for use in the monograph.

Study design, potential confounding, and analysis and reporting[a]

Study	Study design	Potential confounding	Reporting and analysis
Reference	*Rating:* 0, +, ++, +++	*Rating:* 0, +, ++, +++	*Rating:* 0, +, ++, +++
	Rationale	*Rationale*	*Rationale*

[a]Tables will present the rating and rationale for each questions. Depending on the database, a separate table with all the questions may be made for each study.
0 = critical concerns, + = major concerns, ++ = some concern, +++ = minimal or low concern; rationale is the reason for the rating.

Exposure conditions and outcome assessment and measurement[a]

Study	Exposure (chemical and dosing)	Outcome (assessment and measurement)
Reference	*Rating:* 0, +, ++, +++	*Rating:* 0, +, ++, +++
	Rationale	*Rationale*

[a]Tables will present the rating and rationale for each questions. Depending on the database, a separate table with all the questions may be made for each study
0 = critical concerns, + = major concerns, ++ = some concern, +++ = minimal or low concern; rationale is the reason for the rating

Overall study utility[a]

Reference	Study design	Exposure	Outcome	Con-founding	Reporting /analysis	Overall rating
Reference	0, +, ++, +++	0, +, ++, +++	0, +, ++, +++	0, +, ++, +++	0, +, ++, +++	High, moderate, low, inadequate

[a]Tables will present the rating and rationale for each questions.
0 = critical concerns, + = major concerns, ++ = some concern, +++ = minimal or low concern; rationale is the reason for the rating

