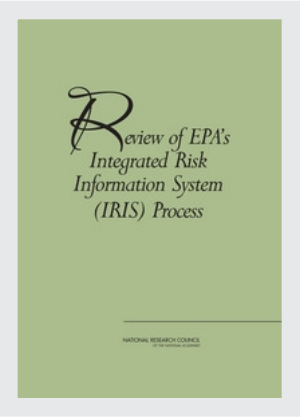


Review of EPA's Integrated Risk Information System (IRIS) Process

DETAILS

170 pages | 6 x 9 | HARDBACK
ISBN 978-0-309-38750-7 | DOI 10.17226/18764

AUTHORS

Committee to Review the IRIS Process; Board on Environmental Studies and Toxicology; Division on Earth and Life Studies; National Research Council

Review of EPA's Integrated Risk Information System (IRIS) Process

Committee to Review the IRIS Process

Board on Environmental Studies and Toxicology

Division on Earth and Life Studies

NATIONAL RESEARCH COUNCIL
OF THE NATIONAL ACADEMIES

THE NATIONAL ACADEMIES PRESS
Washington, D.C.
www.nap.edu

THE NATIONAL ACADEMIES PRESS 500 Fifth Street, NW Washington, DC 20001

NOTICE: The project that is the subject of this report was approved by the Governing Board of the National Research Council, whose members are drawn from the councils of the National Academy of Sciences, the National Academy of Engineering, and the Institute of Medicine. The members of the committee responsible for the report were chosen for their special competences and with regard for appropriate balance.

This project was supported by Contract EP-C-09-003 between the National Academy of Sciences and the U.S. Environmental Protection Agency. Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the authors and do not necessarily reflect the view of the organizations or agencies that provided support for this project.

International Standard Book Number-13: 978-0-309-30414-6
International Standard Book Number-10: 0-309-30414-8

Additional copies of this report are available for sale from the National Academies Press, 500 Fifth Street, NW, Keck 360, Washington, DC 20001; (800) 624-6242 or (202) 334-3313; http://www.nap.edu/.

Printed in the United States of America

THE NATIONAL ACADEMIES

Advisers to the Nation on Science, Engineering, and Medicine

The **National Academy of Sciences** is a private, nonprofit, self-perpetuating society of distinguished scholars engaged in scientific and engineering research, dedicated to the furtherance of science and technology and to their use for the general welfare. Upon the authority of the charter granted to it by the Congress in 1863, the Academy has a mandate that requires it to advise the federal government on scientific and technical matters. Dr. Ralph J. Cicerone is president of the National Academy of Sciences.

The **National Academy of Engineering** was established in 1964, under the charter of the National Academy of Sciences, as a parallel organization of outstanding engineers. It is autonomous in its administration and in the selection of its members, sharing with the National Academy of Sciences the responsibility for advising the federal government. The National Academy of Engineering also sponsors engineering programs aimed at meeting national needs, encourages education and research, and recognizes the superior achievements of engineers. Dr. C. D. Mote, Jr., is president of the National Academy of Engineering.

The **Institute of Medicine** was established in 1970 by the National Academy of Sciences to secure the services of eminent members of appropriate professions in the examination of policy matters pertaining to the health of the public. The Institute acts under the responsibility given to the National Academy of Sciences by its congressional charter to be an adviser to the federal government and, upon its own initiative, to identify issues of medical care, research, and education. Dr. Harvey V. Fineberg is president of the Institute of Medicine.

The **National Research Council** was organized by the National Academy of Sciences in 1916 to associate the broad community of science and technology with the Academy's purposes of furthering knowledge and advising the federal government. Functioning in accordance with general policies determined by the Academy, the Council has become the principal operating agency of both the National Academy of Sciences and the National Academy of Engineering in providing services to the government, the public, and the scientific and engineering communities. The Council is administered jointly by both Academies and the Institute of Medicine. Dr. Ralph J. Cicerone and Dr. C. D. Mote, Jr., are chair and vice chair, respectively, of the National Research Council.

www.national-academies.org

COMMITTEE TO REVIEW THE IRIS PROCESS

Members

JONATHAN M. SAMET (*Chair*), University of Southern California
SCOTT BARTELL, University of California, Irvine
LISA BERO, University of California, San Francisco
ANN BOSTROM, University of Washington
KAY DICKERSIN, Johns Hopkins School of Public Health, MD
DAVID C. DORMAN, North Carolina State University
DAVID L. EATON, University of Washington
JOE G. GARCIA, University of Arizona
MIGUEL HERNÁN, Harvard School of Public Health, MA
JAMES S. HOUSE, University of Michigan
MARGARET M. MACDONELL, Argonne National Laboratory, IL
RICHARD P. SCHEINES, Carnegie Mellon University, PA
LEONARD M. SIEGEL, Center for Public Environmental Oversight, CA
ROBERT B. WALLACE, University of Iowa College of Public Health
YILIANG ZHU, University of South Florida

Staff

ELLEN K. MANTUS, Project Director
KERI STOEVER, Research Associate
NORMAN GROSSBLATT, Senior Editor
MIRSADA KARALIC-LONCAREVIC, Manager, Technical Information Center
RADIAH ROSE, Manager, Editorial Projects
IVORY CLARKE, Senior Program Assistant

Sponsors

U.S. ENVIRONMENTAL PROTECTION AGENCY

BOARD ON ENVIRONMENTAL STUDIES AND TOXICOLOGY[1]

Members

ROGENE F. HENDERSON (*Chair*), Lovelace Respiratory Research Institute, Albuquerque, NM
PRAVEEN AMAR, Clean Air Task Force, Boston, MA
RICHARD A. BECKER, American Chemistry Council, Washington, DC
MICHAEL J. BRADLEY, M.J. Bradley & Associates, Concord, MA
JONATHAN Z. CANNON, University of Virginia, Charlottesville, VA
GAIL CHARNLEY, HealthRisk Strategies, Washington, DC
DOMINIC M. DI TORRO, University of Delaware, Newark, DE
DAVID C. DORMAN, North Carolina State University, Raleigh, NC
CHARLES T. DRISCOLL, JR., Syracuse University, Syracuse, NY
WILLIAM H. FARLAND, Colorado State University, Fort Collins, CO
LYNN R. GOLDMAN, George Washington University, Washington, DC
LINDA E. GREER, Natural Resources Defense Council, Washington, DC
WILLIAM E. HALPERIN, Rutgers University, Newark, NJ
STEVEN P. HAMBURG, Environmental Defense Fund, New York, NY
ROBERT A. HIATT, University of California, San Francisco, CA
PHILIP K. HOPKE, Clarkson University, Potsdam, NY
SAMUEL KACEW, University of Ottawa, Ontario, ON, Canada
H. SCOTT MATTHEWS, Carnegie Mellon University, Pittsburgh, PA
THOMAS E. MCKONE, University of California, Berkeley, CA
TERRY L. MEDLEY, E.I. du Pont de Nemours & Company, Wilmington, DE
JANA MILFORD, University of Colorado, Boulder, CO
MARK A. RATNER, Northwestern University, Evanston, IL
JOAN B. ROSE, Michigan State University, East Lansing, MI
GINA M. SOLOMON, California Environmental Protection Agency, Sacramento, CA
PETER S. THORNE, University of Iowa, Iowa City, IA
JOYCE S. TSUJI, Exponent Environmental Group, Bellevue, WA

Senior Staff

JAMES J. REISA, Director
DAVID J. POLICANSKY, Scholar
RAYMOND A. WASSEL, Senior Program Officer for Environmental Studies
ELLEN K. MANTUS, Senior Program Officer for Risk Analysis
SUSAN N.J. MARTEL, Senior Program Officer for Toxicology
MIRSADA KARALIC-LONCAREVIC, Manager, Technical Information Center
RADIAH ROSE, Manager, Editorial Projects

[1]This study was planned, overseen, and supported by the Board on Environmental Studies and Toxicology.

OTHER REPORTS OF THE
BOARD ON ENVIRONMENTAL STUDIES AND TOXICOLOGY

Review of the Environmental Protection Agency's State-of-the-Science Evaluation of Nonmonotonic Dose–Response Relationships as They Apply to Endocrine Disruptors (2014)
Assessing Risks to Endangered and Threatened Species from Pesticides (2013)
Science for Environmental Protection: The Road Ahead (2012)
Exposure Science in the 21st Century: A Vision and A Strategy (2012)
A Research Strategy for Environmental, Health, and Safety Aspects of Engineered Nanomaterials (2012)
Macondo Well–Deepwater Horizon Blowout: Lessons for Improving Offshore Drilling Safety (2012)
Feasibility of Using Mycoherbicides for Controlling Illicit Drug Crops (2011)
Improving Health in the United States: The Role of Health Impact Assessment (2011)
A Risk-Characterization Framework for Decision-Making at the Food and Drug Administration (2011)
Review of the Environmental Protection Agency's Draft IRIS Assessment of Formaldehyde (2011)
Toxicity-Pathway-Based Risk Assessment: Preparing for Paradigm Change (2010)
The Use of Title 42 Authority at the U.S. Environmental Protection Agency (2010)
Review of the Environmental Protection Agency's Draft IRIS Assessment of Tetrachloroethylene (2010)
Hidden Costs of Energy: Unpriced Consequences of Energy Production and Use (2009)
Contaminated Water Supplies at Camp Lejeune—Assessing Potential Health Effects (2009)
Review of the Federal Strategy for Nanotechnology-Related Environmental, Health, and Safety Research (2009)
Science and Decisions: Advancing Risk Assessment (2009)
Phthalates and Cumulative Risk Assessment: The Tasks Ahead (2008)
Estimating Mortality Risk Reduction and Economic Benefits from Controlling Ozone Air Pollution (2008)
Respiratory Diseases Research at NIOSH (2008)
Evaluating Research Efficiency in the U.S. Environmental Protection Agency (2008)
Hydrology, Ecology, and Fishes of the Klamath River Basin (2008)
Applications of Toxicogenomic Technologies to Predictive Toxicology and Risk Assessment (2007)
Models in Environmental Regulatory Decision Making (2007)
Toxicity Testing in the Twenty-first Century: A Vision and a Strategy (2007)
Sediment Dredging at Superfund Megasites: Assessing the Effectiveness (2007)
Environmental Impacts of Wind-Energy Projects (2007)
Scientific Review of the Proposed Risk Assessment Bulletin from the Office of Management and Budget (2007)
Assessing the Human Health Risks of Trichloroethylene: Key Scientific Issues (2006)
New Source Review for Stationary Sources of Air Pollution (2006)
Human Biomonitoring for Environmental Chemicals (2006)
Health Risks from Dioxin and Related Compounds: Evaluation of the EPA Reassessment (2006)
Fluoride in Drinking Water: A Scientific Review of EPA's Standards (2006)
State and Federal Standards for Mobile-Source Emissions (2006)
Superfund and Mining Megasites—Lessons from the Coeur d'Alene River Basin (2005)
Health Implications of Perchlorate Ingestion (2005)
Air Quality Management in the United States (2004)
Endangered and Threatened Species of the Platte River (2004)
Atlantic Salmon in Maine (2004)

Endangered and Threatened Fishes in the Klamath River Basin (2004)
Cumulative Environmental Effects of Alaska North Slope Oil and Gas Development (2003)
Estimating the Public Health Benefits of Proposed Air Pollution Regulations (2002)
Biosolids Applied to Land: Advancing Standards and Practices (2002)
The Airliner Cabin Environment and Health of Passengers and Crew (2002)
Arsenic in Drinking Water: 2001 Update (2001)
Evaluating Vehicle Emissions Inspection and Maintenance Programs (2001)
Compensating for Wetland Losses Under the Clean Water Act (2001)
A Risk-Management Strategy for PCB-Contaminated Sediments (2001)
Acute Exposure Guideline Levels for Selected Airborne Chemicals (seventeen volumes, 2000-2014)
Toxicological Effects of Methylmercury (2000)
Strengthening Science at the U.S. Environmental Protection Agency (2000)
Scientific Frontiers in Developmental Toxicology and Risk Assessment (2000)
Ecological Indicators for the Nation (2000)
Waste Incineration and Public Health (2000)
Hormonally Active Agents in the Environment (1999)
Research Priorities for Airborne Particulate Matter (four volumes, 1998-2004)
The National Research Council's Committee on Toxicology: The First 50 Years (1997)
Carcinogens and Anticarcinogens in the Human Diet (1996)
Upstream: Salmon and Society in the Pacific Northwest (1996)
Science and the Endangered Species Act (1995)
Wetlands: Characteristics and Boundaries (1995)
Biologic Markers (five volumes, 1989-1995)
Science and Judgment in Risk Assessment (1994)
Pesticides in the Diets of Infants and Children (1993)
Dolphins and the Tuna Industry (1992)
Science and the National Parks (1992)
Human Exposure Assessment for Airborne Pollutants (1991)
Rethinking the Ozone Problem in Urban and Regional Air Pollution (1991)
Decline of the Sea Turtles (1990)

Copies of these reports may be ordered from the National Academies Press
(800) 624-6242 or (202) 334-3313
www.nap.edu

Preface

In 2011, the National Research Council (NRC) released a report that reviewed the draft health assessment on formaldehyde produced by the US Environmental Protection (EPA) for its Integrated Risk Information System (IRIS). The report recommended improvements in the formaldehyde assessment and in the program responsible for producing the assessment. Congress directed EPA to implement the report's recommendations and then asked the NRC to review the changes that EPA was making (or proposing to make) in response to the recommendations.

In the present report, the Committee to Review the IRIS Process first provides an overview of some general issues associated with IRIS assessments. It then addresses evidence identification and evaluation for IRIS assessments and discusses evidence integration for hazard evaluation and methods for calculating reference values and unit risks. It concludes with some overall recommendations and considerations for future directions.

This report has been reviewed in draft form by persons chosen for their diverse perspectives and technical expertise in accordance with procedures approved by the NRC Report Review Committee. The purpose of the independent review is to provide candid and critical comments that will assist the institution in making its published report as sound as possible and to ensure that the report meets institutional standards of objectivity, evidence, and responsiveness to the study charge. The review comments and draft manuscript remain confidential to protect the integrity of the deliberative process. We thank the following for their review of this report: Gary Ginsberg, Connecticut Department of Public Health; William Griffith, University of Washington; Thomas Hartung, Johns Hopkins Bloomberg School of Public Health; Gunnar Johanson, Karolinska Institute; Roderick Little, University of Michigan; Malcolm Macleod, University of Edinburgh; Peter McClure, SRC Environmental Science Center; Ana Navas-Acien, Johns Hopkins Bloomberg School of Public Health; Joseph Rodricks, ENVIRON; Ivan Rusyn, University of North Carolina at Chapel Hill; Christopher Schmid, Brown University; Kurt Straif, International Agency for Research on Cancer; and Joyce Tsuji, Exponent Environmental Group, Inc.

Although the reviewers listed above have provided many constructive comments and suggestions, they were not asked to endorse the conclusions or recommendations, nor did they see the final draft of the report before its release. The review of the report was overseen by the review coordinator, Danny Reible, Texas Tech University, and the review monitor, Mark Cullen, Stanford University. Appointed by the NRC, they were responsible for making certain that an independent examination of the report was carried out in accordance with institutional procedures and that all review comments were carefully considered. Responsibility for the final content of the report rests entirely with the committee and the institution.

The committee gratefully acknowledges the following for their presentations to the committee during open sessions: Richard Atkinson, St. George's University of London; David Budescu, Fordham University; Weihsueh Chiu, Vincent Cogliano, Glinda Cooper, Lynn Flowers, Samantha Jones, and Kenneth Olden, EPA; Chris Frey, North Carolina State University; Steve Goodman, Stanford University; Thomas Hartung and Karen Robinson, Johns Hopkins University; Jay Kadane, Carnegie Mellon University; Tim Lash, Emory University; George Leikauf, University of Pittsburgh; Malcolm MacLeod, University of Edinburgh; Lorenz Rhomberg,

Gradient Corporation; Joe Rodricks, ENVIRON; David Schwartz, University of Colorado; Kristina Thayer, NTP; Duncan Thomas, University of Southern California; Rusty Thomas, The Hamner Institutes for Health Sciences; Thomas Wallsten, University of Maryland; Tracey Woodruff, University of California, San Francisco; and Lauren Zeise, California EPA. The committee members also thank the staff of EPA for being so helpful in answering their numerous questions throughout the study process.

The committee is grateful for the assistance of the National Research Council staff in preparing this report. Staff members who contributed to the effort are Ellen Mantus, project director; Keri Stoever, research associate; James Reisa, director of the Board on Environmental Studies and Toxicology; Norman Grossblatt, senior editor; Mirsada Karalic-Loncarevic, manager of the Technical Information Center; Radiah Rose, manager of editorial projects; and Ivory Clarke, senior program assistant.

I especially thank the members of the committee for their efforts throughout the development of this report.

Jonathan M. Samet, *Chair*
Committee to Review the IRIS Process

Abbreviations

ADH1	Alcohol Dehydrogenase
ADME	Absorption, Distribution, Metabolism, and Excretion
BEIR	Biological Effects of Ionizing Radiation
BMD	Benchmark Dose
BMDL	Benchmark Dose Lower-Confidence Limit
BMDS	Benchmark Dose Software
BUGS	Bayesian Inference Using Gibbs Sampling
CAAC	Chemical Assessment Advisory Committee
CAST	Chemical Assessment Support Team
CCRIS	Chemical Carcinogenesis Research Information System
CNS	Central Nervous System
CONSORT	Consolidated Standards of Reporting Trials
DNA	Deoxyribonucleic Acid
ECVAM	European Centre for the Validation of Alternative Methods
ED	Effective Dose
ELISA	Enzyme-Linked Immunosorbent Assay
EPA	US Environmental Protection Agency
FDA	US Food and Drug Administration
GENETOX	Genetic Toxicology
GLP	Good Laboratory Practice
GMP	Good Manufacturing Practice
GRADE	Grading of Recommendations Assessment, Development and Evaluation
HERO	Health & Environmental Research Online
HIV	Human Immunodeficiency Virus
HSDB	Hazardous Substances Data Bank
IARC	International Agency for Research on Cancer
ICCVAM	Interagency Coordinating Committee on the Validation of Alternative Methods
IHAD	Integrated Hazard Assessment Database
IOM	Institute of Medicine
IRIS	Integrated Risk Information System
LED	Lower Effective Dose
LOAEL	Lowest Observed-Adverse-Effect Level
MNP	Netherlands Environmental Assessment Agency
NOAEL	No-Observed-Adverse-Effect Level
NRC	National Research Council
NTP	National Toxicology Program
OHAT	Office of Hazard Assessment and Toxicology
OPP	Office of Pesticide Programs
PBPK	Physiologically Based Pharmacokinetic
POD	Point of Departure
PRISM	Planning Tool for Resource Integration, Synchronization, and Management
QC	Quality Control
RIVM	Netherlands National Institute for Public Health and the Environment
RfC	Reference Concentration

RfD	Reference Dose
RTECS	Registry of Toxic Effects of Chemical Substances
SES	Socioeconomic Status
TSCA	Toxic Substance Control Act
TSCATS	Toxic Substance Control Act Test Submissions
UF	Uncertainty Factor
WOE	Weight of Evidence

Contents

BOXES, FIGURES, AND TABLES

BOXES

FIGURES

TABLES

Review of EPA's Integrated Risk Information System (IRIS) Process

Summary

The Integrated Risk Information System (IRIS) is a program within the US Environmental Protection Agency (EPA) that is responsible for developing toxicologic assessments of environmental contaminants. IRIS assessments contain hazard identifications and dose-response assessments of various chemicals that cover cancer and noncancer outcomes. Although the program was created to increase consistency among toxicologic assessments within the agency, other federal agencies, various state and international agencies, and other organizations have come to rely on IRIS assessments for setting regulatory standards, establishing exposure guidelines, and estimating risks to exposed populations. Over the last decade, the National Research Council (NRC) has been asked to review some of the more complex and challenging IRIS assessments, including those of formaldehyde, dioxin, and tetrachloroethylene. In 2011, an NRC committee released its review of the IRIS formaldehyde assessment. Like other NRC committees that had reviewed IRIS assessments, the formaldehyde committee identified deficiencies in the specific assessment and more broadly in some of EPA's general approaches and specific methods. Although the committee focused on evaluating the IRIS formaldehyde assessment, it provided general suggestions for improving the IRIS process and a roadmap for its revision in case EPA decided to move forward with changes to the process.

After release of the formaldehyde report, Congress held several hearings to examine the IRIS program. The House Report (112-151) that accompanied the Consolidated Appropriations Act of 2012 (Public Law 112-74) stated that "EPA shall incorporate, as appropriate, based on chemical-specific datasets and biological effects, the recommendations…of the National Research Council's Review of the Environmental Protection Agency's Draft IRIS Assessment of Formaldehyde into the IRIS process." To ensure that EPA adequately considers the recommendations, Congress requested that NRC assess the scientific, technical, and process changes being implemented or planned by EPA and recommend modifications or additional changes as appropriate to improve the scientific and technical performance of the IRIS program. This committee, the Committee to Review the IRIS Process, was convened by NRC as a result of that request. In addition to reviewing the changes in the IRIS program, the committee was asked to review current methods for evidence-based reviews and recommend approaches for weighing scientific evidence for chemical hazard and dose-response assessments. The present report provides the committee's review and recommendations, which are organized around the general depiction of the IRIS assessment process shown in Figure S-1.

SYSTEMATIC REVIEW

In 2011, the same year that the NRC formaldehyde report was released, the Institute of Medicine (IOM) released a report that recommended standards for systematic review.[1] As defined by IOM, systematic review "is a scientific investigation that focuses on a specific question

[1]IOM (Institute of Medicine). 2011. Finding What Works in Health Care: Standards for Systematic Reviews. Washington, DC: The National Academies Press.

and uses explicit, prespecified scientific methods to identify, select, assess, and summarize the findings of similar but separate studies." Although the IOM report was written in the context of comparative-effectiveness research, which aims to determine the most appropriate evidence-based course of action in the clinical setting, systematic-review methods have been used for decades in fields as varied as agriculture and education. The materials and examples provided by EPA to the present committee indicate that the agency is also incorporating systematic-review principles as it implements changes in the IRIS process. The committee agrees with EPA that the systematic-review standards provide an approach that would substantially strengthen the IRIS process, and the committee uses them as a reference point to evaluate the changes that EPA has made.

In evaluating the literature, NRC reports, and EPA documents, the committee found that *systematic review* and *weight-of-evidence analysis* have historically been described in various ways, and the terms are sometimes used interchangeably; this vagueness in use of terminology results in some confusion as to what the terms mean in practice. In the context of IRIS, the committee has defined systematic review as including protocol development, evidence identification, evidence evaluation, and an analytic summary of the evidence (see Figure S-1). The committee views weight-of-evidence analysis as a judgment-based process for evaluating the strength of evidence to infer causation. However, it found that the phrase as used in practice has become too vague and is of little scientific use. An IRIS assessment must come to a judgment about whether a chemical is hazardous to human health and must do so by integrating a variety of lines of evidence. Therefore, the committee found the term *evidence integration* to be more useful and more descriptive of the process that occurs after completion of systematic reviews.

GENERAL ISSUES

The NRC formaldehyde report made several general recommendations concerning the IRIS process, including improving the clarity of the assessments by rigorous editing to reduce redundancies, inconsistencies, and text volume; describing assessment methods more fully; enhancing quality-control processes for assessments; standardizing review and evaluation approaches; and ensuring appropriate expertise on the various chemical-assessment teams. In response to the recommendations, EPA has implemented a new document structure that streamlines the assessments, added a standard preamble to all assessments that describes the IRIS process and its underlying principles, drafted a handbook that provides a more detailed description of the IRIS process, formed chemical assessment support teams (CASTs) to oversee the assessment-development process and ensure consistency among assessments, established tracking procedures, and implemented several initiatives to increase stakeholder input.

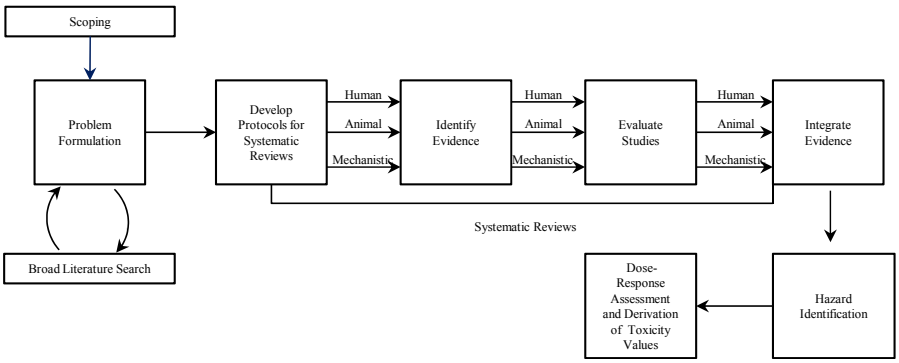


FIGURE S-1 Systematic review in the context of the IRIS process. The committee views public input and peer review as integral parts of the IRIS process, although they are not specifically noted in the figure.

Overall, the changes that EPA has proposed and implemented to various degrees constitute substantial improvements in the IRIS process. If current trajectories are maintained, inconsistencies identified in the present report are addressed, and objectives still to be implemented are successfully completed, the IRIS process will become much more effective and efficient in achieving the program's basic goal of developing assessments that provide an evidence-based foundation for ensuring that chemical hazards are assessed and managed optimally.

Specifically, the present committee finds that the new document structure improves the organization of and streamlines the assessments and reduces redundancies. EPA's use of evidence tables and graphic displays has also reduced text volume and enhanced clarity and transparency. The new approaches bring IRIS assessments much more into line with the state of practice for systematic reviews. The preamble is a useful statement, which will presumably be updated as methods and procedures are modified and updated, but it does not substitute for an overview that indicates how the general principles in the preamble have been applied in any given assessment. The handbook is critical for providing consistency among the assessment teams and contributors, and the final version should be peer-reviewed to ensure that the document is on target and provides the needed guidance.

The committee is encouraged by the efforts to strengthen the overall scientific expertise in the assessment process through the addition of the CASTs and recommends that IRIS assessments clearly identify the members of all teams involved in the development of any given assessment. To strengthen the process further, experts from outside EPA and the government might be needed to fill gaps in expertise in specific areas. Experts should be engaged when needed to augment teams and to conduct peer review of the draft and final assessments.

Finally, the committee applauds EPA initiatives to involve stakeholders in the IRIS process earlier and more fully. Those initiatives are likely to improve assessment quality and to strengthen the program's credibility. However, not all stakeholders who have an interest in the IRIS process have the same scientific or financial resources to provide timely comments, and expanded opportunities for stakeholder involvement might lead to a further imbalance of public input. Therefore, similar to other EPA technical-assistance programs, EPA should consider ways to provide technical assistance to under-resourced stakeholders to help them to develop and provide input into the IRIS process.

PROBLEM FORMULATION AND PROTOCOL DEVELOPMENT

As noted, EPA is incorporating principles of systematic review as it revises the IRIS process. Critical elements of conducting a systematic review include formulating the specific question that will be addressed (problem formulation) and developing the protocol that specifies the methods that will be used to address the question (protocol development). Although the NRC formaldehyde report did not provide any specific recommendations regarding those elements, the present committee found that some discussion of them is warranted.

A major challenge for EPA in the problem-formulation step is to determine what adverse outcomes should be evaluated in a specific IRIS assessment. The committee suggests a three-step process for conducting problem formulation. First, with the support of an information specialist who is trained in conducting systematic reviews, EPA should perform a broad literature search designed to identify possible health outcomes associated with the chemical under investigation. The broad search should not be confused with the comprehensive literature search that is conducted for evidence identification in a systematic review (see Figure S-1); some EPA materials do not sufficiently distinguish between the two. Second, a table should be constructed to guide the formulation of specific questions that would be the subjects of specific systematic reviews. The table could be organized by the lines of evidence typically available to EPA (human, animal, and mechanistic studies) and the various health outcomes to investigate. Third, the table should be examined to determine which outcomes warrant a systematic review and how to define the systematic-review question, such as, Does exposure to chemical X result in neurotoxic effects?

Decisions as to which outcomes should be further evaluated by systematic reviews require careful consideration of numerous factors, and the decision process should be documented and reviewed by relevant experts.

After the systematic-review questions are specified, protocols for conducting the systematic reviews to address the questions should be developed. A protocol makes the methods and the process of the review transparent, can provide the opportunity for peer review of the methods, and stands as a record of the review. It also minimizes bias in evidence identification by ensuring that inclusion of studies in the review does not depend on the studies' findings. Any changes made after the protocol is in place should be transparent, and the rationale for each should be stated. EPA should include protocols for all systematic reviews conducted for a specific IRIS assessment as appendixes to the assessment.

EVIDENCE IDENTIFICATION

The NRC formaldehyde report provided several suggestions aimed at improving EPA's approach to evidence identification, including establishing standard protocols, developing a template to describe the search approach, and using a database to capture study information and relevant quantitative data. Overall, the present committee finds that EPA has been responsive to those suggestions and has substantially improved its approach to evidence identification. Although the agency could not have been expected to incorporate the 2011 IOM standards for systematic review, the preamble, draft handbook, and recent IRIS assessments demonstrate that EPA is well on the way to adopting a more rigorous approach to evidence identification that, when fully implemented, is anticipated to meet standards for systematic review. A few specific findings and recommendations to strengthen the evidence-identification process are highlighted here.

First, searching for and identifying evidence are arguably critical steps in a systematic review, and using a standardized search strategy and reporting format is essential for evidence identification. Protocols for IRIS assessments should include a line-by-line description of the search strategy for each systematic-review question addressed in the assessment that is written in collaboration with information specialists trained in systematic-review methodology. The protocol should also explicitly state the inclusion and exclusion criteria for studies and provide the date of the search, the publication dates searched, and the roles of the various team members.

Second, replicability and quality control are critical for data management. Thus, EPA should have an information specialist trained in systematic-review methodology who reviews the proposed evidence-identification section of the protocol. The committee also encourages the use of at least two reviewers who work independently to screen and select studies, pending an evaluation of validity and reliability that might indicate whether multiple reviewers are needed. The multiple independent reviewers would need to use standardized procedures and forms.

Third, although the basic principles underlying the 2011 IOM standards are most likely relevant to IRIS assessments, EPA is encouraged to perform or support research that examines the applicability of the standards to the hazard and dose-response assessments underlying IRIS assessments.

EVIDENCE EVALUATION

The NRC formaldehyde report provided several recommendations regarding evidence evaluation. Briefly, the recommendations focused on standardizing the presentation of studies and evidence and on evaluating the studies with standardized approaches. In response, EPA now provides checklists in the preamble that indicate how the agency will assess the quality of epidemiologic and experimental studies. Additional details are provided in the draft handbook. EPA correctly identifies important study attributes that can be used to judge study quality but does not

describe how it will assess risk of bias in the identified studies. The committee notes that assessing the quality of the study is not equivalent to assessing the risk of bias in the study. An assessment of study quality evaluates the extent to which the researchers conducted their research to the highest possible standards and how a study is reported. Risk of bias is related to the internal validity of a study and reflects study-design characteristics that can introduce a systematic error (or deviation from the true effect) that might affect the magnitude and even the direction of the apparent effect. An assessment of risk of bias is a key element in systematic-review standards; potential biases must be assessed to determine how confidently conclusions can be drawn from the data.

The committee emphasizes the importance of assessing risk of bias for all study types. Although several approaches are described in the present report, the committee is not recommending the adoption of any specific approach. For a scientifically defensible method, however, EPA should select assessment tools for which empirical evidence links an assessment item with an associated risk of bias. Standardized methods might need to be developed, and EPA might need to conduct or support research on the development and evaluation of empirically based instruments for assessing bias in human, animal, and mechanistic studies relevant to chemical-hazard identification. It might want to consider pooling data across IRIS assessments to determine whether, among various contexts, candidate risk-of-bias items are associated with overestimates or underestimates of effect.

Incorporating risk-of-bias assessments into the IRIS assessment process might take some time, and approaches will depend on the complexity and extent of data on a chemical and the resources available to EPA. An important limitation of all existing tools for assessing study methods is that research reports might not include sufficient details to enable assessment. Consequently, EPA might be hampered by differences in reporting standards for some scientific literature, although the committee expects reporting of toxicology research to improve as risk-of-bias assessments are incorporated into the IRIS process. However, a coordinated effort by government agencies, researchers, publishers, and professional societies will be required to improve the completeness and accuracy of reporting toxicology studies in the near future. Regardless, a risk-of-bias assessment should be conducted on studies that are used by EPA as primary data sources for the hazard identification and dose-response assessment. Whatever approach is adopted, the assessment approach and the results should be fully described and reported in the IRIS assessment.

EVIDENCE INTEGRATION FOR HAZARD IDENTIFICATION

The NRC formaldehyde committee provided several recommendations regarding evidence integration, including reviewing the use of weight-of-evidence guidelines, standardizing an approach to using them, developing uniform language to describe the strength of evidence on noncancer effects, and providing more integrative and transparent discussions of weight of evidence. As in other recommendations, there is an emphasis on transparency and standardization of approach. In response, EPA has provided guidelines in the preamble for what considerations ought to inform the experts who are charged with integrating human, animal, and mechanistic evidence, and it gives extensive guidance on the qualitative categorization that the experts should use, but it articulates no systematic process by which the experts are to come to a conclusion. In the handbook, EPA provides extensive guidelines for synthesizing evidence within each category but no guidelines for integrating evidence among categories. The guidelines and the classification schemes offered for epidemiologic and other studies are reasonable, and similar ones have been used by other organizations with similar aims.

The committee appreciates that EPA's improvements for evidence integration are still being developed but offers some options for moving forward. Several qualitative and quantitative options are available for overall evidence integration. Qualitative options include guided expert judgment, such as the approach used by the International Agency for Research on Cancer

(IARC) in which working groups are used to arrive at overall judgments regarding a chemical's carcinogenicity, and a structured process in which explicit guidelines are developed for qualitative categorization and the process is made as algorithmic as is possible, such as one being developed by the National Toxicology Program (NTP) that is based on the Grading of Recommendations Assessment, Development and Evaluation (GRADE) system. Quantitative options include meta-analysis, probabilistic bias analysis, and Bayesian analysis. Although meta-analysis and probabilistic bias analysis provide quantitative estimates of effect size, the key question in both cases would be whether the effect size can reasonably be inferred to exclude zero (or to exclude being negligible). If so, there is evidence that a hazard exists. If not, there is not adequate evidence to conclude that a hazard exists, although the evidence might suggest a hazard. Bayesian analysis can be used to derive a quantitative judgment, such as "there is at least a 60% chance that chemical X is a carcinogen." Such quantitative judgments could be easily converted into qualitative categorical judgments on the basis of a scale of probabilistic certainty. Quantitative models for evidence integration are powerful tools that can address a wide array of scientific questions, but their clear downside is that model misspecification at any stage can result in incorrect inferences. Nevertheless, they create a rigorous approach that forces analysts to make their assumptions explicit in ways that less formal methods do not. Qualitative and quantitative options are described further in Chapter 6 of this report.

The committee is not recommending a particular approach but suggests that EPA consider which approach among the suggested options best fits its plans for the IRIS program. EPA, however, should continue to improve its evidence-integration process incrementally and enhance the transparency of its process. Thus, it should either maintain its current guided-expert-judgment process but make its application more transparent *or* adopt a structured (or GRADE-like) process for evaluating evidence and rating recommendations along the lines that NTP has taken. If EPA does move to a structured evidence-integration process, it should combine resources with NTP to leverage the intellectual resources and scientific experience in both organizations. Adopting a structured process would have the benefit of transparency. The committee emphasizes that quantitative approaches to integrating evidence will be increasingly needed and useful to EPA, and the agency should seriously consider expanding its ability to perform quantitative modeling for evidence integration.

Regardless of the approach, EPA should develop templates for structured narrative justifications of the evidence-integration process and the conclusion reached. Evidence integration is fundamental in determining whether a chemical poses a hazard. Consequently, the premises and structure of the decision-making process should be as explicit as possible, and the basis for the determination needs to be connected explicitly to the evidence tables produced in the IRIS process.

CALCULATION OF TOXICITY VALUES

In addition to hazard identification, IRIS assessments typically derive toxicity values—reference concentrations, reference doses, and unit risks—that can be used with exposure assessments to derive quantitative risk estimates (see Figure S-1). The NRC formaldehyde committee provided several suggestions regarding this part of the IRIS process, including establishing clear guidelines for study selection, describing and justifying assumptions and models used to determine appropriate points of departure, explaining risk-estimation modeling processes that are used to develop unit-risk estimates, assessing the sensitivity of derived estimates, and adequately documenting the conclusions and estimation of all toxicity values. In response, the preamble provides considerations for selecting studies for deriving toxicity values and describes the process for deriving them. In the draft handbook, EPA has expanded on the study-selection criteria, provided considerations for combining data in dose-response modeling, and discussed data management and quality control for dose-response modeling. More detailed guidance on conducting

dose-response modeling, developing candidate toxicity values, and characterizing confidence and uncertainty in toxicity values has yet to be developed for the draft handbook.

The committee is encouraged by the improvements that EPA has made in the IRIS process for deriving toxicity values, particularly the shift away from choosing one study as the "best" study for deriving a toxicity value and toward deriving and graphically presenting multiple candidate toxicity values. As the program evolves, EPA will need to make the best use of the totality of evidence with increased attention to distinguishing the quality and relevance of studies for assessing human dose-response relationships. That will require EPA to develop clear criteria for judging the relative merits of individual mechanistic, animal, and epidemiologic studies for estimating human dose-response relationships. Although subjective judgment remains an inherent feature of deriving toxicity values, EPA should develop formal methods for combining the results of multiple studies and selecting the final IRIS values with an emphasis on achieving a transparent and replicable process. EPA could also improve documentation of dose-response information by clearly presenting two dose-response values: a central estimate (such as a maximum likelihood estimate or a posterior mean) and a lower-bound estimate for a point of departure from which a final toxicity value is derived.[2] Reporting both values provides information on statistical uncertainty, such as sampling variation, and makes available to the risk assessor the full range of information. Finally, EPA should develop guidelines for uncertainty analysis and communication in the context of IRIS to support the consistent and transparent treatment of uncertainties.

FUTURE DIRECTIONS

The committee commends EPA for the improvements that it has made in the IRIS assessment-development process and expects the revisions when completed to result in a transformation of the IRIS program. To ensure that the IRIS program provides the best assessments possible, the committee identified three broad areas on which EPA should focus attention. First, the assessment methodology will need to be updated in a continuing, strategic fashion, and EPA should develop a plan for doing so. Specifically, the agency will need to consider how methods relevant to all elements of the process will evolve and how such progress can be tracked and incorporated into the IRIS assessment-development approach. Second, EPA staff, the CASTs, and the Chemical Assessment Advisory Committee should be encouraged to identify inefficiencies in the IRIS process, which should then be addressed systematically by the IRIS program leadership. EPA should continue to pursue development of firm stopping rules for key points throughout the process to guard against delay and should consider working with other agencies to avoid duplication of effort. Third, EPA management needs to evaluate the human and technologic resources that are needed to conduct IRIS assessments and support methodologic research and the implementation of new approaches. If sufficient financial and staff resources are not available to EPA, it will not be able to continue to improve the IRIS program and keep pace with scientific advancements.

Overall, the committee finds that substantial improvements in the IRIS process have been made, and it is clear that EPA has embraced and is acting on the recommendations in the NRC formaldehyde report. The NRC formaldehyde committee recognized that its suggested changes would take several years and an extensive effort by EPA staff to implement. Substantial progress, however, has been made in a short time, and the present committee's recommendations should be seen as building on the progress that EPA has already made.

[2]The lower bound becomes an upper bound for a cancer slope factor but remains a lower bound for a reference value.

1

Introduction

In 2011, the National Research Council (NRC) released the report *Review of the Environmental Protection Agency's Draft IRIS Assessment of Formaldehyde*. The report provided a scientific review of the toxicological review of formaldehyde drafted by the US Environmental Protection Agency (EPA) for its Integrated Risk Information System (IRIS). Chapter 7 of the NRC report also suggested changes in the general process used to develop IRIS assessments and suggested a roadmap for making revisions if EPA decided to do so. In response, EPA announced plans to work with its Science Advisory Board to address the committee's suggestions and recommendations. In 2011 testimony before a subcommittee of the US House of Representatives, the assistant administrator of EPA's Office of Research and Development outlined the approach that EPA planned to take in response to the NRC recommendations.[1] On December 23, 2011, the Consolidated Appropriations Act, 2012 (Public Law 112- 74) was signed into law; the House report (112-151) accompanying the act stated that "EPA shall incorporate, as appropriate, based on chemical-specific datasets and biological effects, the recommendations of Chapter 7 of the National Research Council's Review of the Environmental Protection Agency's Draft IRIS Assessment of Formaldehyde into the IRIS process." To ensure that EPA adequately considers the NRC recommendations, Congress requested that NRC assess the scientific, technical, and process changes that EPA is making. As a result of that request, NRC convened the Committee to Review the IRIS Process, which drafted the present report.

THE INTEGRATED RISK INFORMATION SYSTEM AND THE 2011 FORMALDEHYDE REPORT

EPA created the IRIS program in 1985 to provide information on human health effects that could arise from chronic exposures to environmental contaminants. A primary goal of IRIS is to increase the consistency of assessments being conducted throughout the agency. Accordingly, the IRIS program develops toxicologic assessments of various chemicals that include hazard and dose-response characterizations. The assessments also include toxicity values (reference values and unit risks) that can be used with exposure estimates to derive quantitative risk estimates.

Over the years, federal and state agencies and other entities have come to rely on IRIS assessments for setting regulatory standards and establishing exposure guidelines. IRIS assessments provide information needed to evaluate risks at the local level and are also considered authoritative internationally. In recent years, however, questions have been raised about the scientific basis of reference values and unit risks reported in some IRIS assessments. The IRIS program has also been criticized for the time that it takes the agency to complete assessments. Some assessments have taken more than a decade to develop and have gone through multiple cycles of revision and review, which further delay their acceptance. For example, the develop-

[1]EPA's IRIS Program: Evaluating the Science and Process behind Chemical Risk Assessment, 2011: Hearing before the Subcommittee on Investigations and Oversight of the Committee on Science, Space, and Technology, House of Representatives, 112th Cong., 1st Sess., July 14, 2011.

ment of the most recent draft IRIS assessment of formaldehyde began in October 2004, and the draft was released to the public in June 2010. The US Government Accountability Office concluded in 2008 that "the IRIS database is at serious risk of becoming obsolete because EPA has not been able to routinely complete timely, credible assessments or decrease its backlog of 70 ongoing assessments" (GAO 2008, p.1).

Over the last decade, EPA risk-assessment guidance documents have reiterated EPA's policy of evaluating and integrating evidence with an approach that is consistent, comprehensive, balanced, and reproducible (EPA 2002, 2003, 2004, 2005). However, NRC committees have conducted several reviews of some of the more complex and challenging IRIS assessments in the last decade and have identified methodologic problems and pointed out deficiencies in EPA's approaches. For example, the NRC committee that reviewed the dioxin reassessment found problems with the noncancer assessment and stated that "EPA does not use a rigorous approach for evaluating evidence from studies and the weight of their evidence" (NRC 2006, p. 47). The NRC committee that reviewed the tetrachloroethylene assessment offered similar criticisms and emphasized that "the overall impression is that data are presented to support a positive association between tetrachloroethylene and cancer, and that studies that found no such association are criticized or minimized" (NRC 2010, p. 85).

In June 2010, EPA released the IRIS formaldehyde assessment. Recognizing the complex nature of the assessment and its importance as the basis of risk calculations and regulatory decisions for this high-production chemical, EPA asked NRC to review the assessment and answer questions related specifically to the derivation of the inhalation reference concentration (RfC) and unit risk values. The NRC committee convened to conduct that task released its report in 2011. It identified problems similar to those expressed by earlier NRC committees and concluded that "the draft was not prepared in a consistent fashion; it lacks clear links to an underlying conceptual framework; and it does not contain sufficient documentation on methods and criteria for identifying evidence from epidemiologic and experimental studies, for critically evaluating individual studies, for assessing the weight of evidence, and for selecting studies for derivation of the RfCs and unit risk estimates " (NRC 2011, p. 4). As noted, the committee provided specific recommendations for revision of the IRIS formaldehyde assessment. It also made general suggestions for improvement of the IRIS process and provided a roadmap for its revision as guidance if EPA decided to move forward with changes in the process.

IMPROVEMENTS IN THE INTEGRATED RISK INFORMATION SYSTEM

After release of the NRC formaldehyde report (NRC 2011), Congress held several hearings to examine the objectivity and credibility of IRIS assessments and the program.[2,3] On July 12, 2011, EPA (2011) emphasized its commitment to respond to the recommendations in the NRC formaldehyde report and to improve the IRIS program further. Figure 1-1 highlights EPA's actions and demonstrates its commitment to improve IRIS since release of the NRC formaldehyde report.

As requested by Congress, EPA gave relevant congressional committees a progress report in June 2012 that described a phased approach to implementing the NRC recommendations. EPA noted that its first action was "streamlining documents, increasing transparency and clarity, and using more tables and figures to present information and data in assessments" (EPA 2012a, p. 12). Specific improvements that were described in the EPA progress report included developing

[2]EPA's IRIS Program: Evaluating the Science and Process behind Chemical Risk Assessment, 2011: Hearing before the Subcommittee on Investigations and Oversight of the Committee on Science, Space, and Technology, House of Representatives, 112th Cong., 1st Sess., July 14, 2011.

[3]Chemical Risk Assessment: What Works for Jobs and Economy? 2011: Hearing Before the Subcommittee on Environment and the Economy, Committee on Energy and Commerce, House of Representatives, 112th Cong., 1st Sess., October 6, 2011.

a new document structure with an executive summary to highlight major toxicologic findings and a preamble to outline approaches for identifying and evaluating studies, weighing evidence, selecting studies for deriving toxicity values, and deriving toxicity values. EPA also noted that literature-search strategies and criteria for evaluating studies would be explicitly described in new assessments and that it was developing a framework for reaching conclusions on noncancer effects, although it indicated that this effort would be implemented in a later phase. EPA emphasized, however, that it would be using more systematic approaches for analyzing data. Finally, EPA noted that it was expanding efforts for early peer and stakeholder consultation by hosting public workshops on various issues and that it was forming the Chemical Assessment Advisory Committee (CAAC) under the auspices of its Science Advisory Board. The purpose of the CAAC is to advise the agency on specific assessments and possibly on broader issues. Regarding peer consultation, EPA held a public stakeholder meeting in November 2012 to hear the needs of IRIS users and their views on improvements needed in the IRIS program (EPA 2012b).

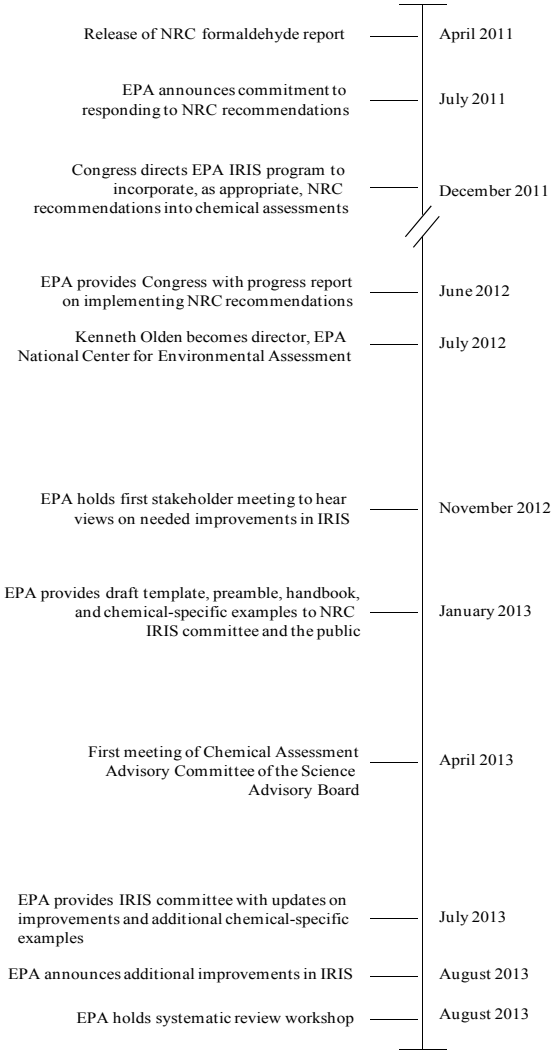


FIGURE 1-1 Timeline of events since release of the NRC report, *Review of the Environmental Protection Agency's Draft IRIS Assessment of Formaldehyde*. The timeline does not include all the meetings that have been held concerning IRIS and chemical-specific assessments.