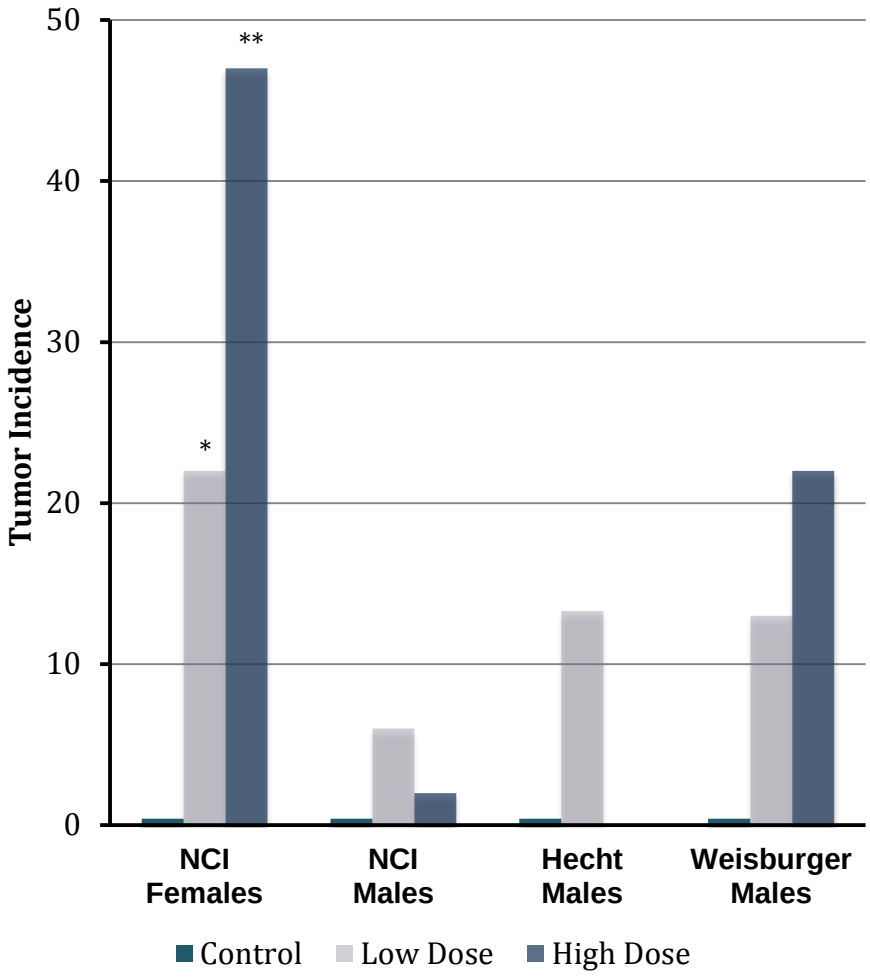
Evidence-based table template

Reference, Model	Dosing regimen	Dose levels	No. of animals	Tumor incidence (%)	Comments, strengths, and limitations
Reference Species (strain) Sex Study duration	Chemical purity Route Conditions			Incidence specific for each cell/tumor type (e.g., hepatocellular adenoma) and sex Trend test	Survival and body weight Strengths Limitations Other comments

Footnote will identify incidence that are significant at different *P* value, historical control information, and statistical methods.

Histograph example of tumor incidence data from studies in experimental animals

The plot below graphs urinary-bladder tumor incidence in rats exposed to *ortho*-toluidine from three different studies (see NTP 2014b). The listing status of *ortho*-toluidine was changed from *reasonably anticipated to be a human carcinogen* to *known to be a human carcinogen* in the 13th RoC.



Urinary-bladder tumor incidence from feeding studies in rats (carcinoma and papilloma combined)

Controls: 0 tumors.
* = $P < 0.001$; ** = $P < 0.0001$.

Part F: Evaluation of Mechanistic and Other Relevant Data

Introduction and Objective

This part of the handbook describes the methods for writing the sections in the RoC monograph that discuss and evaluate mechanistic and other data relevant to the potential carcinogenicity of a candidate substance. The following types of data typically are included:

- disposition (absorption, distribution, metabolism, and excretion [ADME]) and toxicokinetics of the substance in an organism (usually discussed in a separate monograph section)
- genotoxicity and related effects
- toxicity (at the targeted cancer sites) and other measurements of biological reactivity
- mechanisms of cancer formation

Guidance from the RoC listing criteria, key questions, and components of the literature-based assessment are listed below.

RoC listing criteria related to other relevant data

The RoC listing criteria allow for the consideration of other relevant data, although they do not specify any formal criterion that defines the strength (e.g., compelling or convincing) of this type of data. The following criterion for listing a substance in the RoC as *reasonably anticipated to be a human carcinogen* is based on other relevant data:

> "There is less than sufficient evidence of carcinogenicity in humans or laboratory animals; however, the agent, substance, or mixture belongs to a well-defined, structurally related class of substances whose members are listed in a previous Report on Carcinogens as either known to be a human carcinogen or reasonably anticipated to be a human carcinogen, or there is convincing relevant information that the agent acts through mechanisms indicating it would likely cause cancer in humans."

The RoC listing criteria also state that relevant data should be considered in reaching decisions about the carcinogenicity of a substance (e.g., listing status), which applies whether or not a substance is listed as *known to be a human carcinogen* or *reasonably anticipated to be a human carcinogen:*

> "Conclusions regarding carcinogenicity in humans or experimental animals are based on scientific judgment, with consideration given to all relevant information. Relevant information includes, but is not limited to, dose response, route of exposure, chemical structure, metabolism, pharmacokinetics, sensitive sub-populations, genetic effects, or other data relating to mechanism of action or factors that may be unique to a given substance. For example, there may be substances for which there is evidence of carcinogenicity in laboratory animals, but there are compelling data indicating that the agent acts through mechanisms which do not operate in humans and would therefore not reasonably be anticipated to cause cancer in humans."

Key questions

- How is the substance absorbed, distributed, metabolized, and excreted in humans and experimental animals? What are the data for targeted cancer sites?
- What are the data regarding species or sex differences or similarities in ADME or toxicokinetics?
- Is the candidate substance mutagenic and/or genotoxic? If so, what type of damage does it cause? Is there evidence that it is genotoxic in humans or exposed animals?
- Does the substance cause effects that are characteristic of cancer (such as genotoxicity or oxidative stress)?
- What are the potential or proposed key events, modes of action, and/or mechanisms by which the candidate substance causes cancer? What are the strengths and limitations of the evidence?
- What is the evidence (including strengths and limitations) that the proposed mechanisms provide biological plausibility for the effects observed in humans or experimental animals?

Components of the literature-based cancer hazard evaluation

The components in the cancer hazard evaluation of other relevant data are similar to those for the human cancer studies (see Figure D-1) and are listed below. The procedures and considerations for each component are described in Sections 1 through 3. Section 4 provides examples of table templates and graphs.

- Planning and research (see the Introduction)
- Literature identification and selection (Section 1, below)
- Planning and protocol development (see Part A)
- Data extraction and study quality evaluation (Section 2, below)
- Assessment of the evidence (Section 3, below)

1 Identification and Selection of the Relevant Literature

Part B of this handbook discusses general procedures used to identify and select relevant literature for preparing the RoC monograph. This section provides information specific to identifying studies on mechanisms and other related data. Searches are conducted in PubMed and at least one other database (such as Scopus or Web of Science) using search terms for the candidate substance combined with topic search terms (see Table F-1 for examples of MeSH search terms). Search terms for the candidate substance may be chemical synonyms, which are usually identified from National Library of Medicine databases (e.g., ChemIDplus or HSDB).

Table F-1. Examples of MeSH search terms typically used for mechanistic studies and other related data

Topic	MeSH terms used in PubMed
ADME & toxicokinetics	pharmacokinetics; metabolism; activation, metabolic; cytochrome P-450 enzyme system
Mechanisms	mutagenicity tests; mutagen; DNA adducts; DNA damage; chromosomal breakage; chromosomal aberrations; micronucleus tests; sister chromatid exchanges; DNA repair; genomic instability; cell transformation, neoplastic; epigenetic, genetic; reactive oxygen species; oxidative stress; inflammation; immunosuppression; immune evasion; apoptosis; cell proliferation; signal transduction; toxicity [subheading][a]

Note that these are examples of search terms and not the detailed or fully developed search string used in the actual literature search.
[a]Subheading terms for toxicity and adverse effects will be modified by targeted cancer sites for the candidate substance.

Studies or publications are initially included in the evaluation if they meet the following inclusion criteria:

- provide information on disposition or toxicokinetics
- provide information on genotoxicity
- provide information on toxicity affecting the target cells or target cancer sites
- provide information on potential modes of carcinogenic action

2 Data Extraction and Evaluation of Study Quality

The monograph sections that discuss other relevant data are often written in review style, summarizing concepts across the literature and focusing on data from key studies. Moreover, these types of studies are more heterogeneous in nature, which complicates the development of standardized methods for data extraction. Data extraction (whether in Word tables or a database) is done on a case-by-case basis. For end points with many similar types of studies, such as genotoxicity, the data may be extracted into databases or web applications such as Table Builder or HAWC. Examples of the types of genotoxicity data usually extracted into tables are provided in Section 4.

Although no formal study quality criteria or considerations have been developed for *in vitro*, disposition, toxicokinetics, or mechanistic data, study quality is considered in the evaluation of the data based on knowledge of the field or in consultation with technical advisors. Differences in study quality may also explain heterogeneity of the findings. For example, the analytical methods used to measure metabolites may be important in explaining the findings from ADME studies.

In general, many of the questions used in evaluating cancer studies in humans or animals are applicable to evaluating other types of end points (such as genotoxicity) in exposed humans or animals. Moreover, many of the elements (or domains) identified for experimental animal studies may be applicable to evaluating the quality of *in vitro* studies, such as experimental design, the use of appropriate controls, the quality of the exposure conditions, quality of the outcome

measurements, use of appropriate statistical methods and complete reporting, and appropriateness of the experimental model (e.g., cells).