On January 30, 2013, EPA provided materials to the present committee that outlined its actions in response to each of the recommendations offered in Chapter 7 of the NRC formaldehyde report (NRC 2011) and indicated the status of implementation of each change. It also provided the new IRIS document template, the draft preamble, a draft handbook for the IRIS assessment development, and several chemical-specific examples of its new approaches. The committee was later given a fact sheet issued on July 31, 2013 (EPA 2013) that restated EPA's commitment to improving IRIS assessments and that highlighted some of the improvements that it had described previously, such as the new document structure and the use of systematic-review methods. It also noted changes to enhance productivity and transparency, such as focusing staff on fewer assessments and introducing stopping rules for the inclusion of new data or scientific issues. EPA described several other program improvements, including planning and scoping meetings for each assessment; making public at early stages critical pieces of draft assessments, such as literature searches, evidence tables, and dose-response figures; providing a forum to receive public comments; and ensuring opportunities for the public to make comments on draft assessments once released. Finally, the committee was updated on August 20, 2013, on the implementation of several NRC recommendations and was given additional chemical-specific examples of assessment documents. The materials provided to the committee are listed in Table 1-1 and discussed and reviewed in greater depth in the chapters that follow.

SYSTEMATIC REVIEW

In 2011, the Institute of Medicine (IOM) released the report *Finding What Works in Health Care: Standards for Systematic Review*. As defined by the IOM report, systematic review is "a scientific investigation that focuses on a specific question and uses explicit, prespecified scientific methods to identify, select, assess, and summarize the findings of similar but separate studies" (IOM 2011, p. 1). Although the report was written in the context of comparative-effectiveness research, systematic-review methods have been used for decades in a variety of fields, from agriculture to education (Light and Pillemer 1984; Chalmers et al. 2002). Systematic reviews might or might not result in a quantitative summary of the data, such as an effect estimate. A specific approach to summarize quantitatively somewhat more homogeneous information is often referred to as meta-analysis.

TABLE 1-1 Materials Received from the US Environmental Protection Agency

Document	Content	Date Received
Part 1: Status of Implementation of Recommendations	Status report on implementation IRIS toxicologic review template Preamble to IRIS toxicologic reviews Example of directions for contractors Information on Comment Tracker Database Information on scoping for IRIS assessments Draft handbook for IRIS assessment development	January 30, 2013
Part 2: Chemical-Specific Examples	Seven chemical-specific examples of implementation of various NRC recommendations	January 30, 2013
EPA Fact Sheet	Summary of enhancements of IRIS program	July 31, 2013
E-mail from IRIS program director	Responses to NRC committee inquiries Attachments: Tables documenting implementation of NRC recommendations from draft ammonia, trimethylbenzene, and benzo[a]pyrene assessments	August 20, 2013
IRIS Toxicological Review of Benzo[a]pyrene	Draft IRIS assessment of benzo[a]pyrene and supplemental materials	August 20, 2013
IRIS Toxicological Review of Methanol[a]	Draft IRIS noncancer assessment of methanol	September 30, 2013

[a]The IRIS methanol assessment was not sent directly to the committee but was released while the committee was conducting its review and therefore was used as an example of implemented changes.

The materials and examples provided by EPA indicate that the agency is incorporating systematic-review principles as it makes changes in the IRIS process. Figure 1-2 shows systematic review as the present committee envisions its use in the context of the IRIS process. As noted above, one might be able to use meta-analysis to summarize quantitatively the results of the systematic reviews of each data stream (human, animal, and mechanistic). Meta-analysis is also one analytic approach that could be used to integrate evidence across data streams for hazard identification (see Chapter 6) and that could be used for combining data or dose-response estimates of individual studies to derive toxicity values (see Chapter 7).

The terms *weight of evidence (WOE)* and *WOE analysis* are sometimes used interchangeably with *systematic review*, and there is often confusion surrounding the meanings of the terms. However, the committee views a WOE analysis as a judgment-based process for evaluating the strength of evidence to infer causation, that is, as one approach to integrating evidence for hazard identification. The meaning and use of WOE analysis are discussed further in Chapter 6 of this report.

THE COMMITTEE, ITS TASK, AND ITS APPROACH

The committee, which was convened, included experts in epidemiology, toxicology, dose-response modeling, risk assessment, systematic review, and risk communication (see Appendix A for biographic information on the committee). As noted earlier, it was asked to assess the scientific, technical, and process changes that EPA is making in the IRIS process. The verbatim statement of task is provided in Box 1-1.

The committee held six committee meetings to accomplish its task. Open sessions were held during the first and second meetings in which the committee heard from the sponsor on changes being made in the IRIS process and from staff of the National Toxicology Program on changes being made in its chemical-assessment program. During the third meeting, the committee held a workshop that involved participants from academe, government agencies, and private organizations to address approaches used to evaluate and integrate evidence for use in an IRIS assessment (see Appendix B for the workshop agenda). In each open session, interested parties were allowed to address the committee. The committee reviewed materials provided by EPA that document changes that it has made or is planning to make in the IRIS process and materials submitted by interested parties.

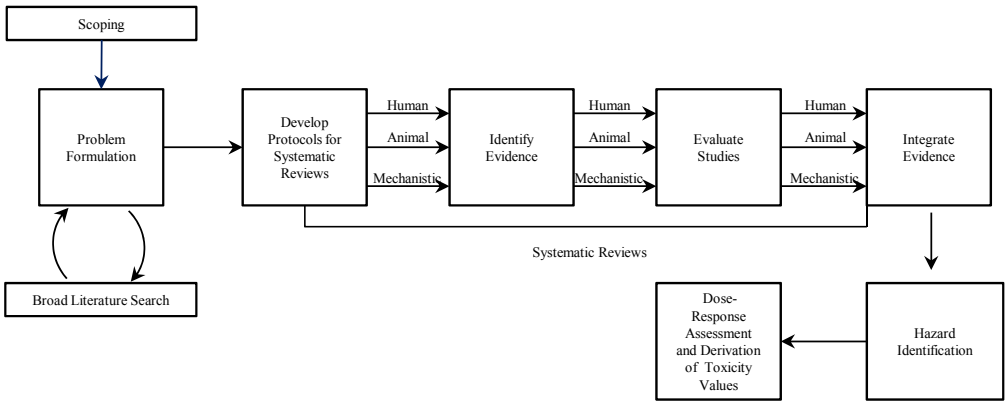


FIGURE 1-2 Systematic review in the context of the IRIS process. The committee views public input and peer review as integral parts of the IRIS process, although those activities are not specifically noted in the figure.

BOX 1-1 Statement of Task

A committee of the National Research Council (NRC) will assess the scientific, technical, and process changes being implemented by the U.S. Environmental Protection Agency (EPA) for its Integrated Risk Information System (IRIS). Specifically, the committee will review the IRIS process and the changes being implemented or planned by EPA and will recommend modifications or additional changes as appropriate to improve the scientific and technical performance of the IRIS program. The committee will focus on the development of the IRIS assessments rather than the review process that follows draft development. Because several reviews of IRIS assessments have expressed concerns about EPA's weight-of-evidence analyses, the committee will review current methods for evidence-based reviews and recommend approaches for weighing scientific evidence for chemical hazard and dose-response assessments.

ORGANIZATION OF REPORT

The present report is organized into eight chapters and three appendixes. Chapter 2 provides an overview of some general issues associated with IRIS assessments. Chapter 3 describes the need for problem formulation to define the systematic-review questions and protocol development to describe the methods used in the systematic reviews. Chapters 4 and 5 address evidence identification and evidence evaluation, respectively. Chapter 6 discusses evidence integration for hazard identification, and Chapter 7 evaluates methods for deriving toxicity values. Chapter 8 presents some overall findings and considerations for future directions. Appendix A provides biographic information on the committee, Appendix B is the agenda of the committee's workshop, and Appendix C provides some background information on Bayesian analysis.

REFERENCES

Chalmers, I., L.V. Hedges, and H. Cooper. 2002. A brief history of research synthesis. Eval. Health Prof. 25(1):12-37.

EPA (U.S. Environmental Protection Agency). 2002. Guidelines for Ensuring and Maximizing the Quality, Objectivity, Utility, and Integrity of Information Disseminated by the Environmental Protection Agency. EPA/260R-02-008. Office of Environmental Information, U.S. Environmental Protection Agency, Washington, DC [online]. Available: http://www.epa.gov/quality/informationguidelines/documents/EPA_InfoQualityGuidelines.pdf [accessed Aug. 13, 2013].

EPA (U.S. Environmental Protection Agency). 2003. A Summary of General Assessment Factors for Evaluating the Quality of Scientific and Technical Information. EPA 100/B-03/001. Science Policy Council, U.S. Environmental Protection Agency, Washington, DC [online]. Available: http://www.epa.gov/stpc/pdfs/assess2.pdf [accessed Aug. 13, 2013].

EPA (U.S. Environmental Protection Agency). 2004. An Examination of EPA Risk Assessment Principles and Practices. EPA/100/B-04/001. Office of the Science Advisor, U.S. Environmental Protection Agency Washington, DC [online]. Available: http://www.epa.gov/osa/pdfs/ratf-final.pdf [accessed Aug. 13, 2013].

EPA (U.S. Environmental Protection Agency). 2005. Guidelines for Carcinogen Risk Assessment. EPA/630/P-03/001F. Risk Assessment Forum, U.S. Environmental Protection Agency, Washington, DC [online]. Available: http://www.epa.gov/raf/publications/pdfs/CANCER_GUIDELINES_FINAL_3-25-05.PDF [accessed Aug. 13, 2013].

EPA (U.S. Environmental Protection Agency). 2011. EPA Strengthens Kay Scientific Database to Protect Public Health. EPA News: July 12, 2011 [online]. Available: http://yosemite.epa.gov/opa/admpress.nsf/48f0fa7dd51f9e9885257359003f5342/a3fcd608381970678525 78cb00666c4d!OpenDocument [accessed Aug. 13, 2013].

EPA (U.S. Environmental Protection Agency). 2012a. EPA's Integrated Risk Information Program, Progress Report and Report to Congress. Office of Research and Development, U.S. Environmental Protection Agency [online]. Available: http://www.epa.gov/iris/pdfs/irisprogressreport2012.pdf [accessed Aug. 13, 2012].

EPA (U.S. Environmental Protection Agency). 2012b. Integrated Risk Information System Program: Summary Report from November 2012 Public Stakeholder Meeting [online]. Available: http://www.epa.gov/iris/publicmeeting/stakeholders-kickoff/ [accessed Aug. 13, 2013].

EPA (U.S. Environmental Protection Agency). 2013. Enhancements to EPA's Integrated Risk Information System Program [online]. Available: http://www.epa.gov/iris/pdfs/irisprocessfactsheet2013.pdf [accessed Aug. 13, 2013].

GAO (U.S. Government Accountability Office). 2008. Toxic Chemicals: EPA's New Assessment Process Will Increase Challenges EPA Faces in Evaluating and Regulating Chemicals. GAO-08-743T. Washington, DC: U.S. Government Accountability Office [online]. Available: http://www.gao.gov/assets/120/119860.pdf [accessed Aug. 13, 2013].

IOM (Institute of Medicine). 2011. Finding What Works in Health Care: Standards for Systematic Reviews. Washington, DC: National Academies Press.

Light, R.J., and D.B. Pillemer. 1984. Summing Up: The Science of Reviewing Research. Cambridge MA: Harvard University Press.

NRC (National Research Council). 2006. Health Risks from Dioxin and Related Compounds: Evaluation of the EPA Reassessment. Washington, DC: National Academies Press.

NRC (National Research Council). 2010. Review of the Environmental Protection Agency's Draft IRIS Assessment of Tetrachloroethylene. Washington, DC: National Academies Press.

NRC (National Research Council). 2011. Review of the Environmental Protection Agency's Draft IRIS Assessment of Formaldehyde. Washington, DC: National Academies Press.

2

General Process Issues

In response to recommendations provided in the 2011 National Research Council (NRC) report *Review of the Environmental Protection Agency's Draft IRIS Assessment of Formaldehyde*, the US Environmental Protection Agency (EPA) has begun to revise its process for developing toxicologic assessments under its Integrated Risk Information System (IRIS). Materials provided to the present committee (see Table 1-1) indicate that many NRC recommendations have been implemented, and others are yet to be implemented.

Some recommendations from the NRC formaldehyde report dealt with specific steps of the IRIS process, and others dealt more generally with the overall process. The present chapter focuses on the general recommendations from that report, assesses EPA's response to them, and provides further suggestions and guidance for refining what has been implemented. It also discusses two general issues concerning the IRIS process: increasing efficiency and using expert judgment. The chapters that follow review and assess EPA's response to the recommendations specific to various steps of the IRIS process, from framing the assessment through identifying, evaluating, and integrating the evidence to deriving toxicity values (see Figure 1-2).

GENERAL RECOMMENDATIONS IN THE 2011 NATIONAL RESEARCH COUNCIL FORMALDEYDE REPORT

The EPA report *Status of Implementation of Recommendations* (EPA 2013a) summarized the general recommendations of the NRC formaldehyde report (NRC 2011) (see Box 2-1). That report's recommendations focused on improving the clarity of the assessments by rigorous editing, reducing text volume, and addressing redundancies and inconsistencies; describing assessment methods more fully; enhancing quality-control processes for assessments; standardizing review and evaluation approaches; and ensuring that the various chemical-assessment teams included appropriate expertise.

RESPONSE TO THE NATIONAL RESEARCH COUNCIL FORMALDEHYDE REPORT

EPA has made multiple changes in the IRIS program to address the general recommendations in the NRC formaldehyde report (NRC 2011). Although not called for by the formaldehyde report, new leadership for the IRIS program is in place, including Kenneth Olden (former director of the National Institute of Environmental Health Sciences) as the director of EPA's National Center for Environmental Assessment and Vincent Cogliano (former section head of the Monograph Section of the International Agency for Research on Cancer [IARC]) as the acting director of the IRIS program. EPA has recently adopted a new document structure for IRIS assessments, drafted a preamble to be included with all IRIS assessments, and instituted several changes to enhance quality-control processes within the IRIS program. Those changes are reviewed in the following sections.

BOX 2-1 General Recommendations on the IRIS Process in the 2011 National Research Council Formaldehyde Report

- To enhance the clarity of the document, the draft IRIS assessment needs rigorous editing to reduce the volume of text substantially and address redundancies and inconsistencies. Long descriptions of particular studies, for example, should be replaced with informative evidence tables. When study details are appropriate, they could be provided in appendices.
- Chapter 1 needs to be expanded to describe more fully the methods of the assessment, including a description of search strategies used to identify studies with the exclusion and inclusion criteria clearly articulated and a better description of the outcomes of the searches…and clear descriptions of the weight-of-evidence approaches used for the various noncancer outcomes. The committee emphasizes that it is not recommending the addition of long descriptions of EPA guidelines to the introduction, but rather clear concise statements of criteria used to exclude, include, and advance studies for derivation of the RfCs [reference concentrations] and unit risk estimates.
- Elaborate an overall, documented, and quality-controlled process for IRIS assessments.
- Ensure standardization of review and evaluation approaches among contributors and teams of contributors; for example, include standard approaches for reviews of various types of studies to ensure uniformity.
- Assess disciplinary structure of teams needed to conduct the assessments.

Source: NRC 2011, pp. 152, 164.

New Document Structure

A major concern of the 2011 NRC formaldehyde report was that IRIS assessments had become too expansive, were poorly organized and edited, and thus were deficient in communicating clearly, cogently, and transparently how conclusions were reached in the assessments. Specifically, in the formaldehyde IRIS assessment, the basis of the conclusions on health outcomes was not clear and transparent, and the rationales for selecting studies for deriving quantitative toxicity values for cancer and noncancer outcomes were not well developed or consistent.

Consequently, the NRC committee that reviewed the formaldehyde assessment offered a number of general suggestions to enhance document clarity, including reducing document length, editing rigorously, eliminating redundancies and inconsistencies, and using evidence tables rather than long narrative descriptions of individual studies. In response, EPA developed and has implemented a new document structure that streamlines the assessments (EPA 2013a) and appropriately organizes them into two broad sections: hazard identification and dose-response analysis. EPA noted that the new organization aligns better with the traditional risk-assessment paradigm. The new document structure also includes an executive summary to highlight major conclusions. Recent IRIS assessments reflect the new document structure (EPA 2013b,c).

The committee agrees that the new document structure, which is reflected in the toxicologic review of benzo[a]pyrene (EPA 2013c), leads to better organized and streamlined assessments and reduces redundancies. It also notes that EPA has embraced the use of evidence tables and graphic displays of study findings to reduce text volume and enhance clarity and transparency. As a result, the descriptions of individual studies have been shortened. That approach brings the IRIS assessments much more in line with the state of practice for systematic reviews.

IRIS Assessment Preamble

Another important concern of the formaldehyde committee was that the assessment methods were not clearly articulated. At the time of that review, EPA typically provided a brief generic introductory chapter that simply referred to EPA guidelines. Accordingly, the formaldehyde committee recommended that EPA explain more fully the methods and approaches used. In response, EPA has developed a preamble to be included with all IRIS assessments that follows the model used by IARC and that "describes the application of existing EPA guidance and the methods and criteria used in developing the assessments" (EPA 2013a, p. 6). It discusses the general scope and elements of the IRIS program; the peer-review process for the IRIS assessments; the identification, selection, and evaluation of studies; integration of evidence; and derivation of toxicity values.

The present committee finds that the preamble is a useful statement, which presumably will be revised as methods and procedures are modified and updated. As appropriate, each IRIS assessment should note the version of the preamble included in the assessment. The broad description of principles and methods in the preamble does not replace, however, the need for a brief description in each IRIS assessment that indicates how the general principles described in the preamble have been applied in the case of that specific assessment. For example, specific health outcomes and populations that are considered might vary from chemical to chemical, and this variation might lead to notable methodologic differences between assessments. Much of the methodologic detail specific to individual assessments might be presented in the protocols for the systematic reviews conducted as part of the IRIS process; these could be provided as appendixes to each assessment (see Chapter 3 for further discussion).

Initiatives to Improve the Overall Process, Quality Control, and Documentation

To improve the overall process, quality control, and documentation, EPA has developed or adopted several new initiatives, including development of guidance for chemical-assessment teams, institution of chemical-assessment support teams (CASTs), development of information-management tools, initiation of scoping exercises, and expansion of stakeholder engagement (EPA 2013a). The following sections review and evaluate each initiative.

Guidance for Chemical-Assessment Teams

The chemical-assessment team for each IRIS assessment is a multidisciplinary team. The teams often include contractors who provide technical and analytic support, such as conducting literature searches or performing dose-response analyses. (The committee discusses the CASTs, which have a broader role, below.) A concern of the NRC formaldehyde committee was that all team members use a standard and consistent approach in the assessment-development process. In response, EPA is developing instructions for contractors and has provided an example of instructions for conducting dose-response modeling to the present committee (EPA 2013a, Appendix C). EPA is also developing a handbook to guide the development of IRIS assessments (EPA 2013a, Appendix F); however, the committee does not fully understand relationships among the various documents and is concerned that the existence of several guidance documents might cause confusion. A better approach might be to develop a single handbook that can be used by members of the chemical-assessment team regardless of whether they are EPA staff or contractors. In any case, the purpose of and relationships among the various guidance documents—preamble, contractor instructions, and handbook—should be clearly stated.

Institution of Chemical-Assessment Support Teams

As discussed, IRIS assessments involve multifaceted and interdisciplinary chemical-assessment teams that have general expertise and expertise specific to the chemicals in question. In the past, the teams have been supported by discipline-specific work groups that helped to ensure methodologic consistency among IRIS assessments. In late 2011, EPA formally implemented an initiative to establish three CASTs, each composed of two senior scientists, a senior statistician, and a staff scientist to serve as a rapporteur (EPA 2013a). Every IRIS assessment is assigned to one CAST, and the CAST meets with the chemical-assessment team associated with the IRIS assessment to which it has been assigned. The three CASTs then meet weekly to discuss issues that have arisen in their chemical-specific meetings. The CAST initiative was developed to meet several objectives: to provide a forum for problem-solving; to ensure that appropriate scientific expertise is available to each team; to identify problems or issues early in the process, particularly ones that need program-wide discussion; to increase objectivity and consistency among assessments; to monitor the implementation of recommendations of the NRC formaldehyde report; and to assist in responding to peer review and public comments and in documenting and communicating decisions (EPA 2013a).

This initiative addresses the clear need for continuing and consistent expert oversight of individual assessments and the overall IRIS program. The committee endorses EPA's efforts and suggests two advances that will strengthen the overall scientific expertise and leadership in IRIS assessments further. First, the draft IRIS assessments need to identify more clearly and explicitly the members of all teams involved in the work. The identity of CAST members is not included in the report template provided to the committee (EPA 2013a, Appendix A) or in recent draft assessment reports (see, for example, EPA 2013c), and, more important, the roles of team members in various reports are not identified at least at the level now required for authors and other contributors in most biomedical journals. Second, the draft and final reports are subject to considerable peer review within EPA, by federal agencies, and by external reviewers. However, the committee recommends that expert judgment from outside EPA and the government be involved to fill critical gaps in expertise; this could be accomplished with the new Chemical Assessment Advisory Committee (CAAC) of the EPA Science Advisory Board. For example, CAAC members could provide specific expertise for a chemical being assessed by periodically reviewing activities at critical stages of each IRIS assessment and interacting with the chemical-assessment team and the CAST assigned to the assessment. The need for increased expert judgment in the IRIS process is discussed further in the section "Using Expert Judgment in the IRIS Process."

Development of Information-Management Tools

In its status report to the committee, EPA described several information-management tools that should promote quality control in the IRIS process (EPA 2013a, Appendix D). First, EPA has developed the Comment Tracker Database, which captures peer-review and public comments and EPA's responses to them. It facilitates project management by allowing chemical managers to make clear assignments of comments to various team members and facilitates identification of comments that will require substantial time or resources to address. The database should also help to identify recurrent comments (or themes) and therefore prompt a more in-depth review of the comments.

Second, EPA is developing a Cross-Chemical Comparisons Database, which will allow EPA to search for common topics or issues that arise in different chemical assessments. The database should allow EPA staff to determine how scientific issues have been addressed in various chemical assessments, to identify issues that are raised repeatedly by reviewers and stakeholders, and to compare comments provided at various stages of the IRIS process and identify possible inconsistencies.

Third, EPA has invested in a substantial expansion of its Health and Environmental Research Online (HERO) database that provides access to the scientific literature identified for IRIS chemicals, where available, and highlights results of broad literature searches. Some chemicals include a "LitFlow" link that provides visual highlights of the broad search, including reference counts from database searches and additional search strategies and links to the references that were identified, considered, and excluded and to the primary sources of health-effects data.

The committee finds that the management tools that EPA has described should help with quality assurance and management of the IRIS process, but a systematic approach to using the accumulated information will need to be developed to strengthen the process further. If some management tools are not available to the public, EPA will face a challenge in maintaining full transparency.

Initiation of Scoping for IRIS Assessments

EPA (2013a) identifies the need for a scoping process to ensure that each assessment is as informative and useful as possible for the various groups that will use IRIS assessments. EPA (2013a, Appendix E) is consistent with the risk-assessment guidance provided in the report *Science and Decisions: Advancing Risk Assessment* (NRC 2009) in describing the scoping process as one that seeks input from EPA program and regional offices to identify the information and the level of detail needed to inform their decisions. For example, Are some exposure routes or durations of concern? Are some specific life stages, exposure windows, or groups of particular concern? The desired "outcome of the scoping process is a statement that outlines the focus of the assessment, the nature of the hazard characterization needed, and a clear indication of issues that are beyond the scope of the IRIS assessment" (EPA 2013a, p. E-3). The committee notes that scoping is different from problem formulation, which is an early step in the systematic-review process that explicitly defines what is to be evaluated in the assessment and how it is to be evaluated. (See Chapter 3 for a discussion of problem formulation.)

Expansion of Stakeholder Engagement

EPA (2013a, p. 9) acknowledges the importance of stakeholder engagement, lists opportunities that it has provided for stakeholder input throughout the IRIS process, and indicates its plan for expanded stakeholder engagement. EPA's initiatives are consistent with recommendations from past NRC reports, which have repeatedly called for robust stakeholder involvement in environmental decision-making. For example, NRC (2009, p. 13) stated that "greater stakeholder involvement is necessary to ensure that the process is transparent and that risk-based decision-making proceeds effectively, efficiently, and credibly. Stakeholder involvement needs to be an integral part of the risk-based decision-making framework, beginning with problem formulation and scoping." The present committee agrees that early and continuing stakeholder involvement not only will increase the likelihood that EPA will address the concerns of diverse stakeholders but should strengthen the quality of IRIS assessments.

In considering initiatives to expand stakeholder involvement, the various stakeholder groups should be recognized. Stakeholder groups are often identified as having opposing viewpoints regarding chemical risk assessment. Nongovernment organizations, such as environmental advocacy groups, often represent people who might be exposed to the substances that IRIS reviews. Those groups generally seek more conservative or protective health standards and call for the rapid completion of IRIS assessments. They have repeatedly expressed a concern that the public might be exposed to substances that threaten health and safety because assessments have not been completed in a timely manner (Sass and Rosenberg 2011; Denison 2012). In contrast, other organizations and individuals represent industrial and government entities that produce, use, and release chemicals, some of which are toxic. Those stakeholders typically express a con-

cern that scientifically unjustifiably conservative toxicity values will prove costly and provide relatively little additional protection of public health, so they often argue for less protective standards or urge more study before IRIS assessments are completed. In reality, the array of stakeholders who are interested in the IRIS process is much broader. For example, risk assessors, policy setters, and other public-health officials need toxicity values on IRIS to be timely, up to date, protective of public health (including sensitive populations), informative about all relevant end points, and transparent about uncertainties that might result in underestimation or overestimation of the actual hazard. Furthermore, the scientific community—toxicologists, epidemiologists, and other groups of scientists—generally favors the incorporation of valid scientific research and information into the public-policy process, plays a key role in development of toxicity-testing methods (Ashby 2003), and is often at the forefront of identifying toxic hazards.

A well-designed process of stakeholder engagement in the development of each assessment should keep all stakeholders informed, provide suitable opportunities for diverse input, and promote the smooth and timely completion of the draft assessment. Stakeholder involvement today begins with the nomination of substances for review, but not all potential stakeholders are likely to be aware of this opportunity. The IRIS program lists planned reviews in the *Federal Register* (78 Fed. Reg. 48674 [2013]) and on its Web site. The committee suggests that EPA publish an IRIS workplan at least once a year. Furthermore, there should be a clear and readily accessible process for parties outside EPA to suggest new chemical assessments and revisions of completed assessments on the basis of new evidence.

As noted earlier, EPA (2013a) indicated opportunities for expanded stakeholder input. In July 2013, it outlined a process that includes three public meetings (EPA 2013d,e). First, EPA promises to conduct a "public meeting focused on identifying the scientific information available for the chemical under assessment" after an internal planning and scoping meeting. In January 2013, EPA included public stakeholders and other federal agencies in planning meetings for the inorganic arsenic assessment. The committee supports EPA's plan to release a draft planning and scoping summary before such meetings and a final summary afterward (V. Cogliano, EPA, personal commun., August 20, 2013). However, the final summary should be completed quickly—the committee notes that the summary of a November 2012 stakeholder meeting was not released until August 2013 (EPA 2013f).

The second public meeting is slated to occur during the draft-development process. EPA plans to "release the literature search and a search strategy, evidence tables, exposure-response figures, and, as appropriate, information on anticipated key scientific issues for the chemical" (EPA 2013d). EPA noted that "out of consideration for stakeholders with limited resources, they will be released at one time. This way, stakeholders do not have to spend time and money retrieving and extracting data from hundreds of papers…The public will be asked whether any important data were missed" (V. Cogliano, EPA, personal commun., August 20, 2013). EPA clearly is proceeding with these types of meetings: one was held in December 2013, at which preliminary materials developed for the draft assessments of ethyl *tert*-butyl ether, *tert*-butyl alcohol, and hexahydro-1,3,5-trinitro-1,3,5-triazine were discussed.[1] The committee finds that EPA's willingness to engage in early discussion with its stakeholders is a major step forward, and it urges EPA to maintain some flexibility in the process. For example, if public discussion leads to the discovery or selection of important new studies, EPA might wish to develop new evidence tables for discussion at the next meeting. That flexibility would be beneficial if important new data were brought forward. The committee, however, is not suggesting that EPA routinely add new steps to the assessment-development process.

Finally, after EPA completes each draft assessment and coordinates with other federal agencies and the executive branch, there will be—as there were before EPA announced process

[1]See http://www.epa.gov/iris/publicmeeting/iris_bimonthly-dec2013/index.htm.

enhancements—a formal public comment period and a public meeting to receive public comments, both of which have been and will be announced in the *Federal Register*.

Even in the face of expanded transparency and enhanced stakeholder engagement, there is concern about the uneven participation of the first two principal stakeholder groups. Most public comments on draft IRIS assessments have come from industry or parties representing the interests of entities that produce, use, and release possibly toxic substances. Indeed, almost all the public input—written and oral—received by the present committee has come from trade organizations. Furthermore, from January 2011 to October 2013, over 100 distinct substantive comments were submitted to the IRIS program.[2] Representatives of entities that produce, use, or release the studied substances submitted over 80 comments. In that period, only a few comments were submitted by public-interest organizations concerned with the environment. Comments submitted by concerned citizens or entities apparently representing them contained little or no specific scientific information that might influence the IRIS program's findings.

Industry representation and input constitute an important element of stakeholder participation, and its comments are often cogent and constructive. Some industry stakeholders also have the resources for initiating and quickly completing literature reviews and research that might be relevant to a particular assessment (for example, studies of formaldehyde dosimetry). However, other key stakeholders have fewer resources and are not generally organized and staffed to provide comments or detailed scientific input. Thus, their important perspectives and voices might be less well represented to EPA. Therefore, the committee encourages EPA to continue the additional efforts to ensure that the full breadth of perspectives on the IRIS process and specific IRIS assessments are made available to the agency.

One way to ensure broad stakeholder input would be to provide technical assistance to enable under-resourced stakeholders to develop and provide input to the IRIS program; this could be modeled after other EPA technical-assistance programs. For example, EPA's Superfund program has a long history of providing technical assistance in the form of grants and more recently direct consultation to neighbors of sites on the National Priorities List (EPA 2012a). The grants generally improve the process of remedial decision-making by ensuring that the affected public understands both the characterization and the remediation of hazardous-waste contamination and by making it easier for such people to provide constructive input (EPA 2012b).

Assessment of Overall Quality

As discussed above, EPA (2013a) has addressed several issues in managing the overall quality of the IRIS process. One important step is the development of the draft handbook (EPA 2013a, Appendix F), which provides detailed guidance for various steps of the IRIS process. EPA (2013a, Appendix C) is also drafting guidance for contractors. As noted above, the committee concludes that the best approach might be to provide a single detailed guidance document for all those involved in the development of IRIS assessments. Multiple guidance documents could create confusion and inconsistencies. If multiple guidance documents are developed, EPA should provide specific explanations of the applications of the various documents, and the relationships among them should be clearly stated.

It is important to ensure that work done by all contributors to the process meets quality standards. For example, the selection of studies for initial consideration and for further review (see Chapter 4) is critical, as is the evaluation of the risk of bias in individual studies (see Chapter 5). Although the draft handbook (EPA 2013a, Appendix F) provides guidance that is generally informative and useful, it fails to define specific procedures for estimating and evaluating the reliability and validity of processes that are central to the hazard-identification part of the pro-

[2]See http://www.regulations.gov/#!searchResults;rpp=25;po=0;s=EPA%252BIRIS;fp=true;ns=true.

cess, such as identifying, selecting, and evaluating evidence. Accordingly, the processes need empirical investigation. For example, there needs to be an assessment of whether the process for identifying studies is replicable (that is, whether the protocol leads to the same body of studies when repeated, possibly by different teams) and of whether the process for study selection is valid (that is, whether the protocol, as applied in routine practice, leads to the same studies that would be selected by an expert team). Both reliability and validity should be assessed in empirical studies of reasonable size. Such data would provide information for evaluating the quality of the IRIS process and improving it through better selection or training of those who select and rate the relevance of the studies and who abstract data for the systematic review. Conducting studies that will lead to standards for rater and abstracter reliability and validity should in the long run be cost-effective by reducing error. Such methodologic studies would not need to be performed for each assessment once standards are established and tracking implemented to ensure that the standards are met. The committee notes that no explicit measures or systems are provided by EPA for assessing the reliability and validity of contractor work as compared with similar work by EPA staff or external "gold standards."

The committee applauds EPA's initiatives that are designed to enhance the quality of the IRIS process and highlights below several quality-control measures that could be implemented or developed further from EPA's initial efforts.

- Develop explicit timelines for the various components of IRIS assessments. The committee recognizes that IRISTrack on the EPA Web site provides some information, but the information is often too general or incomplete (for example, "TBD" [to be determined] is listed as a completion date for many chemicals).
- Develop explicit guidelines for external researchers and laboratories that are providing useful data for the IRIS process, such as raw data to facilitate reanalysis by EPA. Additional guidance regarding laboratory protocols might be needed for some types of data, such as high-throughput data.
- Implement a periodic strategic planning process that allows the IRIS program to identify long-term needs and goals with a 3- to 10-year horizon.
- Provide IRIS staff with opportunities for continuing training to help to ensure the application of current hazard-assessment and dose-response practices as these practices evolve.

The committee notes that there might be quality-control processes promulgated in other federal agencies that could be exploited to improve quality management, and EPA might want to investigate other similar federal programs.

INCREASING EFFICIENCY IN THE IRIS PROCESS

EPA has a monumental task in safeguarding the public's health; the IRIS program is an integral part of that effort. The agency must navigate within the constraints of its resources between the necessity of making scientifically valid and informative assessments of the health consequences of chemical exposures and the need to do so in a timely and cost-effective manner. Given the challenges facing EPA, organizational, managerial, and scientific efficiency becomes critical, particularly in light of the constraint of inevitably shrinking resources. Thus, promoting efficiency in the IRIS program is paramount.

The committee shares EPA's view that establishing transparent, consistent processes that include opportunities for stakeholder input might reduce delays and promote efficiency in the IRIS process. Furthermore, as noted by EPA (2013d), implementing stopping rules that establish flexible cutoff points for the acceptance of new studies and data should improve productivity in the IRIS process. Participants in the IRIS stakeholder meeting in November 2012 suggested implementing stopping rules to reduce delays (EPA, 2013f, p. 9). The committee emphasizes that

any stopping rules for a particular assessment would need to be grounded in general principles regarding what constitutes pivotal evidence and a reasonable period of delay. EPA should add appropriate text to the preamble, such as "An assessment might be delayed while awaiting potentially pivotal evidence from further analysis or follow-up of a critical epidemiologic study or from a critical animal study."

Several additional suggestions that might be of future use to the IRIS program to promote efficiencies in both the short and long terms are provided below.

- Enhance interactions with other agencies or organizations, within and outside government, to identify existing information and chemical evaluations that might be used, if the external methods are sound and appropriate, instead of recreating them. Avoiding duplication of effort is an important efficiency-promoting activity.
- Continue to expand efforts to develop computer systems, such as the HERO database, that facilitate storage and annotation of information relevant to IRIS's mission. Whenever possible, interagency efforts should be considered to enhance efficiency further and reduce duplication of effort.
- Continue development of automated literature and screening procedures, sometimes referred to as text-mining. Such approaches offer the possibility of recurrent, automated literature searching for relevant papers or related papers. Text-mining tools are available from the US National Library of Medicine (Lu 2011) and are also being developed by EPA.
- Promote within EPA a research program that studies the best way to use and incorporate data that are being generated from new in vitro, in silico, and high-throughput toxicity testing into the IRIS process.

USING EXPERT JUDGMENT IN THE IRIS PROCESS

The name *Integrated Risk Information System* might suggest to some a rather mechanical and automated system of assessing the health consequences of chemical exposure. The name might also imply that it is desirable to make any information-gathering and assessment process as free of human judgment, and hence potential human error, as possible. However, all steps of the IRIS process, especially the evidence integration and conclusions reached, are necessarily laden with human judgment, as are most scientific endeavors.

Expert judgment is often used to resolve problems in the face of uncertainty, but its application to toxicity assessments is often poorly described and often considered a "black box." As noted by Pronk et al. (2012), "the lack of explicit rules makes it difficult to determine the consistency of expert-based decisions over time" or among assessments. In response, the committee sees two approaches that EPA could take to address its concerns regarding the use of expert judgment in the IRIS process. First, EPA can further systematize expert-judgment procedures in the IRIS process and establish their validity and reliability through methodologic research. Second, EPA can ensure balance and expertise in such judgments through expert peer review. Both quality-control mechanisms are necessary in the IRIS process, and EPA appears to be already pursuing them to some extent.

Developing expertise in a specific domain is considered to take at least a decade of dedicated training and study (Ericsson et al. 1993). Experts understand the relationship of concepts within their specific domains and are able to take an organized approach to identifying the elements of a problem and what is known and not known regarding it (Larkin et al. 1980; Chi et al. 1988; Ericsson and Chamess 1994). It follows that in tasks requiring expert judgment discussed later in this report—for example, designing search strategies (see Chapter 4), identifying confounders (see Chapter 5), integrating the evidence for hazard identification (see Chapter 6), and determining how to translate knowledge into prior distributions for analysis in a Bayesian model (see Chapters 6 and 7)—it is essential that appropriate domain-specific expertise be identified

and included with recognition that experts in different fields will probably be required, depending on the task.

As noted in EPA (2011), expert judgment can be susceptible to cognitive biases, although less than lay judgment (Gilovich et al. 2002; Koehler et al. 2002), and how information is presented has the potential to alter judgments. For expert-judgment elicitations, one needs to formulate clear questions, to develop formal protocols, and to summarize and share relevant evidence with the experts (see, for example, EPA 2011, pp. 13-14). As with any such elicitation, the structure of expert judgment in group settings in the context of the IRIS program (as described in Chapter 6 of this report) deserves close attention.

Several steps of the IRIS process require competent professional expert judgment, and the committee concludes that there needs to be a stronger role throughout the IRIS process for expert judgment derived from broadly expert and representative panels, perhaps, as suggested above, as an adjunct to the new CAAC. Integrating the evidence and deriving toxicity values especially should be recognized as requiring a high level of expert judgment to make the conclusion reached and the values derived as valid, reliable, and reputable as possible. At the same time, there is a need for more systematic procedures for assessing the reliability and validity of aspects of the IRIS process—such as literature searching, screening, and evaluation—that are most amenable to the development and application of systematic procedures that can be cost-effectively implemented by competent professional staff. The history of subjectivity in science, the arts, and esthetics goes back a long way and still causes tension in scientific discourse (Shapin 2011; Klempe 2012). The only tentative solution is to describe as accurately as possible the methods by which scientific and policy decisions are made, by whom, and with what expertise.

FINDINGS AND RECOMMENDATIONS

Finding: The committee is impressed and encouraged by EPA's progress, recognizing that the implementation of the recommendations in the NRC formaldehyde report is still in process. If current trajectories are maintained and objectives still to be implemented are successfully brought to fruition, the IRIS process will become much more effective and efficient in achieving its basic goal of developing human-health assessments that can provide the scientific foundation for ensuring that risks posed to public health by chemicals are assessed and managed optimally.

Recommendation: EPA needs to complete the changes in the IRIS process that are in response to the recommendations in the NRC formaldehyde report and specifically complete documents, such as the draft handbook, that provide detailed guidance for developing IRIS assessments. When those changes and the detailed guidance, such as the draft handbook, have been completed, there should be an independent and comprehensive review that evaluates how well EPA has implemented all the new guidance. The present committee is completing its report while those revisions are still in progress.

Finding: Although it is clear that quality control (QC) of the IRIS assessment process is critical for the outcome of the program, the documents provided do not sufficiently discuss the QC processes or provide guidelines that adequately separate the technical methods from the activities of QC management and program oversight. For example, the role of the CASTs in the QC process is not specifically described.

Recommendation: EPA should provide a quality-management plan that includes clear methods for continuing assessments of the quality of the process. The roles of the various internal entities involved in the process, such as the CASTs, should be described. The assessments should be used to improve the overall process and the performance of EPA staff and contractors.

Recommendation: When extracting data for evidentiary tables, EPA should use at least two reviewers to assess each study independently for risk of bias. The reliability of the independent coding should be calculated; if there is good agreement, multiple reviewers might not be necessary.

Finding: The current scoping process for obtaining input from within the agency is clear, but opportunities for stakeholder input from outside EPA early in the process are less clear.

Recommendation: EPA should continue its efforts to develop clear and transparent processes that allow external stakeholder input early in the IRIS process. It should develop communication and outreach tools that are tailored to meet the needs of the various stakeholder groups. For example, EPA might enhance its engagement with the scientific community through interactions at professional-society meetings, advertised workshops, and seminars. In contrast, greater use of social media might help to improve communications with environmental advocacy groups and the public.

Finding: EPA has taken steps to expand opportunities for stakeholder input and discussion that are likely to improve assessment quality. However, not all stakeholders with an interest in the IRIS process have the resources to provide timely comments.

Recommendation: Similar to other EPA technical-assistance programs, EPA should consider ways to provide technical assistance to under-resourced stakeholders to help them to develop and provide input to the IRIS program.

Finding: Promoting efficiency in the IRIS program is paramount given the constraint of inevitably shrinking resources. Thus, the committee agrees with EPA that stopping rules are needed given that the process for some IRIS assessments has become too long as revisions are repeatedly made to the assessments to accommodate new evidence and review comments.

Recommendation: The stopping rules should be explicit and transparent, should describe when and why the window for evidence inclusion should be expanded, and should be sufficiently flexible to accommodate truly pivotal studies. Such rules could be included in the preamble.

Recommendation: Regarding promotion of efficiencies, EPA should continue to expand its efforts to develop computer systems that facilitate storage and annotation of information relevant to the IRIS mission and to develop automated literature and screening procedures, sometimes referred to as text-mining.

Finding: The draft handbook and other materials are useful but lack explicit guidance as to the methods and nature of the use of expert judgment throughout the full scope of the assessment-development process, from literature searching and screening through integrating evidence to analyzing the dose-response relationship and deriving final toxicity values.

Recommendation: More details need to be provided on the recognition and applications of expert judgment throughout the assessment-development process, especially in the later stages of the process. The points at which expert judgment is applied should be identified, those applying the judgment should be listed, and consideration should be given to harmonizing the use of expert judgment at various points in the process.

REFERENCES

Ashby, J. 2003. The leading role and responsibility of the international scientific community in test development. Toxicol. Lett. (140-141):37-42.

Chi, M.T.H., R. Glaser, and M.J. Farr, eds. 1988. The Nature of Expertise. Hillsdale, NJ: Erlbaum.

Denison, R.A. 2012. EDF Comments for EPA IRIS Stakeholder Panel. Environmental Defense Fund, November 13, 2012 [online]. Available: http://www.epa.gov/iris/publicmeeting/stakeholders-kickoff/Denison_IRIS_Stakeholder_Comments.pdf [accessed Nov. 20, 2013].

EPA (U.S. Environmental Protection Agency). 2011. Expert Elicitation Task Force White Paper. Prepared for the Science and Technology Policy Council. August 2011 [online]. Available: http://www.epa.gov/stpc/pdfs/ee-white-paper-final.pdf [accessed Nov. 20, 2013].

EPA (U.S. Environmental Protection Agency). 2012a. Technical Assistance Grants [online]. Available: http://www.epa.gov/superfund/community/tag/ [accessed Nov. 20, 2013].

EPA (U.S. Environmental Protection Agency). 2012b. Superfund Technical Assistance: Giving Communities an Informed Voice [online]. Available: http://www.epa.gov/superfund/accomp/news/tag.htm [accessed Nov. 20, 2013].

EPA (U.S. Environmental Protection Agency). 2013a. Part 1. Status of Implementation of Recommendations. Materials Submitted to the National Research Council, by Integrated Risk Information System Program, U.S. Environmental Protection Agency, January 30, 2013 [online]. Available: http://www.epa.gov/iris/pdfs/IRIS%20Program%20Materials%20to%20NRC_Part%201.pdf [accessed Nov. 13, 2013].

EPA (U.S. Environmental Protection Agency). 2013b. Toxicological Review of Methanol (Noncancer) (CAS No. 67-56-1) in Support of Summary Information on the Integrated Risk Information System (IRIS). EPA/635/R-11/001Fa. U.S. Environmental Protection Agency, Washington, DC. September 2013 [online]. Available: http://www.epa.gov/iris/toxreviews/0305tr.pdf [accessed Nov. 13, 2013].

EPA (U.S. Environmental Protection Agency). 2013c. Toxicological Review of Benzo[a]pyrene (CAS No. 50-32-8) in Support of Summary Information on the Integrated Risk Information System (IRIS), Public Comment Draft. EPA/635/R13/138a. National Center for Environmental Assessment, Office of Research and Development, U.S. Environmental Protection Agency, Washington, DC. August 2013 [online]. Available: http://cfpub.epa.gov/ncea/iris_drafts/recordisplay.cfm?deid=66193 [accessed Nov. 13, 2013].

EPA (U.S. Environmental Protection Agency). 2013d. Enhancements to EPA's Integrated Risk Information System Program [online]. Available: http://www.epa.gov/iris/pdfs/irisprocessfactsheet2013.pdf [accessed Aug. 13, 2013].

EPA (U.S Environmental Protection Agency). 2013e. IRIS Process Flow Chart [online]. Available: http://www.epa.gov/iris/pdfs/IRIS_PROCESS_FLOW_CHART.PDF [accessed Nov. 20, 2013].

EPA (U.S. Environmental Protection Agency). 2013f. Integrated Risk Information System Program: Summary Report from November 2012 Public Stakeholder Meeting [online]. Available: http://www.epa.gov/iris/pdfs/Summary%20Report%20Nov2012%20IRIS%20Public%20Stakeholder%20Mtg.pdf [accessed Nov. 20, 2013].

Ericsson, K.A., and N. Chamess. 1994. Expert performance: Its structure and acquisition. Am. Psychol. 49(8):725-747.

Ericsson, K.A., R.T. Krampe, and C. Tesch-Römer. 1993. The role of deliberate practice in the acquisition of expert performance. Psychol. Rev. 100(3):363-406.

Gilovich, T., D. Griffin, and D. Kahneman, eds. 2002. Heuristics and Biases: The Psychology of Intuitive Judgment. Cambridge: Cambridge University Press.

Klempe, S.H. 2012. Psychology-tensions between objectivity and subjectivity. Integr. Psychol. Behav. Sci. 46(3):373-379.

Koehler, D.J., L. Brenner, and D. Griffin. 2002. The calibration of expert judgment: Heuristics and biases beyond the laboratory. Pp. 686-715 in Heuristics and Biases: The Psychology of Intuitive Judgment, T. Gilovich, D. Griffin, and D. Kahneman, eds. Cambridge: Cambridge University Press.

Larkin, J.H., J. McDermott, D.P. Simon, and H.A. Simon. 1980. Models of competence in solving physics problems. Cognitive Sci. 4(4):317-345.

Lu, Z. 2011. PubMed and beyond: A survey of web tools for searching biomedical literature. Database 2011:baq036. doi: 10.1093/database/baq036.

NRC (National Research Council). 2009. Science and Decisions: Advancing Risk Assessment. Washington, DC: National Academies Press.

NRC (National Research Council). 2011. Review of the Environmental Protection Agency's Draft IRIS Assessment of Formaldehyde. Washington, DC: National Academies Press.

Pronk, A., P.A. Stewart, J.B. Coble, H.A. Katki, D.C. Wheeler, J.S. Colt, D. Baris, M. Schwenn, M.R. Karagas, A. Johnson, R. Waddell, C. Verrill, S. Cherala, D.T. Silverman, and M.C. Friesen. 2012. Comparison of two expert-based assessments of diesel exhaust exposure in a case–control study: programmable decision rules versus expert review of individual jobs. Occup. Environ. Med. 69(10):752-758.

Sass, J., and D. Rosenberg. 2011. The Delay Game: How the Chemical Industry Ducks Regulation of the Most Toxic Substances. Natural Resources Defense Council, September 2011 [online]. Available: http://www.nrdc.org/health/files/IrisDelayReport.pdf [accessed Nov. 20, 2013].

Shapin, S. 2011. The sciences of subjectivity. Soc. Stud. Sci. 42(2):170-184.