3 Evaluation of the Evidence

As mentioned above, evidence from other relevant data (such as structure-activity relationships and mechanistic data) contributes to the cancer hazard evaluation and a listing recommendation. In general, the evaluation of other relevant data considers whether there are *convincing* data demonstrating biologically plausible mechanisms or modes of action (which fulfills a criterion for listing) for cancer end points reported in humans and/or in experimental animals, or *compelling* data that the agent acts through mechanism(s) that do not operate in humans (which could result in a decision not to list a substance). Typically, the mechanisms by which a substance causes cancer are not completely known; however, mechanistic data have played a major role in the listing of several substances in the RoC (such as ethylene oxide, diazoaminobenzene, dyes metabolized to benzidine, neutrons, and 2,3,7,8-tetrachlorodibenzo-*p*-dioxin). Examples of structurally related classes of chemicals listed in the RoC *as reasonably anticipated to be a human carcinogens* include dyes metabolized to 3,3′-dimethoxybenzidine and dyes metabolized to 3,3′-dimethylbenzidine. Classes can also be defined by similar biological activity or a similar mode of action.

The evaluation of other relevant data includes a discussion of ADME, genotoxicity (as a distinct end point), toxicological effects at targeted cancer sites, whether the substance causes effects that are characteristic of carcinogens (see Section 3.2), proposed modes of action, in general and for targeted cancer sites, and information related to evaluating classes of similar related compounds. In most RoC monographs, information on ADME and toxicokinetics is discussed in a separate section from the genotoxicity, toxicity, and mechanistic data. The contents and approach for drafting each of these sections is briefly discussed below.

The assessment of the evidence for a specific endpoint is similar to the approaches used for the evaluation of cancer studies in experimental animals and humans. Findings from individual studies are interpreted considering the strengths and limitations of the study quality and sensitivity, and the strengths of the association (e.g., magnitude and dose-response relationships). In studies evaluating multiple chemicals, the potential for false positive from multiple comparisons is considered as well as other information such as biological plausibility and dose-response patterns. Assessment of the evidence across studies considers factors such as consistency of the findings, strength of the association, dose response, coherence, specificity, and biological plausibility.

3.1 ADME and toxicokinetics

The ADME section typically relies on authoritative and secondary reviews, supplemented by key or more recent primary studies. In addition to providing a concise summary of the available data, it (usually in the synthesis) should emphasize information that may be useful for evaluating potential modes of action and biological plausibility. These include (but are not limited to) (1) differences and similarities of ADME and toxicokinetics in humans and various experimental animal models, (2) metabolizing enzymes and effects of polymorphic expression, (3) data (such as absorption and distribution) relevant to local vs. systematic effects and targeted cancer sites, and (4) metabolic pathways and any reactive metabolites.

3.2 Mechanistic and other relevant data

This section typically relies on authoritative and secondary reviews, supplemented by key or more recent primary studies. Review articles may be used to identify potential modes of action, which are supplemented by primary studies key to elucidating these modes of action. Typical topics discussed include (1) genotoxicity, (2) toxicological effects at targeted cancer sites, whether the substance causes effects that are characteristic of carcinogens (see below), (3) proposed mode of actions in general and at targeted cancer sites, and (4) information related to evaluating structure-activity relationships or common mechanisms across a class of related compounds.

An IARC advisory group has identified 10 key characteristics (not mechanisms *per se*) of carcinogens that can be used to facilitate a more structured evaluation of mechanism-related data (Smith *et al.* manuscript in preparation, which are somewhat similar to 15 characteristics discussed by Guyton *et al.* (2009). These include (somewhat modified) the ability of a substance to (1) act as an electrophile either directly or after metabolic activation, (2) be genotoxic, (3) alter DNA repair or cause genomic instability, (4) induce epigenetic alterations, (5) generate free radicals and/or induce oxidative stress, (6) induce chronic inflammation, (7) modulate the immune response (8) modulate receptor-mediated effects, (9) cause immortalization, or (10) alter cell proliferation, cell death, or nutrient supply. These characteristics can provide a framework for identifying literature search terms and screening and organizing the literature and writing the monograph.