3

Problem Formulation and Protocol Development

As discussed in Chapter 1 of this report, the US Environmental Protection Agency (EPA) is incorporating principles of systematic review as it revises the Integrated Risk Information System (IRIS). Critical elements of a systematic review include formulating the specific question that will be addressed and developing the protocol that specifies the methods that will be used to address it. The National Research Council (NRC) report that reviewed the IRIS formaldehyde assessment (NRC 2011) did not provide any specific recommendations regarding those elements, but the present committee found that some discussion of them is warranted given EPA's shift toward adopting systematic-review principles. Therefore, this chapter discusses problem formulation and protocol development as parts of the IRIS process and systematic review as shown in Figure 3-1. The committee distinguishes between the scoping exercise described in Chapter 2 and problem formulation described here. The scoping exercise involves soliciting input from EPA program and regional offices to determine the bounds of the assessment—such as exposure pathways and specific exposed groups to consider—that will help EPA with its decision-making, whereas problem formulation is intended to frame the specific scientific questions for the systematic reviews in the IRIS-assessment process. That distinction is consistent with the NRC report *Science and Decisions: Advancing Risk Assessment* (NRC 2009).

PROBLEM FORMULATION

The risk-assessment paradigm that dates to the 1983 report *Risk Assessment in the Federal Government: Managing the Process* (NRC 1983) and has long been used by EPA has four components: hazard identification, dose-response assessment, exposure assessment, and risk characterization. Two components are encompassed by the IRIS assessment process: identifying potential hazards related to a chemical by using the available literature as a source of information (hazard identification) and characterizing the dose-response relationship (dose-response assessment) (see Figure 3-2). Thus, problem formulation in the IRIS process is restricted to scientific questions that pertain only to those two elements of the risk-assessment paradigm. Although the committee's review of the problem-formulation step focuses mainly on searching available literature, seeking stakeholder input and advice is an integral part of the process and should not be minimized.

Evidence Used for IRIS Assessments

As indicated in Figure 3-2, evidence typically used by EPA for IRIS assessments comes from human studies, animal studies, and mechanistic studies. Those study types are briefly discussed below to set the stage for further discussion in the present report.

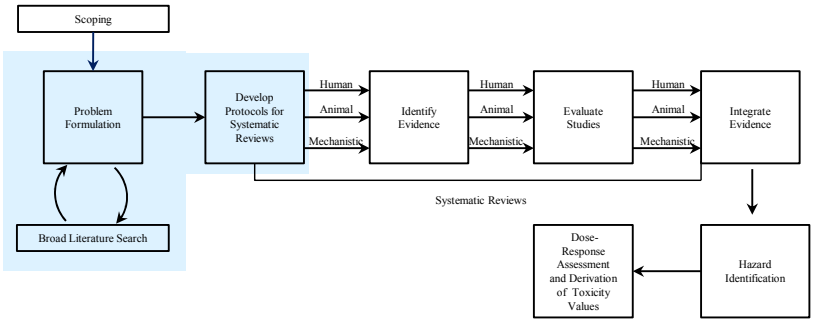


FIGURE 3-1 The IRIS process; problem formulation and protocol development are highlighted. The committee views public input and peer review as integral parts of the IRIS process, although they are not specifically noted in the figure.

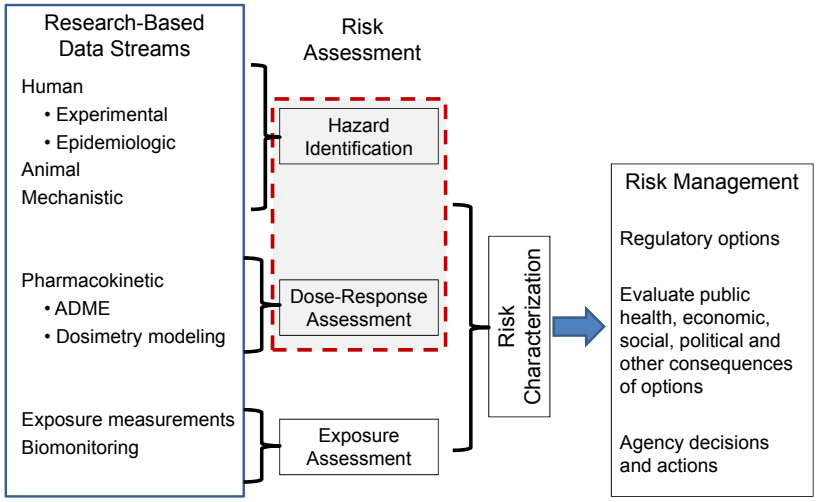


FIGURE 3-2 The risk-assessment, risk-management paradigm. The box outlined in red shows the information included in IRIS assessments. Source: Adapted from NRC 1983.

Human Studies

Human studies of health can be divided into ones that control exposure (experimental studies) and ones that do not (observational studies). Experimental human studies with potentially toxic chemicals are performed infrequently, so most human studies of adverse outcomes are observational epidemiologic studies in which exposure is not controlled, but rather the consequences of inadvertent human exposures.

Broadly speaking, most observational epidemiologic study designs can be categorized as cross-sectional, cohort, or case-control. In cross-sectional studies, measurements of a variety of factors are recorded at a particular time; cross-sectional studies that are considered in an IRIS assessment typically involve assessment of one or more health outcomes in relation to current or past exposure. In cohort studies, persons exposed or nonexposed to a given factor are observed over a period for the onset of health effects related to the exposure. A case-control study compares persons who have a given disease (cases) with those who do not have the disease (controls) with regard to their history of exposure. Each design has appropriate analytic strategies, and each has its own strengths and weaknesses. There are variations of each approach and other designs as

well. Epidemiologic study designs are well described in standard epidemiologic references (see, for example, Rothman et al. 2012).

The general difficulty of observational epidemiologic studies, regardless of design, is that exposure is not randomized but rather is determined by where people live or work, what they eat, what social group they belong to, or a host of other factors that can affect disease risk. As a result, associations between exposure and disease risk can occur even if the exposure does not cause the disease. And it is possible that no association is measured when the exposure does cause disease because confounding factors act to reduce or even cancel the effect of the exposure that is being investigated or the study is too small (underpowered) to see the effect against the background rate of disease.

Despite the inherent weaknesses, epidemiologic studies present a number of advantages for chemical risk assessment. For example, the exposure-effect relationships studied are in the target species, humans; the exposure-effect relationships can be studied in heterogeneous human populations, and it is possible to study the interactions between a chemical exposure and other factors, such as genes and lifestyle; and they provide data on relevant exposure conditions and routes of exposure (Nachman et al. 2011).

For some agents, data might be available from controlled experimental exposures of small groups of people. Of necessity, such studies are limited to brief exposures that are expected to cause no lasting harm and to acute responses. For example, volunteers have been exposed to formaldehyde and other gases. Beyond assessing short-term responses, such studies might be used to improve understanding of dosimetry and to assess biomarkers. They can be a useful bridge to the findings of animal studies.

Animal Studies

Using laboratory animals, primarily rats and mice, to assess chemical hazard remains an essential component of toxicologic and chemical risk assessments (Beyer et al. 2011). Animal studies often provide critical qualitative and quantitative information on the types of adverse effects to expect in humans and some general indication of the amount, frequency, and timing of exposure or dose that could be associated with a particular adverse outcome. Animal testing can be divided into two broad approaches: identification of toxic hazards (discussed here) and mechanistic studies, including pharmacokinetic studies (discussed in the next section).

In vivo animal tests are used to determine the nature of adverse responses (toxicity) that a chemical can produce and then to characterize the dose-response relationship between a chemical and particular types of adverse responses; each type of response will have its own dose-response relationship. Many animal-testing protocols have been developed for specific hazard end points, such as acute organ toxicity, reproductive and developmental (including teratogenic) effects, carcinogenesis, neurologic and behavioral effects, immune-system effects, and eye and skin irritation (see, for example, Eaton and Gilbert 2013). Such tests have rigorous design elements and, if used for regulatory purposes, must follow good laboratory practice guidelines. Many national regulatory agencies—such as EPA, the US Food and Drug Administration (FDA), and the US Department of Agriculture—and international regulatory agencies—such as the European Commission and the Japanese Ministry of Agriculture, Forestry, and Fisheries—have developed such toxicity-testing protocols and have expended substantial efforts to harmonize guidelines (Ertz and Preu 2008). Regardless of the particular end point or type of study being conducted, the ultimate purpose is usually the same: to identify toxic hazards and to characterize the shape of the dose-response curve for a given end point. The doses identified in animal studies with specific response levels can then be modeled to predict minimal response levels, often referred to as benchmark doses (BMDs) at a specified level of response, such as 5% (BMD_5) and 10% (BMD_{10}) (Filipsson et al. 2003; Davis et al. 2011). The values derived from the animal studies can then be used to derive toxicity values (reference concentrations, reference doses, and unit risk values) when suitable data from human studies are unavailable for this purpose.

Animal experiments have at least one important advantage over human studies: exposure can be experimentally controlled. Several other sources of variation besides exposure can also be controlled, and the ability to control exposure and other factors eliminates most of the risk of confounding. However, the use of experimental animal data for predicting human health risk is subject to multiple sources of uncertainty, especially uncertainty regarding the relevance of animal-model findings to humans. Species differences in response to toxic chemicals can be highly variable, and reliance on animal studies alone for predicting human health risks can lead to false positives and false negatives. For example, benzene and arsenic were identified as human carcinogens on the basis of epidemiologic data at a time when animal data failed to identify the carcinogenic risks; later refinement of animal models and a better mechanistic understanding of how these chemicals cause cancer have made it possible to explain the reasons for the disparate results of early studies.

Mechanistic Studies

For purposes of this report, mechanistic data come from a wide variety of studies that are not intended to identify an adverse outcome. The committee notes that it is using the term *mechanism of action* (or simply *mechanism*) in this report rather than *mode of action* simply for ease of reading; it recognizes that these terms can have different meanings. This third source of experimental data includes in vitro and in vivo laboratory studies directed at the cellular, biochemical, and molecular mechanisms that explain how a chemical produces particular adverse effects. These studies increasingly take advantage of new "-omics" tools, such as proteomics and metabolomics, to identify early biomarkers of effect. In vitro studies that use cells and tissues derived from humans and animals can provide information on the relative sensitivity of human and animal cells and can identify critical differences in how a chemical is metabolized and eliminated from the body.

Another broad class of mechanistic data is related to the toxicokinetics of a chemical. Physiologically based pharmacokinetic (PBPK) models are increasingly used by EPA and other agencies to support risk assessments (Lipscomb et al. 2012). PBPK models integrate mechanistic absorption, metabolism, distribution, and excretion (ADME) data and can be used to predict the time course of a parent chemical, metabolites, or biomarkers in the exposed organism under various exposure conditions. Thus, they can provide critical insights into potential differences in the dose-response relationship between species or between different groups within a species (for example, sex, race, or ethnicity differences related to genomic variation). PBPK models can also be used to support quantitative extrapolation of in vitro to in vivo data (Yoon et al. 2012) and are used by EPA and others to support extrapolations between species, exposure routes (for example, inhalation to oral), and exposure durations (Barton et al. 2007; Kenyon 2012). The recent methanol IRIS assessment (EPA 2013a) is an excellent example of how PBPK modeling and related mechanistic data can be used to understand species differences in response to different exposures.

Use of mechanistic information in the IRIS process has been focused on supporting the biologic plausibility of in vivo observations in animal or human studies. In some cases, in vitro results can substantially influence hazard identification and dose-response assessment. For example, EPA's Guidelines for Carcinogen Risk Assessment (EPA 2005) require that a chemical that is associated with an excess incidence of cancers in animal bioassays or human epidemiologic observations be treated as a genotoxic carcinogen if there are largely positive in vitro mutagenesis or genotoxicity studies. That classification will result in low-dose linear extrapolations in dose-response modeling. In contrast, under those same EPA guidelines, a chemical for which in vitro mutagenesis or genotoxicity assays are largely negative might be classified as a nongenotoxic carcinogen; such carcinogens are often modeled by using nonlinear approaches at low doses. Mechanistic data, including data from in vitro studies, can also be used in interpreting discrepancies between results of human and in vivo animal studies. For example, chronic animal

bioassays of the widely used dietary sweetener saccharin found an increased incidence of bladder carcinomas in rats, although extensive use in human populations failed to identify any risks. Mechanistic studies demonstrated that the bladder carcinogen was secondary to a phenomenon peculiar to male rats, and the FDA later removed saccharin from the list of potential food carcinogens. Likewise, EPA has developed guidance documents to evaluate chemicals that induce accumulation of the low-molecular-weight protein alpha2u-globulin (α-2U) in the male rat kidney (EPA 1991). Renal accumulation of α-2U in male rats initiates a chain of events that lead to renal tubule tumor formation. Unlike male rats, female rats and other laboratory mammals do not accumulate α-2U in the kidney and do not develop renal tubule tumors. Humans appear to respond more like female rats than like male rats; thus, the male rat in this case is not a good model for evaluating human risk (Rodgers and Baetcke 1993; McClellan 1996). Conversely, studies might also show that human responses that lead to increased susceptibility to a risk are different from animal responses. For example, the human teratogen thalidomide failed to induce phocomelia and other birth defects in laboratory rats and mice at equivalent doses (Collins 2006).

For a chemical hazard evaluation, there might be hundreds of in vitro and other mechanistic studies of a given chemical and only one or a few in vivo animal or human epidemiologic studies. Although EPA would be unlikely to initiate an IRIS review of a chemical on which the only available data are from in vitro or mechanistic studies, a well-designed systematic review of all the mechanistic information available is an important element of the IRIS process for chemicals on which in vivo animal or human epidemiologic data are available. Kushman et al. (2013) describe a process for conducting systematic reviews of mechanistic data in human-health assessments. Using diethylhexylphthalate as an example, they provide a process that includes all the basic elements of systematic review (defined literature search, inclusion and exclusion criteria, and evidence tables) for evaluation of mechanistic data.

Development of Systematic-Review Questions

A major challenge in the problem-formulation step is determining what adverse outcomes are of potential concern.[1] To identify the potentially relevant outcomes associated with exposure to a given chemical, the IRIS chemical-assessment team needs to conduct an initial broad search of the literature and toxicology databases by using the procedures described in the draft handbook for IRIS assessments (EPA2013b, Appendix F). The initial search provides the foundation for constructing well-defined questions and for constructing the protocol for each targeted systematic review for a particular outcome. The thorough and systematic literature search that is conducted for each systematic review (as described in Chapter 4) should not be confused with the broad literature search conducted for problem formulation. The recently revised IRIS process that is described in the preamble of each assessment (see, for example, EPA 2013c) should differentiate better the sequence of steps taken to survey the literature, develop the focused questions for each identified putative outcome, and identify and assess the evidence that addresses the questions.

The committee suggests the following process for conducting problem formulation under the assumptions that the outcomes are broadly defined and that all putative toxicological outcomes are considered.

[1]The committee is using the term *outcome* to refer to a disease phenotype—for example, various cancer types, asthma, or diabetes—or specific tissue or organ system damage or dysfunction, such as liver damage, kidney damage, perturbed neurologic function, or altered reproductive function. The adverse outcome might be identified by functional end points (for example, altered liver function or metabolic changes), anatomic end points (for example, histopathologic changes or fetal resorptions), or behavioral end points.

Step 1: With the support of an information specialist trained in conducting systematic reviews, a broad literature search should be designed and performed to identify possible outcomes associated with the chemical under investigation. The term *information specialist* (or *informationist*) was developed and is commonly used in the context of clinical medicine (Davidoff and Florance 2000; Whitmore et al. 2008; Grefsheim et al. 2010). In the context of the IRIS process, an information specialist would be a person trained in toxicology and risk assessment, able to interact with the chemical-assessment team, and having expertise in information science and systematic-review methods.

Step 2: A table (see Table 3-1) might be constructed to guide the formulation of specific questions that would then be the subjects of specific systematic reviews. Each study that is identified in the initial search would be included in one or more of the cells in the table. As noted, for simplicity, the committee is not distinguishing between the terms *mechanism of action* and *mode of action* and is using *mechanism of action* (or simply *mechanism*) throughout the report.

Step 3: The completed table would document which toxicological outcomes have been examined scientifically and warrant formulation of a specific research question and a systematic review of the available evidence. For example, if the search identifies articles that examined the mutagenicity of chemical X in animals, the articles would be listed in the row labeled "Genotoxicity or mutagenesis" under the column labeled "Animal (in vivo) studies." The articles would lead to a research question (problem formulation): "Is there scientific evidence that chemical X is mutagenic in animals or humans?" The research question would then be addressed by a systematic review, which would require a separate formal search for evidence (see Chapter 4).

TABLE 3-1 Outcomes for Consideration in Problem Formulation

Outcome	Human (in vivo) Studies[a]	Animal (in vivo) Studies	In vitro, Mechanistic Studies
Genotoxicity or mutagenesis			
Oncogenesis			
Reproductive			
Developmental, teratogenesis			
Pharmacokinetics			
Neurologic and sensory systems			
Hepatic			
Renal			
Gastrointestinal			
Endocrine			
Metabolic disease			
Respiratory			
Cardiovascular			
Hematopoietic			
Immunologic			
Musculoskeletal			
Dermal			
Other			

[a]Human in vivo studies might embody an array of experimental designs, including controlled human exposure studies in chambers, case reports, and epidemiologic studies, including ecologic, cohort, cross-sectional, and case-control studies.

As noted earlier, the approach recommended by the committee appears to be similar to the revised process used in the draft IRIS assessment for benzo[a]pyrene (EPA 2013c). In that document, the literature search (summarized in Figure LS-1) identified 700 potentially relevant publications among 21,000 hits in the search and categorized them by target organ and outcome. The summary of the search did not include the number of articles identified for the study categories listed in Table 3-1 above (human, animal, and mechanistic), and this omission is acceptable in this early step. However, the chemical-assessment team would be expected to expand on that classification for the systematic reviews.

The completed table (Table 3-1) constitutes the basic starting point for careful definition of the hazard-specific questions that can be subjected to systematic review. Each hazard-specific question should specify (1) the specific chemical, process, or mixture being evaluated (and possibly the sources and pathways of exposure), (2) the general types of studies of interest (for example, in vitro, animal in vivo, human clinical, and epidemiologic studies), and (3) the outcomes of interest and the organ system potentially affected.

Questions should be formulated for systematic review if an outcome is deemed to be of possible importance regardless of the amount of evidence that is thought to exist. That approach allows the hazard-identification process to be structured to minimize the chances of incorrectly assuming that a risk does not exist (a false negative). Even if the evidence on a hazard-specific question appears minimal initially, a systematic review can be undertaken if the question is deemed worthy of investigation. Decisions as to which specific outcomes should be further evaluated by specific systematic reviews require careful consideration of numerous factors, including whether the potential outcome is likely to occur at doses encountered by the general population and what the significance of the outcome will be if the potential association suggested in the screening review is real. Expert judgment will play an important role in this step of the IRIS process. The decision process for determining which outcomes should be subjected to a systematic review should be carefully described, and the description should be subject to peer review by experts along with the rest of the IRIS document. The committee recognizes that some chemicals will have numerous "positive" end points and a large database of studies and that multiple systematic reviews could be resource-intensive and challenging to complete in a timely matter. In such circumstances, EPA might need to establish an additional prescreening process to ensure that efforts are focused on the most relevant public-health end points. Additional guidelines should be established that will ensure consistency in the approach for all chemicals.

The committee notes that the preamble appears to merge scoping and problem formulation by stating that "the IRIS Program discusses the scope with other EPA programs and regions to ensure that the assessment will meet their needs. Then a public meeting on problem formulation invites discussion of the key issues and the studies and analytic approaches that might contribute to their resolution" (EPA 2013c, p xiv). It does not explicitly describe how the question is formulated, and it is unclear whether the process described is used for specific systematic-review questions for each relevant outcome. A properly formulated question is important in setting eligibility criteria for the review and designing the literature search strategy. Once the questions for the systematic reviews are specified, the protocol for each review can be developed.

PROTOCOL DEVELOPMENT

When the systematic-review questions have been specified, a protocol for each review should be developed. A protocol makes the methods and the process of the review transparent, can provide the opportunity for peer review of the methods, and stands as a record of the review. The protocol also minimizes bias in evidence identification by ensuring that inclusion of studies in the review does not depend on the findings of the studies. Any changes made after the protocol is in place should be documented and justified in the final report. Box 3-1 lists the common

BOX 3-1 Systematic-Review Protocol Elements

A. Systematic review question (for example, is benzo[a]pyrene exposure of adult animals associated with neurotoxic effects?)
B. Methods
    1. Inclusion and exclusion criteria for studies:
        a. Types of studies or participants (for example, experimental animal, observational human, or in vitro mechanistic).
        b. Types of exposures (for example, oral or inhalation).
        c. Types of outcome (for example, neurotoxic or developmental).
    2. Search methods for identification of studies.
    3. Assessment of risk of bias and other methodologic characteristics of included studies.
    4. Data-collection methods.
    5. Analysis.

elements of a systematic-review protocol. The plan for completing each step should be described in the protocol. Further discussion of protocol elements is provided where appropriate in the chapters that follow.

FINDINGS AND RECOMMENDATIONS

Finding: The materials provided to the committee by EPA describe the need for carefully constructed literature searches but do not provide sufficient distinction between an initial survey of the literature to identify putative adverse outcomes of interest and the comprehensive literature search that is conducted as part of a systematic review of an identified putative outcome.

Recommendation: EPA should establish a transparent process for initially identifying all putative adverse outcomes through a broad search of the literature. The agency should then develop a process that uses guided expert judgment to identify the specific adverse outcomes to be investigated, each of which would then be subjected to systematic review of human, animal, and in vitro or mechanistic data.

Recommendation: For all literature searches, EPA should consult with an information specialist who is trained in conducting systematic reviews.

Finding: A protocol is an essential element of a systematic review. It makes the methods and the process of the review transparent, can provide the opportunity for peer review of the methods, and stands as a record of the review.

Recommendation: EPA should include protocols for all systematic reviews conducted for a specific IRIS assessment as appendixes to the assessment.

REFERENCES

Barton, H.A., W.A. Chiu, R. Woodrow Setzer, M.E. Andersen, A.J. Bailer, F.Y. Bois, R.S. Dewoskin, S. Hays, G. Johanson, N. Jones, G. Loizou, R.C. Macphail, C.J. Portier, M. Spendiff, and Y.M. Tan. 2007. Characterizing uncertainty and variability in physiologically based pharmacokinetic models: State of the science and needs for research and implementation. Toxicol. Sci. 99(2):395-402.

Beyer, L.A., B.D. Beck, and T.A. Lewandowski. 2011. Historical perspective on the use of animal bioassays to predict carcinogenicity: Evolution in design and recognition of utility. Crit. Rev. Toxicol. 41(4):321-338.

Collins, T.F. 2006. History and evolution of reproductive and developmental toxicology guidelines. Curr. Pharm. Des. 12(12):1449-1465.

Davidoff, F., and V. Florance. 2000. The informationist: A new health profession? Ann. Intern. Med. 132(12):996-998.

Davis, J.A., J.S. Gift, and Q.J. Zhao. 2011. Introduction to benchmark dose methods and U.S. EPA's benchmark dose software (BMDS) version 2.1.1. Toxicol. Appl. Pharmacol. 254(2):181-191.

Eaton, D.L., and S.G. Gilbert. 2013. Principles of toxicology. Pp. 13-48 in Casarett and Doull's Toxicology: The Basic Science of Poisons, 8th Ed., C.D. Klaassen, ed. New York: McGraw-Hill.

EPA (U.S. Environmental Protection Agency). 1991. Alpha2u-Globulin: Association with Chemically Induced Renal Toxicity and Neoplasia in the Male Rat. EPA/625/3-91/019F. Risk Assessment Forum, U.S. Environmental Protection Agency, Washington, DC [online]. Available: http://www.epa.gov/raf/publications/alpha2u-globulin.htm [accessed December 17, 2013].

EPA (U.S. Environmental Protection Agency). 2005. Guidelines for Carcinogen Risk Assessment. EPA/630/P-03/001F. Risk Assessment Forum, U.S. Environmental Protection Agency, Washington, DC. March 2005 [online]. Available: http://www.epa.gov/raf/publications/pdfs/CANCER_GUIDELINES_FINAL_3-25-05.PDF [accessed October 3, 2013].

EPA (U.S. Environmental Protection Agency). 2013a. Toxicological Review of Methanol (Noncancer) (CAS No. 67-56-1) in Support of Summary Information on the Integrated Risk Information System (IRIS). EPA/635/R-11/001Fa. U.S. Environmental Protection Agency, Washington, DC. September 2013 [online]. Available: http://www.epa.gov/iris/toxreviews/0305tr.pdf [accessed November 13, 2013].

EPA (U.S. Environmental Protection Agency). 2013b. Part I: Status of Implementation of Recommendations. Materials Submitted to the National Research Council, by Integrated Risk Information System Program, U.S. Environmental Protection Agency, January 30, 2013 [online]. Available: http://www.epa.gov/IRIS/iris-nrc.htm [accessed October 22, 2013].

EPA (U.S. Environmental Protection Agency). 2013c. Toxicological Review of Benzo[a]pyrene (CAS No. 50-32-8) in Support of Summary Information on the Integrated Risk Information System (IRIS), Public Comment Draft. EPA/635/R13/138a. National Center for Environmental Assessment, Office of Research and Development, U.S. Environmental Protection Agency, Washington, DC. August 2013 [online]. Available: http://cfpub.epa.gov/ncea/iris_drafts/recordisplay.cfm?deid=66193 [accessed October 22, 2013].

Ertz, K., and M. Preu. 2008. International GLP: A critical reflection on the harmonized global GLP standard from a test facility viewpoint. Ann. Ist Super Sanita. 44(4):390-394.

Grefsheim, S.F., S.C. Whitmore, B.A. Rapp, J.A. Rankin, R.R. Robison, and C.C. Canto. 2010. The informationist: Building evidence for an emerging health profession. J. Med. Libr. Assoc. 98(2):147-156.

Filipsson, A.F., S. Sand, J. Nilsson, and K. Victorin. 2003. The benchmark dose method--review of available models, and recommendations for application in health risk assessment. Crit. Rev. Toxicol. 33(5):505-542.

Kenyon, E.M. 2012. Interspecies extrapolation. Methods Mol. Biol. 929:501-520.

Kushman, M.E., A.D. Kraft, K.Z. Guyton, W.A. Chiu, S.L. Makris, and I. Rusyn. 2013. A systematic approach for identifying and presenting mechanistic evidence in human health assessment. Regul. Toxicol. Pharmacol. 67(2):266-277.

Lipscomb, J.C., S. Haddad, T. Poet, and K. Krishnan. 2012. Physiologically-based pharmacokinetic (PBPK) models in toxicity testing and risk assessment. Adv. Exp. Med. Biol. 745:76-95.

McClellan, R.O. 1996. Reducing uncertainty in risk assessment by using specific knowledge to replace default options. Drug Metab. Rev. 28(1-2):149-179.

Nachman, K.E., M.A. Fox, M.C. Sheehan, T.A. Burke, J.V. Rodricks, and T.J. Woodruff. 2011. Leveraging epidemiology to improve risk assessment. Open Epidemiol. J. 4:3-29.

NRC (National Research Council). 1983. Risk Assessment in the Federal Government: Managing the Process. Washington, DC: National Academy Press.

NRC (National Research Council). 2009. Science and Decisions: Advancing Risk Assessment. Washington, DC: National Academies Press.

NRC (National Research Council). 2011. Review of the Environmental Protection Agency's Draft IRIS Assessment of Formaldehyde. Washington, DC: National Academies Press.

Rodgers, I.S., and K.P. Baetcke. 1993. Interpretation of male rat renal tubule tumors. Environ. Health Perspect. 101(suppl. 6):45-52.

Rothman, K.J., T.L. Lash, and S. Greenland. 2012. Modern Epidemiology, 3rd Ed. Philadelphia, PA: Lippincott Williams & Wilkins.

Whitmore, S.C, S.F. Grefsheim, and J.A. Rankin. 2008. Informationist programme in support of biomedical research: A programme description and preliminary findings of an evaluation. Health Info Libr. J. 25(2):135-141.

Yoon, M., J.L. Campbell, M.E. Andersen, and H.J. Clewell. 2012. Quantitative in vitro to in vivo extrapolation of cell-based toxicity assay results. Crit. Rev. Toxicol. 42(8):633-652.

4

Evidence Identification

This chapter addresses the identification of evidence pertaining to questions that are candidates for systematic reviews as described in Chapter 3 (see Figure 4-1). Systematic reviews for US Environmental Protection Agency (EPA) Integrated Risk Information System (IRIS) assessments, as for any topic, should be based on comprehensive, transparent literature searches and screening to enable the formulation of reliable assessments that are based on all relevant evidence. EPA has substantially improved and documented its approach for identifying evidence in response to the report *Review of the Environmental Protection Agency's Draft IRIS Assessment of Formaldehyde* (NRC 2011) and other criticisms and advice. As a way to encourage further progress, the present committee compares recent EPA materials and assessments with the guidelines developed for systematic review by the Institute of Medicine (IOM 2011) and offers specific suggestions for improving evidence identification in the IRIS process. The committee makes this comparison with the IOM guidelines because they are derived from several decades of experience, are considered a standard in the clinical domain, and are thought to be applicable to the IRIS process.

CONSIDERATION OF BIAS IN EVIDENCE IDENTIFICATION

Systematic reviews of scientific evidence are preferable to traditional literature reviews partly because of their transparency and adherence to standards. In addition, the systematic-review process gathers all the evidence without relying on the judgment of particular people to select studies. Nonetheless, systematic reviews are prone to two types of bias: bias present in the individual studies included in a review and bias resulting from how the review itself was conducted (meta-bias). Meta-bias cannot be identified by examining the methods of an individual study because it stems from how a systematic review is conducted and from factors that broadly affect a body of research.

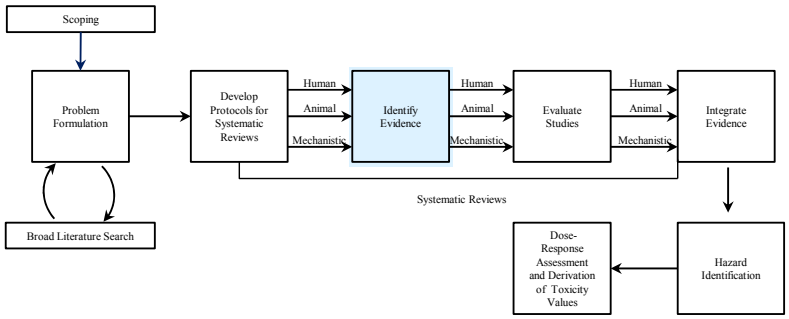


FIGURE 4-1 The IRIS process; the evidence-identification step is highlighted. The committee views public input and peer review as integral parts of the IRIS process, although they are not specifically noted in the figure.

It might be argued that the most important form of meta-bias that threatens the validity of findings of a systematic review results from the differential reporting of study findings on the basis of their strength and direction. Since the early focus on publication bias or the failure to publish at all because of potential implications of study findings, investigators have come to recognize that reporting biases encompass a wide array of behaviors, including selective outcome reporting. Reporting biases have been repeatedly documented in studies that show that research with statistically significant results is published more often (sometimes more often in English-language journals) and more quickly, and in journals that have higher citation frequencies than research whose results are not statistically significant (Dickersin and Chalmers 2011). The reporting bias related to the publication of research with statistically significant results might also be exacerbated by increased publication pressures (Fanelli 2010). Reporting biases have been shown to be associated with all sorts of sponsors and investigators; for example, industry-supported studies in health sciences have been shown to be particularly vulnerable to distortion of findings or to not being reported at all (Lundh et al. 2012). Moreover, an investigator's failure to submit (as opposed to selectivity on the part of the editorial process) appears to be the main reason for failure to publish (Chalmers and Dickersin 2013). There is evidence that reporting biases might also be a subject of concern in laboratory and animal research (Sena et al. 2010; Korevaar et al. 2011; ter Riet et al. 2012). The potential for reporting biases is one reason to search the gray (unpublished) literature. Specifically, the gray literature might be less likely to support specific hypotheses than literature sources that might be biased toward publication of "positive" results.

Systematic review does not identify the presence of reporting biases themselves. However, a comprehensive search will include the types of studies particularly prone to reporting biases, such as industry-supported studies in the health sciences. A failure to find studies in such categories that are particularly prone to reporting bias should raise concern that reporting bias is present. In addition, a systematic review provides the opportunity to compare findings among different groups of funders and investigators and to identify any indication of meta-bias.

A second type of meta-bias is information bias, which occurs when data on the groups being compared (for example, animals exposed at different doses or control vs exposed animals) are collected differentially (nonrandom misclassification). Such bias can affect a whole body of literature. Incorrect information can also be collected in error (random misclassification) without a direction of the bias. Random misclassification is understandably undesirable in toxicity assessments.

This chapter specifically addresses two steps that are critical for minimizing meta-bias: performing a comprehensive search for all the evidence, including unpublished findings, and screening and selecting reports that address the systematic-review question and meet eligibility criteria specified in the protocol. Error can also arise when data are abstracted from studies during the review process. Systematic-review methods should be structured to maximize the accuracy of the data extracted from the identified studies in the systematic review. Therefore, the committee addresses an additional step in the systematic-review process in this chapter: extracting the data from studies included in the IRIS review (see Table 4-1, section on IOM Standard 3.5).

RECOMMENDATIONS ON EVIDENCE IDENTIFICATION IN THE NATIONAL RESEARCH COUNCIL FORMALDEYDE REPORT

The National Research Council (NRC) formaldehyde report (NRC 2011) recommended that EPA adopt standardized, documented, quality-controlled processes and provided specific recommendations related to evidence identification (see Box 4-1). Implementation of the recommendations is addressed in the following section of this chapter. Detailed findings and recommendations are provided in Table 4-1, and general findings and recommendations are provided at the conclusion of the chapter.

BOX 4-1 Recommendations on Evidence Identification in the 2011 National Research Council Formaldehyde Report

- Establish standard protocols for evidence identification.
- Develop a template for description of the search approach.
- Use a database, such as the Health and Environmental Research Online (HERO) database, to capture study information and relevant quantitative data.

Source: NRC 2011, p. 164.

EVALUATION OF ENVIRONMENTAL PROTECTION AGENCY RESPONSE TO THE NATIONAL RESEARCH COUNCIL FORMALDEHYDE REPORT

Relatively little research has been conducted specifically on the issue of evidence identification related to hazard identification and dose-response assessments for IRIS assessments. To capitalize on recent efforts in this regard, the committee used the standards established by IOM to assess the effectiveness of health interventions as the foundation of its evaluation of the IRIS process. The standards are presented in *Finding What Works in Healthcare: Standards for Systematic Review* (IOM 2011). The general approaches and concepts underlying systematic reviews for evidence-based medicine should be relevant to the review of animal studies and epidemiologic studies (Mignini and Khan 2006). In animal studies, the work on testing of various methods of evidence identification is in the early stages (Leenaars et al. 2012; Briel et al. 2013; Hooijmans and Ritskes-Hoitinga 2013).

The IOM standards relied on three main sources: the published methods of the Cochrane Collaboration (Higgins and Green 2008), the Centre for Reviews and Dissemination of the University of York in the United Kingdom (CRD 2009), and the Effective Health Care Program of the Agency for Healthcare Research and Quality in the United States (AHRQ 2011). Those standards reflect the input of experts who were consulted during their development. Although the IOM standards for conducting systematic reviews focus on assessing the comparative effectiveness of medical or surgical interventions and the evidence supporting the standards is based on clinical research, the committee considers the approach useful for a number of aspects of IRIS assessments because the underlying principles are inherent to the scientific process (see Hoffman and Hartung 2006; Woodruff and Sutton 2011; Silbergeld and Scherer 2013). Some analysts, however, have noted challenges associated with implementing the IOM standards (Chang et al. 2013; IOM 2013).

Table 4-1 summarizes elements of the IOM standards for identifying information in systematic reviews in evidence-based medicine, presents the rationale or principle behind each element, and indicates the status of the element as reflected in materials submitted to the committee that document changes in the IRIS program (EPA 2013a,b) as described in Chapter 1 (see Table 1-1). Two chemical-specific examples (EPA 2013c,d,e) are included in Table 4-1 to assess the intent and application of the new EPA strategies as reflected in the draft preamble (EPA 2013a, Appendix B) and the draft handbook (EPA 2013a, Appendix F). The committee interprets those portions of the draft preamble and handbook that address the literature search and screening as constituting draft standard approaches for evidence identification. It assumes that the preamble summarizes the methods used in IRIS assessments and that the handbook is a detailed record of methods that are intended to be applied in evidence assessments. In other words, the committee assumes that people who are responsible for performing systematic reviews to support upcoming IRIS assessments will rely on the handbook as it continues to evolve.

TABLE 4-1 Comparison of EPA Draft Materials with IOM Systematic-Review Standards for Evidence Identification

IOM Standard (IOM 2011) and Rationale	Draft IRIS Preamble (EPA 2013a, Appendix B)	Draft IRIS Handbook (EPA 2013a, Appendix F)	Draft IRIS Benzo[a]pyrene Assessment (EPA 2013c)	Draft IRIS Ammonia Assessment (EPA 2013d,e)	Considerations for Further Development
3.1 Conduct a comprehensive systematic search for evidence					
3.1.1 Work with a librarian or other information specialist trained in performing systematic reviews to plan the search strategy (p. 266). **Rationale**: As with other aspects of research, specific skills and training are required to navigate a wide range of bibliographic databases and electronic information sources.	Not mentioned.	The initial steps of the systematic review process involve formulating specific strategies to identify and select studies related to each key question (p. F-2). EPA refers to tapping HERO resources (which include librarian expertise) and advises consulting a librarian early (to develop search terms) and often. Nevertheless, the outline suggests that their search process begins with literature collection. EPA acknowledges that the process developed for evidence-based medicine is generally applied to narrower, more focused questions and nonetheless provides a strong foundation for IRIS assessments; EPA notes that IRIS addresses assessment-specific questions. However, the materials do not describe information specialists trained in systematic reviews.	Not mentioned.	Not mentioned.	Begin by referencing the key role played by information specialists who have expertise in systematic reviews in planning the search strategy and their role as members of the IRIS team throughout the evidence-identification process.

(Continued)

TABLE 4-1 Continued

IOM Standard (IOM 2011) and Rationale	Draft IRIS Preamble (EPA 2013a, Appendix B)	Draft IRIS Handbook (EPA 2013a, Appendix F)	Draft IRIS Benzo[a]pyrene Assessment (EPA 2013c)	Draft IRIS Ammonia Assessment (EPA 2013d,e)	Considerations for Further Development
3.1.2 Design the search strategy to address each key research question (p. 266). **Rationale**: The goal of the search strategy is to maximize both sensitivity (the proportion of all eligible articles that are correctly identified) and precision (the proportion of all articles identified by the search that are eligible). With multiple research questions, a single search strategy is unlikely to cover all questions posed with any precision.	Not mentioned.	No mention of key research questions. Step 1 of the proposed process sets the goal of identifying primary studies. Step 1b (p. F-6) on selecting search terms specifies "the appropriate forms of the chemical name, CAS number, and if relevant, major metabolite(s)." EPA also describes the possible addition of secondary search strategies that include key words for end points, the possibility of other more targeted end points, and the use of filters and analysis of small samples of review results to assess relevance (pp. F-6–F-7).	Not mentioned. The section titles imply the general search questions (for example, developmental effects, reproductive effects, immunotoxicity, other toxicity, and carcinogenicity), but they are not listed anywhere explicitly.	No mention of key research questions.	In the protocol, describe the role of key research questions and their relationship to the search strategies. Do not omit any helpful information related to this standard; rather, include this information in the same section as appropriate.
3.1.3 Use an independent librarian or other information specialist to peer review the search strategy (p. 267). **Rationale:** This part of the evidence review requires peer review like any other part. Given the specialized skills required, a person with similar skills would be expected to serve as peer reviewer.	Not mentioned.	Not mentioned.	Not mentioned.	Not mentioned.	Add a review of the search strategy by an independent information specialist (that is, one who did not design the protocol), who is trained in evidence identification for systematic reviews to strengthen the search process.

3.1.4 Search bibliographic databases (p. 267). **Rationale:** A single database is typically not sufficient to cover all publications (journals, books, monographs, government reports, and others) for clinical research. Databases for reports published in languages other than English and for the gray literature could also be searched.	The literature search follows standard practices and includes the PubMed and ToxNet databases of the National Library of Medicine, Web of Science, and other databases listed in EPA's HERO system. Searches for information on mechanisms of toxicity are inherently specialized and might include studies of other agents that act through related mechanisms (p. B2).	Step 1A describes specific databases for IRIS reviews (Table F-1), including PubMed, Web of Science, Toxline, TSCATS, PRISM, and IHAD, several of which are accessible through the EPA HERO interface. EPA identifies the HERO interface, directly searching the named databases, or supervising the search process conducted by contractors.	Table LS-1 and Table C-1 (Appendix C) outline the online databases searched. There is not 100% agreement between the tables.	Appendix D in Supplement provides search strings for some of but not all the databases listed in Table LS-1 as searched. Table LS-1 provides keywords used for bibliographic databases.	Systematically and regularly assess the relevance and usefulness of the identified databases (PubMed, Web of Science, Toxline, TSCATS, PRISM, IHAD, and others) for finding primary studies. Ensure that the search process conducted by contractors follows specific (detailed) guidelines for systematic literature reviews established by EPA (considering the elements outlined here) and that the contractor searches regularly undergo peer review or outside assessment.
3.1.5 Search citation indexes (p. 267). **Rationale**: Citation indexes are a good way to ensure that eligible reports were not missed.	See 3.1.4—The literature search includes Web of Science.	EPA mentions citation indexes, such as Web of Science, but a suggestion to search them and how is not specified (p. F-7).	The preamble mentions that Web of Science is searched; not mentioned otherwise.	The preamble mentions that searching the Web of Science is standard practice, but it is not mentioned in text otherwise.	Document specific guidance or protocols for searching citation databases (for example, to ensure that searches look for citations to the identified literature).

(Continued)

TABLE 4-1 Continued

IOM Standard (IOM 2011) and Rationale	Draft IRIS Preamble (EPA 2013a, Appendix B)	Draft IRIS Handbook (EPA 2013a, Appendix F)	Draft IRIS Benzo[a]pyrene Assessment (EPA 2013c)	Draft IRIS Ammonia Assessment (EPA 2013d,e)	Considerations for Further Development
3.1.6 Search literature cited by eligible studies (p. 268). **Rationale:** The literature cited by eligible studies (for example, references provided in a journal article or thesis) is a good way to ensure eligible reports were not missed.	Not mentioned.	EPA appropriately discusses this as a strategy.	References from previous EPA assessments and others were also examined.	References from other EPA assessments were also examined.	
3.1.7 Update the search at intervals appropriate to the pace of generation of new information for the research question being addressed (p. 268). **Rationale**: Given that new articles and reports are being generated in an ongoing manner, searches would be updated regularly to reflect new information relevant to the topic.	Not mentioned.	EPA appropriately discusses this step.	A comprehensive literature search was last conducted in February 2012. Appendix C gives dates of all searches as February 14, 2012.	Search was first conducted through March 2012 and updated in March 2013. Search string in Appendix D-1 should provide exact dates included in search in addition to the date when the search was performed.	Develop standardized processes for updating the literature searches to enable efficient updates on a regular basis, for example, during key stages of development for IRIS assessments.
3.1.8 Search subject-specific databases if other databases are unlikely to provide all relevant evidence (p. 268). **Rationale:** If other databases are unlikely to be	See entry 3.1.4—The literature search includes ToxNet of the National Library of Medicine and other databases listed in EPA's HERO.	EPA recommends searching "regulatory resources and other websites" for additional resources.	Table LS-1: Pubmed, Toxline, Toxcenter, TSCATS, ChemID, Chemfinder, CCRIS, HSDB, GENETOX, and RTECS; listed as searched. Appendix Table C-1: Pubmed,	Appendix D provides search strings for four subject-specific databases. Although Table LS-1 provides additional database names,	Consider and specify other databases beyond those listed in handbook (Appendix F, Figure F-1) and EPA (2013b, Figure

comprehensive, search a variety of other sources to cover the missing areas.			Toxline, Toxcenter, TSCATS, TSCATS2, and TSCA recent notices; does not mention ChemID, Chemfinder, CCRIS, HSDB, GENETOX, and RTECS.	search strings are not provided.	1-1). For example, consider additional resources from the set identified on the HERO website.
3.1.9 Search regional bibliographic databases if other databases are unlikely to provide all relevant evidence (p. 269). **Rationale**: Many countries have their own databases and either because of language or other regional factors the reports are not necessarily also present in US-based databases	Not mentioned.	Currently, non-English language is considered a criterion for excluding studies, and foreign language databases are not included in the discussion of search strategies.	Not mentioned.	Not mentioned.	Assess (conduct research to determine) whether studies in non-English-language countries are examining topics relevant to IRIS assessments. Revisit findings periodically to assess the effects of including or excluding non-English-language studies.
3.2	**Take action to address potentially biased reporting of research results**				
3.2.1 Search gray literature databases, clinical trial registries, and other sources of unpublished information about studies (p. 269). **Rationale:** Negative or null results, or undesirable results, might be published in difficult to access sources.	Not mentioned.	EPA recommends searching "regulatory resources and other websites" for additional resources.	Not mentioned.	Not mentioned.	Consider searching other gray literature databases beyond those listed in handbook (Appendix F, Table F-1) and other sources of unpublished information about studies (also see IOM Standard 3.1.8 above).

(Continued)

TABLE 4-1 Continued

IOM Standard (IOM 2011) and Rationale	Draft IRIS Preamble (EPA 2013a, Appendix B)	Draft IRIS Handbook (EPA 2013a, Appendix F)	Draft IRIS Benzo[a]pyrene Assessment (EPA 2013c)	Draft IRIS Ammonia Assessment (EPA 2013d,e)	Considerations for Further Development
3.2.2 Invite researchers to clarify information about study eligibility, study characteristics, and risk of bias (p. 269). **Rationale:** Rather than classify identified studies as missing critical information, it is preferable to ask the investigators directly for the information.	Not mentioned.	Not mentioned.	Not mentioned.	Not mentioned.	As needed, request additional information needed from investigators to determine eligibility, study characteristics, and other information.
3.2.3 Invite all study sponsors and researchers to submit unpublished data, including unreported outcomes, for possible inclusion in the systematic review (p. 270). **Rationale:** So as to include all relevant studies and data in the review, ask sponsors and researchers for information about unpublished studies or data.	EPA posts the results of the literature search on the IRIS Web site and requests information from the public on additional studies and current research. EPA also considers studies received through the IRIS Submission Desk and studies (typically unpublished) submitted under the Toxic Substances Control Act or the Federal Insecticide, Fungicide, and Rodenticide Act. Material submitted as Confidential Business Information is considered only if it includes health and safety data that can be publicly released. If a study that might be critical for the conclusions of the assessment has not been peer-reviewed, EPA will have it peer-reviewed.	EPA endorses requesting public scrutiny of the list of identified studies from the initial literature search and requests reviews of the list by independent scientists active in research on the topic to ensure that all relevant studies are identified (pp. F-3, F-7). It is also noteworthy that EPA duly identifies the importance of tracking why studies later identified were missed in the initial literature search.	Section 3.1 of the Preamble to the benzo[a]pyrene report states that unpublished health and safety data submitted to the EPA are also considered as long as the data can be publicly released. Per Figure LS-1, the American Petroleum Institute submitted 30 references, but it is not clear whether all study sponsors and researchers were invited to submit unpublished data.	Section 3.1 of the Preamble to the ammonia report states that EPA considers studies submitted to the IRIS Submission Desk and through other means. Many of them are unpublished. Section 3.1 of the Preamble describes inviting the public to comment on the literature search and suggest additional or current studies that might have been missed in the search.	Create a structured process for inviting study sponsors and researchers to submit unpublished data.

3.2.4 Hand search selected journals and conference abstracts (p. 270). **Rationale:** Hand searching of sources most likely provides relevant up-to-date information and contributes to the likelihood of comprehensive identification of eligible studies.	Not mentioned.	Not mentioned.	Not mentioned.	Not mentioned.	Assess (conduct research to determine) whether the IOM standard suggesting hand-searching of journals and conference abstracts is applicable and useful to the EPA task.
3.2.5 Conduct a web search (p. 271). **Rationale:** Web searches, even when broad and relatively untargeted, can contribute to the likelihood that all eligible studies have been identified.	Not mentioned.	As noted for IOM Standard 3.2.1, EPA recommends searching regulatory and other Web sites.	Not mentioned.	Not mentioned.	Assess (conduct research to determine) whether Web searches are likely to turn up additional useful information and, if so, determine which Web sites would be appropriate.
3.2.6 Search for studies reported in languages other than English if appropriate (p. 271). **Rationale:** There is limited evidence that negative, null, or undesirable findings might be published in languages other than English.	Not mentioned.	As noted for IOM Standard 3.1.9, studies published in languages other than English are currently excluded from review and non-English-language databases are not included in the discussion of search strategies.	Non-English-language articles were excluded, per Figure LS-1.	Not mentioned.	Assess (conduct research to determine) whether to search for studies reported in languages other than English for IRIS assessments and revisit question periodically.