The evaluation should discuss the strengths and limitations of the data supporting these types of alterations or effects, as well as for any better-delineated modes of action, adverse outcome pathways, or structured mechanisms. It is anticipated that future evaluations will include more comprehensive types of data (such as Tox 21 data) and approaches (such as adverse outcome pathway software).

4 Examples of Table Templates and Figures

The studies of other relevant data (with the exception of genotoxicity data) are more heterogeneous in nature, and their descriptions and results do not easily fit into structured templates. Examples of tables for genotoxicity data and examples of figures for other relevant data used in previous RoC monographs are provided below.

4.1 Table templates

For genotoxicity, given a sufficient database, separate tables (from either primary studies or secondary reviews) are usually created for different types of experimental systems (e.g., bacteria, yeast or prokaryotes, *in vitro* studies in mammalian and human cells, *in vivo* studies in experimental animals, and studies of exposed humans). Within each experimental system, studies are organized by end point and then chronologically. The types of information included are shown in the templates below.

***In vitro* genotoxicity studies**

Study	End point	Test system	Concentration (LEC or HIC)[a]	Cytotoxicity (% survival)[b]	Results[c]		Evaluation
					–S9	+S9	
Reference	e.g., DNA strand breaks	e.g., lymphocytes					Strengths Limitations Conclusions

[a] Lowest effective concentration or highest ineffective concentration.
[b]If provided.
[c]Depending on the candidate substance, these can be data from a primary study or the numbers of studies with positive and negative results from reviews.

***In vivo* genotoxicity studies**

Study	End point	Species/sex	Exposure	Results[a]	Evaluation
Reference	e.g., DNA strand breaks	e.g., F344/N rats males	Route Dose		Strengths Limitations Conclusions

[a]Depending on the candidate substance, these can be data from a primary study or the numbers of studies with positive and negative results from reviews.

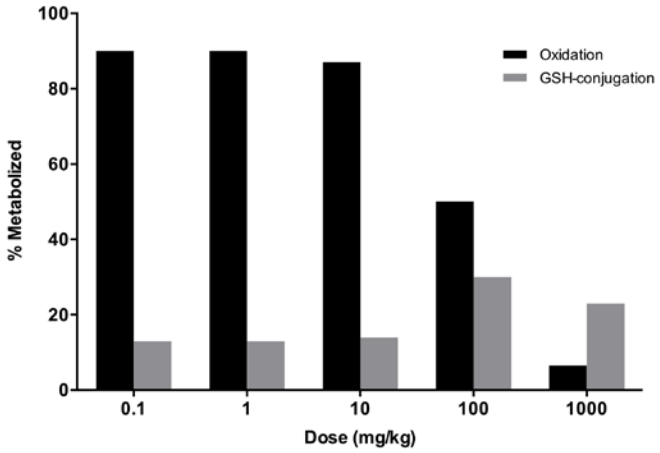
Template summarizing genotoxicity evidence across studies

End point[a]	Bacteria	*In vitro*		*In vivo*	
		Rodents	Humans	Rodents	Humans[b]
DNA adducts					
Gene mutations					
DNA damage/strand breaks					
Sister chromatid exchanges					
Chromosomal aberrations					
Micronuclei					
Aneuploidy					
Inhibition of DNA synthesis					
Unscheduled DNA synthesis					

Evidence across studies is usually summarized as positive, mostly positive, inconclusive, mostly negative, or negative and can be represented by +/– symbols.
[a]Example end points.
[b]Typically, from biomonitoring and molecular epidemiology studies.