(Continued)

TABLE 4-1 Continued

IOM Standard (IOM 2011) and Rationale	Draft IRIS Preamble (EPA 2013a, Appendix B)	Draft IRIS Handbook (EPA 2013a, Appendix F)	Draft IRIS Benzo[a]pyrene Assessment (EPA 2013c)	Draft IRIS Ammonia Assessment (EPA 2013d,e)	Considerations for Further Development
3.3 Screen and select studies					
3.3.1 Include or exclude studies based on the protocol's pre-specified criteria (p. 272). **Rationale**: On the basis of the study question, inclusion and exclusion criteria for the review would be set a priori, before reviewing the search results (see 3.3.5) so as to avoid results-based decisions.	Exposure route is a key design consideration for selecting pertinent experimental animal studies or human clinical studies. Exposure duration is also a key design consideration for selecting pertinent experimental animal studies. Short-duration studies involving animals or humans might provide toxicokinetic or mechanistic information. Specialized study designs are used for developmental and reproductive toxicity (p. B-3).	EPA specifically mentions that "casting a wide net" is a goal of the search process and that results might not address the question of interest. A two- or three-stage process is suggested (review title and abstract, then full text, or screen title and abstract in separate steps) for relevance. Table F-5 specifies excluding duplicates, studies for which only abstracts are available, and examples of criteria that might be defined for excluding studies.	Protocol not provided so unable to judge whether criteria are prespecified. Sections 3.1, 3.2, and 3.3 of the preamble to the benzo[a]pyrene report provide information on types of studies included, and Figure LS-1 provides reasons for report exclusions.	Protocol not provided so unable to judge whether criteria are prespecified. Sections 3.1, 3.2, and 3.3 of the preamble to the ammonia report provide information on types of studies included, and Figure LS-1 provides reasons for report exclusions.	Provide inclusion and exclusion criteria in IRIS assessment protocol, and use these criteria in figure describing "study selection" flow.
3.3.2 Use observational studies in addition to randomized controlled trials to evaluate harms of interventions (p. 272). **Rationale:** Predetermine study designs that will be eligible for each study question.	Cohort studies, case-control studies, and some population-based surveys provide the strongest epidemiologic evidence; ecologic studies (geographic correlation studies) that relate exposures and effects by geographic area; case reports of high or accidental exposure provide information on rare effects or relevance of results from animal testing (p. B-3).	In Step 1, literature search, it is recommended that articles be sorted into categories (for example, experimental studies of animals and observational studies of humans). Later, in 2B, the Appendix says that studies could include acute-exposure animal experiments, 2-year bioassays, experimental-chamber studies of humans, observational epidemiologic studies, in vitro studies, and many other types of designs. No restriction by study design is intended.	Described in Section 3.2 of preamble to the benzo[a]pyrene report.	Described in Section 3.2 of preamble to the ammonia report.	Not applicable.

3.3.3 Use two or more members of the review team, working independently, to screen and select studies (p. 273). **Rationale:** Because reporting is often not clear or logically placed, having two independent reviewers is a quality-assurance approach.	Not mentioned.	It appears that the handbook does not require independence of the screeners on the basis of this statement: "Review of the title and abstract, and in some cases, the full text of the article, should be conducted by two reviewers. If a contractor is used for this step, one of the reviewers should be an EPA staff member…One strategy for accomplishing this task is to have one member do the initial screening and sorting of the database, with the second member responsible for checking the accuracy of each of the resulting group (i.e., assuring that the reason for exclusion applies to each study in this group)" (pp. F-11, F13).	Not mentioned.	Not mentioned.	Two or more members of the team should work independently to screen and select studies, and inter-rater reliability should be assessed.
3.3.4 Train screeners using written documentation; test and retest screeners to improve accuracy and consistency (p. 273). **Rationale:** Training and documentation are standard quality-assurance approaches.	Not mentioned.	The handbook includes minimal discussion of training: "The two reviewers need to assure they have the same interpretation of the meaning of each category. For large databases especially, this may involve working through selected batches of 50-100 citations as 'training' exercises" (lines 6-7, p. F-13).	Not mentioned.	Not mentioned.	Provide written documentation and formally train screeners; specify testing procedures for screeners to improve their accuracy and consistency.
3.3.5 Use one of two strategies to select studies: 1) read all full-text articles identified in the search or 2) screen titles and abstracts of all articles and then read the	Not mentioned.	Figure F-1 suggests that titles and abstracts are screened first, and then possibly full text of relevant articles is screened. The handbook also says, however, "in some situations, a three-stage process	Not mentioned.	A preliminary manual screen of titles or abstracts was conducted by a toxicologist. A more detailed	Clearly document screening and selection process. Until research confirms a different approach, ensure

(Continued)

TABLE 4-1 Continued

IOM Standard (IOM 2011) and Rationale	Draft IRIS Preamble (EPA 2013a, Appendix B)	Draft IRIS Handbook (EPA 2013a, Appendix F)	Draft IRIS Benzo[a]pyrene Assessment (EPA 2013c)	Draft IRIS Ammonia Assessment (EPA 2013d,e)	Considerations for Further Development
full-text of articles identified in initial screening (p. 273). **Rationale:** Data are not clear, even for clinical intervention questions, regarding which method is best, although 2) appears to be more common.		may be more efficient, with an initial screen based on title, followed by screening based on abstract, followed by full text screening. There is not a 'right' or 'wrong' choice; however, whichever you choose, be sure to document the process you use" (pp. F-10, F11).		review of the reports identified was conducted by a person not described.	that screeners follow a process that reflects the concepts underlying the IOM standards.
3.3.6 Taking account of the risk of bias, consider using observational studies to address gaps in the evidence from randomized clinical trials on the benefits of interventions (p. 274). **Rationale:** Rather than exclude evidence where it is sparse, it might be necessary to use data from studies using design more susceptible to bias than a preferred design.	See entry 3.2.2.	Not applicable because all types of study designs are potentially eligible (and randomized clinical trials are not conducted for IRIS assessments).	Not applicable because all types of study designs are potentially eligible (and randomized clinical trials are not conducted for IRIS assessments).	Not applicable because all types of study designs are potentially eligible (and randomized clinical trials are not conducted for IRIS assessments).	Not applicable.
3.4 Document the search					
3.4.1 Provide a line-by-line description of the search strategy, including the date of search for each database, web browser, etc. (p. 274).	Each assessment specifies the search strategies, keywords, and cutoff dates of its literature searches.	The handbook supports careful documentation of the search strategy and provides Tables F-3 and F-4 as examples of the types of information that would be retained. No specific statement is	Table LS-1 provides more database names than Appendix C but does not provide search	Table LS-1 provides more database names than Appendix D but does not provide search	Document, line-by-line, a description of the search strategy, including the dates included in the search of each

Rationale: Appropriate documentation of the search processes ensures transparency of the methods used in the review, and appropriate peer review by information specialists.		made about documenting a line-by-line search strategy.	strings for them. Appendix C: Table C-1 provides search strategies for more than four databases searched with date of search, but exact dates for what was included in search are not provided—for example, for PubMed, "Date range 1950's to 2/14/2012."	strings for them. Appendix D: Table D-1 provides search strings for four subject-specific databases, but exact dates for what was included in search are not provided—for example, for PubMed, Appendix D states "Date range: 1950's to present."	database and the date of the search for each database and any Web searches.
3.4.2 Document the disposition of each report identified, including reasons for their exclusion if appropriate (p. 275). **Rationale:** The standard supports creation of a flow chart that describes the sequence of events leading to identification of included studies, and it also supports assessment of the sensitivity and precision of the searches a posteriori.	Not mentioned.	Some support is given to documenting the reasons for excluding each citation at the full-text review stage. "In these situations, the citation should be 'tagged' into the appropriate exclusion category" (p. F-16).	Summary data are provided in Figure LS-1.	The disposition of identified citations is summarized in the study-selection figure but is otherwise not mentioned. The disposition of articles identified in the search is documented in HERO.	Consider a more explicit statement in the handbook regarding documenting the disposition of each report identified by the search. Flowcharts can also be used to illustrate dispositions by category, similar to the LitFlow diagram in HERO.
3.5	**Manage data collection**				
3.5.1 At a minimum, use two or more researchers, working independently, to extract quantitative or other critical data from each	Not mentioned.	This item is not fully described in the process of data collection: "Ideally, two independent reviewers would independently identify the relevant	Not mentioned.	Not mentioned.	Ensure the quality of the data collected. For example, at a minimum, use two or more researchers

(Continued)

TABLE 4-1 Continued

IOM Standard (IOM 2011) and Rationale	Draft IRIS Preamble (EPA 2013a, Appendix B)	Draft IRIS Handbook (EPA 2013a, Appendix F)	Draft IRIS Benzo[a]pyrene Assessment (EPA 2013c)	Draft IRIS Ammonia Assessment (EPA 2013d,e)	Considerations for Further Development
study. For other types of data, one individual could extract the data while the second individual independently checks for accuracy and completeness. Establish a fair procedure for resolving discrepancies—do not simply give final decision-making power to the senior reviewer (p. 275). **Rationale:** Because reporting is often not clear or logically placed, having two independent reviewers is a quality-assurance approach. The evidence supporting two independent data extractors is limited and so some reviewers prefer that one person extracts and the other verifies, a time- saving approach. Discrepancies would be decided by discussion so that each person's viewpoint is heard.		methodological details, and then compare their results and interpretations and resolve any differences" (p. F-21).			working independently to extract quantitative and other critical data from each study document. For other types of data, one person could extract the data while a second independently checks for accuracy and completeness. Establish a fair procedure for resolving discrepancies—do not simply give final decision-making power to the senior reviewer (per the IOM wording, p. 275).

3.5.2 Link publications from the same study to avoid including data from the same study more than once (p. 276). **Rationale:** There are numerous examples in the literature where two articles reporting the same study are thought to represent two separate studies.	Not mentioned.	It is acknowledged implicitly that there can be more than one publication per study, but there are no specific instructions about linking the publications from a single study together.	Not mentioned.	Not mentioned.	Create an explicit mechanism for linking multiple publications from the same study to avoid including duplicate data.
3.5.3 Use standard data extraction forms developed for the specific systematic review (p. 276). **Rationale:** Standardized data forms are broadly applied quality-assurance approaches.	Not mentioned.	An example worksheet is provided for observational epidemiologic studies and items to be extracted from the articles for animal toxicologic studies. A structured form may be useful for recording the key features needed to evaluate a study. An example form is shown in Figure F-3; details of such a form will need to be modified based on the specifics of the chemical, exposure scenarios, and effect measures under study.	Data-extraction forms are not described, and it is not known whether forms were used; evidence tables and exposure-response arrays provide a structured format for data-reporting.	Data-extraction forms are not described, and it is not known whether forms were used; evidence tables and exposure-response arrays provide a structured format for data-reporting.	Create, pilot-test, and use standard data-extraction forms (see also 3.5.4 below).
3.5.4 Pilot-test the data extraction forms and process (p. 276). **Rationale:** Pre-testing of the data collection forms and processes are broadly applied quality-assurance approaches.	Not mentioned.	Not mentioned.	Not mentioned.	Not mentioned.	Create, pilot-test, and use standard data-extraction forms (see 3.5.3 above).

In general, EPA has been responsive to the recommendations from the NRC formaldehyde report. As discussed in Chapter 1, the timing of the publication of the IOM standards was such that EPA could not have been expected to have incorporated the standards into its assessments to date. Nevertheless, comparison of statements made in the draft preamble (EPA 2013a, Appendix B) and draft handbook (EPA 2013a, Appendix F) with the 2011 IOM standards demonstrates that EPA has not only responded to the recommendations made in the NRC formaldehyde report but is well on the way to meeting the general systematic-review standards for identifying and assessing evidence.

Thus, the table is useful primarily for pointing out where further standardization might be helpful, not as a test and demonstration of whether IOM standards have been met. Sometimes the information that the committee sought is not mentioned in the sources examined but is present in other sources, for example, in explanatory materials provided on the EPA IRIS Web site and in chemical-specific links on the EPA Health and Environmental Research Online (HERO) Web site. After discussion, the committee elected to retain "not mentioned" in the table because the information sought was not mentioned in the documents reviewed even though it might have been noted elsewhere. A key goal was to see whether the information appeared where the average reader might expect to find it, notably in documents describing the methods used in developing IRIS assessments. For transparency, there should be no difficulty in accessing all aspects of review methods.

In addition, the subset of documents reflected in the table does not represent all the materials available. Because EPA's transition to a systematic process for reviewing the evidence is evolving, the committee expects that more recent documents will reflect an increasingly standardized and comprehensive response. The committee had to halt its examination of recent example documents in September 2013 so that the present report could be drafted, with the understanding that some elements that appear undeveloped in Table 4-1 have been addressed in materials released more recently.

Establish Standard Strategies for Evidence Identification

The IOM standards for finding and assessing individual studies include five main elements: searching for evidence, addressing possible reporting biases, screening and selecting studies, documenting the search, and managing data. As Table 4-1 shows, in most instances, the draft preamble (EPA 2013a, Appendix B) focuses on principles and does not address specific elements of the IOM standard. Because identifying evidence for IRIS involves all five elements reflected in the IOM standards, a concise preamble would not be expected to serve as a stand-alone roadmap for evidence-identification methods in IRIS assessments.

The draft handbook (EPA 2013a, Appendix F), however, should include that level of detail and does cover the IOM standards more completely, although some gaps exist. To address the gaps, the committee recommends expanding the handbook as itemized in Table 4-1. In general, EPA might find it helpful to include a table of standards in the handbook (perhaps repeated in the preamble) and to adopt the wording in Table 4-1 for each standard (for example, from IOM) or to modify the wording to be specific to the IRIS case, as appropriate.

As an overarching recommendation, the committee encourages EPA to include standard approaches for evidence identification in IRIS materials and to incorporate them consistently in the various materials. For components that are intentionally less detailed, such as the preamble, the committee encourages EPA to refer the reader elsewhere, notably to relevant parts of the handbook, for those interested in additional detail. The handbook serves as a valuable complement to the preamble, but without pointers to more detailed resources the average reader might not understand the relationship between the two documents or be aware that detailed strategies or standards exist.

Develop a Template for Description of the Search Strategies

EPA has provided the committee with a substantial set of tables, figures, and examples that demonstrate marked progress in implementing the recommendations from the NRC formaldehyde report. In reviewing the materials provided (EPA 2013a,b), the committee did not see evidence that a consistent search template was being used. The preamble (EPA 2013a, Appendix B) and the handbook (EPA 2013a, Appendix F) are helpful with regard to illustrating the overall structure and flow of the evidence-identification process. For example, Figures F-1 and F-2 in EPA (2013a, Appendix F) and Figure 1-1 in EPA (2013b) illustrate the literature-search documentation for ethyl *tert*-butyl ether. The committee recognizes that the process of developing and refining materials for the IRIS process is still going on and that representations of the search approach have probably continued to evolve. However, materials provided show that the approach is not yet specified consistently and in equivalent detail among the various documents. For example, Figure 1-1 in EPA (2013b) includes Proquest—a step that involves reviewing references cited in papers identified by the search—whereas the preamble (EPA 2013a, Appendix B) and Figure F-1 in the handbook (EPA 2013a, Appendix F) do not. It is also unclear whether inconsistencies are deliberate (and thus desirable) and related to the specific IRIS assessment being undertaken or are unintentional (and perhaps undesirable). For example, the preamble specifies searching "other databases listed in EPA's HERO system" (p. B-2) and a number of other, mostly unpublished sources, whereas Figure F-1 specifies "OPP databases" and also refers to searching other sources.

The draft materials provided to the committee do not yet appear to include some quality-control and procedural guidelines identified in the IOM standards (see Table 4-1) that are relevant to identifying the evidence. In particular, the materials do not consider whether prespecified research questions were used to guide the evidence identification (see Chapter 3 for a discussion of the development of the research question). The committee encourages EPA to consider prespecifying research questions when establishing the standard template for evidence identification to ensure that a search reflects the research goals appropriately. The committee commends EPA's collaboration with the National Toxicology Program of the National Institute of Environmental Health Sciences in this regard and encourages incorporation of insights gained into the IRIS process.

Use a Database to Capture Study Information and Relevant Quantitative Data

The NRC formaldehyde report recommended that EPA use a database, such as HERO, to serve as a repository for documents supporting its toxicity assessments. The HERO database was developed to support scientific assessments for the national ambient air quality standards, notably integrated science assessments for the six criteria pollutants. EPA responded to the recommendation with a substantial expansion of HERO to support IRIS (EPA 2013f,g). The extensive effort has involved incorporating more than 160,000 references relevant to IRIS since 2011, and updating has continued to today. For example, from August to September 2013, nearly 2,400 references were added to the IRIS set in HERO.

The committee encourages EPA to adapt HERO or create a related database to contain data extracted from the individual documents. Although it is not yet evident in the draft preamble or handbook (EPA 2013a), the HERO Web site (EPA 2013f) suggests such an adaptation. In describing what data HERO provides, the Web site (EPA 2013g) states "for 'key' studies: objective, quantitative extracted study data [future enhancement]." It further states that "HERO revisions are planned to broaden both the features and scope of information included. Future directions include additional data sets, environmental models, and services that connect data and models."

The committee recognizes that EPA has expanded the HERO database to capture information about documents relevant to IRIS assessments. Searching HERO, in addition to other databases, will be increasingly useful to identify relevant studies for IRIS assessments. As noted, the committee encourages EPA to expand HERO or build a complementary database into which data extracted from the documents in the HERO database can be entered. By creating a data repository for study information and relevant quantitative data, EPA will be able to accumulate and evaluate evidence among IRIS assessments. Such a repository of identified data (see Goldman and Silbergeld 2013) would further enhance the process and its consistent application for IRIS, as well as enhancing data-sharing (for example, see related discussion in IOM 2013).

COMMENTS ON BEST PRACTICES FOR EVIDENCE IDENTIFICATION

IOM (2011) standards as highlighted in Table 4-1 capture recent best practices. Searching for and identifying the evidence is arguably the most important step in a systematic review. Accordingly, a standardized search strategy and reporting format are essential for evidence identification. As discussed in Chapter 3, the protocol frames an answerable question or questions that will be addressed by the assessment, states the eligibility criteria for inclusion in the assessment, and describes in detail how the relevant evidence will be identified. Searches should always be well documented with an expected format, as described in the articles on search filters for Embase (deVries et al. 2011) and for PubMed (Hooijmans et al. 2010). As noted above, the committee could not always find some of the critical information in the draft materials that it reviewed, although it has found that more recent IRIS assessments and preliminary materials for upcoming assessments reflect increasing standardization (for example, see, EPA 2013h,i), which is commendable. Standardizing the search strategy and reporting format would aid the reader of IRIS assessments and would facilitate an evaluation of how well the standards and concepts set forth in the preamble and handbook are being applied. In addition, standardization would help to minimize unnecessary duplication, overlaps, and inconsistencies among various IRIS assessments. An example of format and documentation issues related to searching for animal studies can be found in Leenaars et al. (2012).

The IOM standards also emphasize the role of various specialists in the review process, including information specialists (also referred to as informationists) and topic-specific experts. Those screening the studies and abstracting the data also need explicit training, and typically topic-specific experts are involved at this step. The roles of all team members should be identified in the protocol.

The evidence supporting the IOM standards is likely to be useful in the IRIS domain, but it would be appropriate for EPA to perform research that examines evidence specifically applicable to epidemiology and toxicity evaluations underlying IRIS assessments. For example, a targeted research effort could address the question of whether it is useful and necessary to search the gray literature—research literature that has not been formally published in journal articles, such as conference abstracts, book chapters, and theses—and the non-English-language literature in systematic reviews for IRIS assessments. Given how quickly methods for systematic reviews are evolving, including databases and indexing terms, methodologic research related to systematic reviews for IRIS assessments should be kept current to ensure that standards are up to date and relevant.

FINDINGS AND RECOMMENDATIONS

The findings and recommendations that follow are broad recommendations on evidence identification; specific suggestions or considerations for each step in the process are provided in Table 4-1.

Finding: EPA has been responsive to recommendations in the NRC formaldehyde report regarding evidence identification and is well on the way to adopting a more rigorous approach to evidence identification that would meet standards for systematic reviews. This finding is based on a comparison of the draft EPA materials provided to the committee with IOM standards.

Recommendation: The trajectory of change needs to be maintained.

Finding: Current descriptions of search strategies appear inconsistently comprehensive, particularly regarding (a) the roles of trained information specialists; (b) the requirements for contractors; (c) the descriptions of search strategies for each database and source searched; (d) critical details concerning the search, such as the specific dates of each search and the specific publication dates included; and (e) the periodic need to consider modifying the databases and languages to be searched in updated and new reviews. The committee acknowledges that recent assessments other than the ones that it reviewed might already address some of the indicated concerns.

Recommendation: The current process can be enhanced with more explicit documentation of methods. Protocols for IRIS assessments should include a section on evidence identification that is written in collaboration with information specialists trained in systematic reviews and that includes a search strategy for each systematic-review question being addressed in the assessment. Specifically, the protocols should provide a line-by-line description of the search strategy, the date of the search, and publication dates searched and, as noted in Chapter 3, explicitly state the inclusion and exclusion criteria for studies.

Recommendation: Evidence identification should involve a predetermined search of key sources, follow a search strategy based on empirical research, and be reported in a standardized way that allows replication by others. The search strategies and sources should be modified as needed on the basis of new evidence on best practices. Contractors who perform the evidence identification for the systematic review should adhere to the same standards and provide evidence of experience and expertise in the field.

Finding: One problem for systematic reviews in toxicology is identifying and retrieving toxicologic information outside the peer-reviewed public literature.

Recommendation: EPA should consider developing specific resources, such as registries, that could be used to identify and retrieve information about toxicology studies reported outside the literature accessible by electronic searching. In the medical field, clinical-trial registries and US legislation that has required studies to register in ClinicalTrials.gov have been an important step in ensuring that the total number of studies that are undertaken is known.

Finding: Replicability and quality control are critical in scientific undertakings, including data management. Although that general principle is evident in IRIS assessments that were reviewed, tasks appear to be assigned to a single information specialist or review author. There was no evidence of the information specialist's or reviewer's training or of review of work by others who have similar expertise. As discussed in Chapter 2, an evaluation of validity and reliability through inter-rater comparisons is important and helps to determine whether multiple reviewers are needed. This aspect is missing from the IOM standards.

Recommendation: EPA is encouraged to use at least two reviewers who work independently to screen and select studies, pending an evaluation of validity and reliability that might indicate that multiple reviewers are not warranted. It is important that the reviewers use standardized procedures and forms.

Finding: Another important aspect of quality control in systematic reviews is ensuring that information is not double-counted. Explicit recognition of and mechanisms for dealing with multiple publications that include overlapping data from the same study are important components of data management that are not yet evident in the draft handbook.

Recommendation: EPA should engage information specialists trained in systematic reviews in the process of evidence identification, for example, by having an information specialist peer review the proposed evidence-identification strategy in the protocol for the systematic review.

Finding: The committee did not find enough empirical evidence pertaining to the systematic-review process in toxicological studies to permit it to comment specifically on reporting biases and other methodologic issues, except by analogy to other, related fields of scientific inquiry. It is not clear, for example, whether a reporting bias is associated with the language of publication for toxicological studies and the other types of research publications that support IRIS assessments or whether any such bias (if it exists) might be restricted to specific countries or periods.

Recommendation: EPA should encourage and support research on reporting biases and other methodologic topics relevant to the systematic-review process in toxicology.

Finding: The draft preamble and handbook provide a good start for developing a systematic, quality-controlled process for identifying evidence for IRIS assessments.

Recommendation: EPA should continue to document and standardize its evidence-identification process by adopting (or adapting, where appropriate) the relevant IOM standards described in Table 4-1. It is anticipated that its efforts will further strengthen the overall consistency, reliability, and transparency of the evidence-identification process.

REFERENCES

AHRQ (Agency for Healthcare Research and Quality). 2011. The Effective Health Care Program Stakeholder Guide. Publication No. 11-EHC069-EF [online]. Available: http://www.ahrq.gov/research/findings/evidence-based-reports/stakeholderguide/stakeholdr.pdf [accessed December 4, 2013].

Briel, M., K.F. Muller, J.J. Meerpohl, E. von Elm, B. Lang, E. Motshall, V. Gloy, F. Lamontagne, G. Schwarzer, and D. Bassler. 2013. Publication bias in animal research: A systematic review protocol. Syst. Rev. 2:23.

Chalmers, I., and K. Dickersin. 2013. Biased under-reporting of research reflects biased under-submission more than biased editorial rejection. F1000 Research 2(1).

Chang, S.M., E.B. Bass, N. Berkman, T.S. Carey, R.L. Kane, J. Lau, and S. Ratichek. 2013. Challenges in implementing The Institute of Medicine systematic review standards. Syst. Rev. 2(1):69.

CRD (Centre for Reviews and Dissemination). 2009. Systematic Reviews: CRD's Guidance for Undertaking Reviews in Health Care. York, UK: York Publishing Services, Ltd [online]. Available: http://www.york.ac.uk/inst/crd/pdf/Systematic_Reviews.pdf [accessed December 4, 2013].

de Vries, R.B., C.R. Hooijmans, A. Tillema, M. Leenaars, and M. Ritskes-Hoitinga. 2011. A search filter for increasing the retrieval of animal studies in Embase. Lab Anim. 45(4):268-270.

Dickersin, K., and I. Chalmers. 2011. Recognizing, investigating and dealing with incomplete and biased reporting of clinical research: From Francis Bacon to the World Health Organisation. J. R. Soc. Med. 104(12):532-538.

EPA (U.S. Environmental Protection Agency). 2013a. Part 1. Status of Implementation of Recommendations. Materials Submitted to the National Research Council, by Integrated Risk Information System Program, U.S. Environmental Protection Agency, January 30, 2013[online]. Available: http://www.epa.gov/IRIS/iris-nrc.htm [accessed October 22, 2013].

EPA (U.S. Environmental Protection Agency). 2013b. Part 2. Chemical-Specific Examples. Materials Submitted to the National Research Council, by Integrated Risk Information System Program, U.S. Envi-

ronmental Protection Agency, January 30, 2013 [online]. Available: http://www.epa.gov/iris/pdfs/IRIS%20Program%20Materials%20to%20NRC_Part%202.pdf [accessed October 22, 2013].

EPA (U.S. Environmental Protection Agency). 2013c. Toxicological Review of Benzo[a]pyrene (CAS No. 50-32-8) in Support of Summary Information on the Integrated Risk Information System (IRIS), Public Comment Draft. EPA/635/R13/138a. National Center for Environmental Assessment, Office of Research and Development, U.S. Environmental Protection Agency, Washington, DC. August 2013 [online]. Available: http://cfpub.epa.gov/ncea/iris_drafts/recordisplay.cfm?deid=66193 [accessed Nov. 13, 2013].

EPA (U.S. Environmental Protection Agency). 2013d. Toxicological Review of Ammonia (CAS No. 7664-41-7), In Support of Summary Information on the Integrated Risk Information System (IRIS). Revised External Review Draft. EPA/635/R-13/139a. National Center for Environmental Assessment, Office of Research and Development, U.S. Environmental Protection Agency, Washington, DC. August 2013 [online]. Available: http://cfpub.epa.gov/ncea/cfm/recordisplay.cfm?deid=254524 [accessed October 14, 2013].

EPA (U.S. Environmental Protection Agency). 2013e. Toxicological Review of Ammonia (CAS No. 7664-41-7), In Support of Summary Information on the Integrated Risk Information System (IRIS), Supplemental Information. Revised External Review Draft. EPA/635/R-13/139b. National Center for Environmental Assessment, Office of Research and Development, Washington, DC [online]. Available: http://cfpub.epa.gov/ncea/cfm/recordisplay.cfm?deid=254524 [accessed October 14, 2013].

EPA (U.S. Environmental Protection Agency). 2013f. Health and Environmental Research Online (HERO): The Assessment Process [online]. Available: http://hero.epa.gov/index.cfm?action=content.assessment [accessed October 14, 2013].

EPA (U.S. Environmental Protection Agency). 2013g. Health and Environmental Research Online (HERO): Basic Information [online]. Available: http://hero.epa.gov/index.cfm?action=content.basic [accessed October 14, 2013].

EPA (U.S. Environmental Protection Agency). 2013h. Preliminary Materials for the Integrated Risk Information System (IRIS) Toxicological Review of tert-Butyl Alcohol (tert-Butanol) [CASRN 75-65-0]. EPA/635/R-13/107. National Center for Environmental Assessment, Office of Research and Development, Washington, DC. July 2013 [online]. Available: http://www.epa.gov/iris/publicmeeting/iris_bimonthly-oct2013/t-butanol-litsearch_evidence-tables.pdf [accessed October 14, 2013].

EPA (U.S. Environmental Protection Agency). 2013i. Systematic Review of the tert-Butanol Literature (generated by HERO) [online]. Available: http://www.epa.gov/iris/publicmeeting/iris_bimonthly-oct2013/mtg_docs.htm [accessed October 14, 2013].

Fanelli, D. 2010. Do pressures to publish increase scientists' bias? An empirical support from U.S. States data. PLoS One 5(4):e10271.

Goldman, L.R., and E.K. Silbergeld. 2013. Assuring access to data for chemical evaluations. Environ. Health Perspect. 121(2):149-152.

Higgins, J.P.T., and S. Green, eds. 2008. Cochrane Handbook for Systematic Reviews of Interventions. Chichester, UK: John Wiley & Sons.

Hoffman, S., and T. Hartung. 2006. Toward an evidence-based toxicology. Hum. Exp. Toxicol. 25(9):497-513.

Hooijmans, C.R., and M. Ritskes-Hoitinga. 2013. Progress in using systematic reviews of animal studies to improve translational research. PLOS Med. 10(7):e1001482.

Hooijmans, C.R., A. Tillema, M. Leenaars, and M. Ritskes-Hoitinga. 2010. Enhancing search efficiency by means of a search filter for finding all studies on animal experimentation in PubMed. Lab Anim. 44(3):170-175.

IOM (Institute of Medicine). 2011. Finding What Works in Health Care: Standards for Systematic Reviews. Washington, DC: National Academies Press.

IOM (Institute of Medicine). 2013. Sharing Clinical Research Data: Workshop Summary. Washington, DC: National Academies Press.

Korevaar, D.A., L. Hooft, and G. ter Riet. 2011. Systematic reviews and meta-analyses of preclinical studies: Publication bias in laboratory animal experiments. Lab. Anim. 45(4):225-230.

Leenaars, M., C.R. Hooijmans, N. van Veggel, G. ter Riet, M. Leeflang, L. Hooft, G.J. van der Wilt, A. Tillema, and M. Ritskes-Hoitinga. 2012. A step-by-step guide to systematically identify all relevant animal studies. Lab Anim. 46(1):24-31.

Lundh, A., S. Sismondo, J. Lexchin, O.A. Busuioc, and L. Bero. 2012. Industry sponsorship and research outcome. Cochrane Database Syst. Rev. (12):Art. MR000033. doi: 10.1002/14651858.MR000033.pub2.

Mignini, L.E., and K.S. Khan. 2006. Methodological quality of systematic reviews of animal studies: A survey of reviews of basic research. BMC Med. Res. Methodol. 6:10. doi:10.1186/1471-2288-6-10.

NRC (National Research Council). 2011. Review of the Environmental Protection Agency's Draft IRIS Assessment of Formaldehyde. Washington, DC: National Academies Press.

Sena, E.S., H.B. van der Worp, P.M. Bath, D.W. Howells, and M.R. Macleod. 2010. Publication bias in reports of animal stroke studies leads to major overstatement of efficacy. PLoS Biol. 8(3):e1000344.

Silbergeld, E., and R.W. Scherer. 2013. Evidence-based toxicology: Strait is the gate, but the road is worth taking. ALTEX 30(1):67-73.

ter Riet, G., D.A. Korevaar, M. Leenaars, P.J. Sterk, C.J.F. Van Noorden, L.M. Bouter, R. Lutter, R.P.O. Elferink, and L. Hooft. 2012. Publication bias in laboratory animal research: A survey on magnitude, drivers, consequences and potential solutions. PLOS ONE 7(9):e43404.

Woodruff, T.J., and P. Sutton. 2011. An evidence-based medicine methodology to bridge the gap between clinical and environmental health sciences. Health Aff. (Millwood) 30(5):931-937.

5

Evidence Evaluation

This chapter focuses on a critical part of the systematic-review process: the assessment of the individual studies that are selected for inclusion in a review. As depicted in Figure 5-1, this step comes after the comprehensive search for and identification of relevant studies and precedes the integration of human, animal, and mechanistic evidence from the systematic reviews. The chapter first reviews the recommendations on evidence evaluation from the National Research Council (NRC) formaldehyde report and the US Environmental Protection Agency (EPA) responses to them. Best practices for evaluating clinical and epidemiologic studies, animal toxicology studies, and mechanistic studies in the systematic-review process are then discussed. Drawing on approaches developed for systematic reviews in clinical practice and public health, the committee emphasizes the need for EPA to assess the "risk of bias" in individual studies. Accordingly, the best-practice section defines the terms, notes the possibility that bias could arise throughout the conduct and reporting of a study, and discusses how to review studies within a risk-of-bias framework. Needs for further developing best practices are identified, and the committee's findings and recommendations are provided at the conclusion of the chapter.

RECOMMENDATIONS ON EVIDENCE EVALUATION FROM THE NATIONAL RESEARCH COUNCIL FORMALDEYDE REPORT

The earlier formaldehyde report observed that "ultimately, the quality of the studies reviewed and the strength of evidence provided by the studies for deriving RfCs [reference concentrations] and unit risks need to be clearly presented" (NRC 2011a, p. 155). To that end, the report provided several recommendations for evaluating the evidence; these are provided in Box 5-1. Briefly, the recommendations focus on standardizing the presentation of the reviewed studies and evaluating the studies with standardized approaches that consider confounding and other biases.

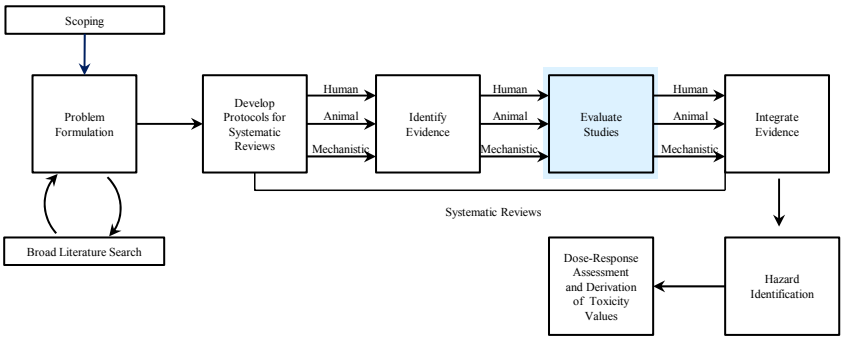


FIGURE 5-1 The IRIS process; the evidence-evaluation step is highlighted. The committee views public input and peer review as integral parts of the IRIS process, although they are not specifically noted in the figure.

BOX 5-1 Recommendations from the National Research Council Formaldehyde Report on Evidence Evaluation

- All critical studies need to be thoroughly evaluated with standardized approaches that are clearly formulated and based on the type of research, for example, observational epidemiologic or animal bioassays. The findings of the reviews might be presented in tables to ensure transparency.
- Standardize the presentation of reviewed studies in tabular or graphic form to capture the key dimensions of study characteristics, weight of evidence, and utility as a basis for deriving reference values and unit risks.
- Standardized evidence tables for all health outcomes need to be developed. If there were appropriate tables, long text descriptions of studies could be moved to an appendix or deleted.
- Develop templates for evidence tables, forest plots, or other displays.
- Establish protocols for review of major types of studies, such as epidemiologic and bioassay.

Source: NRC 2011a, pp. 152 and 165.

EVALUATION OF ENVIRONMENTAL PROTECTION AGENCY RESPONSE TO THE NATIONAL RESEARCH COUNCIL FORMALDEHYDE REPORT

As discussed in Chapter 1 (see Table 1-1), the committee used the EPA reports *Status of Implementation of Recommendations* (EPA 2013a) and *Chemical-Specific Examples* (EPA 2013b) to evaluate EPA's progress in implementing the recommendations in the NRC formaldehyde report (NRC 2011a). The committee also reviewed the draft IRIS assessment for benzo[a]pyrene (EPA 2013c) to gauge EPA's progress in implementing recommendations for evidence evaluation. The draft preamble for IRIS assessments (EPA 2013a, Appendix B) provides a checklist that EPA will use to assess the quality of epidemiologic and experimental studies (see Box 5-2). EPA (2013a, p. F-44) also states that those and other considerations are "consistent with guidelines for systematic reviews that evaluate the quality and weight of evidence." On the basis of those comments, the committee assumes that EPA intends to adopt principles associated with systematic review and with analysis of studies for risk of bias.

EPA (2013a, Appendix F) also includes a draft handbook for IRIS assessment development. The handbook mentions the following approach for epidemiologic studies: "to the extent possible, you want to assess not just the 'risk of bias,' but also the likelihood, direction, and magnitude of bias" (EPA 2013a, p. F-23). The draft handbook provides a table that outlines general considerations for evaluating features of epidemiologic studies. Features that would be assessed for quality and potential risk of bias include study design, study population and target-population setting, participation rate and follow-up, comparability between exposed and control populations, exposure-assessment methods, outcome measures, and data presentation, such as statistical analyses. The handbook also develops similar criteria for assessing the quality of animal studies. Animal-study features that would be assessed for quality and potential risk of bias include study design, exposure quality, test animals, end-point evaluation, data presentation and analysis, and reporting.

Several factors, including blinding and sampling bias, are discussed in the handbook, but several important risk-of-bias and quality-control elements, including attrition and statistical-power calculations, are not explicitly considered. The handbook emphasizes evaluation of how a study is reported (as opposed to how it is conducted). EPA describes the use of tables to present

BOX 5-2 Aspects to Consider in Evaluating Study Quality as Listed in the Preamble for IRIS Assessments

Epidemiologic Studies

- Documentation of study design, methods, population characteristics, and results.
- Definition and selection of the study group and comparison group.
- Ascertainment of exposure to the chemical or mixture.
- Ascertainment of disease or health effect.
- Duration of exposure and follow-up and adequacy for assessing the occurrence of effects.
- Characterization of exposure during critical periods.
- Participation rates and potential for selection bias as a result of the achieved participation rates.
- Measurement error…and other types of information bias.
- Potential confounding and other sources of bias addressed in the study design or in the analysis of results.

Experimental Studies

- Documentation of study design, animals or study population, methods, basic data, and results.
- Nature of the assay and validity for its intended purpose.
- Characterization of the nature and extent of impurities and contaminants of the administered chemical or mixture.
- Characterization of dose and dosing regimen (including age at exposure) and their adequacy to elicit adverse effects, including latent effects.
- Sample sizes and statistical power to detect dose-related differences or trends.
- Ascertainment of survival, vital signs, disease or effects, and cause of death.
- Control of other variables that could influence the occurrence of effects.

Source: EPA 2013a, Appendix B.

the results of the study-quality evaluations (if there are robust datasets). Some examples provide qualitative descriptors for quality factors being assessed, for example, robust, moderate, or poor; or + to indicate that "criteria [are] not completely met or potential issues identified, but [they are] unlikely to directly affect study interpretation"; and ++ to indicate that "criteria [are] determined to be completely met" (EPA 2013a, Appendix F, p. F-32). Evidence tables presented in the draft IRIS assessment for benzo[a]pyrene (EPA 2013c, Tables 1-1 through 1- 9 and 1-11 through 1-16) describe the populations, exposures, and outcomes of each study and the results. However, there is no assessment of the risk of bias in the studies evaluated, so it is unclear how EPA will meet its goal of assessing the direction and magnitude of bias in epidemiologic or animal studies. There is also no description of quality-assurance measures for the collection of assessment data.

Overall, EPA (2013a) has identified relevant study attributes to consider in evaluating study quality and indicates its intention to adopt risk-of-bias analyses. However, its considerations ignore some important elements that should be covered; accordingly, the following section focuses heavily on best practices for evaluating risk of bias in individual studies.

BEST PRACTICES FOR EVALUATING EVIDENCE FROM INDIVIDUAL STUDIES

Best practices for evaluating evidence from individual studies have developed in a number of fields. Here, following the general recommendation of the NRC formaldehyde report and the direction that EPA is taking with its revisions, the committee focuses on the best practices as developed for systematic reviews in clinical medicine and public health. As described above, although EPA has identified and is assessing important characteristics of the quality of human and animal studies, it has not historically conducted the assessments in a consistent and standardized way for studies included in IRIS assessments.

The report *Finding What Works in Health Care: Standards for Systematic Reviews* (IOM 2011) provides a useful roadmap for conducting systematic reviews, including the process of evaluating individual studies (see Chapter 4). Key elements of the Institute of Medicine (IOM) standards for study evaluation include an assessment of the relevance of the study's populations, interventions, and outcomes for the systematic-review questions. After studies and outcomes are assessed for relevance, the IOM standards recommend a systematic assessment of the risk of bias (see below) according to predefined criteria. The potential biases must be assessed to determine how confident one can be in the conclusions drawn from the data.

The IOM standards are directed primarily toward the evaluation of evidence that compares the benefits and harms associated with alternative methods for preventing, diagnosing, treating for, and monitoring a clinical condition or for improving the delivery of care (that is, comparative-effectiveness research). Although the reviews conducted by the IRIS program have a distinctly different objective, many of the guiding principles identified by IOM can be applied to the IRIS program. For example, a systematic evaluation of research evidence for selection bias, dose-response associations, plausible confounders that could bias an estimated effect, and the strength of observed associations is relevant for both comparative-effectiveness research and toxicologic assessment.

Evaluation of Risk of Bias and Study Quality in Human Clinical and Epidemiologic Research

The validity of scientific evidence from a particular study has multiple determinants, from the initial formulation of a study hypothesis to the reporting of findings. In its assessment of findings that will be used from one or more studies in an IRIS assessment, EPA needs to address potential threats to the validity of evidence. The threats are generally well recognized by researchers but are referred to in different terms among fields. Here, the committee adopts the terminology that is used in systematic reviews of the medical literature.

The concept of risk of bias is central to the evaluation of studies for systematic reviews of clinical evidence (Higgins and Green 2008). The term is now widely used by those conducting such reviews; for example, it is extensively discussed in the *Cochrane Handbook for Systematic Reviews of Interventions* (Higgins and Green 2008). The term, which might be unfamiliar to those in the field of toxicology, is defined by IOM as "the extent to which flaws in the design and execution of a collection of studies could bias the estimate of effect for each outcome under study" (IOM 2011, p. 165). Bias is generally defined as error that reduces validity, and risk of bias refers to the potential for bias to have occurred. A study might be in compliance with regulatory standards but still have an important risk of bias. For example, a study might adhere to all rules regarding the ethical treatment of animals (for example, appropriate housing, diet, and pain management) but have a bias due to lack of blinding of outcome measurement. The committee notes that risk of bias is not the same as imprecision (Higgins and Green 2008). *Bias* refers to systematic error, and *imprecision* refers to random error; smaller studies are less precise, but they might be less biased than larger ones.

Risk of bias has been empirically linked to biased effect estimates, but the direction (overestimate or underestimate) of the effect cannot be determined a priori on the basis of the specific type of risk of bias. For example, inadequate randomization (a risk of bias) in a drug study has been associated with overestimates of efficacy measures and underestimates of harm measures. In controlled human clinical trials that test drug efficacy, studies that have a high risk of bias—such as those lacking randomization, allocation concealment, or blinding of participants, personnel, and outcome assessors—tend to produce larger treatment-effect sizes and thus falsely inflate the efficacy of the drug being evaluated compared with studies that include those design features (Schulz et al. 1995; Schulz and Grimes 2002a,b).[1] Biased human studies that assess the harm of drugs are less likely than nonbiased studies to report statistically significant adverse effects (Nieto et al. 2007). Those results reflect the need to consider bias related to funding source that might arise from systematic influences on the design and conduct of a study and the extent to which the full results and analyses of the study are published (Lundh et al. 2012).

Study quality and risk of bias are not equivalent concepts, and there is a difference between assessing risk of bias and assessing other methodologic characteristics of a particular study (Higgins and Green 2008). Assessment of study quality involves an investigation of the extent to which study authors conducted their research to high standards—for example, by following a well-documented protocol with trained study staff, by conducting an animal study according to good laboratory practices (GLP), or by complying with human-subjects guidelines for a clinical study. Assessment of study quality also evaluates how a study is reported—for example, whether the study population is described sufficiently. Risk of bias involves the internal validity of a study and reflects study-design characteristics that can introduce a systematic error or deviation from a true effect that might affect the magnitude and even the direction of the apparent effect (Higgins and Green 2008). [2]

Table 5-1 identifies for the present report the various forms of bias within a study that underlie risk-of-bias assessments. The committee recognizes that terminology related to bias is not standardized and varies among fields and textbooks. It has not proposed specific definitions, but to ensure consistency throughout this report, it uses the terminology and descriptions of how biases arise that are presented in Table 5-1. The terms in the table are broad but cover the major types of bias that are relevant to IRIS assessments. EPA will need to give consideration to the terminology and definitions that it will use.

The Cochrane Handbook provides a classification scheme that includes selection, performance, detection, attrition, and reporting bias (Higgins and Green 2008). That scheme is oriented toward randomized clinical trials. In observational studies, the potential for confounding is a critical aspect of study design, data collection, and data analysis that needs to be assessed; in randomized clinical trials, confounding is addressed through randomization. Selection bias reflects differences in participant characteristics at baseline or that arise during follow-up and reflects patterns of participation that distort the results from those that would be found if the full

[1]The committee defines randomization, allocation concealment, and blinding as follows: randomization is a process that ensures that test subjects are randomly assigned to treatment groups, allocation concealment is a process that ensures that the person allocating subjects to treatment groups is unaware of the treatment groups to which the subjects are being assigned, and blinding (or masking) is a set of procedures that keeps the people who perform an experiment, collect the data, or assess the outcome measures unaware of the treatment allocation.

[2]The committee distinguishes internal validity, which is related to the bias of an individual study design, from external validity, which is the degree to which the results of the study can be generalized to settings or groups other than those used in the given study. Although external validity is relevant for determining whether a study should be included in a systematic review, it is not relevant for assessing bias within a study.

TABLE 5-1 Types of Biases and Their Sources

Type of Bias	Sources
Randomized Studies	
Selection	Systematic differences between exposed and control groups in baseline characteristics that result from how subjects are assigned to groups
Performance	Systematic differences between exposed and control groups with regard to how the groups are handled
Detection[a]	Systematic differences between exposed and control groups with regard to how outcomes are assessed
Attrition or exclusion	Systematic difference between exposed and control groups in withdrawal from the study or exclusion from analysis
Observational Studies	
Confounding and selection	Differences in the distribution of risk factors between exposed and nonexposed groups—can occur at baseline or during follow-up
Measurement	Mismeasurement of exposures, outcomes, or confounders—can occur at any time during the study
Randomized or Observational Studies	
Reporting	Selective reporting of entire studies, outcomes, or analyses

[a]Detection bias includes measurement errors.
Source: Adapted from Higgins and Green 2008, Table 8.4, p. 8.7.

population could be observed. Problems with measurements in observational studies might also affect outcomes (detection bias in randomized clinical trials), exposures, potential confounders, and modifying factors. Such measurement problems might be systematic or random.

A number of methods can be used to minimize bias. The biases—and methods used to reduce them—are the same for human and animal studies. For example, randomized studies of humans or rodents could be at risk of selection or exclusion bias. There is empirical evidence that some methodologic characteristics can protect against specific biases. For example, in randomized clinical trials, selection bias can be minimized by randomization and concealment of allocation. In observational studies, confounding that arises from differences between exposed and nonexposed groups at enrollment or from differences that develop during follow-up in a prospective study can sometimes be addressed by using statistical techniques to adjust for group differences that can be measured. Selection bias that arises from difference at baseline or patterns of dropout during follow-up is not readily addressed by modeling.

The NRC formaldehyde report (NRC 2011a) noted a number of study characteristics that should be included in a template for evaluating observational epidemiologic studies (see Box 5-3). The recommended evaluation template includes items for assessing bias in design (the internal validity of a study) and the generalizability of the study (the external validity). As noted above, a number of assessment tools have been developed to assess the different types of biases for different study designs and data streams and are discussed in the following sections.

Cochrane Approach

The Cochrane Handbook for Systematic Reviews of Interventions includes a tool for systematically assessing the risk of bias in individual studies of the causal efficacy of a health intervention (Higgins and Green 2011). The tool is designed to handle randomized trials that assign human exposures at random and asks reviewers to assess whether the treatment and control groups are comparable (random allocation and allocation concealment), whether the participants or subjects were blind to their treatment, whether there was detection bias (whether knowledge of exposure condition affected the measurement of outcome), and whether there was attrition bias (whether dropout was associated with treatment), reporting bias, or "other sources of bias."

BOX 5-3 Considerations for a Template for Evaluating an Epidemiologic Study

- Approach used to identify the study population and the potential for selection bias.
- Study population characteristics and the generalizability of findings to other populations.
- Approach used for exposure assessment and the potential for information bias, whether differential (nonrandom) or nondifferential (random).
- Approach used for outcome identification and any potential bias.
- Appropriateness of analytic methods used.
- Potential for confounding to have influenced the findings.
- Precision of estimates of effect.
- Availability of an exposure metric that is used to model the severity of adverse response associated with a gradient of exposures.

Source: NRC 2011a, p. 158.

The Cochrane treatment of risk of bias in nonrandomized studies (Higgins and Green 2011, Section 13.5.1.1) is extremely limited, but a new tool is being developed. It states that the sources of bias remain the same but that in some cases statistical techniques are used to control or adjust for potential confounding. It provides little help in systematic review of observational studies, which predominate in human research on the risks posed by chemicals.

The Cochrane Handbook (Higgins and Green 2011, Section 8.3.1) states that "the Collaboration's recommended tool for assessing risk of bias is neither a scale nor a checklist. It is a domain-based evaluation, in which critical assessments are made separately for different domains." Cochrane discourages using a numerical scale because calculating a score involves choosing a weighting for the subcomponents, and such scaling generally is nearly impossible to justify (Juni et al. 1999). Furthermore, a study might be well designed to eliminate bias, but because the study failed to report details in the publication under review, it will receive a low score. Most scoring systems mix criteria that assess risk of bias and reporting. However, there is no empirical basis for weighting the different criteria in the scores. Reliability and validity of the scores often are not measured. Furthermore, quality scores have been shown to be invalid for assessing risk of bias in clinical research (Juni et al. 1999). The current standard in evaluation of clinical research calls for reporting each component of the assessment tool separately and not calculating an overall numeric score (Higgins and Green 2008).

The Cochrane tool does include sources of risk of bias that are empirically based. For example, empirical studies of clinical trials show that inadequate or unclear concealment of allocation results in greater heterogeneity of results and effect sizes that are up to 40% larger than those of studies that contain adequate concealment of allocation (Schulz et al. 1995; Schulz and Grimes 2002b). In an empirical evaluation of the literature, Odgaard-Jensen et al. (2011) also found that both randomization and concealment of allocation can be associated with the estimated effect; results of randomized and nonrandomized studies differed, although variably, and nonblinded studies tended to provide larger effect estimates.

Ottawa-Newcastle Tool

The Ottawa-Newcastle tool for observational studies is an alternative to the Cochrane tool and, because it is intended for epidemiologic studies, is more relevant to the human studies eval-

uated in IRIS assessments.[3] Developed primarily for case-control and cohort studies, the tool contains separate coding manuals for each. It is relatively short and easy to implement and has been used to evaluate risk of bias in studies of the association between coronary heart disease and hormone-replacement therapy in postmenopausal women. However, the Ottawa-Newcastle tool focuses on only three dimensions of a study: selection of the sample, comparability of the study groups, and outcome assessment (for cohort studies) or exposure assessment (for case-control studies).

National Toxicology Program Tool

The National Toxicology Program (NTP) in collaboration with the Office of Health Assessment and Translation (OHAT) of the National Institute of Environmental Health Sciences has recently constructed a method for systematic reviews to "assess the evidence that environmental chemicals, physical substances, or mixtures...cause adverse health effects and [provide] opinions on whether these substances may be of concern given what is known about current human exposure levels" (NTP 2013, p. 1). NTP assesses the risk of bias in individual studies by using questions related to five categories—selection, performance, detection, attrition or exclusion, and selective reporting bias (the latter term, used by OHAT, is equivalent to reporting bias)—that are similar to those used by the Cochrane Collaboration and adapted for environmental studies by the Navigation Guide Work Group (Woodruff and Sutton 2011). For each study outcome, risk of bias is assessed on a four-point scale: definitely low, probably low, probably high, or definitely high.

Evaluation of Tools for Assessing Risk of Bias in Human Studies

All the existing tools applicable to assessing risk of bias in a study of human risks require a substantial amount of expert judgment to rate a study's effectiveness in controlling or adjusting for potential bias with methods other than randomization, allocation concealment, or blinding. One challenge in adapting systematic-review methods for environmental-epidemiology studies is the formal consideration of potential confounding and its consequences. Standard reviews and meta-analyses of epidemiologic studies typically include identification of key confounding variables; the failure to adjust for which is thought to result in bias. For example, a review of epidemiologic studies of maternal caffeine consumption and low birth weight would usually exclude (or stratify) studies that did not adequately adjust for maternal smoking as a confounding variable because maternal smoking is a risk factor for low birth weight and a known correlate of caffeine consumption.

Confounding is such an important bias in environmental epidemiology that it is often the primary consideration in assessing a study. Statistical techniques can be used to adjust for confounding in observational studies,[4] but they require expert knowledge to identify, measure, and model the confounders correctly and resources to measure the confounders. It is often helpful to represent one's understanding of potential confounding in a causal diagram (Greenland et al. 1999). In practice, however, it is unlikely that the etiology of any disease is sufficiently understood to create a causal diagram that fully and correctly accounts for all possible confounding factors. A reviewer of observational studies of a particular chemical and adverse health outcome might start by drawing several possible causal diagrams, including all known and suspected etio-

[3]The Ottawa-Newcastle tool (Wells et al. 2013) was developed at the Ottawa Hospital Research Institute and is available at http://www.ohri.ca/programs/clinical_epidemiology/oxford.asp.

[4]For example, see Hernan and Robins 2008; Hernan et al. 2009; Hernan and Cole 2009; Hernan and Robins 2013; Robins and Hernan 2008; Robins 2000; Robins et al. 2000; Greenland 1989; Greenland et al. 1999; Greenland 2000.

logic factors. Each reviewed study could then be compared with each causal diagram to determine whether control for confounding was adequate given the present state of knowledge regarding the etiology of the health outcome.

Measurement error is another common source of bias in epidemiologic research; it can be systematic or random and can affect data related to exposures, outcomes, confounders, or modifiers (Armstrong 1998). Indeed, chemical-exposure assessment can be so difficult that one environmental-epidemiology textbook devotes an entire chapter to the topic of exposure-measurement error and states that "the quality of exposure data is a major factor in the validity of epidemiological studies" (Baker and Nieuwenhuijsen 2008, p. 258). Difficulties in exposure assessment arise from the time variability of personal exposures to chemicals, and it is often not feasible to assess exposures of each participant in an epidemiologic study. In addition, lifetime exposure is often relevant for cancer, and various estimation approaches are needed to fill gaps in the context of an epidemiologic study; the error that results from the estimation approaches might lead to measurement bias. Thus, many environmental epidemiology studies rely on complex exposure assignments based on a few environmental measurements and occupational records, residential histories, or time-activity surveys. Exposure biomarkers are increasingly used for exposure assessment but are not necessarily superior to other methods, depending on the period covered and their sensitivity and specificity for the exposure of interest. Although biomarkers might provide an accurate snapshot of personal biologic exposure at the time of sampling, their usefulness for epidemiologic-exposure assignment decreases with decreasing biologic half-life and increasing temporal variation in exposure (Handy et al. 2003; Bartell et al. 2004). Exposure biomarkers are also potentially susceptible to reverse causation—whereby an adverse health outcome or a risk factor for it induces a change in the exposure biomarker rather than reverse— and confounding by unmeasured physiologic differences.

Evaluation of Risk of Bias and Study Quality in Animal Toxicologic Research

Animal toxicology studies encompass a variety of experimental designs and end points. The studies can include short- to long-term bioassays that evaluate a wide array of clinical outcomes and more focused studies that evaluate one or more end points of interest. Results from animal toxicology studies are a critical stream of evidence and often the only data available for an IRIS assessment. Despite their importance, how to use animal studies in risk assessments and in regulatory decision-making is a subject of continuing debate (Guzelian et al. 2005; Weed 2005; ECETOC 2009; Adami et al. 2011; Woodruff and Sutton 2011). The NRC formaldehyde report concluded that EPA should develop a template for evaluation of toxicology studies in laboratory animals that would "consider the species and sex of animals studied, dosing information (dose spacing, dose duration, and route of exposure), end points considered, and the relevance of the end points to human end points of concern" (NRC 2011a, p. 159). Experience gained from randomized clinical trials in human and veterinary medicine suggests that systematic reviews that assess animal toxicology studies for quality and risk of bias would improve the quality of IRIS reviews. The committee notes that efforts have been made over the last decade to apply principles of systematic review, such as those made by the Evidence-based Toxicology Collaboration (EBTC 2012).

When the body of evidence is considered collectively, an evaluation of risk of bias in data derived from toxicologic studies can help to determine sources of inconsistency. Broad sources of bias considered earlier for human studies—including selection, performance, attrition, detection, and reporting bias—apply equally well to animal studies. A priori determination of suitable risk-of-bias questions tailored for animal studies and a variety of acceptable responses can help to guide the review process.

Because empirical studies of risk of bias in animal toxicology studies are few, the committee's recommendations are based primarily on information derived from other types of research. On the basis of experience with risk of bias in controlled human clinical trials (Schulz et al. 1995;

Schulz and Grimes 2002a,b), one can anticipate that estimates of treatment effect might be influenced in animal studies that lack randomization, blinding, and allocation concealment and that such animal studies would therefore have a high risk of bias. Reviews of human clinical studies have shown that study funding sources and financial ties of investigators are associated with research outcomes that are favorable for the sponsors (Lundh et al. 2012). Favorable research outcomes were defined as increased effect sizes in drug-efficacy studies and decreased effect sizes in studies of drug harm. One study (Krauth et al. 2014) has demonstrated funding bias in preclinical studies of statins. Although selective reporting of outcomes is considered an important source of bias in clinical studies (Rising et al. 2008; Hart et al. 2012) and one study (Tsilidis et al. 2013) suggests that there is selective outcome reporting in animal studies of neurologic disease, further research is needed to determine the importance of biases—for example, related to funding and selective reporting—in the animal-toxicology literature.

Some empirical evidence shows that several risk-of-bias criteria are applicable to animal studies. For example, lack of randomization, blinding, specification of inclusion and exclusion criteria, statistical power, and use of clinically relevant animals has been shown to be associated with a risk of inflated estimates of the effects of an intervention (Bebarta et al. 2003; Crossley et al. 2008; Minnerup et al. 2010; Sena et al. 2010; Vesterinen et al. 2011). Because a detailed discussion of all possible risk-of-bias criteria was beyond the scope of the present report, the committee focused its attention on sources of bias in animal studies that could affect the scientific rigor of a systematic review that uses animal toxicology studies as an evidence stream. The committee briefly discusses the importance of randomization, allocation concealment, and blinding. Although GLPs contain those practices, many high-quality studies are conducted outside the regulatory framework and hence not directly subject to GLPs. Moreover, a recent systematic review of tools for assessing animal toxicology studies identified 30 distinct tools for which the number of assessed criteria ranged from two to 25 (Krauth et al. 2013). The most common criteria were randomization (25 of 30, 83%) and investigator blinding (23 of 30, 77%).

Regarding randomization, several approaches are used in toxicology to assign animals to treatment groups randomly, including computer-based algorithms, random-number tables, and card assignment (Martin et al. 1986). Conversely, having study staff select animals from their cages "at random" poses a risk of conscious or unconscious manipulation and does not lead to true randomization (van der Worp et al. 2010). Proper randomization includes not only generating a truly random-number sequence but ensuring that the person who is allocating animals is unaware of whether the next animal will be assigned to the control group or the treatment group. If allocation is not concealed, animals can be differentially assigned to control and treatment groups on the basis of characteristics other than their random-number assignment. Allocation concealment therefore helps to prevent selection bias, and it has been shown empirically that lack of randomization or of allocation concealment in animal studies biases research outcomes by altering effect sizes (Bebarta et al. 2003; Sena et al. 2007; Macleod et al. 2008; Vesterinen et al. 2011). For example, Macleod et al. (2008) published a systematic review of the effects of the free-radical scavenger disufenton sodium in animal experiments on focal cerebral ischemia. Studies that included methods for randomization and allocation concealment and studies that used a more clinically relevant animal model (spontaneously hypertensive rats) reported lower efficacy than other studies. However, a similar meta-analysis of experimental studies of stroke performed by Crossley et al. (2008) did not find that lack of randomization was associated with a bias although lack of concealment was so associated.

As noted earlier, blinding or masking is a set of procedures that keeps the people who perform an experiment, collect the data, or assess the outcome measures unaware of the treatment allocation. If treatment allocation is known, that knowledge could affect decisions regarding the supply of additional care and the withdrawal of animals from an experiment and affect how outcomes are assessed. In blinded studies, those involved in the study will not be influenced by knowing treatment allocation, so performance, detection, and attrition bias will be prevented (Sargeant et al. 2010). Thus, blinding is used to prevent bias toward fulfilling the expectations of

the investigator on the basis of knowledge of the treatment (Kaptchuk 2003). There is substantial evidence that lack of blinding in a variety of types of animal studies is associated with exaggerated effect sizes (Bebarta et al. 2003; Sena et al. 2007; Vesterinen et al. 2011). Crossley et al. (2008) found that studies that did not use blinded investigators and studies that used healthy animals instead of animals that had relevant clinical comorbidities reported greater effect sizes. The review of animal drug interventions by Bebarta et al. (2003) showed that studies that lacked blinding or randomization in their design were more likely to find positive outcomes than studies that included blinding and randomization.

In contrast with allocation concealment, blinding might not always be possible throughout a toxicology experiment, for example, when the agent being tested imparts a visible change in the outward appearance of an animal. However, blinding of outcome assessment is almost always possible, and there are many ways to achieve it, such as providing study personnel with coded animal numbers (which blinds the study personnel to treatment assignment). Coded data can be analyzed by a statistician who is independent of the rest of the research team. The blinding of study personnel to treatment groups is not without controversy. For example, several groups and individuals have recommended blinded evaluation of histopathology slides in animal toxicology studies (EFSA 2011; Holland and Holland 2011) whereas others have argued against this approach (Neef et al. 2012). The committee notes that the terms *single-blinded*, *double-blinded*, and *triple-blinded* are often used to describe blinding in human clinical trials, but such terms can be ambiguous (Devereaux et al. 2001), and, unlike human study subjects, animals cannot be blinded to treatment assignment. Therefore, it is preferable to state which members of the study team were blinded and how. In general, it is insufficient to state that staff members were blinded to treatment groups. The method of blinding should be described in publications or in an accessible protocol to allow readers to assess the validity of the blinding procedures.

Another type of bias can occur in animal studies and can be reduced with proper procedures. *Exclusion bias* refers to the systematic difference between treatment and control groups in the number of animals that were included in and completed the study. Data on whether all animals in a study are accounted for and use of intention-to-treat analysis (analyzing animals in the groups to which they were assigned) can reduce exclusion bias (Marshall et al. 2005).

Several studies have shown that assessing risk of bias in animal studies is challenging because of inconsistent standards for reporting procedures in preclinical animal experiments (Lamontagne et al. 2010; Faggion et al. 2011). Similarly, several systematic reviews related to disease in animals have noted "a lack of reporting of group-allocation methods, blinding, and details related to intervention protocols, outcome assessments, and statistical analysis methods in some published veterinary clinical trials" (O'Connor et al. 2006; Wellman and O'Connor 2007; Burns and O'Connor 2008; Sargeant et al. 2010, p. 580). Those deficiencies have led to the development of reporting guidelines for randomized controlled trials in livestock and food safety (Sargeant et al. 2010). The reporting guidelines have been endorsed by several veterinary journals, and their adoption is expected to improve the quality of reporting of livestock-based randomized clinical trials (Sargeant et al. 2010). Because of the similarities between many animal toxicology studies and livestock clinical trials, the committee also considered those guidelines during its deliberations. Common deficiencies noted in many animal studies included lack of sample-size calculations, sufficient sample sizes, appropriate animal models, randomized treatment assignment, conflict-of-interest statements , and blinded procedures for drug administration, induction of injury, and outcome assessment (Knight 2008; Sargeant et al. 2010).

Because of the importance of many of those factors for study quality, they are included in GLPs that apply to animal studies as required by EPA or the US Food and Drug Administration (FDA) for product registration. The GLPs form a framework for study design, conduct, and oversight that reduces the risk of bias that can be associated with the adequacy of temperature, humidity, and other environmental conditions; experimental equipment and facilities; animal care; health status of animals; animal identification; separation from other test systems; and presence of contaminants in feed, soil, water, or bedding. The GLP regulations also require that "the test,

control, and reference substances be analyzed for identity, strength, purity, and composition, as appropriate for the type of study" (EPA 1999) and that the solubility and stability of the substances be determined. The GLP regulations also consider the need for blinding and randomization of the animal test system. In some tools used to assess study quality (Klimisch et al. 1997), tests conducted and reported according to accepted test guidelines (of the European Union, EPA, FDA, the Organisation for Economic Co-operation and Development) and in compliance with GLP principles have the highest grade of reliability.

Other considerations of quality of animal toxicology studies, such as the choice of the species for study, are important. Commonly used mammalian orders include rodents (such as mice, rats, hamsters, and guinea pigs), carnivores (such as dogs), lagomorphs (such as rabbits), primates (such as monkeys), and artiodactyla (such as pigs and sheep). Other vertebrates—including birds, amphibians, and fish—are also widely used in toxicologic research. Invertebrate models, including such flies as *Drosophila melanogaster* and such nematodes as *Caenorhabditis elegans*, are increasingly used, especially for mechanistic toxicology studies. Not all animal models are relevant for assessing human health effects, so understanding of the toxicologic end point of interest, concordance between animal and human responses, pharmacokinetic data, and mechanistic information is often key to evaluating the importance of results of an animal study for human health. Animal models that are irrelevant to the end point being studied are often excluded a priori from the review.

One of the most important elements in the design and interpretation of animal toxicology studies is selection of the appropriate doses for investigation (Rhomberg et al. 2007). Details of the test chemical and its administration are also quality concerns. For toxicologic treatments, the investigators should at least provide the chemical name, the concentration, and the delivery matrix and describe how the doses were selected and administered. Additional information, such as particle size and distribution, is often needed for inhalation studies. High-quality studies should include multiple dose groups, including a high-dose group that increases a study's statistical power to detect effects that might be rare and other groups to allow characterization of the dose-response relationship for observed adverse effects (Rhomberg et al. 2007).

The committee recognizes that many toxicologists are unfamiliar with the concept of risk of bias and that underreporting of its determinants probably affects the toxicology literature. Recent calls for reporting criteria for animal studies (NC3Rs 2010; NRC 2011b; Landis et al. 2012) recognize the need for improved reporting of animal research. Reporting of animal research is likely to improve if risk-of-bias assessments become more common, particularly after use in IRIS assessments.

Evaluation of Risk of Bias and Study Quality in Mechanistic Research

As described in Chapter 3, EPA often relies on a third evidence stream in reaching conclusions about the risk associated with a chemical hazard. In the broadest terms, mechanistic studies can include animal or human pharmacokinetic studies, nonanimal alternatives to hazard identification, and cell-based in vitro assays that evaluate responses of interest. Mechanistic studies can also include high-throughput assays that exploit technologic advances in molecular biology and bioinformatics. Those studies often evaluate cellular pathways that are thought to lead to such adverse health effects as carcinogenicity, genotoxicity, reproductive and developmental toxicity, neurotoxicity, and immunotoxicity in humans. Data are increasingly generated from mechanistic toxicology studies, so it is important for EPA to develop a framework that will facilitate regulatory acceptance of the results of such studies.

As in other experiments, risk of bias should be considered in evaluating mechanistic toxicology data. One challenge that EPA faces is the lack of critically evaluated risk-of-bias assessment tools. Empirically, many of the same risk-of-bias considerations that apply to bioassay studies could apply equally well to animal-based pharmacokinetic studies. Pharmacokinetic data can be used to extrapolate dose among species and treatment groups and to quantify interindivid-

ual variability. They are also used to assess the quality of individual studies, consistency among outcomes, and differences among population groups, such as susceptible populations. Like bioassays, pharmacokinetic studies need to specify test-chemical stability, chemical exposure route, and exposure-measurement methods. A recent review of mechanistic studies used a systematic approach to identify studies, extract information, and summarize study findings but fell short of assessing risk of bias in the studies (Kushman et al. 2013).

The committee encourages EPA to advance its methods for using in vitro studies in hazard assessment. Several criteria should be considered in assessing in vitro toxicology studies for risk of bias and toxicologic relevance. Relevance should be determined in several domains, including cell systems used, exposure concentrations, metabolic capacity, and the relationship between a measured in vitro response and a clinically relevant outcome measure. Few tools are available for assessing risk of bias in in vitro studies. Because of the nascent status of this field, the committee can provide only provisional recommendations for EPA to consider.

FDA's Good Manufacturing Practice (GMP) regulations describe how to ensure performance and consistency of in vitro methods when they are approved for commercial use by FDA, including criteria for setting performance standards for each assay that falls under FDA authority as a device or kit for medical purposes.[5] The provisions in the regulations can also be applied to many in vitro tests that are used in toxicology (Gupta et al. 2005). Likewise, GLP principles discussed earlier could be applied to in vitro tests (OECD 2004; Gupta et al. 2005). Useful guidelines, such as Good Cell Culture Practice (Bal-Price and Coecke 2011), have also been developed to define minimum standards in cell and tissue culture.

Another approach that could be used by EPA is to consider criteria developed by the Interagency Coordinating Committee on the Validation of Alternative Methods (ICCVAM) to validate in vitro and other alternative (nonanimal) toxicity assays. Test methods validated by ICCVAM for regulatory risk-assessment purposes generally include (a) a scientific and regulatory rationale for the test method, (b) an understanding of the relationship of the test method's end points to the biologic effect of interest, (c) a description of what is measured and how, (d) test performance criteria (for example, the use of positive and negative controls), (e) a description of data analysis, (f) a list of the species to which the test results can be applied, (g) and a description of test limitations, including classes of materials that the test cannot accurately assess. Joint efforts between ICCVAM and the European Centre for the Validation of Alternative Methods (ECVAM) are evaluating criteria for the validation of toxicogenomics-based test systems (Corvi et al. 2006). Potential sources of bias that have been empirically identified during the joint ICCVAM-ECVAM effort include data quality, cross-platform and interlaboratory variability, lack of dose- and time-dependent measurements that examine the range of biologic variability of gene responses for a given test system, unknown concordance with known toxicologic outcomes (phenotypic anchoring), and poor microarray and instrumentation quality. The following items were deemed necessary for the use of microarray-based toxicogenomics in regulatory assessments: (a) microarrays should be made according to GMP principles; (b) "specifications and performance criteria for all instrumentation and method components should be available"; (c) "all quality-assurance and quality-control…procedures should be transparent, consistent, comparable, and reported"; (d) "arrays should have undergone sequence verification, and the sequences should be publicly available"; and (e) all data should be exportable in a microarray and gene expression compatible format (Corvi et al. 2006, p. 423). The extent of bias associated with each of those items is not known.

The test chemicals used in in vitro systems should also include known positive and known negative agents. Whenever possible, chemicals or test agents should be coded to reduce bias. Other problems that might be associated with study quality and risk of bias involve the character-

[5]See 21 CFR 210 and 211 [2003] Good Manufacturing Practices and 21 CFR 800 [2003] Good Manufacturing Practices Device.

ization of the source, the growing conditions of the cultured cell, and the genetic materials used in microarrays; documentation of cell-culture practices, such as incubator temperatures; and definition of reagents and equipment. The limitations of the test systems should be understood and described. Important limitations that could affect how study results are interpreted include the inability to replicate the metabolic processes relevant to chemical toxicity that occur in vivo in an in vitro test system.

Faggion (2012) recently assessed existing guidelines for reporting in vitro studies in dentistry and presented a method for reporting risk of bias that was based on the Consolidated Standards of Reporting Trials (CONSORT) checklist for reporting randomized clinical trials. Important elements in the checklist included (a) the scientific background and rationale for the study; (b) the interventions used for each group, including how and when they were administered, with sufficient detail to enable replication; (c) outcome measures, including how and when they were assessed; (d) sample-size calculations; (e) how the random-allocation sequence was generated and implemented; (g) blinding of investigators responsible for treatment and others responsible for outcome assessment; (h) statistical methods; and (i) trial limitations, such as sources of potential bias or imprecision.

Reporting inadequacies similar to those mentioned earlier have been noted in the literature describing results of in vitro studies. Watters and Goodman (1999) compared the rigor with which tissue-culture and cell-culture in vitro studies were reported and randomly selected clinical studies published in the same anesthesia journals. They focused on basic aspects of study design and reporting that might lead to bias, including sample size, randomization, and reporting of exclusions and withdrawals. They found that reporting of those aspects in the in vitro studies occurred at much lower rate than that of clinical studies and thus hindered interpretation of reported tissue-culture and cell-culture studies.

EVALUATING EVIDENCE FROM INDIVIDUAL STUDIES

Assessment of the risk of bias and other methodologic characteristics of relevant studies is a critical part of the systematic-review process. The committee acknowledges that these assessments will take time and effort; the resources required will depend on the complexity and amount of data available on a given chemical. However, the use of standard risk-of-bias criteria by trained coders can be efficient. Koustas et al. (in press) and Johnson et al. (in press) found that risk-of-bias assessment, data analysis, and evaluation of quality and strength of evidence took about 2-3 additional months.

The risk-of-bias assessment can be used to exclude studies from a systematic review or can be incorporated qualitatively or quantitatively into the review results. The plan for incorporating the risk-of-bias assessment into a systematic review should be specified a priori in the review protocol. Various options are discussed in the sections that follow.

Exclusion of Studies

Individual studies are most commonly excluded from a systematic review if they do not meet the inclusion criteria for the review with respect to the population examined, the exposures measured, or the outcomes assessed. Less often, individual studies are excluded because they did not meet a particular threshold of risk of bias or other methodologic criteria. Some studies that entail a substantial risk of bias or that have severe methodologic shortcomings ("fatal flaws") could be excluded from consideration. Examples of such exclusion criteria include instability of the test compound, inappropriate animal models, inadequate or no controls (or comparison group), or invalid measures of exposure or outcome. EPA needs a transparent process and clear criteria for excluding certain studies that have important deficiencies. If studies are excluded because of risk of bias, the reasons for the exclusion should be described.

Incorporating the Assessment of a Risk of Bias in the Review

There are two main options for presenting an assessment of the risk of bias and methodologic characteristics of studies that will be included in the IRIS assessment. First, if a meta-analysis has been conducted, a sensitivity analysis can be conducted to determine the effects of risk of bias on the meta-analytic result. For example, a sensitivity analysis could be conducted to determine whether a meta-analysis result is affected by excluding studies that have a high risk of bias. All studies would be included in the meta-analysis, and then the meta-analysis would be recalculated by including only studies that meet a threshold for low risk of bias. If the summary estimate does not change, the estimate is not sensitive to the risk of bias of the studies. If the included studies are very heterogeneous or if all included studies have a high risk of bias, it might be inappropriate to calculate a quantitative summary statistic.

The second option is to describe the results of individual study evaluations qualitatively. Qualitative presentation of individual study evaluations can include narrative descriptions or graphic presentations. Tables summarizing the evidence could include a column that summarizes the risk-of-bias assessment in text and provides a ranking (for example, one, two, or three stars). Graphic displays could show the evaluation of each individual item assessed for each study. Two displays from a Cochrane review are shown in Figures 5-2 and 5-3. As described in later chapters of the present report, the risk-of-bias assessment can be included in the process for selecting studies for calculating toxicity values or in the uncertainty analysis.

FINDINGS AND RECOMMENDATIONS

Finding: The checklist developed by EPA that is presented in the preamble and detailed in the draft handbook addresses many of the concerns raised by the NRC formaldehyde report. EPA has also developed broad guidance for the assessment of the quality of observational studies of exposed human populations and, to a smaller extent, animal toxicology studies. It has not developed criteria for the evaluation of mechanistic toxicology studies. Still lacking is a clear picture of the assessment tools that EPA will develop to assess risk of bias and of how existing assessment tools will be adapted.

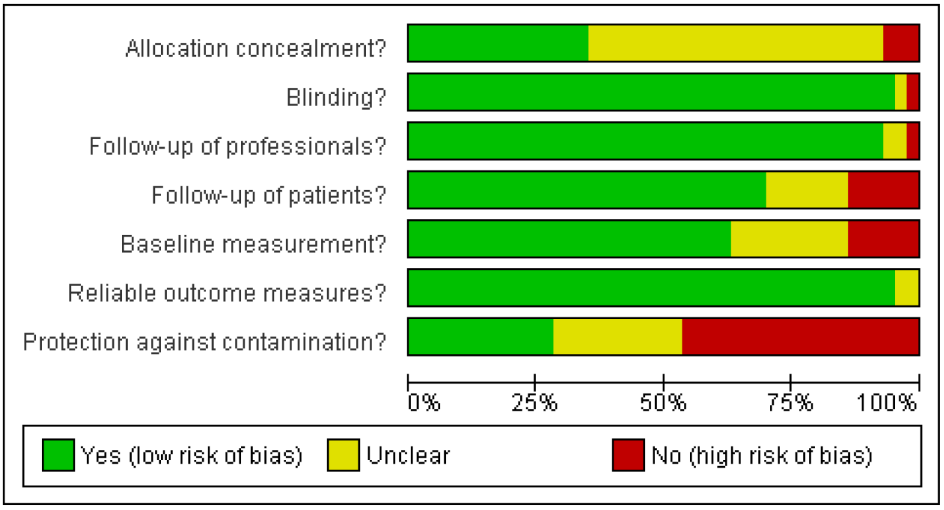


FIGURE 5-2 Sample graphic display of risk-of-bias evaluations. Source: Nkansah et al. 2010. Reprinted with permission; copyright 2010, *Cochrane Database of Systematic Reviews*.

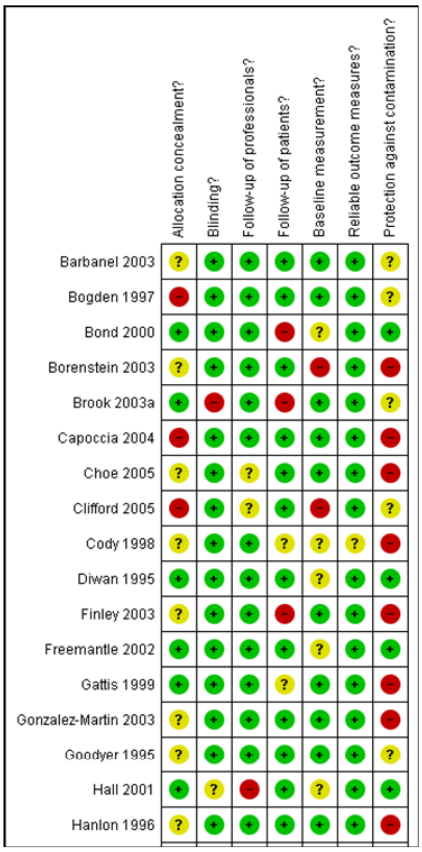


FIGURE 5-3 Truncated graph of the risk-of-bias summary that shows review authors' judgments about each risk-of-bias item for each included study. Green: yes (low risk of bias); yellow: unclear; red: no (high risk of bias). Source: Nkansah et al. 2010. Reprinted with permission; copyright 2010, *Cochrane Database of Systematic Reviews*.

Recommendation: To advance the development of tools for assessing risk of bias in different types of studies (human, animal, and mechanistic) used in IRIS assessments, EPA should explicitly identify factors, in addition to those discussed in this chapter, that can lead to bias in animal studies—such as control for litter effects, dosing, and methods for exposure assessment—so that these factors are consistently evaluated for experimental studies. Likewise, EPA should consider a tool for assessing risk of bias in in vitro studies.

Finding: The development of standards for evaluating individual studies for risk of bias is most advanced in human clinical research. Even in that setting, the evidence base to support the standards is modest, and expert guidance varies. Furthermore, many of the individual criteria included in risk-of-bias assessment tools, particularly for animal studies and epidemiologic studies, have not been empirically tested to determine how the various sources of bias influence the results of individual studies. The validity and reliability of the tools have also not been tested.

Finding: Thus, the committee acknowledges that incorporating risk-of-bias assessments into the IRIS process might take additional time; the ability to do so will vary with the complexity and extent of data on each chemical and with the resources available to EPA. However, the use of standard risk-of-bias criteria by trained coders has been shown to be efficient.

Recommendation: When considering any method for evaluating individual studies, EPA should select a method that is transparent, reproducible, and scientifically defensible. Whenever possible, there should be empirical evidence that the methodologic characteristics that are being assessed in the IRIS protocol have systematic effects on the direction or magnitude of the outcome. The methodologic characteristics that are known to be associated with a risk of bias should be included in the assessment tool. Additional quality-assessment items relevant to a particular systematic-review question could also be included in the EPA assessment tool.

Recommendation: EPA should carry out, support, or encourage research on the development and evaluation of empirically based instruments for assessing bias in human, animal, and mechanistic studies relevant to chemical-hazard identification. Specifically, there is a need to test existing animal-research assessment tools on other animal models of chemical exposures to ensure their relevance and generalizability to chemical-hazard identification. Furthermore, EPA might consider pooling data collected for IRIS assessment to determine whether, among various contexts, candidate risk-of-bias items are associated with overestimates or underestimates of effect.

Recommendation: Although additional methodologic work might be needed to establish empirically supported criteria for animal or mechanistic studies, an IRIS assessment needs to include a transparent evaluation of the risk of bias of studies used by EPA as a primary source of data for the hazard assessment. EPA should specify the empirically based criteria it will use to assess risk of bias for each type of study design in each type of data stream.

Recommendation: To maintain transparency, EPA should publish its risk-of-bias assessments as part of its IRIS assessments. It could add tables that describe the assessment of each risk-of-bias criterion for each study and provide a summary of the extent of the risk of bias in the descriptions of each study in the evidence tables.

Finding: The nomenclature of the various factors that are considered in evaluating risk of bias is variable and not well standardized among the scientific fields relevant to IRIS assessments. Such terminology has not been standardized for IRIS assessments.

Recommendation: EPA should develop terminology for potential sources of bias with definitions that can be applied during systematic reviews.

Finding: Although reviews of human clinical studies have shown that study funding sources and financial ties of investigators are associated with research outcomes that are favorable for the sponsors, less is known about the extent of funding bias in animal research.

Recommendation: Funding sources should be considered in the risk-of-bias assessment conducted for systematic reviews that are part of an IRIS assessment.

Finding: An important weakness of all existing tools for assessing methodologic characteristics of published research is that assessment requires full reporting of the research methods. EPA might be hampered by differences in traditions of reporting risk of bias among fields in the scientific literature.

Recommendation: EPA should contact investigators to obtain missing information that is needed for the evaluation of risk of bias and other quality characteristics of included studies. The committee expects that, as happened in the clinical literature in which additional reporting standards for journals were implemented (Turner et al. 2012), the reporting of toxicologic research will eventually improve as risk-of-bias assessments are incorporated into the IRIS program.

However, a coordinated approach by government agencies, researchers, publishers, and professional societies will be needed to improve the completeness and accuracy of the reporting of toxicology studies in the near future.

Finding: EPA has not developed procedures that describe how the evidence evaluation for individual studies will be incorporated, either qualitatively or quantitatively, into an overall assessment.

Recommendation: The risk-of-bias assessment of individual studies should be carried forward and incorporated into the evaluation of evidence among data streams.

REFERENCES

Adami, H.O., S.C. Berry, C.B. Breckenridge, L.L. Smith, J.A. Swenberg, D. Trichopoulos, N.S. Weiss, and T.P. Pastoor. 2011. Toxicology and epidemiology: Improving the science with a framework for combining toxicological and epidemiological evidence to establish causal inference. Toxicol. Sci. 122(2):223-234.

Armstrong, B.G. 1998. Effect of measurement error on epidemiological studies of environmental and occupational exposures. Occup. Environ. Med. 55(10):651–656.

Baker, D., and M.J. Nieuwenhuijsen. 2008. Environmental Epidemiology. Study Methods and Application. New York: Oxford University Press.

Bal-Price, A., and S. Coecke. 2011. Guidance on Good Cell Culture Practice (GCCP). Pp. 1-25 in Cell Culture Techniques, M. Aschner, C. Sunol, and A. Bal-Price, eds. Springer Protocols, Neuromethods Vol. 56. New York: Humana Press.

Bartell, S.M., W.C. Griffith, and E.M. Faustman. 2004. Temporal error in biomarker based mean exposure estimates for individuals. J. Expo. Anal. Environ. Epidemiol. 14(2):173-179.

Bebarta, V., D. Luyten, and K. Heard. 2003. Emergency medicine animal research: Does use of randomization and blinding affect the results? Acad. Emerg. Med. 10(6):684-687.

Burns, M.J., and A.M. O'Connor. 2008. Assessment of methodological quality and sources of variation in the magnitude of vaccine efficacy: A systematic review of studies from 1960 to 2005 reporting immunization with Moraxella bovis vaccines in young cattle. Vaccine 26(2):144-152.

Corvi, R., H.J. Ahr, S. Albertini, D.H. Blakey, L. Clerici, S. Coecke, G.R. Douglas, L. Gribaldo, J.P. Groten, B. Haase, K. Hamernik, T. Hartung, T. Inoue, I. Indans, D. Maurici, G. Orphanides, D. Rembges, S.A. Sansone, J.R. Snape, E. Toda, W. Tong, J.H. van Delft, B. Weis, and L.M. Schechtman. 2006. Meeting report: Validation of toxicogenomics-based test systems: ECVAM-ICCVAM/NICEATM considerations for regulatory use. Environ. Health Perspect. 114(3):420-429.

Crossley, N.A., E. Sena, J. Goehler, J. Horn, B. van der Worp, P.M. Bath, M. Macleod, and U. Dirnagl. 2008. Empirical evidence of bias in the design of experimental stroke studies: A metaepidemiologic approach. Stroke 39(3):929-934.

Devereaux, P.J., B.J. Manns, W.A. Ghali, H. Quan, C. Lacchetti, V.M. Montori, M. Bhandari, and G.H. Guyatt. 2001. Physician interpretations and textbook definitions of blinding terminology in randomized controlled trials. JAMA 285(15):2000-2003.

EBTC (Evidence–based Toxicology Collaboration). 2012. Methods Work Group [online]. Available: http://www.ebtox.com/methods-work-group/ [accessed Feb. 14, 2014].

ECETOC (European Centre for Ecotoxicology and Toxicology of Chemicals). 2009. Framework for the Integration of Human and Animal Data in Chemical Risk Assessment. Technical Report No. 104. ECETOC, Brussels, Belgium [online]. Available: http://www.ecetoc.org/uploads/Publications/documents/TR%20104.pdf [accessed December 6, 2013].

EFSA (European Food Safety Authority). 2011. Scientific Opinion EFSA Guidance on Repeated-Dose 90-Day Oral Toxicity Study on Whole Food/Feed in Rodents. Draft for public consultation. Scientific Committee, EFSA, Parma, Italy [online]. Available: http://www.efsa.europa.eu/en/consultations/call/110707.pdf [accessed December 6, 2013].

EPA (U.S. Environmental Protection Agency). 2013a. Part 1. Status of Implementation of Recommendations. Materials Submitted to the National Research Council, by Integrated Risk Information System Program, U.S. Environmental Protection Agency, January 30, 2013 [online]. Available: http://www.

epa.gov/iris/pdfs/IRIS%20Program%20Materials%20to%20NRC_Part%201.pdf [accessed Nov. 13, 2013].

EPA (U.S. Environmental Protection Agency). 2013b. Part 2. Chemical-Specific Examples. Materials Submitted to the National Research Council, by Integrated Risk Information System Program, U.S. Environmental Protection Agency, January 30, 2013 [online]. Available: http://www.epa.gov/iris/pdfs/IRIS%20Program%20Materials%20to%20NRC_Part%202.pdf [accessed October 22, 2013].

EPA (U.S. Environmental Protection Agency). 2013c. Toxicological Review of Benzo[a]pyrene (CAS No. 50-32-8) in Support of Summary Information on the Integrated Risk Information System (IRIS), Public Comment Draft. EPA/635/R13/138a. National Center for Environmental Assessment, Office of Research and Development, U.S. Environmental Protection Agency, Washington, DC. August 2013 [online]. Available: http://cfpub.epa.gov/ncea/iris_drafts/recordisplay.cfm?deid=66193 [accessed Nov. 13, 2013].

EPA (U.S. Environmental Protection Agency). 1999. Audit for Determining Compliance of Studies with GLP Standards Requirements. Standard Operating Procedure SOP No. GLP-C-02 [online]. Available: http://www.epa.gov/compliance/resources/policies/monitoring/fifra/sop/glp-c-02.pdf [accessed December 11, 2013].

Faggion, C.M., Jr. 2012. Guidelines for reporting pre-clinical in vitro studies on dental materials. J. Evid. Based Dent. Pract. 12(4):182-189.

Faggion, C.M., Jr., N.N. Giannakopoulos, and S. Listl. 2011. Risk of bias of animal studies on regenerative procedures for periodontal and peri-implant bone defects - a systematic review. J. Clin. Periodontol. 38(12):1154-1160.

Finley, P.R., H.R. Rens, J.T. Pont, S.L. Gess, C. Louie, S.A. Bull, J.Y. Lee, and L.A. Bero. 2003. Impact of a collaborative care model on depression in a primary care setting: A randomized controlled trial. Pharmacotherapy 23(9):1175-1185.

Greenland, S. 1989. Modeling and variable selection in epidemiologic analysis. Am. J. Public Health 79(3):340-349.

Greenland, S. 2000. An introduction to instrumental variables for epidemiologists. Int. J. Epidemiol. 29(4):722-729.

Greenland, S., J. Pearl, and J.M. Robins. 1999. Causal diagrams for epidemiologic research. Epidemiology 10(1):37-48.

Gupta, K., A. Rispin, K. Stitzel, S. Coecke, and J. Harbell. 2005. Ensuring quality of in vitro alternative test methods: Issues and answers. Regul. Toxicol. Pharmacol. 43(3):219-224.

Guzelian, P.S., M.S. Victoroff, N.C. Halmes, R.C. James, and C.P. Guzelian. 2005. Evidence-based toxicology: A comprehensive framework for causation. Hum. Exp. Toxicol. 24(4):161-201.Hall, L., M. Eccles, R. Barton, N. Steen, and M. Campbell. 2001. Is untargeted outreach visiting in primary care effective? A pragmatic randomized controlled trial. J. Public Health Med. 23(2):109-113.

Handy, R.D., T.S. Galloway, and M.H. Depledge. 2003. A proposal for the use of biomarkers for the assessment of chronic pollution and in regulatory toxicology. Ecotoxicology 12(1-4):331-343.

Hart, B., A. Lundh, and L. Bero. 2012. Effect of reporting bias on meta-analyses of drug trials: Reanalysis of meta-analyses. BMJ 344:d7202.

Hernán, M.A., and S.R. Cole. 2009. Causal diagrams and measurement bias. *Am. J. Epidemiol.* 170(8):959-962.

Hernán, M.A., and J.M. Robins. 2008. Observational studies analyzed like randomized experiments: Best of both worlds. *Epidemiology* 19(6):789-792.

Hernán, M.A., and J.M. Robins. 2013. *Causal Inference.* New York: Chapman & Hall/CRC [online]. Available: http://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/ [accessed December 6, 2013].

Hernán, M.A., M. McAdams, N. McGrath, E. Lanoy, and D. Costagliola. 2009. Observation plans in longitudinal studies with time-varying treatments. *Stat. Methods Med. Res.* 18(1):27-52.

Higgins, J.P.T., and S. Green, eds. 2008. Cochrane Handbook for Systematic Reviews of Interventions. Chichester, UK: John Wiley & Sons.

Higgins, J.P.T., and S. Green, eds. 2011. Cochrane Handbook for Systematic Reviews of Interventions, Version 5.1.0. The Cochrane Collaboration [online]. Available: http://handbook.cochrane.org/ [accessed December 11, 2013].

Holland, T., and C. Holland. 2011. Analysis of unbiased histopathology data from rodent toxicology studies or, are these groups different enough to ascribe it to treatment? Toxicol. Pathol. 39(4):569-575.

IOM (Institute of Medicine). 2011. Finding What Works in Health Care: Standards for Systematic Reviews. Washington, DC: National Academies Press.

Johnson, P.I., P. Sutton, D.S. Atchley, E. Koustas, J. Lam, S. Sen, K.A. Robinson, D.A. Axelrad, and T.J. Woodruff. In press. The Navigation Guide - Evidence-based medicine meets environmental health: Systematic review of human evidence for PFOA effects on fetal growth. Environmental Health Perspectives.

Juni, P., A. Witschi, R. Bloch, and M. Egger. 1999. The hazards of scoring the quality of clinical trials for meta-analysis. JAMA 282(11):1054-1060.

Kaptchuk, T.J. 2003. Effect of interpretive bias on research evidence. BMJ 326(7404):1453-1455.

Klimisch, H.J., M. Andreae, and U. Tillmann. 1997. A systematic approach for evaluating the quality of experimental toxicological and ecotoxicological data. Regul. Toxicol. Pharmacol. 25(1):1-5.

Knight, A. 2008. Systematic reviews of animal experiments demonstrate poor contributions toward human healthcare. Rev. Recent Clin. Trials 3(2):89-96.

Koustas, E., J. Lam, P. Sutton, P.I. Johnson, D.S. Atchley, S. Sen, K.A. Robinson, D.A. Axelrad, and T.J. Woodruff. In press. The Navigation Guide - Evidence-based medicine meets environmental health: Systematic review of non-human evidence for PFOA effects on fetal growth. Environmental Health Perspectives.

Krauth, D., T. Woodruff, and L. Bero. 2013. A systematic review of instruments for assessing risk of bias and other methodological criteria of published animal studies: A systematic review. Environ. Health Perspect. 121(9):985-992.

Krauth, D., A. Anglemyer, R. Philipps, and L. Bero. 2014. Nonindustry-sponsored preclinical study on statins yield greater efficacy estimates than industry-sponsored studies: A meta-analysis. PLOS Biol. 12(1):e1001770.

Kushman, M.E., A.D. Kraft, K.Z. Guyton, W.A. Chiu, S.L. Makris, and I. Rusyn. 2013. A systematic approach for identifying and presenting mechanistic evidence in human health assessments. Regul. Toxicol. Pharmacol. 67(2):266-277.

Lamontagne, F., M. Briel, M. Duffett, A. Fox-Robichaud, D.J. Cook, G. Guyatt, O. Lesur, and M.O. Meade. 2010. Systematic review of reviews including animal studies addressing therapeutic interventions for sepsis. Crit. Care Med. 38(12):2401-2408.

Landis, S.C., S.G. Amara, K. Asadullah, C.P. Austin, R. Blumenstein, E.W. Bradley, R.G. Crystal, R.B. Darnell, R.J. Ferrante, H. Fillit, R. Finkelstein, M. Fisher, H.E. Gendelman, R.M. Golub, J.L. Goudreau, R.A. Gross, A.K. Gubitz, S.E. Hesterlee, D.W. Howells, J. Huguenard, K. Kelner, W. Koroshetz, D. Krainc, S.E. Lazic, M.S. Levine, M.R. Macleod, J.M. McCall, R.T. Moxley, III, K. Narasimhan, L.J. Noble, S. Perrin, J.D. Porter, O. Steward, E. Unger, U. Utz, and S.D. Silberberg. 2012. A call for transparent reporting to optimize the predictive value of preclinical research. Nature 490(7419):187-191.

Lundh, A, S. Sismondo, J. Lexchin, O.A. Busuioc, and L. Bero. 2012. Industry sponsorship and research outcome. Cochrane Database Syst. Rev. 12:Art. MR000033. doi: 10.1002/14651858.MR000033.pub2.

Macleod, M.R., H.B. van der Worp, E.S. Sena, D.W. Howells, U. Dirnagl, and G.A. Donnan. 2008. Evidence for the efficacy of NXY-059 in experimental focal cerebral ischaemia is confounded by study quality. Stroke 39(10):2824-2829.