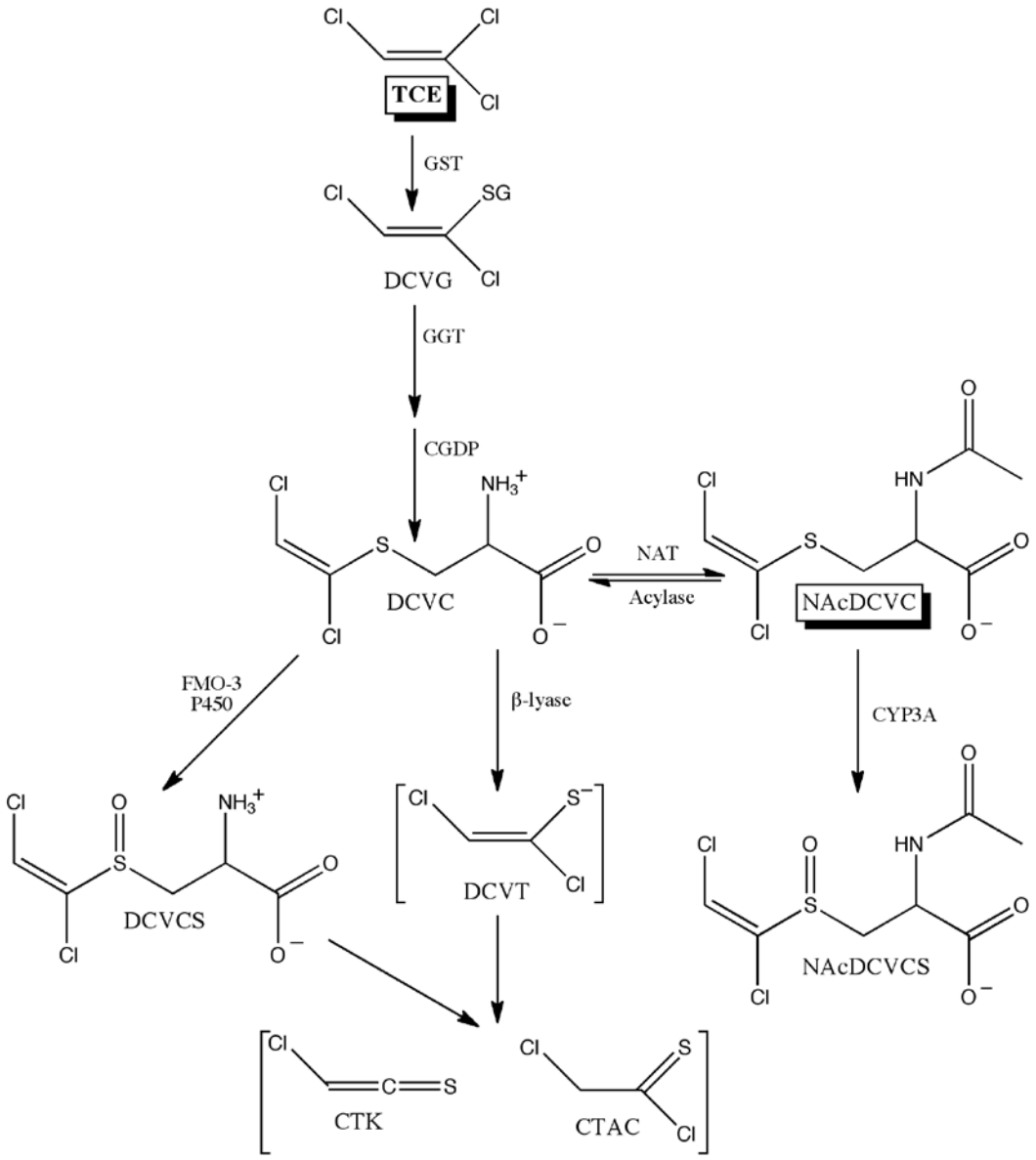
4.2 Examples of figures for mechanistic data

Figures for metabolic pathways or metabolic activation diagrams are usually adapted from the published literature. Where feasible, mechanistic data should be visualized as tables, graphs, or other types of figures. Examples of figures for ADME and metabolic pathways are provided below; see other RoC monographs for more examples (NTP 2013, 2014a,b, 2015).



PBPK model predictions for the fraction of trichloroethylene intake that is metabolized under continuous inhalation exposure in humans

Figure 1-4 from the RoC Monograph on Trichloroethylene (NTP 2015).