Marshall, J.C., E. Deitch, L.L. Moldawer, S. Opal, H. Redl, and T. van der Poll. 2005. Preclinical models of shock and sepsis: What can they tell us? Shock 24(suppl. 1):1-6.

Martin, R.A., A. Daly, C.J. DiFonzo, and F.A. de la Iglesia. 1986. Randomization of animals by computer program for toxicity studies. J. Environ. Pathol. Toxicol. Oncol. 6(5-6):143-152.

Minnerup, J., H. Wersching, K. Diederich, M. Schilling, E.B. Ringelstein, J. Wellmann, and W.R. Schäbitz. 2010. Methodological quality of preclinical stroke studies is not required for publication in high-impact journals. J. Cereb. Blood Flow Metab. 30(9):1619-1624.

NC3Rs (National Centre for the Replacement, Refinement and Reduction of Animals in Research). 2010. The ARRIVE Guidelines. Animal Research: Reporting of in Vivo Experiments [online]. Available: http://www.nc3rs.org.uk/downloaddoc.asp?id=1206&page=1357&skin=0 [accessed Feb. 14, 2014].

Neef, N., K.J. Nikula, S. Francke-Carroll, and L. Boone. 2012. Regulatory forum opinion piece: Blind reading of histopathology slides in general toxicology studies. Toxicol. Pathol. 40(4):697-699.

Nieto, A., A. Mazon, R. Pamies, J.J. Linana, A. Lanuza, F.O. Jimenez, A. Medina-Hernandez, and F.J. Nieto. 2007. Adverse effects of inhaled corticosteroids in funded and nonfunded studies. Arch. Intern. Med. 167(19):2047-2053.

Nkansah, N., O. Mostovetsky, C. Yu, T. Chheng, J. Beney, C.M. Bond, and L. Bero. 2010. Effects of outpatient pharmacists' non-dispensing roles on patient outcomes and prescribing patterns. Cochrane Database Syst. Rev. 2010(7):Art. No. CD000336. doi: 10.1002/14651858.CD000336.pub2.

NRC (National Research Council). 2011a. Review of the Environmental Protection Agency's Draft IRIS Assessment of Formaldehyde. Washington, DC: National Academies Press.

NRC (National Research Council). 2011b. Guidance for the Description of Animal Research in Scientific Publications. Washington, DC: National Academies Press.

NTP (National Toxicology Program). 2013. Draft OHAT Approach for Systematic Review and Evidence Integration for Literature-Based Health Assessments – February 2013. U.S. Department of Health and Human Services, National Institute of Health, National Institute of Environmental Health Sciences, Division of the National Toxicology Program [online]. Available: http://ntp.niehs.nih.gov/NTP/OHAT/EvaluationProcess/DraftOHATApproach_February2013.pdf [accessed December 11, 2013].

O'Connor, A.M., N.G. Wellman, R.B. Evans, and D.R. Roth. 2006. A review of randomized clinical trials reporting antibiotic treatment of infectious bovine keratoconjunctivitis in cattle. Anim. Health Res. Rev. 7(1-2):119-127.

Odgaard-Jensen, J., G.E. Vist, A. Timmer, R. Kunz, E.A. Akl, H. Schuemann, M. Briel, A.J. Nordmann, S. Pregno, and A.D. Oxman. 2011. Randomization to protect against selection bias in healthcare trials. Cochrane Database Syst. Rev. 2011(4):Art. No. MR0000012. doi: 10,1002/14651858. MR000012. pub3.

OECD (Organization for Economic Cooperation and Development). 2004. Advisory Document of the Working Group on Good Laboratory Practice: The Application of the Principles of GLP to In vitro Studies. OECD Series on Principles of Good Laboratory Practice and Compliance Monitoring No. 14. ENV/JM/MONO(2004)26. OECD, Paris, France [online]. Available: http://www.bfr.bund.de/cm/349/Glp14.pdf [accessed December 6, 2013].

Rhomberg, L.R., K. Baetcke, J. Blancato, J. Bus, S. Cohen, R. Conolly, R. Dixit, J. Doe, K. Ekelman, P. Fenner-Crisp, P. Harvey, D. Hattis, A. Jacobs, D. Jacobson-Kram, T. Lewandowski, R. Liteplo, O. Pelkonen, J. Rice, D. Somers, A. Turturro, W. West, and S. Olin. 2007. Issues in the design and interpretation of chronic toxicity and carcinogenicity studies in rodents: Approaches to dose selection. Crit. Rev. Toxicol. 37(9):729-837.

Rising, K., P. Bacchetti, and L. Bero. 2008. Reporting bias in drug trials submitted to the Food and Drug Administration: Review of publication and presentation. PLoS Med. 5(11):e217.

Robins J.M. 2000. Marginal structural models versus structural nested models as tools for causal inference. Pp. 95-134 in *Statistical Models in Epidemiology: The Environment and Clinical Trials, M.E.* Halloran, and D. Berry, eds. The IMA Volume in Mathematics and its Applications No. 116. New York: Springer.

Robins, J.M., and M.A. Hernán. 2008. Estimation of the causal effects of time-varying exposures. Pp. 553-600 in *Advances in Longitudinal Data Analysis*, G. Fitzmaurice, M. Davidian, G. Verbeke, and G. Molenberghs, eds. New York: Chapman and Hall/CRC Press.

Robins, J.M., M.A. Hernán, and B. Brumback. 2000. Marginal structural models and causal inference in epidemiology. Epidemiology 11(5):550-560.

Sargeant, J.M, A.M. O'Connor, I.A. Gardner, J.S. Dickson, M.E. Torrence, I.R. Dohoo, S.L. Lefebvre, P.S. Morley, A. Ramirez, and K. Snedeker. 2010. The REFLECT statement: Reporting guidelines for randomized controlled trials in livestock and food safety: Explanation and elaboration. J. Food Prot. 73(3):579-603.

Schulz, K.F., and D.A. Grimes. 2002a. Blinding in randomised trials: Hiding who got what. Lancet 359(9307):696-700.

Schulz, K.F., and D.A. Grimes. 2002b. Allocation concealment in randomised trials: Defending against deciphering. Lancet 359(9306):614-618.

Schulz, K.F., I. Chalmers, R.J. Hayes, D.G. Altman. 1995. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. JAMA 273(5):408-412.

Sena, E., H.B. van der Worp, D. Howells, and M. Macleod. 2007. How can we improve the pre-clinical development of drugs for stroke? Trends Neurosci. 30(9):433-439.

Sena, E.S., C.L. Briscoe, D.W. Howells, G.A. Donnan, P.A. Sandercock, and M.R. Macleod. 2010. Factors affecting the apparent efficacy and safety of tissue plasminogen activator in thrombotic occlusion models of stroke: Systematic review and meta-analysis. J. Cereb. Blood Flow Metab. 30(12):1905-1913.

Turner, L., L. Shamseer, D.G. Altman, K.F. Schulz, and D. Moher. 2012. Does use of the CONSORT statement impact the completeness of reporting of randomised controlled trials published in medical journals? A Cochrane Review. Syst. Rev. 1:60. doi:10.1186/2046-4053-1-60.

Tsilidis, K.K., O.A. Panagiotou, E.S. Sena, E. Aretouli, E. Evangelou, D.W. Howells, R. Al-Shahi Salman, M.R. Macleod, and J.P. Ioannidis. 2013. Evaluation of excess significance bias in animal studies of neurological diseases. PLOS Biol. 11(7):e1001609.

van der Worp, H.B., D.W. Howells, E.S. Sena, M.J. Porritt, S. Rewell, V. O'Collins, and M.R. Macleod. 2010. Can animal models of disease reliably inform human studies? PLoS Med. 7(3):e1000245.

Vesterinen, H.M., K. Egan, A. Deister, P. Schlattmann, M.R. Macleod, and U. Dirnagl. 2011. Systematic survey of the design, statistical analysis, and reporting of studies published in the 2008 volume of the Journal of Cerebral Blood Flow and Metabolism. J. Cereb. Blood Flow Metab. 31(4):1064-1072.

Watters, M.P., and N.W. Goodman. 1999. Comparison of basic methods in clinical studies and in vitro tissue and cell culture studies reported in three anaesthesia journals. Br. J. Anaesth. 82(2):295-298.

Weed, D.L. 2005. Weight of evidence: A review of concept and methods. Risk Anal. 25(6):1545-1557.

Wellman, N.G., and A.M. O'Connor. 2007. Meta-analysis of treatment of cattle with bovine respiratory disease with tulathromycin. J. Vet. Pharmacol. Ther. 30(3):234-241.

Wells, G.A., B. Shea, D. O'Connell, J. Petersen, V. Welch, M. Losos, and P. Tugwell. 2013. The Newcastle-Ottawa Scale (NOS) for Assessing the Quality of Nonrandomized Studies in Meta-Analyses [online]. Available: http://www.ohri.ca/programs/clinical_epidemiology/oxford.htm [accessed December 9, 2013].

Woodruff, T.J., and P. Sutton. 2011. An evidence-based medicine methodology to bridge the gap between clinical and environmental health sciences. Health Aff. (Millwood) 30(5):931-937.

6

Evidence Integration for Hazard Identification

Hazard identification is a well-recognized term in the risk-assessment field and was codified in the 1983 NRC report *Risk Assessment in the Federal Government: Managing the Process* (NRC 1983). In the present report, hazard identification is understood to answer the qualitative scientific question, Does exposure to chemical X cause outcome Y in humans? Evidence integration is understood to be the process of combining different kinds of evidence relevant to hazard identification. In a typical assessment developed for the Integrated Risk Information System (IRIS), for example, evidence integration might involve observational epidemiologic studies, experimental studies of animals and possibly humans, in vitro mechanistic studies, and perhaps other mechanistic knowledge. If the answer to the qualitative question for a given outcome is affirmative, the US Environmental Protection Agency (EPA) produces a *quantitative* estimate of toxicity by using selected studies to characterize the dose-response relationship with some estimate of uncertainty to yield a reference dose (RfD), a reference concentrations (RfC), or a unit risk value for the given outcome. This chapter focuses on the qualitative question of hazard identification (that is, the hazard-identification process, see Figure 6-1). Chapter 7 considers the quantitative process that follows hazard identification.

In this chapter, the committee first discusses some concerns about current terminology, next addresses the kinds of evidence that must be combined, and then outlines some organizing principles for integrating evidence. A review of the approach that EPA has recently taken and its responsiveness to the recommendations of the NRC formaldehyde report (NRC 2011) follows. Options for integrating evidence are then discussed with a focus first on qualitative approaches and then on quantitative approaches. The final section of the chapter provides the committee's findings and recommendations, which are offered in light of consideration of how EPA might best increase transparency and implement a process that is feasible within its time and resource constraints and that is ultimately scientifically defensible.

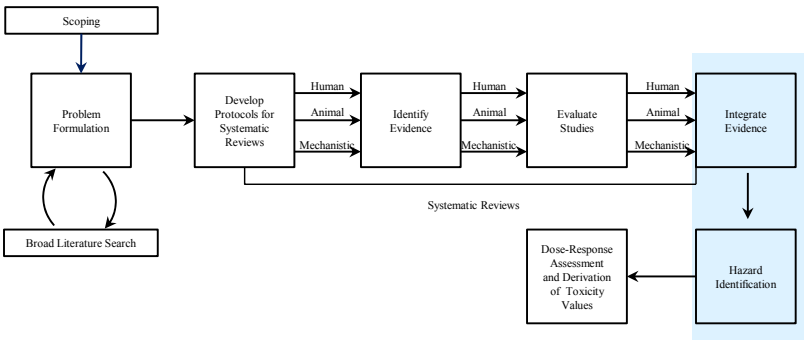


FIGURE 6-1 The IRIS process; the hazard-identification process is highlighted. The committee views public input and peer review as integral parts of the IRIS process, although they are not specifically noted in the figure.

TERMINOLOGY

An early challenge faced by this committee was to determine how the phrase *weight of evidence* (WOE) is used by EPA and others. The term is often used by EPA in the context of a WOE "narrative." In the case of a carcinogenic risk assessment, the narrative consists of a short summary that "explains what is known about an agent's human carcinogenic potential and the conditions that characterize its expression" (EPA 2011). In EPA's *Guidelines for Carcinogen Risk Assessment*, the WOE narrative "explains the kinds of evidence available and how they fit together in drawing conclusions, and it points out significant issues/strengths/limitations of the data and conclusions" (EPA 2005, p. 1-12). Current guidelines for evidence integration are given in Section 5 of the preamble for IRIS toxicological reviews, and guidelines for writing a WOE narrative are given specifically in Section 5.5 (EPA 2013a, pp. B-5 to B-9). Guidance has also been provided in some of the outcome-specific guidelines (for example, EPA 2005).

Rhomberg et al. (2013), in a review article surveying best practices for WOE frameworks or analyses, describe WOE as encompassing *all* of causal inference. They state that "in the broadest sense, almost all of scientific inference about the existence and nature of general causal processes entails WOE evaluation" (Rhomberg et al. 2013, p. 755). They then describe the wide array of meanings attached to the phrases *systematic review* and *weight of evidence* as follows:

> Some terms are used differently in different frameworks. In particular, to some practitioners, the term "systematic review" refers specifically to the systematic assembly of evidence (for example, by using explicit inclusion and exclusion criteria or by using standard tabulation and study-by-study quality evaluation), while "weight of evidence" refers to the subsequent integration and interpretation of these assembled selected studies/data as they are brought to bear on the causal questions of interest. To others, "systematic review" refers to the whole process from data assembly through evaluation, interpretation, and drawing of conclusions; for still others, this whole suite of processes is subsumed under WoE. …when we refer to "WoE frameworks," we mean approaches that have been developed for taking the process all the way from scoping of the assessment and initial identification of relevant studies through the drawing of appropriate conclusions.

The present committee found that the phrase *weight of evidence* has become far too vague as used in practice today and thus is of little scientific use. In some accounts, it is characterized as an oversimplified balance scale on which evidence supporting hazard is placed on one side and evidence refuting hazard on the other. That analogy neglects to account for the total weight on either side (that is, the scope of evidence available) and captures only where the balance stands. Others characterize WOE as a single scale, and different kinds of evidence have different weights. For example, a single human study with low risk of bias might be considered as providing the same evidential weight as three well-conducted animal studies. The weights might be adjusted according to the quality of the study design. This analogy neglects to account for the "weight for" vs the "weight against" hazard.

Perhaps the overall idea of the WOE for hazard should combine both characterizations. It is evident, however, that its use in the literature and by scientific agencies, including EPA, is vague and varied. The present committee found the phrase *evidence integration* to be more useful and more descriptive of what is done at this point in an IRIS assessment—that is, IRIS assessments must come to a judgment about whether a chemical is hazardous to human health and must do so by integrating a variety of evidence. In this chapter, therefore, the committee uses the phrase *evidence integration* to refer to the process that occurs after assessment of all the individual lines of evidence (see Figure 6-1).

As described in previous chapters, the committee uses the phrase *systematic review* to describe a process that ends before evidence integration and hazard identification (Figure 6-1). After hazard identification, the IRIS process turns to dose-response assessment and derivation of toxicity values. By defining systematic review as a process that ends *before* hazard identification,

the committee is *not* implying that the process by which IRIS conducts hazard identification and dose-response assessment is or should be nonsystematic; it simply ensures that the committee's use of the phrase *systematic review* is clear and consistent with current literature.

Finally, the committee makes a distinction between *data* and *evidence*. Although it is common to use the two somewhat interchangeably, they are not synonymous. As the report *Ethical and Scientific Issues in Studying the Safety of Approved Drugs* (IOM 2012, p. 122) states:

> The Compact Oxford English Dictionary [Oxford Dictionaries 2011] defines data as "facts and statistics collected together for reference or analysis" and evidence as "the available body of facts or information indicating whether a belief or proposition is true."

Data become evidence for or against a claim of hazard only after some sort of statistical or scientific inference.

EVALUATING STRENGTHS AND WEAKNESS OF EVIDENCE

As discussed in Chapter 3, evidence on hazard can come from human studies, animal studies, mechanistic studies, background knowledge, and a host of other sources. Each source has its relative strengths and weaknesses, and Table 6-1 highlights some of the important ones. In using integrative approaches, those considering the evidence should take their strengths and weaknesses into account.

ORGANIZING PRINCIPLES FOR INTEGRATING EVIDENCE

One challenge that EPA and other regulatory agencies face when attempting to establish guidelines for integrating evidence is that the amount and quality of the various types of evidence can vary substantially from one chemical to another. For example, a small number of environmental contaminants—such as arsenic, dioxins, polychlorinated biphenyls, and formaldehyde—have extensive human data, often from relatively well-designed cohort studies, substantial animal data from several animal models, and mechanistic information. On a larger number of chemicals, there are few or no high-quality human data, but there are a small number of animal studies and some in vitro mechanistic studies. For the great majority of the chemicals in the environment that might cause harm, however, there are virtually no human or animal data, although there might be some scientific knowledge relevant to a chemical's potential toxicity or putative mechanism (often inferred from structurally similar compounds).[1]

That variation in the evidence base invites different organizing principles by which evidence could be combined into a single judgment. One option is to organize the evidence around potential mechanisms by which a chemical might cause harm. As models of chemical action improve, it might become possible to predict the toxicity of a chemical reasonably accurately merely by using sophisticated models of its interaction with human cells and tissues. As it is clearly infeasible to generate human or animal data on the more than 80,000 chemicals in commercial use in the United States, that approach might be the only option for the great majority of chemicals, and it is the approach proposed in the NRC report *Toxicity Testing in the 21st Century: A Vision and a Strategy* (NRC 2007). In fact, EPA's strategic plan for evaluating chemical toxicity provides a framework for the agency to incorporate the new scientific paradigm into future toxicity-testing and risk-assessment practices (EPA 2009).

[1]As noted in Chapter 3, the committee is using the term *mechanism of action* (or *mechanism*) in this report rather than *mode of action* simply for ease of reading; it recognizes that these terms can have different meanings.

TABLE 6-1 Common Strengths and Weaknesses of Human Epidemiologic (HE), Experimental Animal (EA), and Mechanistic (MECH) Studies for Hazard Identification

Source of Uncertainty	Strength	Weakness
Interspecies extrapolations	HE: Not applicable, because not needed.	HE: Not applicable, because not needed.
	EA: Can use multiple species, and this provides a broad understanding of species differences.	EA: Inherent weakness when interspecies extrapolation from animals to humans is required.
	MECH: Can identify cellular, biochemical, and molecular pathways that are similar or different in humans and the test species and thus lend strength to the veracity of the extrapolation.	MECH: For a given chemical, multiple mechanisms might be involved in a given end point, and it might not be evident how different mechanisms interact in different species to cause the adverse outcome.
Intraspecies extrapolation	HE: Often able to study effects in heterogeneous populations.	HE: Many studies involve occupational cohorts, which do not reflect the general population.
	EA: Effects seen during different life stages (such as pregnancy and lactation) can be evaluated. Use of transgenic animals can provide important mechanistic data.	EA: Often rely on a few strains in which animal genetics, life stage, diet, and initial health state are controlled.
	MECH: Observed differences between strains of a common test species (such as Fisher 344 rats and Sprague-Dawley rats) might be readily explained by different pathways. Comparison with human in vitro mechanistic data might allow better selection of the most appropriate animal model for predicting human response.	MECH: Putative mechanism of the adverse outcome might not be known, and mechanistic data might not reveal the basis of differences within a species.
High-dose to low-dose extrapolation	HE: Often better suited for considering actual range of population exposures.	HE: Occupational exposure is often higher than that seen in the general human population.
	EA: Wide range of exposures is possible, and this allows better estimation of quantitative dose-response relationships.	EA: Exposures used are often orders of magnitude higher than those seen in the general human population.
	MECH: Dose-related differences in ADME properties and pharmacodynamic processes might be used to adjust for differences in rate of response between high and low doses.	MECH: The ultimate molecular target for toxicity might not be known at low or high doses, so mechanism might not accurately predict high-dose to low-dose extrapolations.

Acute to chronic extrapolation (temporal considerations)	HE: Might closely mimic exposure durations seen in the general population.	HE: Occupational exposure durations are often shorter (years vs lifetime; 8 hr/day vs 24 hr/day) than those seen in the general human population.
	EA: Wide range of exposure durations is possible.	EA: Highly dependent on study design.
	MECH: Provides invaluable information on whether a product or effect can accumulate on repeated exposure and whether repair pathways or adaptive responses can lead to outcomes that are significantly different between single and repeated exposures.	MECH: If mechanism differs between acute or chronic response, the information on one might not be informative of the other.
Route-to-route extrapolation	HE: Often involve route of exposure relevant to the general human population.	HE: Data might be available on only one route of exposure.
	EA: Can involve route of exposure relevant to the general human population.	EA: Often uses an exposure method that requires extrapolation of data (for example, diet to drinking water).
	MECH: Pharmacokinetic differences (ADME, PBPK) might facilitate more accurate identification of target-tissue dose from different exposure pathways.	MECH: Mechanism might be tissue-specific and therefore route-dependent as the route determines the initially exposed tissue.
Other considerations	HE: Can evaluate cumulative exposures and health effects.	HE: Long lag time to identify some effects. Increased potential for exposure and outcome misclassification and confounding. Variable cost.
	EA: Shorter animal lifespans allow for more rapid evaluation of hazards. Reduced misclassification of exposures and outcomes. Allows examination of full spectrum of toxic effects.	EA: Multiple extrapolations required. Variable cost.
	MECH: Conservation of fundamental biologic pathways (such as cell-cycle regulations, apoptosis, and basic organ-system physiology) might allow quick and inexpensive identification of potential adverse effects of a new chemical in the absence of human or animal in vivo data.	MECH: Identification of relevant pathways in producing the toxic response might be difficult because of the lack of understanding of pathobiologic processes.

Organizing evidence around mechanism for chemicals on which only some human or animal data are available, however, seems inappropriate. Consider the Food and Drug Administration (FDA) and drug safety. If FDA were required to organize drug safety around mechanism, it would be nearly impossible to regulate many important drugs because the mechanism is often not understood, even for drugs that have been studied extensively. For example, it is known that estrogen plus progestin therapy causes myocardial infarctions on the basis of randomized clinical trials even though the mechanism is not understood (Rossouw et al. 2002). Randomized clinical trials are so successful partly because they bypass the need for mechanistic information and provide an indication of efficacy. Similarly, epidemiologic studies that identify unintended effects are often credible because explanations of an observed association other than a causal effect are implausible. For example, the associations between statins and muscle damage and between thalidomide and birth defects are widely accepted as causal; mechanistic information played a minor role in the determination, if any. The history of science is replete with solid causal conclusions in advance of solid mechanistic understanding.

A second option is to organize the case for hazard around the kinds of evidence either actually or potentially available. Different kinds of evidence have more or less direct relevance to the determination of hazard and can often be indirectly relevant by virtue of bearing on the relevance of other kinds of evidence. As discussed previously, each major type of evidence has inherent strengths and weaknesses, and the three major lines of research used by the IRIS program produce complementary findings. For example, mechanistic knowledge can often be informative about the relevance of animal-model data, as exemplified in the approaches of EPA and the International Agency for Research on Cancer (IARC).

In considering which kind of evidence is more or less important in driving a conclusion about hazard, human studies are historically taken to be more important than animal studies. For example, the EPA guidelines for cancer risk assessment state that classification of a chemical as a human carcinogen is reached when there is "convincing epidemiologic evidence of a *causal* association between human exposure and cancer" (EPA 2005, p. 2-54; emphasis added). According to the guidelines, the determination can be made irrespective of the strength of the animal data. In other words, in cases in which extensive human data strongly support a causal association between exposure and disease and the studies are judged to have a relatively low risk of bias, the human evidence can outweigh animal and other evidence, no matter what it is. Furthermore, a judgment of "carcinogenic to humans" can be justified when human studies show only an association (not a causal association) if they are buttressed by extensive animal evidence and mechanistic evidence that support a conclusion of causation (EPA 2005, p. 2-54).

When human data are nonexistent, are mixed, or consistently show no association and an animal study finds a positive association, the importance of mechanistic data is increased. Fundamental toxicologic questions related to dose, exposure route, exposure duration, timing of exposure, pharmacokinetics, pharmacodynamics, and mechanisms would then play an even more important role in determining the relevance of positive in vivo animal data, especially when the human data are negative or inconclusive.

A final option for organizing evidence integration might be called an alternative interpretation, which Rhomberg et al. (2013, p. 755) argue is desirable and will improve transparency:

> A WoE evaluation is only useful and applicable to constructive scientific debate if the logic behind it is made clear; with that, it is often necessary to take the reader through alternative interpretations of the data so that the various interpretations can be compared logically. This approach does not eliminate the need for scientific judgment, and often may not lead to a definitive choice of one interpretation over the other, but it will clearly lay out the logic for how one weighs the evidence for and against each interpretation. Only in this way is it possible to have constructive scientific debate about potential causality that is focused on an organized, logical "weighing" of the evidence.

The alternative interpretations are implied to arise from different potential mechanisms, but the committee does not view mechanisms as central to this sort of organizing framework. Rather, the pattern of human, animal, and mechanistic evidence analyzed might be explained in various ways. For example, if human data show little association between exposure and disease but animal data provide consistent evidence of toxicity, one explanation might be that the chemical is toxic to animals but not to humans because of some difference in metabolic response. Another explanation might be that the human data were consistently underpowered statistically, rife with measurement difficulties, or taken from populations exposed to low doses of the chemical. The organizing principle for integrating the evidence in this case would be to consider each explanation and describe the evidence (of any type) that supports it and refutes it.

Whether one organizes evidence around mechanism, kinds of evidence, or alternative interpretations, it has to be integrated into a single judgment on hazard. Integrating evidence rationally requires an implicit or explicit set of guidelines. The guidelines for integration are often called a framework, which is defined as a clear process or a clear set of guidelines for evidence integration. Such frameworks range from ones that involve a rigid, algorithmic integration process to ones that provide loose guidelines and allow experts substantial freedom in applying them.

It seems impossible and undesirable to build a scientifically defensible framework in which evidence is integrated in a completely explicit, fixed, and predefined recipe or algorithm. There are no empirical data on the basis of which fixed weighting schemes are more likely to produce true answers than other schemes, and getting such data is far off. Furthermore, substantial expert judgment in making categorizations according to such schemes is unavoidable. On the other hand, simply putting a group of experts into a room and asking them to consider the evidence in its totality and to emerge with a decision seems equally undesirable and endangers transparency. To ensure transparency, it seems desirable to have an articulated framework within which to consider the relevance of different evidence to the causal question of hazard identification. Various options for evidence integration are considered further below.

THE BRADFORD HILL GUIDELINES

Common considerations (or quasicriteria) used in many frameworks in which various bodies of evidence must be integrated to reach a causal decision are the "Hill criteria for causality," a set of guidelines first articulated by Austin Bradford Hill in 1965 to deal with the problem of integrating evidence on environmental exposure and disease, particularly with respect to smoking and lung cancer. EPA states that "in general, IRIS assessments integrate evidence in the context of Hill (1965)" (EPA 2013a, p. 13). Hill's guidelines are meant as considerations in assessing the move from association to causation (causal association). They include strength of association, consistency, specificity, temporality, biologic gradient, plausibility, coherence, experimental evidence, and analogy. The Hill criteria are widely regarded as useful (Glass et al. 2013) and, as noted in the IRIS preamble, explicitly constitute the basis on which EPA should evaluate the overall evidence on each effect (EPA 2013a, Appendix B, p. B-5).

The Hill guidelines, however, are by no means rigid guides to reaching "the truth." Rothman and Greenland (2005) used a series of examples to illustrate why the Hill criteria cannot be taken as either necessary or sufficient conditions for an association to be raised to a causal association. They provide counterexamples to each of Hill's criteria, some from the very example—smoking— that Hill considered in his 1965 article. For example, they note that although the association between smoking and cardiovascular disease is comparatively weak, as is the association between second-hand smoke and lung cancer, both relationships are now considered causal (Rothman and Greenland 2005). They further note that examples of strong associations that are not causal also abound, such as birth order and Down syndrome. There are many examples of causal inference in which there is no known mechanism. Therefore, although the guidelines can

usefully inform an evidence-integration narrative, Rothman and Greenland caution against using the Hill guidelines as "checklist criteria"—a warning that the present committee considers appropriate.

CURRENT ENVIRONMENTAL PROTECTION AGENCY APPROACH TO INTEGRATING EVIDENCE: THE AGENCY'S RESPONSE TO RECOMMENDATIONS IN THE NATIONAL RESEARCH COUNCIL FORMALDEHYDE REPORT

The 2011 NRC formaldehyde report made several recommendations for evidence integration in IRIS assessments (see Box 6-1). As in the other recommendations, there is an emphasis on transparency and standardization of approach.

The draft preamble (EPA 2013a, Appendix B) and the draft handbook (EPA 2013a, Appendix F) contain the most recent guidelines on evidence integration for IRIS assessments. Whereas the preamble and the handbook provide reasonably extensive guidelines on evidence integration *within* evidence streams, the preamble does not provide guidelines for evidence integration *among* evidence streams (only what hazard descriptors should be used), and instructions for evidence integration have yet to be written for the handbook. Therefore, this section discusses the guidelines that EPA has outlined and how evidence integration has been carried out and described in recent IRIS assessments of methanol and benzo[a]pyrene (EPA 2013b,c). Potential revisions that EPA might want to consider are provided. The committee recognizes that the methanol and benzo[a]pyrene assessments do not reflect all changes that EPA has made or plans to make to the IRIS process in response to the recommendations in the NRC formaldehyde report.

Two guiding principles are apparent in the committee's review of the current IRIS process. First, as of fall 2013, EPA still relies on a *guided expert judgment* process (discussed below). EPA (2013a, p. 14) states that hazard identification requires a critical weighing of the available evidence, but this process "is not to be interpreted as a simple tallying of the number of positive and negative studies" (EPA 2002, p. 4-12). EPA (2013a, p. 14) further states that "hazards are identified by an informed, expert evaluation and integration of the human, animal, and mechanistic evidence streams." Second, overall conclusions regarding causality are to be reached and justified according to the Hill criteria (EPA 2013a).

BOX 6-1 Recommendations on Evidence Integration from 2011 National Research Council Formaldehyde Report

- Strengthened, more integrative and more transparent discussions of weight of evidence are needed. The discussions would benefit from more rigorous and systematic coverage of the various determinants of weight of evidence, such as consistency.
- Review use of existing weight of evidence guidelines.
- Standardize approach to using weight of evidence guidelines.
- Conduct agency workshops on approaches to implementing weight of evidence guidelines.
- Develop uniform language to describe strength of evidence on noncancer effects.
- Expand and harmonize the approach for characterizing uncertainty and variability.
- To the extent possible, unify consideration of outcomes around common modes of action rather than considering multiple outcomes separately.

Source: NRC 2011, pp. 152, 165.

Section 5 of the IRIS preamble articulates guidelines for "evaluating the overall evidence of each effect" (EPA 2013a, Appendix B). Rather than giving an explicit process for evaluating the overall evidence, the preamble states that "causal inference involves scientific judgment, and the considerations are nuanced and complex" (EPA 2013a, p. B-5). It also describes evidence integration *within* each kind of evidence stream—evidence in humans, evidence in animals, and mechanistic data to identify adverse outcome pathways and mechanisms of action—before combining different kinds of evidence. For evidence in humans, IRIS assessments are to "select a standard descriptor" from among the following (EPA 2013a, p. B-6):

- "Sufficient epidemiologic evidence of an association consistent with causation."
- "Suggestive epidemiologic evidence of an association consistent with causation."
- "Inadequate epidemiologic evidence to infer a causal association."
- "Epidemiologic evidence consistent with no causal association."

No detailed process is suggested for arriving at a classification other than relying on expert judgment that is based on the aspects listed above. A subset of the Hill guidelines is offered as relevant for integrating the evidence in animals. For integration of mechanistic evidence, IRIS assessments are to consider the following three questions (EPA 2013a, pp. B-7 to B-8):

1. "Is the hypothesized mode of action sufficiently supported in test animals?"
2. "Is the hypothesized mode of action relevant to humans?"
3. "Which populations or lifestages can be particularly susceptible to the hypothesized mode of action?"

For *overall* evidence integration, an IRIS assessment must answer the causal question, "Does the agent cause the adverse effect?" (EPA 2013a, p. B-8). It then must summarize the overall evidence with a "narrative that integrates the evidence pertinent to causation" (EPA 2013a, p. B-8). The narrative should target a qualitative categorization, and two examples are offered in the IRIS preamble. The first is taken directly from the EPA guidelines for carcinogen risk assessment (EPA 2005, Table 6-2). The second is taken from EPA's integrated science assessments for the criteria pollutants (EPA 2010, see Table 6-3).

In summary, the draft IRIS preamble (EPA 2013a, Appendix B) gives guidelines as to what considerations ought to inform the experts' integration of human, animal, and mechanistic evidence, and it gives extensive guidance on the qualitative categorization that the experts should use, but it articulates no systematic process by which the experts are to come to a conclusion. The draft handbook (EPA 2013a, Appendix F) gives extensive guidelines for synthesizing evidence within each stream but no guidelines for integrating evidence among streams. The guidelines and the summary descriptors offered for epidemiologic and other studies are reasonable, and similar ones have been used in many other organizations that have similar aims and problems, such as IARC and the National Toxicology Program (NTP).

Draft IRIS Assessment of Methanol

A recent IRIS assessment of Methanol (EPA 2013b) includes a section (Section 4.6, Synthesis of Major Noncancer Effects) that provides a summary of the dose-related effects that have been observed after subchronic or chronic methanol exposure. EPA (2013b, p. 4-77) provides the following conclusion in the summary:

> Taking all of these findings into consideration reinforces the conclusion that the most appropriate endpoints for use in the derivation of an inhalation RfC for methanol are associated with developmental neurotoxicity and developmental toxicity. Among an array of findings indicating developmental neurotoxicity and developmental malformations and anomalies that have been observed in the fetuses and pups of exposed dams, an increase in

the incidence of cervical ribs of gestationally exposed mice (Rogers et al., 1993b) and a decrease in the brain weights of gestationally and lactationally exposed rats (NEDO, 1987) appear to be the most robust and most sensitive effects.

TABLE 6-2 Categories of Carcinogenicity

Category	Conditions
Carcinogenic to humans	There is convincing epidemiologic evidence of a causal association (that is, there is reasonable confidence that the association cannot be fully explained by chance, bias, or confounding); or there is strong human evidence of cancer or its precursors, extensive animal evidence, identification of key precursor events in animals, and strong evidence that they are anticipated to occur in humans.
Likely to be carcinogenic to humans	The evidence demonstrates a potential hazard to humans but does not meet the criteria for *carcinogenic*. There may be a plausible association in humans, multiple positive results in animals, or a combination of human, animal, or other experimental evidence.
Suggestive evidence of carcinogenic potential	The evidence raises concern for effects in humans but is not sufficient for a stronger conclusion. This descriptor covers a range of evidence, from a positive result in the only available study to a single positive result in an extensive database that includes negative results in other species.
Inadequate information to assess carcinogenic potential	No other descriptors apply. Conflicting evidence can be classified as inadequate information if all positive results are opposed by negative studies of equal quality in the same sex and strain. Differing results, however, can be classified as suggestive evidence or as likely to be carcinogenic.
Not likely to be carcinogenic to humans	There is robust evidence for concluding that there is no basis for concern. There may be no effects in both sexes of at least two appropriate animal species; positive animal results and strong, consistent evidence that each mode of action in animals does not operate in humans; or convincing evidence that effects are not likely by a particular exposure route or below a defined dose.

Source: EPA 2013a, pp. B-8 to B-9.

TABLE 6-3 Categories of Evidential Weight for Causality

Category	Conditions
Causal relationship	Sufficient evidence to conclude that there is a causal relationship. Observational studies cannot be explained by plausible alternatives, or they are supported by other lines of evidence, for example, animal studies or mechanistic information.
Likely to be a causal relationship	Sufficient evidence that a causal relationship is likely, but important uncertainties remain. For example, observational studies show an association but co-exposures are difficult to address or other lines of evidence are limited or inconsistent; or multiple animal studies from different laboratories demonstrate effects and there are limited or no human data.
Suggestive of a causal relationship	At least one high-quality epidemiologic study shows an association but other studies are inconsistent.
Inadequate to infer a causal relationship	The studies do not permit a conclusion regarding the presence or absence of an association.
Not likely to be a causal relationship	Several adequate studies, covering the full range of human exposure and considering susceptible populations, are mutually consistent in not showing an effect at any level of exposure.

Source: EPA 2013a, p. B-9.

EPA goes on to use those and other studies to develop candidate RfCs. Although the discussion often provides details concerning the decision-making process used by EPA with more transparency than previous IRIS assessments, what remains somewhat lacking is an explicit description of the integrative approach used by EPA to combine data streams.

More specifically, the report notes that informative human studies of methanol are limited to acute exposures, but "the relatively small amount of data for subchronic, chronic, or in utero human exposures are inconclusive. However, a number of reproductive, developmental, subchronic, and chronic toxicity studies have been conducted in mice, rats, and monkeys" (EPA 2013b, p. xxiv). The report also notes, however, that the "enzymes responsible for metabolizing methanol are different in rodents and primates" (EPA 2013b, p. xxii), but then remarks that several PBPK models have been developed to account for these differences. Even though reproductive and developmental end points are identified as hazards in humans, the report notes that there is "insufficient evidence to determine if the primate fetus is more sensitive or less sensitive than rodents to the developmental or reproductive effects of methanol" (EPA 2013b, p. xxv). Interspecies differences are clearly important in methanol. Some central nervous system toxicities, such as blindness, have been observed in humans but not rodents; the differences are most likely due to species differences in the rate of elimination of formic acid that is formed by the oxidation of methanol. Section 4.7 of the IRIS assessment includes a discussion of noncancer mechanisms and the uncertainties in how such mechanisms are shared between humans and rodents. It ultimately concludes by saying that "the effects observed in rodents are considered relevant for the assessment of human health" (EPA 2013b, p. xxvi).

The narrative is informative, detailed, and accessible. The issues are clear, but the narrative does not include any systematic discussion of evidence integration that uses the Hill criteria or any others, such as the ones listed in Table 6-3. Although the interspecies evidence is complicated (and in this case crucial), the overall evidence-integration statement is as follows (EPA 2013b, p. xxv):

> Taken together, however, the NEDO (1987) rat study and the Burbacher et al. (2004a; 2004b; 1999a; 1999b) monkey study suggest that prenatal exposure to methanol can result in adverse effects on developmental neurology pathology and function, which can be exacerbated by continued postnatal exposure.

Draft IRIS Assessment of Benzo[a]pyrene

In August 2013, EPA released the draft *Toxicological Review of Benzo[a]pyrene* (EPA 2013c). The draft assessment shows that the IRIS program has taken several additional steps toward addressing the recommendations in the 2011 NRC formaldehyde report. In the executive summary, EPA concludes that benzo[a]pyrene is carcinogenic; that noncarcinogenic effects might include developmental, reproductive, and immunological effects; that animal studies clearly demonstrate these effects; and that human studies show associations between DNA adducts that are biomarkers of exposure and these effects. "Overall, the human studies report developmental and reproductive effects that are generally analogous to those observed in animals, and provide qualitative, supportive evidence for hazards associated with benzo[a]pyrene exposure" (EPA 2013c, p. xxxiii).

In Section 1 (Hazard Identification) of the IRIS assessment, an accessible and detailed narrative describes the human, animal, and mechanistic evidence on developmental, reproductive, and immunotoxicity. In Section 1.2, an explicit narrative describes the evidence on noncancer effects (1.2.1) and then on cancer (1.2.2). For noncancer outcomes, the Hill criteria are not mentioned, nor is there a qualitative categorization for any end point of the sort described in the preamble. Yet the narrative is clear and describes the evidence in a way that roughly matches the conditions given in Table 6-3. Section 1.2.2, which describes the evidence on carcinogenicity, is

extremely clear and follows closely EPA's 2005 *Guidelines for Carcinogen Risk Assessment*. It includes separate assessments of the human, animal, and mechanistic evidence according to those guidelines and includes evidence tables (for example, Table 1-18, p. 1-87) that connect evidential categorization with the supporting studies.

In summary, EPA is clearly moving toward implementing the recommendations of the NRC formaldehyde report but needs to continue to improve the narratives for noncancer outcomes to bring them into line with the preamble or the narrative for carcinogenicity. The two draft assessments that the committee reviewed are not consistent in their approach.

Evaluation of Agency Response

The purpose of this section was to assess how the IRIS process for evidence integration has evolved in response to the recommendations in the NRC formaldehyde report. First, the committee discussed how the guidelines for evidence integration have evolved. A preamble has been created that broadly describes how evidence-integration narratives might be structured to follow the Hill criteria. The committee discusses in more detail in the following section options for improving the guidelines included in such a preamble.

The recent IRIS assessments for methanol and benzo[a]pyrene include the new preamble, and in both cases, the evidence-integration narrative for cancer and noncancer outcomes is clear and informative. In both cases, however, the narrative for noncancer outcomes does not follow the guidelines as given in the preamble or other recent IRIS guidelines. In future assessments, EPA might either change the guidelines to follow more closely the kinds of narratives given for methanol and benzo[a]pyrene or ensure that the narratives more closely follow the guidelines included.

OPTIONS FOR MOVING FORWARD

In this section, the committee describes several options for improving the IRIS process for evidence integration. The options are divided into qualitative approaches and quantitative approaches. A qualitative approach is a process whose output is a categorization of the overall evidence, for example, "the evidence suggests that it is *likely* that chemical X is immunotoxic." A quantitative approach is a process that produces a quantitative output, for example, "the evidence suggests that there is at least a 75% chance that chemical X is immunotoxic."

Qualitative Approaches for Integrating Evidence

Several approaches have been taken by the scientific and regulatory communities for integrating diverse evidence for hazard identification. Most commonly, the target is a qualitative categorization of the overall evidence as to whether an agent is a health hazard, that is, Can it cause cancer or some other adverse outcome, and with what degree of certainty can the judgment as to causation be made? For example, the 2004 *Surgeon General's Report—The Health Consequences of Smoking* (DHHS 2004) and the 2006 Institute of Medicine (IOM) Committee on Asbestos: Selected Health Effects (IOM 2006) uses a four-level categorization for the overall strength of evidence on causation: "sufficient to infer a causal relationship," "suggestive but not sufficient to infer a causal relationship," "inadequate to infer the presence or absence of a causal relationship," and "suggestive of no causal relationship." Tables 6-2 and 6-3 provide qualitative categorizations used by EPA.

There are several methods for making a qualitative categorization of the evidence on hazard. One method can be described generally as guided expert judgment, of which there are many varieties. An alternative method for reaching a qualitative categorization of evidence can be described as a structured process. Those processes are described in the following sections.

Guided Expert Judgment

Recently, the IRIS program has undergone several organizational changes that should improve the expertise available to develop assessments. EPA has created disciplinary workgroups that have expertise in epidemiology, reproductive and developmental toxicity, neurotoxicity, toxicokinetic modeling, and biostatistics. Multiple disciplinary workgroups contribute to each assessment, and each disciplinary workgroup reviews the studies and develops the conclusions regarding all assessments on which there are studies in their fields. The new workflow is meant to ensure that all sections of an assessment are developed by appropriate experts, and use of the workgroups contributes to quality control and consistency among assessments. The workgroups have been initially formed by drawing on expertise available in EPA's National Center for Environmental Assessment. A long-term goal is to broaden the expertise available to the IRIS program by including other scientists from within and outside EPA. In addition, EPA has established a Systematic Review Implementation Group whose primary purpose is to coordinate the implementation of systematic review in the IRIS program. Evidence integration done internally by EPA experts is (or can be) an extremely efficient method in arriving at sensible conclusions. The process in which EPA experts integrate evidence via a set of guidelines through some form of internal deliberation, however, does not lend itself in any obvious way to being transparent or reproducible.

If EPA wants to achieve more transparency within the basic process of evidence integration that it now follows, one option is to expand its current practice by following the IARC model. IARC recruits a working group of experts on a particular substance from epidemiology (cancer in humans), toxicology (cancer in experimental animals), mechanistic or biologic disciplines, and exposure science. A two-stage process is then used by which the experts in each field are asked to come to consensus on a qualitative categorization of the evidence in their field (except exposure). In the second stage, the full and diverse group of experts is asked to integrate their group findings into an overall judgment of a chemical's carcinogenicity (IARC 2006; Meek et al. 2007). IARC has a general scheme for bringing together the human and animal data and for modifying the resulting classification on the basis of the findings of mechanistic research.

The overall IARC process and the judgments required are guided by an extensive preamble, (IARC 2011), but the process relies on expert judgment rather than on a structured approach to weighing and combining evidence. Experts who have acknowledged conflicts of interest are not allowed on the panels but can be present to observe without participating in formal deliberations, although on occasion they might be asked to respond to specific questions. According to the IARC (2011) guidelines,

> In the spirit of transparency, Observers with relevant scientific credentials are welcome to attend IARC Monographs meetings. Observers can play a valuable role in ensuring that all published information and scientific perspectives are considered. The chair may grant Observers an opportunity to speak, generally after they have observed a discussion. Observers do not serve as meeting chair or subgroup chair, draft any part of a Monograph, or participate in the evaluations.

Thus, it is consistent with IARC that EPA could allow a wide variety of experts at least to observe the deliberations, including experts who have a potential conflict of interest and those representing key commercial or government concerns. Another option would be for EPA to provide more systematic requirements for the write-up and involve targeted scientific review of that part of the assessment.

Structured Processes

One approach for using a structured process to integrate evidence is the Grading of Recommendations Assessment, Development and Evaluation (GRADE) system, which is being used with

increasing frequency in the development of clinical-practice guidelines and health-technology assessments (Guyatt et al. 2011a,b,c,d,e). A version of GRADE is also being used by NTP in its Office of Health Assessment and Translation to make judgments about whether a chemical is hazardous to humans (NTP 2013). GRADE is a system for rating the quality of evidence and the strength of recommendations. The quality of the evidence on each outcome being assessed is evaluated according to the following approach: evidence is downgraded if there is risk of bias, inconsistency, indirectness, imprecision, and publication bias in or among studies; and evidence is upgraded if there is a large effect size, a dose-response relationship, and elimination of all plausible confounding. The strength of the recommendations is rated on the basis of the evidence evaluation. The highest-quality, most upgraded evidence results in a strong recommendation, whereas the lowest-quality evidence results in a weak recommendation.

The main advantage of GRADE over other qualitative approaches for integrating evidence is that the judgments made to evaluate and integrate bodies of evidence are systematic and transparent. In practice, consensus approaches or multiple coders working independently are used to assess each GRADE criterion. The reasons for each conclusion or recommendation are then clearly summarized. The final product of a GRADE assessment is a qualitative, tabular summary of the evidence with a quality rating (0+ to 4+) for each outcome.

GRADE criteria for assessing the quality of evidence are closely aligned with the Hill criteria for establishing causality (Schünemann et al. 2011). The strength of association assessed with Hill is accounted for by upgrading or downgrading the evidence according to the specific GRADE criteria. Upgrading reflects scientific confidence that a causal relationship exists. The Hill criterion for strength of association means that the stronger the association, the more confidence in causation.[2] As shown in Table 6-4, the Hill criteria and the GRADE approach consider inconsistency of the evidence, indirectness of evidence, the magnitude of the effect, and the dose-response relationship. The Hill criteria require a prospective temporal relationship between the exposure and outcome and note that experimental evidence strengthens causation. Similarly, the GRADE criteria give greater weight to randomized clinical trials, although such studies are rarely available for environmental chemicals. Notable exceptions are some irritant gases and a few metals (such as trivalent chromium, cobalt, and selenium) and organic chemicals (such as perchlorate) that have been used or investigated for medicinal purposes.

In developing recommendations for clinical-practice guidelines, GRADE is usually applied to randomized clinical trials (of efficacy of clinical interventions) and observational studies (of harm). However, a GRADE-like approach has recently been applied to the integration of evidence from different data streams for the assessment of chemical hazards. Woodruff and Sutton (2011) have proposed the *Navigation Guide,* which specifies the study question, selects the evidence, rates the quality of individual studies, and rates the quality of evidence and the strength of a recommendation. That final step uses a GRADE-like approach to rate the quality of the overall body of evidence— including human, animal, and in vitro studies—on the basis of a priori and clearly stated criteria and integrates the quality assessment with information about exposure to develop a recommendation. NTP has adopted a similar approach for combining evidence from different data streams (NTP 2013). NTP develops a confidence rating for the body of evidence on a particular outcome by considering the strengths and weaknesses of the entire collection of studies that are relevant for a particular question. As in GRADE, the initial confidence in a recommendation is determined by the strength of the study design for assessing causality independently of the risk of bias in an individual study. However, unlike GRADE, which rates randomized controlled trials with higher confidence than epidemiologic studies, the NTP approach

[2]This statement assumes that all important confounding factors have been controlled. Strong associations in poorly controlled studies are unreliable, and such associations might become weaker after adjustment for confounding factors.