Glutathione-dependent metabolic pathways of trichloroethylene

Figure 1-2 in the RoC Monograph for Trichloroethylene (NTP 2015).

Part G: Evidence Integration to Reach a Preliminary Listing Recommendation

Introduction and Objectives

The last step in the cancer hazard evaluation process is to integrate the evidence from the cancer studies in humans and animals with the evidence from mechanistic and other relevant data and apply the RoC listing criteria to reach a preliminary listing recommendation. This step is usually captured in the final section of the RoC monograph.

Listing recommendation: RoC listing criteria

The RoC listing criteria for the two listing categories are briefly summarized as follows:

- ***Known to be a human carcinogen*:** Sufficient evidence of carcinogenicity from studies in humans.
- ***Reasonably anticipated to be a human carcinogen:***
    - Limited evidence of carcinogenicity from studies in humans, or
    - sufficient evidence of carcinogenicity from studies in experimental animals, or
    - the substance belongs to a structurally related class of substances that are listed in the RoC, or there is convincing relevant information that the agent acts through a mechanism indicating that it would likely cause cancer in humans.

Conclusions regarding carcinogenicity in humans or experimental animals are based on scientific judgment, with consideration of all relevant data. (See Part F for the complete and actual language.)

Approach

The approach to documenting the level-of-evidence conclusions for each type of evidence stream (e.g., cancer studies in humans and animals), integrating across evidence streams, and reaching a preliminary listing conclusion may vary according to the nature of the specific candidate substance. As discussed in Part F (Mechanistic and Other Relevant Data), convincing mechanistic data has played a role in the listing of several substances in the RoC. Compelling mechanistic data that the mechanisms of urinary-bladder tumor formation in rats were not relevant to humans led to the removal of saccharin from the RoC (NTP 2014c).

Depending on the database, this section may be organized by type of evidence stream or by cancer site.

In most cases, level-of-evidence conclusions for cancer studies in experimental animals and humans are reached in the corresponding monograph sections. However, in cases where mechanism may play a greater role, the decision may be reached in the evidence integration section. For example, the RoC listing criteria for sufficient and limited evidence in humans allows for consideration of studies in tissues from exposed humans (i.e., mechanistic studies). Similarly, conclusions related to mechanistic data (such as whether the mechanism for a specific cancer site is relevant to humans) may influence external validity considerations in the evaluation of the animal carcinogenicity data (see Part E, Cancer Studies in Experimental

Animals, Section 4.4). Ideally, integration of evidence should use mechanistic data (where available) to help interpret human cancer studies or experimental studies, in addition to integrating level-of-evidence conclusions. For example, mechanistic data may help to explain patterns of exposure-response relationships observed in human or animal studies.

For classes of chemicals and substances, this section should integrate data or evidence across different disciplines (reported in the different sections of the monograph), such as physicochemical properties, toxicokinetics, key biological end points or steps related to the proposed mechanism, and findings from cancer studies in humans and experimental animals. The data could be illustrated using an approach such as heat mapping. This integrative approach provides part of the scientific rationale for the preliminary listing recommendation for the class and could potentially be used to reach conclusions about data-poor members of the class.

References

1. Applebaum KM, Malloy EJ, Eisen EA. 2011. Left truncation, susceptibility, and bias in occupational cohort studies. *Epidemiology* 22(4): 599-606.

2. Greenland S, Pearl J, Robins JM. 1999. Causal diagrams for epidemiologic research. *Epidemiol* 10: 37-48.

3. Guyton KZ, Kyle AD, Aubrecht J, Cogliano VJ, Eastmond DA, Jackson M, *et al.* 2009. Improving prediction of chemical carcinogenicity by considering multiple mechanisms and applying toxicogenomic approaches. *Mutat Res* 681(2-3): 230-240.

4. Hill AB. 1965. The environment and disease: association or causation? *Proc R Soc Med* 58(5): 295-300.

5. IARC. 2006. *Preamble to the IARC Monographs.* IARC Monographs on the Evaluation of Carcinogenic Risk of Chemicals to Humans. Lyon, France: International Agency for Research on Cancer. 23 pp. http://monographs.iarc.fr/ENG/Preamble/

6. Macaskill P, Walters, SD, Irwig L 2001. A comparison of methods to detect publication bias in meta-analysis. *Stat Med.* 20(4):641-654.

7. McConnell EE, Solleveld HA, Swenberg JA, Boorman GA. 1986. Guidelines for combining neoplasms for evaluation of rodent carcinogenesis studies. *J Natl Cancer Inst* 76(2): 283-289.

8. NRC. 2014. *Review of the Formaldehyde Assessment in the National Toxicoogy Program 12th Report on Carcinogens*. National Research Council. Washington, DC: National Academies Press. 232 pp.

9. NTP. 2013. *Report on Carcinogens Monograph on 1-Bromopropane.* NIH Publication No. 13-5982. Research Triangle Park: National Toxicology Program. 168 pp. http://ntp.niehs.nih.gov/ntp/roc/thirteenth/monographs_final/1bromopropane_508.pdf.

10. NTP. 2014a. *Report on Carcinogens Monograph on Pentachlorophenol and By-products of Its Syntheesis.* Research Triangle Park: National Toxicology Program. NIH Publication No. 14-5953. 312 pp. http://ntp.niehs.nih.gov/ntp/roc/thirteenth/monographs_final/pentachlorophenol_508.pdf.

11. NTP. 2014b. *Report on Carcinogens Monograph on* ortho-*Toluidine.* Research Triangle Park: National Toxicology Program. NIH Publication No. 14-5954. 236 pp. http://http://ntp.niehs.nih.gov/ntp/roc/thirteenth/monographs_final/otoluidine_508.pdf.

12. NTP. 2014c. *Report on Carcinogens, Thirteenth Edition.* Research Triangle Park, NC: U.S. Department of Health and Human Services, Public Health Service. http://ntp.niehs.nih.gov/pubhealth/roc/roc13/.

13. NTP. 2015. *Report on Carcinogens Monograph on Trichloroethylene.* Research Triangle Park: National Toxicology Program. 406 pp. http://ntp.niehs.nih.gov/ntp/roc/monographs/finaltce_508.pdf

14. Pearce N, Checkoway H, Kriebel D. 2007. Bias in occupational epidemiology studies. *Occup Environ Med* 64: 562-568.

15. Picciotto S, Brown DM, Chevrier J, Eisen EA. 2013. Healthy worker survivor bias: implications of truncating follow-up at employment termination. *Occup Environ Med* 70(10):736-742.

16. Rothman KJ and Greenland S. 2005. Causation and causal inference in epidemiology *Am J Public Health*95 Suppl 1:S144-S150.

17. Rothman K, Greenland S, Lash T. 2008. *Modern Epidemiology*, 3rd ed. New York: Lippincott, Williams, and Wilkins. 851 pp.

18. Smith MT, Guyton KZ, Gibbons CF, Fritz JM, Portier C, Rusyn I, *et al.* (manuscript in preparation). Key characteristics of carcinogens as a basis for organizing data on mechanisms of carcinogenesis.

19. Sterne JAC, Higgins JPT, Reeves BC, eds., on behalf of the development group for ACROBAT-NRSI. 2014. *A Cochrane Risk Of Bias Assessment Tool: for Non-Randomized Studies of Interventions (ACROBAT-NRSI)*, v. 1.0.0. http://www.riskofbias.info.

20. Stayner L, Steenland K, Dosmesci M, Hertz-Picciotto, I. 2003 Attenuation of exposure-response curves in occupational cohort studies at high exposure levels. *Scand J Work Environ Health* 29(4):317-324.

21. Thomas DC. 2009. Chapter 5: Some special-purpose designs. In *Statistical Methods in Environmental Epidemiology*. New York: Oxford University Press. pp. 92-109. (As cited in NRC 2014.)

22. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. 2007. STROBE Initiative. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: guidelines for reporting observational studies**.** *Lancet* 370(9596):1453-1457.