TABLE 6-4 Comparison of Hill, GRADE, *Navigation Guide*, and NTP Criteria for Evaluating and Integrating Evidence

	Hill	GRADE	*Navigation Guide*	NTP
Downgrading confidence or weakening recommendation				
Risk of bias		X	X	X
Inconsistency	X	X	X	X
Indirectness[a]	X	X[b]	X	X
Imprecision		X	X	X
Publication bias		X	X	X
Financially conflicted sources of funding		X[c]	X	
Upgrading confidence or strengthening recommendation				
Large effect	X	X	X	X
Dose-response relationship[d]	X	X	X	X
No plausible confounding		X	X	X
Cross-species, population, or study consistency				X
Serious or rare end points, such as teratogenicity			X	X[b]

[a]Indirectness is the extent to which a study directly addresses the study question (Higgins and Green 2011). Indirectness might arise from the lack of a direct comparison or if some restriction of the study limits generalizability.
[b]Includes Hill criteria of specificity, biologic plausibility, and coherence.
[c]Rated under "other."
[d]A formal dose-response assessment is typically performed, depending on the outcome of the hazard identification. However, at this stage, a potential dose-response relationship provides evidence of a hazard and should be used in a hazard-identification process.

determines the initial confidence on the basis of whether the exposure to the substance is controlled, data indicate that the exposure precedes development of the outcome, individual level (not population aggregate) data are used to assess the outcome, and the study uses a comparison group (NTP 2013). Thus, randomized controlled trials meet the first criterion, and epidemiologic studies are distinguished by how well they meet the remaining three criteria; prospective cohort studies, for example, start at a higher level of confidence than case-control studies. See Table 6-4 for a comparison of the various structured approaches.

Structured assessments like GRADE are useful primarily as a means of systematically documenting the judgments made in evaluating the evidence. This kind of documentation might enhance transparency to the extent that it tracks the details of how the evidence was assessed. The committee emphasizes that structured assessments like GRADE formalize and organize but do not replace expert judgment. Although the idea of adopting a structured-assessment process to enhance transparency is commendable, there is some risk that imposing excessively formal criteria for describing and evaluating evidence could slow the process and produce more complex output without improving the quality of decisions. The criteria in GRADE were developed for the assessment of evidence from clinical studies and might not always be appropriate for evaluating the effects of environmental chemicals. Thus, if EPA decides to adopt a GRADE-like approach, it should take care to customize it to the needs of IRIS, perhaps along the lines currently being developed at NTP.

Quantitative Approaches to Integrating Evidence

In each approach above, evidence is integrated with reliance on expert judgment and the output is qualitative. Although a structured process can use several quasi-formal rules for integrating evidence of different types, the rules are based essentially on scientific intuition and ex-

perience in a given domain. In many settings, integrating the evidence requires estimating a number or a set of numbers that can summarize the information obtained from various sources. For example, in the context of IRIS assessments, one needs to estimate the magnitude of harm potentially caused by a chemical and the uncertainty of the estimate.

A number of quantitative approaches can be used for hazard identification. Three approaches are meta-analysis, probabilistic bias analysis, and Bayesian analysis.[3] In the case of meta-analysis and probabilistic bias analysis, the natural targets of the analyses are not qualitative yes-no questions but rather quantitative estimates of an effect size. In both cases, however, the key question is whether the estimate of the effect size can reasonably be inferred to exclude zero or to be negligible. If so, one can conclude hazard. If not, there is not adequate evidence to conclude hazard, but there might be evidence that suggests hazard. Bayesian models can be used to produce quantitative judgments, for example, "There is at least a 60% chance that chemical X is a human carcinogen." Quantitative judgments are easily converted into qualitative categorical judgments as shown, for example, in Table 6-5. The committee emphasizes that the numbers provided in the table are arbitrary and are meant only as illustration. They are not taken from an existing source, nor do they reflect any recommendation by the committee.

Meta-analysis and probabilistic bias analysis, as they are typically carried out, produce effect-size estimates and confidence intervals around them. Converting an estimate and its accompanying confidence interval into a quantitative judgment about hazard is not as straightforward as it is in a Bayesian analysis, but it can be done. A vast literature and excellent textbooks are devoted to each approach. Here, a brief discussion of the methods and their relative advantages and disadvantages is provided. See Appendix C for a primer on the Bayesian approach.

Meta-Analysis

Meta-analysis is a broad term that encompasses statistical methods of combining data from similar studies. Typically, meta-analysis is used to estimate the effects of an exposure on the risk of an outcome. In its simplest form, a meta-analysis combines the effect estimates from several studies into a single weighted estimate that is accompanied by a 95% confidence interval that reflects the pooled data.

The primary goal of a meta-analysis is to integrate rigorously a set of *similar* studies with respect to a single estimate of the size of an effect and to the uncertainty due to random error. In fixed-effect meta-analysis, investigators assume that all studies are estimating a common causal effect, and the pooled estimate is simply a more precise estimate of the common effect. In random-effects meta-analysis, investigators assume sizable variation in effect size among studies, and the pooled estimate summarizes the mean of the distribution of the individual estimates of effect size. In both cases, investigators are not required to have information or hypotheses about the magnitude of systematic biases. Expert knowledge about the causal mechanisms by which exposure or other variables affect the outcome also is not required. However, it is worth noting that meta-analysis does not correct for or "fix" biases; indeed, it is possible for all studies in a meta-analysis to be biased in the same direction because of confounding or selection effects.

Meta-analysis is typically used as a technique to combine the results of similar randomized clinical trials, but it can be applied to results of epidemiologic studies. Meta-regression (Greenland and O'Rourke 2001) allows pooling of data from epidemiologic studies with some unexplained heterogeneity, and Kaizar (2005, 2011) and Roetzheim et al. (2012) improve on meta-regression for situations in which data are available from randomized clinical trials and epidemiologic studies. Bayesian methods are also used to conduct meta-analyses and are commonly used in network meta-analyses in which many agents are compared simultaneously (Cipriani et al. 2009).

[3]Both meta-analysis and probabilistic bias analysis can be done in a Bayesian framework.

TABLE 6-5 Example Conversion of Quantitative Output to Qualitative Categorical Judgments

Chance that Chemical X is a Carcinogen	Categorical Judgment
> 90%	Carcinogenic in humans
≤ 90% to > 75%	Likely to be carcinogenic in humans
≤ 75% to > 50%	Suggestive evidence of carcinogenicity
≤ 50% to > 5%	Inadequate information
≤ 5%	Not likely to be carcinogenic in humans

Although meta-analytic methods have generated extensive discussion (see, for example, Berlin and Chalmers 1988; Dickersin and Berlin 1992; Berlin and Antman 1994; Greenland 1994; Stram 1996; Stroup et al. 2000; Higgins et al. 2009), they can be useful when there are similar studies on the same question. For example, the 2006 IOM Committee on Asbestos and Selected Cancers (IOM 2006) did a quantitative meta-analysis on asbestos and cancer risk and presented an overall estimate that was derived from the combination of the estimates from the individual studies for each cancer type.

Probabilistic Bias Analysis

In all studies that seek to estimate causal effects, there are two broad sources of uncertainty: systematic bias and random error from sampling. In the famous poll that predicted that Thomas Dewey had beaten Harry Truman in the 1948 presidential election, there was systematic bias related to the sampling and the external validity of the survey; it was a telephone poll, telephone ownership was not ubiquitous at that time, and telephone ownership was heavily skewed toward Dewey supporters. The systematic bias was severe enough to dwarf uncertainty that was due to sample variability. There is still some systematic bias in modern presidential polls, but it is much smaller. When poll results are reported as accurate to within ± 3%, this number represents only variation in the reported number due to sampling variability (random error); it does not include systematic bias. Similarly, the confidence intervals in meta-analysis reflect only uncertainty that is due to random error from sampling. However, the possible presence of systematic bias due to various types of bias discussed in Chapter 5 can be another important source of uncertainty around effect estimates. The uncertainty that is due to systematic bias is well recognized by investigators and is usually a central part of the discussion section of scientific articles.

Methods collectively referred to as quantitative or probabilistic bias analysis produce intervals around the effect estimate that integrate uncertainty that is due to random and systematic sources. If empirical data on the direction and magnitude of systematic biases are unavailable, investigators need to use their expert knowledge to make quantitative assumptions about systematic bias. See the excellent books by Lash et al. (2009) and Rosenbaum (2010) for details.

The Bayesian Approach

Whether the uncertainty in a meta-analysis includes only random sampling error or also includes systematic bias, it is still limited to combining statistical evidence from *similar* studies into a single statistical estimate of effect size. A technique for combining all the available evidence into a single judgment needs to accommodate human studies, animal studies, and mechanistic analyses. One approach for doing so is to build a Bayesian model (Berry and Stangl 1996; Peters et al. 2005; Kadane 2011). The Bayesian approach has been used extensively in evaluating clinical data and in regulatory decision-making (Etzioni and Kadane 1995; Parmigiani 2002; Kadane 2005; DuMouchel 2012) and has several general advantages and disadvantages.

Regarding advantages, the Bayesian model is built to calculate, on the basis of prior knowledge and new data, how likely a hypothesis is to be true or false. It provides an opportunity to include as much rigor in constructing a formal model of evidence integration and uncertainty as one wants, and it comes with a type of theoretical guarantee. If experts are not dogmatic and agree on the fundamental design of a model and update their opinions with a Bayesian model, their opinions will eventually converge.

Because it supports the explicit modeling of all types of uncertainty, not only uncertainty due to sampling variability, a Bayesian model can help to identify the specific gaps in knowledge that make a large difference in overall uncertainty. For example, one might learn from a Bayesian model that measurement error of exposure in a series of epidemiologic studies produces far more uncertainty in a final estimate of toxicity than does uncertainty related to cross-species (rodent to human) extrapolation.

Regarding disadvantages, building a Bayesian model requires the elicitation and modeling of expert opinion. Although a large literature exists on elicitation (see, for example, Chaloner 1996; Kadane and Wolfson 1998), it requires expertise that is not typically possessed by a biostatistician or epidemiologist.

Overall, the Bayesian approach is being adopted by a growing number of scientists and regulatory agencies. For example, the IOM report *Ethical and Scientific Issues in Studying the Safety of Approved Drugs* endorses the Bayesian approach as providing "decision-makers with useful quantitative assessments of evidence" (IOM 2012, p. 159). FDA's Center for Devices and Radiological Health has published explicit guidelines on using Bayesian methods in regulatory decision-making (FDA 2010), and they are used increasingly in legal settings (Kadane and Terrin 1997; Perlin et al. 2009; Woodworth and Kadane 2010).

In the Bayesian approach, probability is typically treated as a degree of belief. Any proposition, (that is, any statement that is either true or false) can be given a degree of belief, including a proposition regarding hazard, that is, that a chemical causes some sort of specific human harm, such as lung cancer or heart disease. If "H" notates a proposition about hazard—exposure to methanol causes blindness—then "~H" notates the opposite—exposure to methanol does not cause blindness.[4] Before seeing any evidence, one might ask a scientist to express his or her "prior" degree of belief in H. Scientist A might say that H is 75% likely, and this would translate to $P_A(H) = 0.75$. Scientist B might say that H is only 40% likely, and this would translate to $P_B(H) = 0.4$. In the Bayesian approach, the goal is to compute the "posterior probability" of H *after* seeing evidence E, which is notated as $P(H \mid E)$. If E favored H, the two scientists' posterior probability might be closer than when they started: $P_A(H \mid E) = 0.85$ and $P_B(H \mid E) = 0.65$.

In hazard identification, which is essentially a qualitative yes-no answer to a causal question, one would use the Bayesian approach to assess the probability of hazard, that is, the degree of belief in a causal proposition, after seeing all the evidence (human, animal, and mechanistic). In dose-response estimation, the target is not a yes-no proposition but rather a more complicated quantity: What is the quantitative dependence of disease response on dose. In the simplest possible case, the relationship might be linear, so that the amount of extra disease burden that one could expect can be expressed with one extra unit of dose exposure: $\text{Disease} = \beta_0 + (\beta_1 \times \text{Dose})$. In this case, the parameter β_1 expresses the dose-response relationship. If β_1 is 0, there is no effect. If β_1 is large, there is a large effect. In a Bayesian analysis for β_1, the output would be

[4]One complication with this approach is that it forces one to collapse all degrees of causation into a single yes–no proposition. It forces one to make the same distinction between causes that have an extremely small effect vs no effect at all and other causes that have a substantial effect vs no effect at all. One solution is to let H stand for a proposition, such as that chemical X is an appreciable or substantial or meaningful cause of harm Y, where the term *appreciable*, *substantial*, or *meaningful* would have to be defined. If one equates an effect-size interval, such as greater than 0.1, with the idea of *appreciable*, a Bayesian analysis can also quantify the probability that a chemical X is an *appreciable* cause of harm Y.

$P(\beta_1 \mid E)$—that is, a probability distribution over all the possible values of β_1—when one has seen the evidence E.

In Figure 6-2, for example, β_1 is shown to range from 0 to 70,000. In the blue "prior" distribution, the mode is about 35,000, and the distribution is wide, demonstrating considerable uncertainty. In the green "posterior" distribution, the mode is below 20,000, and the distribution is much narrower, representing a reduction in uncertainty. Chapter 7 discusses a Bayesian approach to dose-response estimation and its attendant uncertainty in more detail. In the present chapter, the discussion is restricted to a Bayesian approach to hazard identification, which involves a yes-no proposition: Does chemical X cause outcome Y?

To combine evidence from disparate studies, a Bayesian approach needs to model the likelihood of data or evidence from different kinds of studies, given the hypothesis that a chemical is hazardous to humans. The approach must explicitly model the relevance that each kind of evidence has to the overall question of human hazard and how much uncertainty accompanies the modeling assumptions that allow us to relate disparate studies to the common target of human hazard. For example, if one in vivo animal study shows that a chemical poses a hazard, it is relevant to the question of human health only insofar as the animal model for this chemical and this outcome is relevant to humans. Almost every IRIS assessment that involves animal data must deal with the question of whether the animal model is relevant to humans. Rather than incorporate expert opinion about this question informally, a Bayesian hierarchical model can explicitly incorporate data from previous studies about cross-species relevance or mechanistic similarity and use them to derive overall estimates and uncertainties. In the early 1980s, for example, DuMouchel and Harris (1983) showed how to combine human and animal studies of radium toxicity to derive the evidential signal of animal studies of plutonium toxicity in terms of how it bears on the target: the toxicity of plutonium to humans. More recent work by Jones et al. (2009) and Peters et al. (2005) shows how to combine epidemiologic and toxicologic evidence in a Bayesian model. The report *Biological Effects of Ionizing Radiation (BEIR) IV* (NRC 1988), which sought to estimate the carcinogenicity of plutonium in humans, adopted a Bayesian approach, which included an uncertainty analysis that incorporated the variability in the ratios of relative carcinogenicities of different radionuclides among species. That analysis revealed that although there were few human data on plutonium, they could be combined with animal data to estimate carcinogenicity in humans effectively.

In the IRIS assessment of methanol, uncertainty about animal-model relevance plays a large role. Studies show species differences in the rates at which rodents and humans metabolize methanol into formic acid, which produces acidosis and causes lasting CNS damage. Those interspecies differences could be explicitly modeled in a Bayesian model, and the uncertainty estimate would incorporate them.

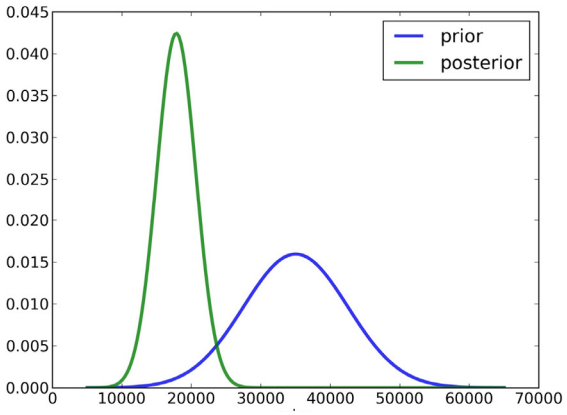


FIGURE 6-2 Bayesian estimate of β_1.

There is similarly uncertainty about the relationship between adult humans, infant humans, and rodents in how they metabolize methanol. Adult humans primarily use alcohol dehydrogenase (ADH1), whereas rodents use ADH1 and catalase to metabolize methanol. It is not known whether infant humans, like rodents, use catalase to metabolize methanol. The uncertainty about methanol metabolism could be included in a Bayesian model, and its effect on overall uncertainty could be computed by incorporating the relevance of rodent studies to developmental toxicity.

Uncertainty in human studies is equally amenable to a Bayesian analysis. Models can explicitly include uncertainty about unmeasured confounding, about measurement error in exposure, and about any other risk of bias in an epidemiologic study.

In principle, Bayesian methods provide a quantitative framework for combining theoretical understanding and evidence from human, animal, and mechanistic studies with data to update model-parameter estimates or the probability that a particular hypothesis is true. Although the Bayesian approach is growing in popularity in many scientific arenas, it is still not perceived as being widely applicable or widely used in public health, partly because the computational demands imposed by the method were prohibitive a decade ago. There also have been many conceptual misunderstandings regarding its subjective nature, and reliably eliciting expert knowledge and converting it into model parameters is difficult and takes special expertise. The computational worries have largely been resolved. Enormous computational advances have taken place over the last 15 years, and several software platforms are available for carrying out sophisticated Bayesian modeling (for example, BUGS[5]). Eliciting expert opinion is time-consuming and in some cases difficult, but there is now a considerable literature on how it should be done and a considerable number of cases in which it has been done successfully (Chaloner 1996; Kadane and Wolfson 1998; Hiance et al. 2009; Kuhnert et al. 2010).

Quantitative models for integrating evidence are powerful tools that can answer a wide array of scientific questions. Their obvious downside is that model misspecification at any level can result in incorrect inferences. Nevertheless, they make rigorous what other techniques have to make heuristic, and they force scientists to make their assumptions explicit in ways that less formal methods do not.

Comparison of Quantitative Methods

Meta-analysis is appropriate for situations in which there are a number of similar statistical studies involving experiments on humans or animals or similar epidemiologic studies. Probabilistic bias analysis is appropriate when the risk of bias in observational studies is substantial, and there is information that makes estimating or at least bounding such bias feasible. A Bayesian analysis seems appropriate when the stakes are high and when the uncertainty is substantial, especially when the evidence is to some degree inconsistent. For example, when a chemical is fairly common in the environment and might have serious health effects and the relevant evidence is difficult to integrate because human studies show little or no association and animal studies show toxicity, a Bayesian analysis can help to weight the evidence provided by both study types and characterize uncertainty appropriately.

A Template for the Evidence-Integration Narrative

No matter what method is used to integrate the different kinds of evidence available for an IRIS assessment, using a template for the evidence-integration narrative could help to make IRIS assessments more transparent. In particular, an evidence-integration narrative can make clear EPA's view on the strength of the case for or against a specific hazard when all the available evidence is taken into account.

[5]See http://www.mrc-bsu.cam.ac.uk/bugs/.

Rather than organize the narrative around a checklist of criteria, such as the Hill criteria, EPA might consider organizing the narrative as an *argument* for or against hazard on the basis of available evidence. It should be qualified by explicitly considering alternative hypotheses, uncertainty, and gaps in knowledge. Elements of the Hill criteria will undoubtedly find their way into such arguments and might even help to organize some of the discussion supporting the argument, but they need not be required topics to be discussed in every evidence narrative.

If the narrative is organized around types of evidence, it might begin by considering the conclusions supported by the human evidence and then consider how the available animal evidence confirms, does not support, or is irrelevant to the conclusions. Mechanistic evidence, if available, should be used in the discussion of the animal evidence to determine whether the animal evidence is relevant to the claim about human hazard. Gaps in knowledge and important uncertainties should be explicitly included.

Both the benzo[a]pyrene and methanol draft IRIS assessments contain narratives that mostly satisfy that sort of template. Both build a case for a variety of different cancer and noncancer end points and leave the reader with a clear sense of the evidence available that is relevant to the end points and thus the strength of the case for each end point. Where the narratives are particularly effective, they explain specifically how different strands of evidence connect. For example, the assessment of methanol explains that CNS toxicity has been observed in humans but not in rodents but then goes on to explain the differences in the rates at which humans and rodents eliminate formic acid, which explain the apparent evidential discrepancy. What is missing and might be desirable is a more systematic discussion of gaps in knowledge and gaps in the evidence.

FINDINGS AND RECOMMENDATIONS

Finding: Critical considerations in evaluating a method for integrating a diverse body of evidence for hazard identification are whether the method can be made transparent, whether it can be feasibly implemented under the sorts of resource constraints evident in today's funding environment, and whether it is scientifically defensible.

Recommendation: EPA should continue to improve its evidence-integration process incrementally and enhance the transparency of its process. It should either maintain its current guided-expert-judgment process but make its application more transparent or adopt a structured (or GRADE-like) process for evaluating evidence and rating recommendations along the lines that NTP has taken. If EPA does move to a structured evidence-integration process, it should combine resources with NTP to leverage the intellectual resources and scientific experience in both organizations. The committee does not offer a preference but suggests that EPA consider which approach best fits its plans for the IRIS process.

Finding: Quantitative approaches to integrating evidence will be increasingly needed by and useful to EPA.

Recommendation: EPA should expand its ability to perform quantitative modeling of evidence integration; in particular, it should develop the capacity to do Bayesian modeling of chemical hazards. That technique could be helpful in modeling assumptions about the relevance of a variety of animal models to each other and to humans, in incorporating mechanistic knowledge to model the relevance of animal models to humans and the relevance of human data for similar but distinct chemicals, and in providing a general framework within which to update scientific knowledge rationally as new data become available. The committee emphasizes that the capacity for quantitative modeling should be developed in parallel with improvements in existing IRIS evidence-integration procedures and that IRIS assessments should not be delayed while this capacity is being developed.

Finding: EPA has instituted procedures to improve transparency, but additional gains can be achieved in this arena. For example, the draft IRIS preamble provided to the committee states that "to make clear how much the epidemiologic evidence contributes to the overall weight of the evidence, the assessment may select a standard descriptor to characterize the epidemiologic evidence of association between exposure to the agent and occurrence of a health effect" (EPA 2013a, p. B-6). A set of descriptor statements was provided, but they were not used in the recent IRIS draft assessments of methanol and benzo[a]pyrene.

Recommendation: EPA should develop templates for structured narrative justifications of the evidence-integration process and conclusion. The premises and structure of the argument for or against a chemical's posing a hazard should be made as explicit as possible, should be connected explicitly to evidence tables produced in previous stages of the IRIS process, and should consider all lines of evidence (human, animal, and mechanistic) used to reach major conclusions.

Finding: EPA guidelines for evidence integration for cancer and noncancer end points are different; the cancer guidelines are more developed and more specific.

Recommendation: Guidelines for evidence integration for cancer and noncancer end points should be more uniform.

REFERENCES

Berlin, J.A., and E.M. Antman. 1994. Advantages and limitations of metaanalytic regressions of clinical trials data. Online J. Curr. Clin. Trials, Document No. 134.

Berlin, J., and T.C. Chalmers. 1988. Commentary on meta-analysis in clinical trials. Hepatology 8(3):690-691.

Berry, D.A., and D.K. Stangl, eds. 1996. Bayesian Biostatistics. New York, NY: Marcel Dekker.

Burbacher, T.M., D. Shen, K. Grant, L. Sheppard, D. Damian, S. Ellis, and N. Liberato. 1999a. Reproductive and Offspring Developmental Effects Following Maternal Inhalation Exposure to Methanol in Nonhuman Primates, Part I. Methanol Disposition and Reproductive Toxicity in Adult Females. Research Report No. 89. Health Effects Institute, Cambridge, MA (as cited in EPA 2013b).

Burbacher, T.M., K. Grant, D. Shen, D. Damian, S. Ellis, and N. Liberato. 1999b. Reproductive and Offspring Developmental Effects Following Maternal Inhalation Exposure to Methanol in Nonhuman Primates, Part II. Developmental Effects in Infants Exposed Prenatally to Methanol. Research Report No. 89. Health Effects Institute, Cambridge, MA (as cited in EPA 2013b).

Burbacher, T.M., D.D. Shen, B. Lalovic, K.S. Grant, L. Sheppard, D. Damian, S. Ellis, and N. Liberato. 2004a. Chronic maternal methanol inhalation in nonhuman primates (*Macaca fascicularis*): Exposure and toxicokinetics prior to and during pregnancy. Neurotoxicol. Teratol. 26(2):201-221(as cited in EPA 2013b).

Burbacher, T.M., K.S. Grant, D.D. Shen, L. Sheppard, D. Damian, S. Ellis, and N. Liberato. 2004b. Chronic maternal methanol inhalation in nonhuman primates (*Macaca fascicularis*): Reproductive performance and birth outcome. Neurotoxicol. Teratol. 26(5):639-650 (as cited in EPA 2013b).

Chaloner, K. 1996. Elicitation of prior distributions. Pp. 141-156 in Bayesian Biostatistics, D.A. Berry, and D.K. Stangl, eds. New York: Marcel Dekker.

Cipriani, A., T.A. Furukawa, G. Salanti, J.R. Geddes, J.P. Higgins, R. Churchill, N. Watanabe, A. Nakagawa, I.M. Omori, H. McGuire, M. Tansella, and C. Barbui. 2009. Comparative efficacy and acceptability of 12 new-generation antidepressants: A multiple-treatments meta-analysis. Lancet 373(9665):746-758.

DHHS (U.S. Department of Health and Human Services). 2004. The Health Consequences of Smoking: A Report of the Surgeon General. U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health, Atlanta, GA [online]. Available: http://www.cdc.gov/tobacco/data_statistics/sgr/2004/index.htm [accessed December 18, 2013].

Dickersin, K., and J.A. Berlin. 1992. Meta-analysis: State-of-the-science. Epidemiol Rev. 14(1):154-176.

DuMouchel, W. 2012. Multivariate Bayesian Logistic Regression for analysis of clinical study safety issues. Stat. Sci. 27(3):319-339.

DuMouchel, W.H., and J.E. Harris. 1983. Bayes methods for combining the results of cancer studies in humans and other species: Rejoinder. J. Am. Stat. Assoc. 78(382):313-315.

EPA (U.S. Environmental Protection Agency). 2002. A Review of the Reference Dose and Reference Concentration Processes. EPA/630/P-02/002F. Risk Assessment Forum, U.S. Environmental Protection Agency. Washington, DC [online]. Available: http://www.epa.gov/raf/publications/pdfs/rfd-final.pdf [accessed December 18, 2013].

EPA (U.S. Environmental Protection Agency). 2005. Guidelines for Carcinogen Risk Assessment. EPA/630/P-03/001F. Risk Assessment Forum, U.S. Environmental Protection Agency, Washington, DC. March 2005 [online]. Available: http://www.epa.gov/raf/publications/pdfs/CANCER_GUIDELINES_FINAL_3-25-05.PDF [accessed October 3, 2013].

EPA (U.S. Environmental Protection Agency). 2009. The U.S. Environmental Protection Agency's Strategic Plan for Evaluating the Toxicity of Chemicals. EPA/100/K-09/001. Office of Science Advisor, Science Policy Council, U.S. Environmental Protection Agency, Washington, DC [online]. Available: http://www.epa.gov/spc/toxicitytesting/docs/toxtest_strategy_032309.pdf [accessed Feb. 21, 2014].

EPA (U.S. Environmental Protection Agency). 2010. Integrated Science Assessment for Carbon Monoxide. EPA/600/R-09/019F. National Center for Environmental Assessment-RTP Division, Office of Research and Development, U.S. Environmental Protection Agency, Research Triangle Park, NC [online]. Available: http://cfpub.epa.gov/ncea/cfm/recordisplay.cfm?deid=218686 [accessed December 18, 2013].

EPA (U.S. Environmental Protection Agency). 2011. Glossary of Key Terms. U.S. Environmental Protection Agency [online]. Available: http://www.epa.gov/ttn/atw/natamain/gloss1.html [accessed October 3, 2013].

EPA (U.S. Environmental Protection Agency). 2013a. Part 1. Status of Implementation of Recommendations. Materials Submitted to the National Research Council, by Integrated Risk Information System Program, U.S. Environmental Protection Agency, January 30, 2013 [online]. Available: http://www.epa.gov/iris/pdfs/IRIS%20Program%20Materials%20to%20NRC_Part%201.pdf [accessed November 13, 2013].

EPA (U.S. Environmental Protection Agency). 2013b. Toxicological Review of Methanol (Noncancer) (CAS No. 67-56-1) in Support of Summary Information on the Integrated Risk Information System (IRIS). EPA/635/R-11/001Fa. U.S. Environmental Protection Agency, Washington, DC. September 2013 [online]. Available: http://www.epa.gov/iris/toxreviews/0305tr.pdf [accessed October 3, 2013].

EPA (U.S. Environmental Protection Agency). 2013c. Toxicological Review of Benzo[a]pyrene (CAS No. 50-32-8) in Support of Summary Information on the Integrated Risk Information System (IRIS), Public Comment Draft. EPA/635/R13/138a. National Center for Environmental Assessment, Office of Research and Development, U.S. Environmental Protection Agency, Washington, DC. August 2013 [online]. Available: http://cfpub.epa.gov/ncea/iris_drafts/recordisplay.cfm?deid=66193 [accessed November 13, 2013].

Etzioni, R.D., and J.B. Kadane. 1995. Bayesian statistical methods in public health and medicine. Annu. Rev. Public Health 16(1):23-41.

FDA (Food and Drug Administration). 2010. Guidance for Industry and FDA Staff: Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials. U.S. Department of Health and Human Services, Food and Drug Administration [online]. Available: http://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm071121.pdf [accessed December 16, 2013].

Glass, T.A., S.N.Goodman, M.A. Hernán, and J.M. Samet. 2013. Causal inference in public health. Ann. Rev. Pub. Health 34: 61-75.

Greenland, S. 1994. A critical look in some popular meta-analytical methods. Am. J. of Epidemiol. 140(3):290-296.

Greenland, S., and K. O'Rourke. 2001. On the bias produced by quality scores in meta-analysis, and a hierarchical view of proposed solutions. Biostatistics 2(4):463-471.

Guyatt, G.H., A.D. Oxman, E.A. Akl, R. Kunz, G. Vist, J. Brozek, S. Norris, Y. Falck-Ytter, P. Glasziou, H. DeBeer, R. Jaeschke, D. Rind, J. Meerpohl, P. Dahm, and H.J. Schunemann. 2011a. GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. J. Clin. Epidemiol. 64(4):383-394.

Guyatt, G.H., A.D. Oxman, R. Kunz, J. Brozek, P. Alonso-Coello, D. Rind, P.J. Devereaux, V.M. Montori, B. Freyschuss, G. Vist, R. Jaeschke, J.W. Williams, Jr., M.H. Murad, D. Sinclair, Y. Falck-Ytter, J. Meerpohl, C. Whittington, K. Thorlund, J. Andrews, and H.J. Schunemann. 2011b. GRADE guidelines 6. Rating the quality of evidence—imprecision. J. Clin. Epidemiol. 64(12):1283-1293.

Guyatt, G.H., A.D. Oxman, R. Kunz, J. Woodcock, J. Brozek, M. Helfand, P. Alonso-Coello, Y. Falck-Ytter, R. Jaeschke, G. Vist, E.A. Akl, P.N. Post, S. Norris, J. Meerpohl, V.K. Shukla, M. Nasser, and

H.J. Schunemann. 2011c. GRADE guidelines: 8. Rating the quality of evidence—indirectness. J. Clin. Epidemiol. 64(12):1303-1310.

Guyatt, G.H., A.D. Oxman, V. Montori, G. Vist, R. Kunz, J. Brozek, P. Alonso-Coello, B. Djulbegovic, D. Atkins, Y. Falck-Ytter, J.W. Williams, Jr., J. Meerpohl, S.L. Norris, E.A. Akl, and H.J. Schunemann. 2011d. GRADE guidelines: 5. Rating the quality of evidence-publication bias. J. Clin. Epidemiol. 64(12):1277-1282.

Guyatt, G.H., A.D. Oxman, G. Vist, R. Kunz, J. Brozek, P. Alonso-Coello, V. Montori, E.A. Akl, B. Djulbegovic, Y. Falck-Ytter, S.L. Norris, J.W. Williams, Jr., D. Atkins, J. Meerpohl, and H.J. Schunemann. 2011e. GRADE guidelines: 4. Rating the quality of evidence-study limitations (risk of bias). J. Clin. Epidemiol. 64(4):407-415.

Hiance, A., S. Chevret, and V. Lévy. 2009. A practical approach for eliciting expert prior beliefs about cancer survival in phase III randomized trial. J. Clin. Epidemiol. 62(4):431-437.

Higgins, J.P.T., and S. Green, eds. 2011. Cochrane Handbook for Systematic Reviews of Interventions, Version 5.1.0. The Cochrane Collaboration [online]. Available: http://handbook.cochrane.org/ [accessed December 11, 2013].

Higgins, J.P., S.G. Thompson, and D.J. Spiegelhalter. 2009. A re-evaluation of random-effects meta-analysis. J. R. Stat. Soc. Ser. A 172(1):137-159.

Hill, A.B. 1965. The environment and disease: Association or causation? Proc. R. Soc. Med. 58(5):295-300

IARC (International Agency for Research on Cancer). 2006. IARC Monographs on the Evaluation of Carcinogenic Risks to Humans: Preamble. Lyon, France: IARC Press [online]. Available: http://monographs.iarc.fr/ENG/Preamble/CurrentPreamble.pdf [accessed October 6, 2013].

IARC (International Agency for Research and Cancer). 2011. Guidelines for Observers at IARC Monograph Meetings [online]. Available: http://monographs.iarc.fr/ENG/Meetings/ObsGuide0111.php [accessed December 18, 2013].

IOM (Institute of Medicine). 2006. Asbestos: Selected Cancers. Washington, DC: National Academies Press.

IOM (Institute of Medicine). 2012. Ethical and Scientific Issues in Studying the Safety of Approved Drugs. Washington, DC: The National Academies Press.

Jones, D.R., J. Peters., J.L. Rushton, A.J. Sutton, and K.R. Abrams. 2009. Interspecies extrapolation in environmental exposure standard setting: A Bayesian synthesis approach. Regul. Toxicol. Pharmacol. 53(3):217-225.

Kadane, J.B. 2005. Bayesian methods for health-related decision making. Stat. Med. 24(4):563-567.

Kadane, J.B. 2011. Principles of Uncertainty. Boca Raton, FL: Chapman and Hall/CRC.

Kadane, J.B., and N. Terrin. 1997. Missing data in the forensic context. J.R. Stat. Soc. A 160(2):351-357.

Kadane, J.B., and L.J. Wolfson. 1998. Experiences in elicitation. J. R. Stat. Soc. D-Sta. 47(1):3-19.

Kaizar, E.E. 2005. Meta-analyses are observational studies: How lack of randomization impacts analysis. Am. J. Gastroenterol. 100(6):1233-1236.

Kaizar, E.E. 2011. Estimating treatment effect via simple cross design synthesis. Stat. Med. 30(25):2986–3009.

Kuhnert, P.M., T.G. Martin, and S.P. Griffiths. 2010. A guide to eliciting and using expert knowledge in Bayesian ecological models. Ecol. Lett. 13(7):900-914.

Lash, T.L., M.P. Fox, and A.K. Fink. 2009. Applying Quantitative Bias Analysis to Epidemiologic Data. New York: Springer.

Meek, M.E., J. Patterson, J.E. Strawson, and R.G. Liteplo. 2007. Engaging expert peers in the development of risk assessments. Risk Anal. 27(6):1609-1621.

NEDO (New Energy Development Organization). 1987. Toxicological Research of Methanol as a Fuel for Power Station: Summary Report on Tests with Monkeys, Rats and Mice. Technical Report. New Energy Development Organization, Tokyo, Japan (as cited in EPA 2013b).

NRC (National Research Council). 1983. Risk Assessment in the Federal Government: Managing the Process. Washington, DC: National Academy Press.

NRC (National Research Council). 1988. Health Risks of Radon and Other Internally Deposited Alpha-emitters (Beir IV). Washington, DC: National Academy Press.

NRC (National Research Council). 2007. Toxicity Testing in the 21st Century: A Vision and a Strategy. Washington, DC: National Academies Press.

NRC (National Research Council). 2011. Review of the Environmental Protection Agency's Draft IRIS Assessment of Formaldehyde. Washington, DC: National Academies Press.

NTP (National Toxicology Program). 2013. Draft OHAT Approach for Systematic Review and Evidence Integration for Literature-Based Health Assessments – February 2013. U.S. Department of Health and Hu-

man Services, National Institute of Health, National Institute of Environmental Health Sciences, Division of the National Toxicology Program [online]. Available: http://ntp.niehs.nih.gov/NTP/OHAT/EvaluationProcess/DraftOHATApproach_February2013.pdf [accessed December 11, 2013].

Oxford Dictionaries. 2011. Oxford English Dictionary online. Oxford University Press (as cited in IOM 2012).

Parmigiani, G. 2002. Modeling in Medical Decision Making: A Bayesian Approach. Chichester: John Wiley & Sons.

Perlin, M.W., J.B. Kadane, and R.W. Cotton. 2009. Match likelihood ratio for uncertain genotypes. Law Prob. Risk 8(3):289-302.

Peters, J.L., L. Rushton, A.J. Sutton, D.R. Jones, K.R. Abrams, and M.A. Mugglestone. 2005. Bayesian methods for the cross-design synthesis of epidemiological and toxicological evidence. J. R. Stat. Soc. C-Appl. Stat. 54(1):159-172.

Rhomberg, L.R., J.E. Goodman, L. Bailey, R.L. Prueitt, N.B. Beck, C. Bevin, M. Honeycutt, N.E. Kaminski, G. Paoli, L.H. Pottenger, R.W. Scherer, K.C. Wise, and R.A. Becker. 2013. A survey of frameworks for best practices in weight of evidence analysis. Crit. Rev. Toxicol. 43(9):753-784.

Roetzheim, R.G., K.M. Freund, D.K. Corle, D.M. Murray, F.R. Snyder, A.C. Kronman, P. Jean-Pierre, P.C. Raich, A.E. Holden, J.S. Darnell, V. Warren-Mears, and S. Patierno. 2012. Analysis of combined data from heterogeneous study designs: An applied example from the patient navigation research program. Clin. Trials 9(2):176-187.

Rogers, J.M., M.L. Mole, N. Chernoff, B.D. Barbee, C.I. Turner, T.R. Logsdon, and R.J. Kavlock. 1993b. The developmental toxicity of inhaled methanol in the CD-1 mouse, with quantitative dose-response modeling for estimation of benchmark doses. Teratology 47(3):175-188 (as cited in EPA 2013b).

Rosenbaum, P.R. 2010. Observational Studies, 2nd Ed. New York: Springer.

Rossouw, J.E., G.L. Anderson, R.L. Prentice, A.Z. LaCroix, C. Kooperberg, M.L. Stefanick, R.D. Jackson, S.A. Beresford, B.V. Howard, K.C. Johnson, J.M. Kotchen, and J. Ockene. 2002. Risks and benefits of estrogen plus progestin in healthy postmenopausal women: Principal results From the Women's Health Initiative randomized controlled trial. JAMA 288(3):321-333.

Rothman, K.J., and S. Greenland. 2005. Causation and causal inference in epidemiology. Am. J. Public Health 95(suppl. 1):S144-S150.

Schünemann, H., S. Hill, G. Guyatt, E.A. Akl, and F. Ahmed. 2011. The GRADE approach and Bradford Hill's criteria for causation. J. Epidemiol. Community Health 65(5):392-395.

Stram, D.O. 1996. Meta-analysis of published data using a linear mixed-effects model. Biometrics 52(2):536-544.

Stroup, D.F., J.A. Berlin, S.C. Morton, I. Olkin, G.D. Williamson, D. Rennie, D. Moher, B.J. Becker, T.A. Sipe, and S.B. Thacker. 2000. Meta-analysis of observational studies in epidemiology: A proposal for reporting. JAMA 283(15):2008-2012.

Woodruff, T.J., and P. Sutton. 2011. An evidence-based medicine methodology to bridge the gap between clinical and environmental health sciences. The Navigation Guide Work Group. Health Aff. (Millwood) 30(5):931-937.

Woodworth, G., and J. Kadane. 2010. Age- and time-varying proportional hazards models for employment discrimination. Ann. Appl. Stat. 4(3):1139-1157.

7

Derivation of Toxicity Values

This chapter addresses the derivation of quantitative indicators of toxicity in the Integrated Risk Information System (IRIS) process (see Figure 7-1). By this stage in the IRIS process, evidence has been collected and evaluated, clearly defined adverse outcomes have been identified, and different streams of evidence have been integrated for hazard identification. The next phase in an IRIS assessment is to quantify the hazards through the computation of toxicity values—reference doses (RfDs), reference concentrations (RfCs), or unit risks—when the dose-response data support such computation. To clarify discussions in this chapter, the committee provides definitions of some terms and concepts used in this phase in Table 7-1.

The derivation of a toxicity value consists of several steps as depicted in Figure 7-2, which is an expansion of the "Dose-Response Assessment and Derivation of Toxicity Values" box in Figure 7-1. The first step involves the evaluation of the human, animal, and in vitro (mechanistic) studies to determine whether the reported dose-response data are sufficient for dose-response modeling and assessment. Multiple studies with several end points are likely to be considered for dose-response assessment. The next step involves conducting a dose-response assessment and determining a point of departure (POD). The POD is used as the starting point for later extrapolations and analyses. Dose-response modeling is conducted when the data are adequate. In those cases, either an effective dose (ED) for cancer effects or a benchmark dose (BMD) for noncancer effects is calculated with lower confidence limits. If the data are inadequate for modeling, as sometimes occurs for noncancer effects, the dose-response assessment determines a no-observed-adverse-effect level (NOAEL) or a lowest observed-adverse-effect level (LOAEL) when a NOAEL cannot be determined. The next step involves calculation of a unit risk for cancer effects or an RfD or RfC (EPA 2002). Reference values are often calculated by using a BMD (or its lower confidence limit) or a NOAEL (or a LOAEL when a NOAEL is unavailable) and then applying one or more uncertainty factors. Although the NOAEL-LOAEL approach remains in practice, the BMD approach is preferred because it provides and uses dose-response information to a greater extent and reduces uncertainty (EPA 2012).

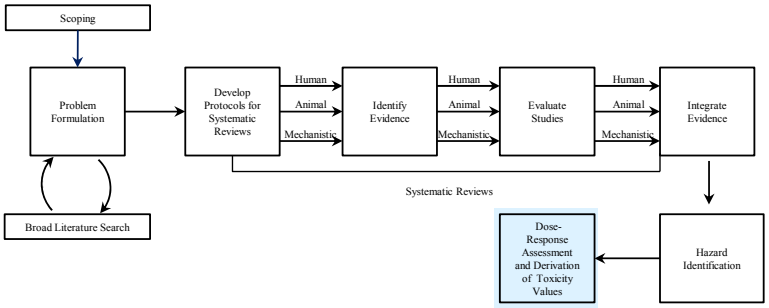


FIGURE 7-1 The IRIS process; the step for dose-response assessment and derivation of toxicity values is highlighted. The committee views public input and peer review as integral parts of the IRIS process, although they are not specifically noted in the figure.

TABLE 7-1 Definitions of Terms Related to Derivation of Toxicity Values[a]

Term	Definition
Dose-response relationship	The relationship between the level of exposure to a chemical (dose) and the probability or magnitude of a biologic response.
Benchmark dose (BMD) or effective dose (ED)	An exposure level determined from a dose-response model that corresponds to a particular level of response, often 1-10% in excess of the control response. The response level is indicated by a subscript; for example, ED_{10} is the effective dose for a given response level that corresponds to a 10% response.
Lower bound benchmark dose (BMDL) or lower bound effective dose (LED)	The lower bound of a confidence interval for the BMD or ED. The response level is indicated by a subscript; for example, LED_{10} is the lower bound of a given response level that corresponds to a 10% response.
No-observed-adverse-effect level (NOAEL)	"The highest exposure level at which there are no biologically significant increases in the frequency or severity of adverse effect between the exposed population and its appropriate control; some effects may be produced at this level, but they are not considered adverse or precursors of adverse effects" (EPA 2013a). In practice, however, it is determined by a statistically significant difference.
Lowest observed-adverse-effect level (LOAEL)	"The lowest exposure level at which there are biologically significant increases in frequency or severity of adverse effects between the exposed population and its appropriate control group" (EPA 2013a). In practice, however, it is determined by a statistically significant difference.
Point of departure (POD)	A BMD or ED or its lower confidence limit or a NOAEL when a BMD is unavailable or a LOAEL when a NOAEL is unavailable. A POD is used as the starting point for later extrapolations and analyses.
Unit risk or slope factor	The increase in the probability of cancer incidence or related risk per unit dose exposure as determined from a POD (effective dose or its lower confidence limit). It is also the slope of an implied linear dose-response relationship below the POD.
Reference dose (RfD)	"An estimate (with uncertainty spanning perhaps an order of magnitude) of a daily oral exposure to the human population (including sensitive subgroups) that is likely to be without an appreciable risk of deleterious effects during a lifetime" (EPA 2013a).
Reference concentration (RfC)	"An estimate (with uncertainty spanning perhaps an order of magnitude) of a continuous inhalation exposure to the human population (including sensitive subgroups) that is likely to be without an appreciable risk of deleterious effects during a lifetime" (EPA 2013a).
Reference value	A reference dose or reference concentration
Central estimate	The "best" estimate of an unknown parameter, such as a BMD. Often determined using maximum likelihood estimation or the posterior mean.
95% confidence interval or bounds	A statistical statement about the most reasonable range of estimates of an unknown parameter, constructed in such a manner that the interval will contain the true value of the parameter with 95% probability when the underlying dataset is replicated.
Lower and upper bound	The two points that define a confidence interval.

[a]Readers might also wish to consult the glossary at http://www.epa.gov/risk_assessment/glossary.htm.

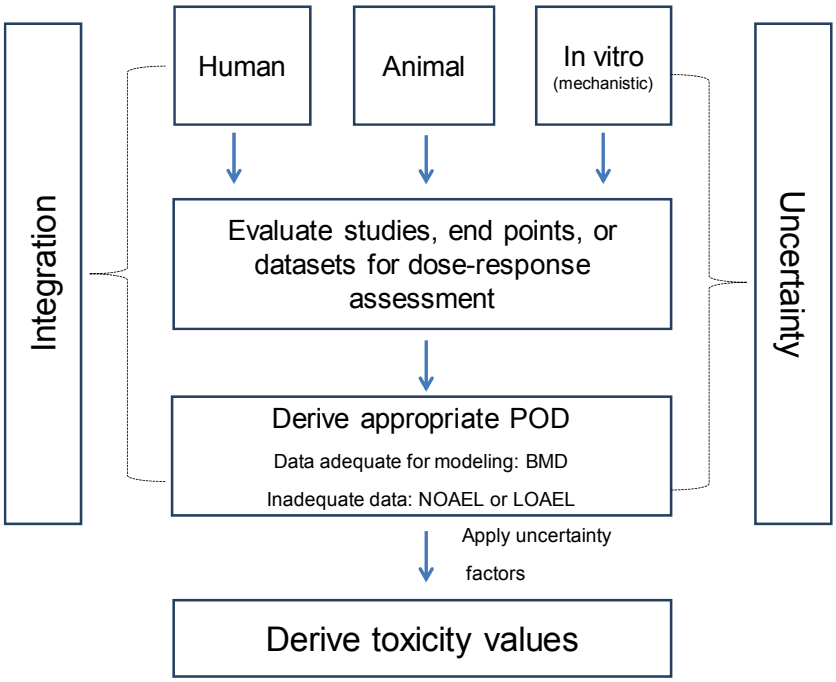


FIGURE 7-2 Derivation of toxicity values. Data integration and uncertainty analysis must be considered in the process.

Toxicity values depend on several factors, including the choice of studies (study design), the health outcomes, the application of dose-response models, the selection of one or more PODs, and the choice of uncertainty factors. Differences among toxicity values are often attributable to *variability* (for example, differences due to species, sex, and age) and *uncertainty* (for example, unknown mechanism of action and choice of dose-response model or POD). Therefore, it is critical to consider systematic approaches to synthesizing and integrating the multitude of toxicity values in light of variability and uncertainty.

Derivation of toxicity values is governed by several Environmental Protection Agency (EPA) guidance documents, including *A Review of the Reference Dose and Reference Concentration Processes* (EPA 2002), *Methods for Derivation of Inhalation Reference Concentrations and Application of Inhalation Dosimetry* (EPA 1994), *Guidelines for Carcinogen Risk Assessment* (EPA 2005a), *Supplemental Guidance for Assessing Susceptibility from Early-Life Exposure to Carcinogens* (EPA 2005b), and *Benchmark Dose Technical Guidance* (EPA 2012). The topic of dose-response assessment was also discussed in the National Research Council (NRC) report *Science and Decisions: Advancing Risk Assessment* (NRC 2009), which identified the need to develop guidance related to the handling of uncertainty and variability and urged development of a unified dose-response assessment framework for chemicals that links the understanding of disease processes, mechanisms, and human heterogeneity in cancer and noncancer outcomes.

This chapter discusses the status of the EPA response to the recommendations in the NRC report *Review of the Environmental Protection Agency's Draft IRIS Assessment of Formaldehyde* (NRC 2011) that have to do with calculation of toxicity values. In addition, the committee describes current practice in deriving toxicity values, suggests some approaches that can help EPA to implement the formaldehyde report recommendations fully, and provides its findings and recommendations.

RECOMMENDATIONS ON CALCULATION OF TOXICITY VALUES IN THE NATIONAL RESEARCH COUNCIL FORMALDEYDE REPORT

The 2011 NRC formaldehyde report recommended that EPA evaluate its methods used in the IRIS process to select studies for the derivation of reference values and unit risks. In particular, the report advocated that EPA "establish clear guidelines for study selection, balance study strengths and weaknesses, weigh human vs experimental evidence, and determine whether combining estimates across studies is warranted" (EPA 2013b, p. 15). The report did not specify the methods that EPA should use to develop guidelines and update its approaches. It also made a number of recommendations related to the calculation of reference values and unit risks (see Box 7-1).

EVALUATION OF ENVIRONMENTAL PROTECTION AGENCY RESPONSE TO THE NATIONAL RESEARCH COUNCIL FORMALDEHYDE REPORT

As described in Chapter 1, the committee reviewed the EPA reports *Status of Implementation of Recommendations* (EPA 2013b) and *Chemical-Specific Examples* (EPA 2013c) and recent draft IRIS assessments for methanol (EPA 2013d) and benzo[a]pyrene (EPA 2013e) to compare progress made against the NRC formaldehyde report recommendations (NRC 2011) regarding calculation of toxicity values. In particular, the committee focused on EPA's response to the major issues noted in the formaldehyde report, which included the need to establish clear guidelines for study selection and to describe, justify, and assess the assumptions and models used in deriving toxicity values.

EPA has made a number of responsive changes in the IRIS program since the publication of the NRC formaldehyde report, including (a) development of a process for study selection that requires transparent documentation of study quality, credibility of the evidence of hazard, and adequacy of quantitative dose-response data for determining a POD; (b) the derivation and graphical presentation of multiple toxicity values; and (c) documentation of the approach for conducting dose-response modeling. The new assessment template (EPA 2013b) includes a streamlined dose-response modeling output and consideration of organ-specific or system-specific and overall toxicity values. EPA has also developed tools and methods for managing data and ensuring quality in dose-response analyses. It stated that its objectives are "to minimize errors, maintain a transparent system for data management, automate tasks where possible, and maintain an archive of data and calculations used to develop assessments" (EPA 2013b, p. 17). The committee encourages EPA to meet the stated goal of having IRIS documents discuss model processes and derivation of toxicity values and associated uncertainties more completely.

Establish Clear Guidelines for Study Selection

The NRC formaldehyde report identified a need for clearly stated criteria for the selection of studies used to derive toxicity values. EPA acknowledges the need for selection criteria by stating that "once these studies have been identified, the basic criterion for selecting a subset for the derivation of toxicity values is whether the quantitative exposure and response data are available to compute a NOAEL, LOAEL or benchmark dose/concentration. When there are many studies, the assessment may focus on those that are more pertinent or of higher quality" (EPA 2013b, p, F-53). EPA provides additional guidance regarding the attributes used to evaluate studies for derivation of toxicity values (see Box 7-2), including balancing strengths and weaknesses and weighing human vs experimental evidence (EPA 2013b, Appendix B, Sections 3-6, and Appendix F, pp. F-53 to F-55). The committee encourages EPA to develop detailed criteria that take

BOX 7-1 Recommendations on Calculation of Toxicity Values in the 2011 National Research Council Formaldehyde Report

- Describe and justify assumptions and models used. This step includes review of dosimetry models and the implications of the models for uncertainty factors; determination of appropriate points of departure (such as benchmark dose, no-observed-adverse-effect level, and lowest observed-adverse-effect level), and assessment of the analyses that underlie the points of departure.
- Provide explanation of the risk-estimation modeling processes (for example, a statistical or biologic model fit to the data) that are used to develop a unit risk estimate.
- Assess the sensitivity of derived estimates to model assumptions and end points selected. This step should include appropriate tabular and graphic displays to illustrate the range of the estimates and the effect of uncertainty factors on the estimates.
- Provide adequate documentation for conclusions and estimation of reference values and unit risks.

Source: NRC 2011, pp. 165-166.

into consideration common technical issues and notes, for example, that group-average exposures as represented by a median or mean of the exposure groups are not as reliable as specific individual exposures. A fitted dose-response model based on a group-average exposure can distort the true underlying dose-response relationship. The committee also encourages EPA to summarize study information in a table, including all end points, datasets for dose-response assessment, and toxicity-value derivations. Careful attention should be paid to study design, including the doses, dose spacing, and number of subjects.

Describe and Justify Assumptions and Models Used to Calculate Toxicity Values

EPA has developed guidance on modeling dose-response data, assessing model fit, selecting suitable models, and reporting POD modeling results (EPA 2005a,b; EPA 2012; EPA 2013b). The draft handbook (EPA 2013b, Appendix F) currently has a placeholder for detailed technical guidance on dose-response models and assumptions. The EPA *Guidelines for Carcinogen Risk Assessment* (Section 3.2, EPA 2005a) provides guidance on many issues raised by the NRC formaldehyde report, including choice of dosimetry, toxicodynamic models vs empirical curve-fitting, and choice of and narrative concerning POD. The present committee encourages EPA to complete the IRIS assessment preamble in accordance with EPA's existing guidelines. It also encourages EPA to include more detailed technical guidance on model assumption, selection, and process in the draft handbook.

Provide Explanation of the Risk-Estimation Modeling Process

Because the draft handbook (EPA 2013b, Appendix F) and the draft preamble (EPA 2013b, Appendix B) are yet to be completed for the risk-estimation modeling process, the committee reviewed chemical-specific examples in EPA (2013c)—specifically, example 6 ("Dose-Response Modeling Output") and example 7 ("Considerations for Selecting Organ/System-Specific or Overall Toxicity Values")—to assess the changes that EPA has made regarding this and related recommendations in the NRC formaldehyde report. The detailed presentation of dose-response modeling output and derivation of toxicity values is helpful. In example 6, EPA

BOX 7-2 Considerations in Deriving Toxicity Values

- *Species*: a preference for the use of human data or mammalian data when human data are unavailable.
- *Relevance of exposure paradigm*: a preference for studies that use an environmentally relevant exposure route, sufficient exposure duration (chronic or subchronic studies when chronic toxicity values are developed), and multiple exposure levels.
- *Potential selection bias*: a preference for studies as appropriate with low risk of selection bias and higher participation and follow-up rates.
- *Potential confounding*: a preference for studies with a design (such as matching procedures) or analysis (such as procedures for statistical adjustment) that adequately address the relevant sources of potential confounding for a given outcome.
- *Exposure measurements*: a preference for studies that evaluate exposure during a biologically relevant time window for the outcome of interest, that use high-quality exposure assessment methods that reduce measurement error, that are not influenced by knowledge of health outcome status, and that include individual exposure measurements.
- *Measurement of health outcome*: a preference for studies using widely accepted, valid, and reliable outcome assessment methods. Measurement or assignment of the outcome should not be influenced by knowledge of exposure status.
- *Power and precision*: EPA evaluates the following factors when choosing studies: numbers of test subjects and doses and appropriate experimental design.

Source: EPA 2013b, p. F-54.

shows how multiple dose-response models (particularly empirical curves) can be fitted to a given dataset and how to use statistical criteria to select a best model and later a toxicity value. Although that approach might remain acceptable under some circumstances, the NRC formaldehyde committee encouraged EPA to move away from that old paradigm and to develop approaches for integrating multiple toxicity values rather than selecting one value or study that appears to be the "best."

Example 6 also shows how EPA uses goodness-of-fit or information criteria in conjunction with the spread of the lower confidence limits on the BMD (BMDLs) to select a preferred model. Specifically, among all models that fit the data reasonably well ($p > 0.1$), the one with the lowest Akaike information criterion[1] is chosen if the corresponding BMDLs are all within a range of a factor of 3; otherwise, the model with the lowest BMDL is selected. Several implications regarding EPA's criteria should be noted. First, although the criteria are easy to implement, EPA should articulate the pros and cons of adopting such model-selection criteria. In particular, if the difference is more than 3-fold among the BMDLs derived from different models fitted to the same dataset, questions arise as to the consistency of the models below the benchmark-response level. In that case, the model that yields the lowest BMDL can be an outlier even though its selection might appear more protective. Choosing the one that has the smallest BMDL could result in selection of the study that has the lowest quality (for example, the smallest sample). Second, goodness-of-fit tests might fail to find a lack of fit because of a small sample but indicate a poor fit when the sample is sufficiently larger. Third, caution must be exercised if those criteria were to be used to compare models fitted to different datasets. Information criteria are designed to differentiate models that have different numbers of parameters but under the same distribution.

[1]The Akaike information criterion "is an estimate of a measure of fit of the model" (Akaike 1974, p. 716).

The information criteria might prefer a simpler model with fewer parameters because the underlying data are insufficient to show that another model with more parameters is statistically better than the simpler model even if the latter might be a "true" model. Fourth, if only one toxicity value is selected, an opportunity to quantify uncertainty associated with the model, the model parameters, and POD could be lost.

Example 6 illustrates dose-response modeling using EPA's BMD software. That example demonstrates that particular model parameters are sometimes set to a default (or boundary) value with no explanation or justification. Example 6 also contains cases in which a saturated model—in which the number of parameters is equal to the number of dose groups—is fitted. It is well established that statistical estimation cannot accommodate a model that has more parameters than distinct data points (dose groups, in this case). Any single statistical criterion for model selection has its limitations and pitfalls. Thus, multiple criteria should be used simultaneously, and all details regarding assumptions and justifications of dose-response modeling should be included in the IRIS assessments.

It should be noted that most of the dose-response models implemented in EPA's BMD software do not accommodate for adjustment for covariates that are independent risk factors or confounders, whereas most epidemiologic studies adjust for multiple covariates. If smoking modifies the risk of the effects of an exposure, EPA's BMD software currently requires separate dose-response models to be fitted for smokers and nonsmokers; this is feasible only if the study is of sufficient size for each group to provide adequate power. However, if there are multiple covariates, including continuous ones, it becomes much less feasible to conduct even a stratified dose-response assessment or toxicity-value estimation specific to each group. There is a clear need for EPA to facilitate model and software development for dose-response modeling of studies that have more complex design than one-generation, single time point settings. Examples include studies in which the exposure is time-dependent or the end point is measured over time (repeated-measurement experiments). EPA's BMD software has implemented a beta version of dose-time-response models for neurobehavioral-toxicity end points. With those models, an RfD can vary greatly depending on when a neurobehavioral end point is observed (Zhu 2005; Zhu et al. 2005a,b). Those models allow the use of random effects to account for both between-subject and within-subject variability.

Although the present committee commends EPA's initiative in deriving an organ-specific, system-specific, or overall toxicity value among multiple candidates for various end points or from various studies, it has some concerns about EPA's approach. The draft handbook (EPA 2013b, Appendix F) lists the following criteria for evaluating each candidate toxicity value: (a) strength of evidence of hazard for the health outcome or end point, (b) attributes previously evaluated in selecting studies for deriving candidate toxicity values, (c) the basis of the POD, (d) other uncertainties in dose-response modeling, and (e) uncertainties due to other extrapolations. On the basis of the criteria, the organ-specific or system-specific toxicity value might be based on a single candidate value that is considered to be the most appropriate for protecting against toxicity to the given organ or system. Alternatively, the value might be based on a derived composite value that is supported by multiple candidate toxicity values that protect against toxicity to the given organ or system. The present committee recommends that the result of the evaluation of individual toxicity values be presented in a tabular form to show which ones meet or fail to meet particular criteria. In example 7, EPA selects a particular RfD "because it is associated with the application of the smaller composite UF [uncertainty factor] and because similar effects were replicated across other studies" (EPA 2013c, p. 45). EPA should also make clear whether and how the criteria are weighed in determining the selected toxicity value to ensure that the process is transparent and consistent. The committee strongly suggests that EPA consider approaches to integration of as much of the evidence as possible rather than selecting a limited segment of the evidence in deriving an organ-specific, system-specific, or an overall toxicity value. The committee discusses the latter point further below.

Dose-response modeling and estimation of toxicity values can be improved when mechanistic data are available. Although incorporating mechanistic evidence into dose-response modeling remains challenging, potential benefits include natural integration of evidence among various end points with the same or similar mechanisms and facilitating extrapolation among species and from higher to lower doses. It has also been suggested that mechanistic information can reveal the functional form of the dose-response relationship, but it is worth noting that it is impossible to determine the correct functional form of a population dose-response curve solely from mechanistic information derived from animal studies and in vitro systems. Consider the threshold model, for example. As described in the NRC report *Assessing Human Health Risks of Trichloroethylene: Key Scientific Issues* (NRC 2006, pp.318-323), the dose threshold separating the low-dose mechanism from the high-dose mechanism is likely to differ among individuals because of widely varied human environments and genetic susceptibilities; this often creates a sigmoidal population dose-response curve even if the dose-response relationship has a clear threshold in a single rodent species or cell line. The committee encourages EPA to develop and apply physiologically based models that incorporate both mechanism and human-population variability into dose-response modeling when feasible.

Assess the Sensitivity of Derived Estimates to Model Assumptions and End Points Selected

As advised in the NRC formaldehyde report (NRC 2011), EPA is adopting the principles of systematic review to select multiple studies, multiple end points, and multiple datasets for dose-response assessment and toxicity-value estimation. The availability of multiple datasets makes it possible to analyze the sensitivity and variability of the toxicity-value estimates to demonstrate uncertainties inherent in study design, population exposed, exposure estimate, mechanism of action, and model choice. The draft IRIS assessments for methanol (EPA 2013d) and benzo[a]pyrene (EPA 2013e) include easy-to-understand tabular and graphic displays that clearly show the PODs for selected end points with corresponding applied uncertainty factors to illustrate the range of the estimates and the effect of uncertainty factors on the estimates.

Provide Adequate Documentation for Conclusions and Estimation of Toxicity Values

Recent EPA IRIS assessments have included extensive detail on published studies, evaluation of the evidence base regarding toxicity, and pharmacokinetic and dose-response modeling. Elements in dose-response analysis for which additional documentation is needed include the decision processes used by EPA to select studies for derivation of an RfC, RfD, or unit risk; the process used to select a particular value for the RfC, RfD, or unit risk from a range of values determined by using separate studies; and the process used to select a response level (typically 1, 5, or 10%) for the POD. Although EPA provides some general guidance regarding those decisions, it is not always clear how it applied the guidance in the selection of final values for any particular chemical assessment. One way to enhance the documentation of the estimates is to use established systematic algorithms, such as meta-analysis, when more than one relevant study is available. Recognizing that subjective judgments are a feature of all systems for combining and evaluating evidence, the committee encourages EPA to be explicit and detailed regarding such judgments.

Additional Progress in Calculating Toxicity Values

EPA has developed standard descriptors to characterize the level of confidence in each reference value on the basis of the likelihood that the value would change with further testing (see Box 7-3). Development of the descriptors is consistent with guidelines for deriving recommendations from systematic reviews that evaluate the quality of evidence.

BOX 7-3 Standard Descriptors to Characterize Level of Confidence

- *High confidence*: The reference value is not likely to change with further testing, except for mechanistic studies that might affect the interpretation of prior test results.
- *Medium confidence*: This is a matter of judgment, between high and low confidence.
- *Low confidence*: The reference value is especially vulnerable to change with further testing.

Source: EPA 2013b, Appendix B, p. B-14.

RELEVANT METHODOLOGIC ISSUES

Overall, the committee considers that EPA has made good progress in implementing the recommendations of the NRC formaldehyde report. Because implementation is a continuing process, the committee provides a brief review of several additional approaches that are relevant for the development of toxicity values that could be considered as the IRIS program continues to evolve. The committee has focused its attention largely on two main subjects: meta-analytic and Bayesian approaches and analysis and communication of uncertainty.

Combining Data for Dose-Response Modeling: Meta-Analytic and Bayesian Approaches

Historically, EPA has often selected a single "best" study to derive RfCs, RfDs, and unit risks. That approach might be preferable when one study is clearly more reliable and valid than all other studies for estimating human dose-response relationships. However, varying study strengths and weaknesses often precludes the identification of any one study as preferred for estimating such relationships. In those cases, it is preferable to use multiple studies to derive the RfCs, RfDs, and unit risks because that approach should provide more reliable estimates of human dose-response relationships than the use of a single study. Moreover, the use of multiple studies takes full advantage of the systematic-review process that led to evidence integration and reduces the potential for bias from selecting the most extreme study.

In recent assessments, EPA has embraced the use of multiple studies, primarily by estimating PODs, RfCs, RfDs, or unit risks separately for all relevant studies and then choosing final toxicity values from within the observed ranges. However, the process that EPA uses to select the final values is still not sufficiently transparent and appears somewhat subjective, and documentation varies among draft IRIS assessments. Formal statistical methods are widely available for combining estimates from multiple studies and might be useful for this step in the IRIS process. For example, meta-analysis (Stroup et al. 2000) and Bayesian hierarchical models (Sutton and Abrams 2001) have been used to combine information from various studies in many kinds of application.[2] Those statistical approaches also can be used to combine toxicity estimates from dose-response studies related to multiple species given exchangeability assumptions (for example, rat and mouse studies are equally relevant surrogates for human dose-response relationships) or with modifications that give more weight to human studies.

[2]A Bayesian hierarchical model can also be used for meta-analysis.

Meta-Analytic Approaches

Information from multiple studies can be combined either at the individual level (pooling all observations from each study into one large data set that is used to fit a dose-response model) or at the aggregate level (collecting only the reported dose-response estimate from each study as in typical meta-analyses). The draft handbook (EPA 2013b, Appendix F) addresses that issue in two sections: "Considerations for Combining Data for Dose-Response Modeling" and "Considerations for Selecting Organ/System-Specific or Overall Toxicity Values." The first section describes criteria for pooling data at the individual level, and the second section indicates that either using a single study or combining aggregate estimates from different studies to produce a composite toxicity value might be acceptable if the methods are documented.

In the first section, on pooled data analysis, EPA includes the following reasons for not combining data sets: heterogeneity in datasets because of differences in laboratory procedures, subject demographics, or route of exposure; and biologic or study-design limitations. Criteria that EPA uses to determine whether data should be combined include whether multiple studies have sufficient quality for deriving PODs, whether common specific outcome measurements are reported, whether common measures of dose or validated physiologically based pharmacokinetic models are available, whether exposure duration and observation duration are comparable, whether there is evidence of homogeneous responses to dose, and whether there is no clear preference for any single study. The criteria used by EPA are relevant, cautionary, and more restrictive than widely accepted approaches for analysis of pooled epidemiologic data (also called individual-level meta-analysis). For example, epidemiologic data from multiple studies that had disparate procedures, exposure duration, and participant demographics can often be combined before dose-response modeling by using appropriate methods and heterogeneity statistics or other empirical evaluations to judge the similarity of dose-response relationships among studies (Steenland et al. 2001; Stroup et al. 2000; Lyman and Kuderer 2005). Indeed, when heterogeneity in dose-response estimates is modest, pooling individual-level data before dose-response modeling has several advantages over meta-analysis and other aggregate methods. The advantages include the ability to use dose-response models other than those used in the original publications, to adjust for a common set of confounding variables, and to evaluate risks in susceptible groups that were not evaluated in the original publications (Blettner et al. 1999). The committee agrees with EPA that pooling data requires careful consideration. Modeling of pooled data often requires specification of all study covariates and the use of random effects to capture remaining study differences. Additional obstacles include reluctance or inability to share data; heterogeneity of study organization, protocols, or data formats; language barriers; missing data; and the need to harmonize information among studies (Schmid et al. 2003).

When individual-level data are unavailable, are not readily obtained and evaluated, or do not meet the criteria for being combined, meta-analysis of aggregate dose-response estimates is a reasonable alternative. Under some conditions, the estimates from aggregate meta-analysis are equivalent to those from pooled individual-level data analysis, and similar evaluation methods are used in both types of analyses (Lyman and Kuderer 2005). As discussed above, using multiple studies to derive dose-response estimates is generally preferable to relying on single studies. Guidelines for these methods are readily available (Blettner et al. 1999; Stroup et al. 2000; Orsini et al. 2012) and could be adapted for EPA's purposes.

Bayesian Hierarchical Models

Bayesian methods are based on the premise that information regarding unknown parameters (in this case, dose-response parameters) can often be obtained apart from the results of any one study (see Appendix C). In the context of dose-response evaluation, the outside information

could include results of mechanistic studies, maximum plausible slope factors based on US cancer prevalence, dose-response relationships for similar or related chemicals, or any other relevant information. In the Bayesian system, the outside information is represented quantitatively in the form of a "prior" probability distribution, which is then formally combined with dose-response data to generate a "posterior" probability distribution that attempts to summarize quantitatively all relevant information regarding the dose-response relationship. In fact, current EPA methods for dose-response characterization fit within a simple Bayesian framework as shown in Figure 7-3. The figure corresponds to a simple model of the dose-response relationship between humans and animals for any particular chemical:

$$\beta_{human} = U \times \beta_{animal}$$

where β_{human} is the dose-response parameter (such as a BMD) in humans, β_{animal} is the same dose-response parameter in animals, and U is the uncertainty ratio of the two parameters. If no human dose-response data are available for a chemical, EPA's calculation (RfD = BMDL/UF) can be viewed as the computation of a lower credible limit for β_{human} by using this Bayesian formulation and the assumption that $\ln(\beta_{animal})$ and $\ln(U)$ are normally distributed. Because traditional uncertainty factors reflect multiplicative uncertainty, they can be represented by lognormal distributions (that is, normal distributions on the log scale). Table 7-2 shows the conversion of common uncertainty factors to log standard deviations.

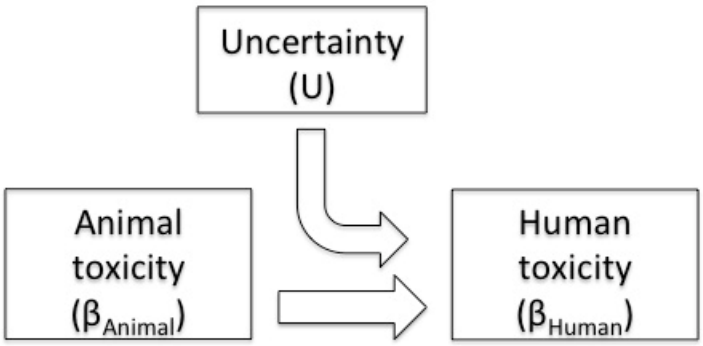


FIGURE 7-3 Simple Bayesian framework for estimating human toxicity from results of an animal study.

TABLE 7-2 Conversion of Traditional EPA Uncertainty Factors to Bayesian Prior Standard Deviations on a Natural Log Scale Using 1-Sided or 2-Sided Confidence Intervals

Uncertainty Factor	95% One-Sided 90% Two-Sided	97.5% One-Sided 95% Two-Sided	99.5% One-Sided 99% Two-Sided
3	0.668	0.561	0.427
5	0.978	0.821	0.625
10	1.400	1.175	0.894
100	2.800	2.350	1.788
300	3.468	2.910	2.214
1,000	4.200	3.524	2.682

For example, consider an RfD calculation that is based on a single animal study, such as the oral RfD of 0.3 mg/kg-day for phenol that is derived from 1-standard deviation decrease in maternal weight gain in rats (EPA 2002). The RfD was calculated on the basis of a BMD of 157 mg/kg-day and a 95% lower confidence limit (BMDL) of 93 mg/kg-day. Dividing the BMDL by an uncertainty factor of 300 results in 0.31 mg/kg-day, rounded to 0.3 mg/kg-day when reported as the RfD. To apply the Bayesian framework shown in Figure 7-3, one first computes ln(BMD) and ln(BMDL), which are 5.06 and 4.53, respectively. If one assumes that the phenol assessment used a two-sided 95% confidence interval to obtain the BMDL, the standard deviation for the confidence interval around the ln(BMD) is (5.06 – 4.53)/1.96 = 0.27. Thus, the prior distribution for $\ln(\beta_{animal})$ has a mean of 5.06 and a standard deviation of 0.27. A prior distribution for ln(U) with a mean of 0 and a standard deviation of 2.91 corresponds to a best estimate of equivalent animal and human toxicity (U = 1) with an uncertainty factor of 300 (see Table 7-2 for conversion between traditional uncertainty factors and Bayesian prior standard deviations). Assuming normal distributions of $\ln(\beta_{animal})$ and ln(U), one can show that $\ln(\beta_{human})$ is normally distributed with a mean of 5.06 and a standard deviation of $\text{sqrt}(0.27^2 + 2.91^2) = 2.92$. Without any human data to inform the dose-response relationship, that distribution for $\ln(\beta_{human})$ would usually be referred to as an induced prior rather than a posterior. The lower bound of the two-sided 95% credible interval for β_{human} is thus $\exp(5.06 - [1.96 \times 2.92]) = 0.51$ mg/kg-day. This example illustrates the ease with which the uncertainty-factor approach could be modernized by using formal Bayesian methods and the Bayesian lower bound of β_{human} in place of the traditional RfD.

If EPA adopted Bayesian methods for dose-response assessment, it would be helpful for it to focus on considerations of *relevance* and *exchangeability* of each study for human toxicity assessment. The two concepts are easily incorporated into Bayesian modeling by grouping exchangeable information in the same stage and sequentially updating each stage in order from least relevant to most relevant. For example, a Bayesian framework for combining data from multiple studies of cancer for a particular chemical could start with an assumption that available high-throughput tests (including structure-activity relationship models, in vitro mutagenicity tests, DNA methylation tests, and nonmammalian studies) are equally likely to reflect human cancer risk (that is, they are "exchangeable") but might be less relevant than data from other available studies, such as mammalian studies or epidemiologic studies. After pharmacokinetic, mechanistic, or default assumptions are used to model the dose-response relationship for each study on a common human-equivalent dose-response scale, estimates from exchangeable studies can be averaged by using familiar meta-analytic methods (frequentist or Bayesian, as in Sutton and Abrams 2001). A similar meta-analysis could be conducted within each class of exchangeable studies. For example, rat and mouse studies that used different strains might be considered exchangeable with each other for the purpose of estimating human-toxicity values but deemed more relevant than the in vitro and nonmammalian studies; a separate meta-analysis of human-equivalent dose-response estimates would then be determined for the rodent studies. Finally, epidemiologic studies with sufficient exposure characterization and adequate confounding control (if any such studies are available) could be grouped as yet another set of exchangeable studies.

As discussed in Chapter 6 of the present report, Bayesian methods also offer a natural framework for combining the three meta-analysis estimates in order of increasing relevance: high-throughput studies, mammalian studies, and human studies. The committee notes that human cell lines are increasingly used for high-throughput studies. If they become more relevant for estimating human risks than animal studies, the order of relevance should be modified accordingly (that is, mammalian studies, high-throughput studies, and human studies).

In the first stage, the high-throughput meta-analysis estimate can be used as a prior mean that can be updated with data from the more relevant mammalian-study meta-analysis estimate. The first-stage posterior mean can then be used as a prior mean that can be updated in the second stage with data from the more relevant human-studies meta-analysis estimate (see Figure 7-4).

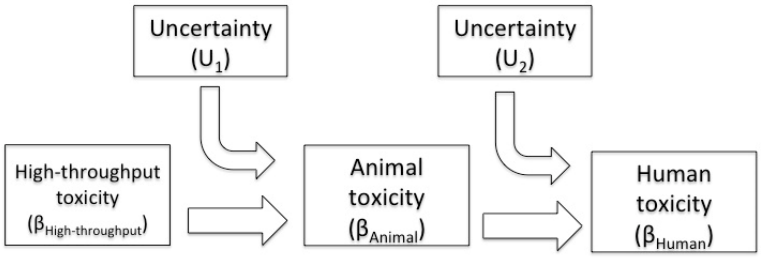


FIGURE 7-4 Bayesian framework for combining studies of different types. High-throughput studies can switch positions in the model with animal studies if they are more relevant to human toxicity.

The model can be written as follows:

$$\beta_{animal} = U_1 \times \beta_{high\text{-}throughput}$$
$$\beta_{human} = U_2 \times \beta_{animal}$$

where β_{human} is the dose-response parameter (such as a BMD) in humans, β_{animal} is the same dose-response parameter in animals, $\beta_{high\text{-}throughput}$ is the same dose-response parameter in high-throughput studies, and U_1 and U_2 are uncertainty ratios of the dose-response parameters that relate high-throughput studies to animals and relate humans to animals, respectively. Prior variances of U_1 and U_2 for each of the two stages of updating would be selected to reflect the general reliability of each type of data source for estimating toxicity in accordance with EPA's guidance on uncertainty-factor selection (Stedeford et al. 2007).

This framework ensures that information at each higher stage of relevance will quickly overcome prior information from previous stages. For example, high-quality epidemiologic studies that used large samples should dominate the posterior dose-response estimates when available, and mammalian studies that used samples of sufficient size should dominate the posterior dose-response estimates when epidemiologic studies are unavailable. More important, however, the framework offers a coherent system for combining dose-response information from disparate studies when no one study group is clearly dominant. In other words, updating a Bayesian estimate at each stage can be seen as providing a weighted average of the prior mean and the data that are entered in the current stage. The weight of the new data, for example, is proportional to the ratio of prior variance to total variance; less weight is given to the component that has a higher variance. For example, application of the Bayesian framework to a situation with extensive high-throughput data but only one small rodent study would result in a posterior dose-response estimate somewhere between that for the rodent study and that suggested by the high-throughput data. A larger sample size in the single small rodent study or the addition of a second rodent study with a similar dose-response estimate would decrease the variance for the meta-analysis estimate and therefore increase the weight given to the rodent studies in computing the posterior dose-response estimate.

As another example, consider a hypothetical scenario in which the ED_{10} for a new chemical is estimated to be 0.1 mg/kg-day and the LED_{10} (two-sided 95% confidence) is 0.07 on the basis of in vitro tests of inflammatory response (such as ELISA) with an uncertainty factor of 1,000 for predicting the ED_{10} for animal studies. On a natural log scale, $\ln(ED_{10})$ is -2.30, $\ln(LED_{10})$ is -2.66, and the standard deviation of U_1 is 3.52 (corresponding to UF = 1,000 in Table 7-2 for 95% confidence). The standard deviation of $\ln(ED_{10})$ from the high-throughput data is therefore (-2.30 – [-2.66])/1.96 = 0.184, assuming that the LED_{10} is the lower bound of the two-sided 95% confidence interval, and the induced prior standard deviation for the $\ln(ED_{10})$ in animals is $\text{sqrt}(0.184^2 + 3.52^2) = 3.52$. If no animal or human data are available and an uncertainty factor of 10 is thought to be appropriate for extrapolation from animals to humans, U_2 has a prior

standard deviation of 1.17 (see Table 7.2), and the induced prior standard deviation of the $\ln(ED_{10})$ in humans is $\text{sqrt}(3.52^2 + 1.17^2) = 3.71$. Therefore, EPA might report a central estimate for the ED_{10} of $\exp(-2.30) = 0.1$ mg/kg-day (the prior median) and a lower bound ED_{10} of $\exp(-2.30 - [1.96 \times 3.71]) = 0.00007$ mg/kg-day (the lower bound on the 95% two-sided credible interval) for humans.

Suppose that two small rat studies become available, and a meta-analysis is conducted that results in a $\ln(ED_{10})$ of -0.161 with a standard error of 0.9. Using a normally distributed prior, the Bayesian analysis at stage 1 (see Figure 7-4) results in a posterior mean that is a simple weighted average of the prior mean and the meta-analysis mean for the animal studies. In that weighted average, the weight given to the animal meta-analysis is the ratio of the prior variance to the total variance ($3.52^2/[3.52^2 + 0.9^2] = 0.94$). Thus, in this example, the Bayesian posterior mean animal $\ln(ED_{10})$ that combines the high-throughput data and results of mammalian studies places 94% weight on the animal studies, and this results in a posterior mean animal $\ln(ED_{10})$ of $[(0.94)(-0.161)] + [(0.06)(-0.230)] = -0.165$ and a posterior standard deviation of $\text{sqrt}([3.52^{-2} + 0.9^{-2}]^{-1}) = 0.872$. If EPA chooses an uncertainty factor of 10 (which corresponds to a prior standard deviation of 1.17 for U_2, according to Table 7-2) for using this posterior animal estimate as a surrogate for human data, the induced prior standard deviation of the $\ln(ED_{10})$ in humans is $\text{sqrt}(0.872^2 + 1.17^2) = 1.46$, and the agency might report a central estimate ED_{10} of $\exp(-0.165) = 0.85$ mg/kg-day and a lower bound ED_{10} of $\exp(-0.165 - [1.96 \times 1.46])$ or 0.048 mg/kg-day.

Finally, suppose that a single human study becomes available and reports a $\ln(ED_{10})$ of -3.0 with a standard error of 0.5. Under the normality assumption, the weight of the human study estimate is $1.46^2/(1.46^2 + 0.5^2) = 90\%$, and this results in a posterior mean $\ln(ED_{10})$ of $(0.90)(-3.0) + (0.10)(-0.165) = -2.72$. The posterior variance is the reciprocal of the sum of the inverse variances of the prior and the human estimates, $(1.46^{-2} + 0.5^{-2})^{-1} = 0.224$. Thus, EPA might report a central estimate ED_{10} of $\exp(-2.72) = 0.066$ mg/kg-day and a lower bound ED_{10} of $\exp(-2.72 - 1.96 \times \text{sqrt}[0.224]) = 0.026$ mg/kg-day.

The results of each stage of updating for this hypothetical example are shown in Table 7-3 with a traditional RfD calculated by using only one data stream (high-throughput, animal, or human) at a time with no intraspecies uncertainty factor. After the two-stage Bayesian updating that combines all three evidence streams, the posterior lower bound for the ED_{10} is 0.026 mg/kg-day compared with traditional RfDs of 0.000007, 0.015, and 0.019 mg/kg-day for the high-throughput, animal, and human studies, respectively. In this example, the Bayesian lower bound ED_{10} is slightly higher than the traditional RfD based on the human study alone because the animal and high-throughput studies suggest that the chemical is less toxic than the human study suggests and because standard error is smaller in the posterior distribution. In situations in which the central estimate of the ED_{10} is lower for the animal studies than for the human studies, the Bayesian lower bound could be less than the traditional RfD that is based on the human studies alone.

Adding a second human study that reports the same $\ln(ED_{10})$ of -3.0 would produce a meta-analysis with the same $\ln(ED_{10})$ but a smaller standard error than that reported for the first study alone; this reflects increased precision of the dose-response estimate. The smaller standard error would result in more than 90% weight for the human data and thus a central estimate $\ln(ED_{10})$ closer to -3.0. It would also result in a smaller posterior variance that would be closer to the human meta-analysis variance. Thus, with more consistent high-quality epidemiologic data, the in vitro data and animal data would have less and less effect on the dose-response estimates.

Although the Bayesian framework described here is relatively simple, it can be expanded to allow additional stages (new categories of exchangeable studies), assumptions other than normality, different dose-response parameters (such as the cancer slope factor), and assessment of the potential effects of bias and other sources of uncertainty on the animal to human dose extrapolation (DuMouchel and Harris 1983; Peters et al. 2005). For example, EPA might wish to group

TABLE 7-3 Summary of Results of the Two-Stage Bayesian Example

	ED_{10}: Central Estimate (mg/kg-day)	ED_{10}: Lower Bound (mg/kg-day)	Traditional RfD (mg/kg-day)
High-throughput studies	0.1	0.07	0.000007
Animal studies	0.85	0.15	0.015
Bayesian posterior—first stage	0.85	0.048	—
Human study	0.05	0.019	0.019
Bayesian posterior—second stage	0.066	0.026	—

epidemiologic studies separately according to study design, such as grouping cross-sectional studies as one category of exchangeable studies and grouping cohort studies and nested case-control studies as a different set of more relevant exchangeable studies. That approach might be particularly useful when cross-sectional designs are deemed to have a higher risk of bias, such as a greater potential for reverse causation noted for some biomarker-based epidemiologic studies (Loccisano et al. 2012). Incorporation of specific mechanisms of action could be handled by updating multiple dose-response parameters at each stage, when each parameter reflects a particular biologic step, such as the amount of aryl hydrocarbon receptor binding or DNA methylation. Bayesian methods are also compatible with quantitative uncertainty analysis because uncertainty distributions can be incorporated as additional prior components, and alternative dose-response model specifications can be incorporated via Bayesian model averaging (Hoeting et al. 1999; Gustafson 2004). Indeed, a Bayesian formulation for computing RfDs and unit risks might help EPA to move forward with quantitative uncertainty analysis in a manner that does not depart radically from existing methods, such as that used in the above examples. As toxicity databases expand, EPA could establish Bayesian priors on the basis of empirical evaluation of the distribution of the human-to-animal dose-response parameter ratios for chemicals in the same class or for all available chemicals in place of default lognormal distributions.

Other, more sophisticated Bayesian approaches have been proposed for combining dose-response estimates for multiple species and multiple chemicals (DuMouchel and Harris 1983; Jones et al. 2009). Those approaches might also be useful to EPA if guidance for selection of appropriate models and priors is developed.

Analysis and Communication of Uncertainty

As discussed earlier, estimation of toxicity values is the culminating step of the IRIS process. The reference values and unit risks draw on data from heterogeneous and dynamic systems that underpin hazard identification, exposure assessment (for epidemiologic studies), and dose-response assessment. Regardless of the studies included, the analytic tools used, and the underlying models assumed, there will always be uncertainties surrounding the final estimates because of incomplete knowledge about the systems involved. Uncertainty can be characterized and reduced by the use of more or better data and should be managed. It should be distinguished from variability, which is the result of inherent differences in susceptibility among humans regarding exposures and related health effects. Variability can be better characterized with more data but cannot be reduced.

How to address uncertainty has been a recurring issue in IRIS assessments and other agency risk-related activities (NRC 2009; 2010; 2012; IOM 2013). The NRC report *Science and Decisions* states (NRC 2009, p. 107) that

> There are different strategies (or levels of sophistication) for addressing uncertainty. Regardless of which level is selected, it is important to provide the decision-maker with information to distinguish reducible from irreducible uncertainty, to separate individual vari-

> ability from true scientific uncertainty, to address margins of safety, and to consider benefits, costs, and comparable risks when identifying and evaluating options.

It further recommends (NRC 2009, p 107) that "to make risk assessment consistent with such an approach, EPA should incorporate formal and transparent treatment of uncertainties in each component of the risk-characterization process and develop guidelines to advise assessors on how to proceed." The present committee agrees with the previous NRC committee and recommends that analysis and communication of uncertainty be an integrated component of IRIS assessments even when a default used in the assessment is consistent with EPA's own guidelines. At a minimum, that approach would include a demonstration of variation in the final toxicity-value estimates under different assumptions, options, models, and methods.

Communication of uncertainty can be mistakenly interpreted by some as indicating poor-quality or insufficient science (Johnson and Slovic 1995; Freudenburg et al. 2008); in some cases, explicit acknowledgments of uncertainty have been misinterpreted as acceptance of lower scientific standards that might weaken a scientific body's authority (Funtowicz and Ravetz 1992). However, experimental research also demonstrates that conveying appropriate information about uncertainty—and in particular balanced characterization of the range rather than one-sided bounds of variation—can be seen as improving transparency (Johnson and Slovic 1995) and can improve risk-management decisions (Joslyn and LeClerc 2012; Joslyn et al. 2013). Furthermore, failure to acknowledge uncertainties leaves EPA vulnerable to attacks on management decisions or policies that are based on the best available science (Brickman et al. 1985) and might be considered unethical by some stakeholders (Smithson 2008).

As EPA revises the IRIS process to make it more efficient, flexible, and transparent, there is an opportunity for the agency to develop a framework for uncertainty analysis and communication and to make uncertainty analysis and communication integral to the IRIS process. In the discussion below, the committee offers suggestions on how the IRIS program can improve uncertainty analysis and communication.

Uncertainties arise in all stages and components of the IRIS process. Those in an earlier stage cascade and propagate to later stages and eventually aggregate to form the overarching uncertainty surrounding a final toxicity estimate. Characterizing this overarching uncertainty fully requires a vertical integration of uncertainties over every stage of the assessment process, including the initial protocol design, study identification and evaluation, dose-response modeling, low-dose extrapolation, cross-species extrapolation, and any other extrapolations that are needed to yield the final toxicity estimate. Omission of one source of uncertainty in a given stage can result in an inaccurate or even distorted characterization of the overall uncertainty. Although it is critical for understanding the uncertainties and their overall effect on a final toxicity estimate, such a vertical integration of uncertainties is rarely done in IRIS assessments (NRC 2009, pp.100-101) partly because of the lack of data, especially for some intermediate steps, and because such resources as in-house expertise and readily applicable tools are insufficient (EPA 2004, p. 34). Uncertainties arising from particular sources or in particular stages might be especially relevant to a specific risk-management decision. Thus, identifying and focusing on uncertainties that contribute the most to the overall uncertainty (have the largest effect on the final toxicity value) is a more practical treatment of uncertainties (NRC 2009; IOM 2013), although it is not always clear which sources contribute the most to the overall uncertainty unless a comprehensive analysis is performed.

Uncertainty analyses have become more common in IRIS assessments, but they are conducted for select individual intermediate stages in the process (such as dose-response modeling and low-dose extrapolation), often in isolation from one another. In the IRIS assessment of dioxin and dioxin-like compounds, for example, EPA examined uncertainty in unit risk estimates for cancer that was due to assumption of different forms of the dose-response model; EPA also examined uncertainty independently for different cancers (NRC 2006). In the IRIS assessment of tetrachloroethylene, EPA examined uncertainty (and variability) in RfC estimates and used mul-

tiple noncancer end points observed in multiple studies (NRC 2009): the uncertainty and variability could be attributed to differences in study design, species, end point, and dose-response form, among others. Those examples demonstrate a horizontal integration of variation associated with one or more factors, sources, or stages (such as model form and dose metric) for several options or combination of assumptions; uncertainties from other sources or stages were not considered. Horizontal integration can characterize the effect of uncertainty from selected sources and is a part of the overall vertical integration of uncertainties (see Figure 7-5).

When it is feasible, overall uncertainty can be characterized through a probabilistic distribution of the toxicity values that corresponds to all possible option combinations or, to a lesser extent, to the range of variation under feasible option combinations that are actually considered. Such a distribution would be ideal, but a range of variation is more commonly estimated in practice. Still, that range of variation simply reflects the observed part of the distribution of the overarching uncertainty (NRC 2010). Within the observed range, one or more toxicity values might be selected for a risk-management decision, and this selection is supported by and can be communicated with the uncertainty analysis. Using a distribution or range-finding approach to assess overall uncertainty offers a systematic approach to combining observed toxicity estimates into one or several groups according to homogeneity criteria with respect to, for example, a common mechanism or similar end points. To that end, the meta-analysis and Bayesian multilevel models discussed earlier are useful tools. Empirical tools, such as graphical displays, are also useful (NRC 2011). For example, Figure 7-6 is a cumulative distribution of 18 RfCs derived from multiple neurotoxicity end points from a collection of epidemiologic studies and laboratory animal experiments (NRC 2009). The smallest RfC and the two largest RfCs stand out, and the rest cluster between 0 and 100. Grouping the toxicity estimates appropriately requires that the systems underpinning the individual toxicity values be comparable or homogeneous regarding such elements as study design, exposure regimen, and health effects. For example, when health end points are plausible for a common mechanism, there is good support for using the variation range of the corresponding toxicity values as the horizontal integration of the overarching uncertainty (NRC 2011). Conversely, caution should be exercised in attempting to combine multiple studies to conduct dose-response modeling of combined data or to group toxicity estimates when the studies are different in design, exposed species, exposure regimen, and generalizability to human populations. If some study designs, species, or exposure regimens are more relevant to human toxicity, it might be better to group the most relevant studies than to group all available studies in the characterization of uncertainty.

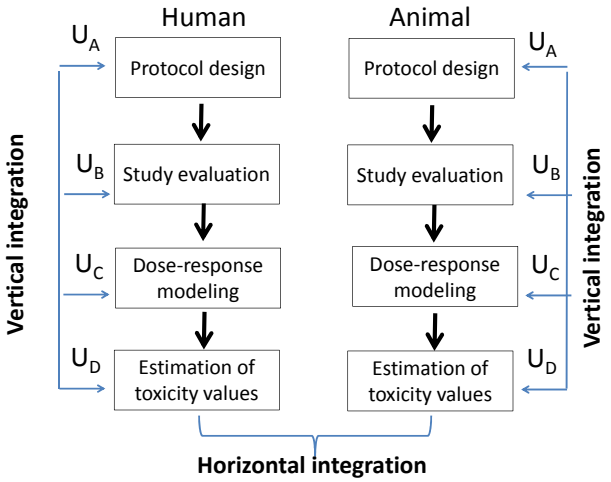


FIGURE 7-5 Characterization of overarching uncertainty requires vertical and horizontal integration of uncertainties in every stage of the assessment process. Note that not all steps are shown in this illustration.

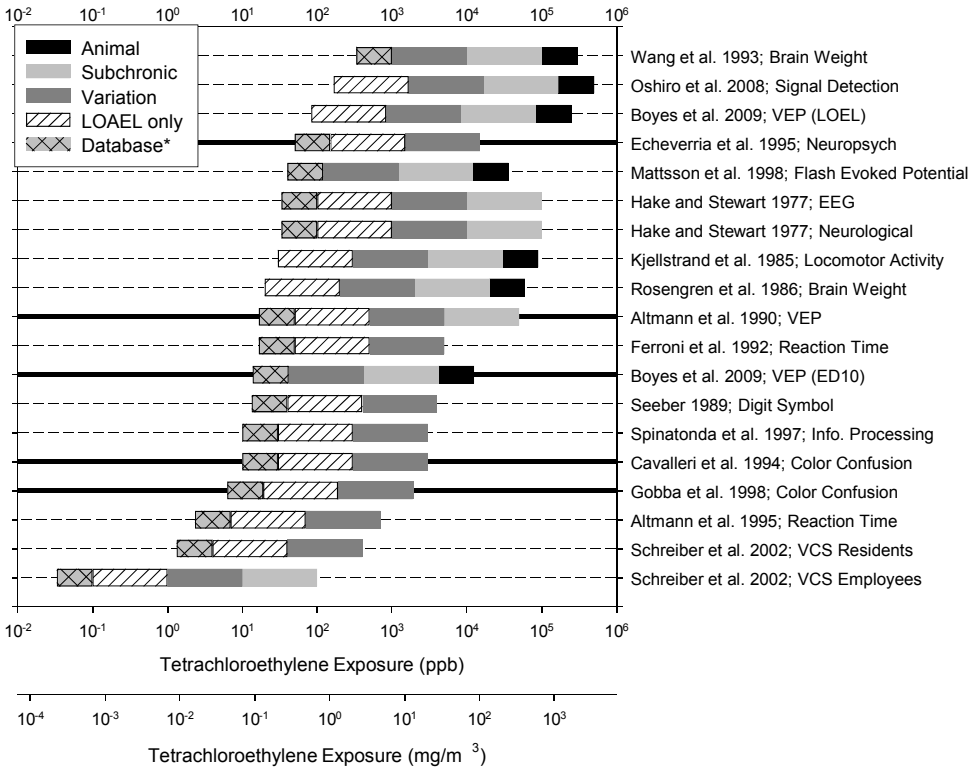


FIGURE 7-6 Cumulative distribution of reference concentrations (RfCs) derived from multiple neurotoxicity end points from a collection of epidemiologic studies and laboratory experiments on humans or animals. The length of a bar represents a 3-fold or 10-fold uncertainty factor, the shade of the bar represents the source of uncertainty as indicated in the figure legend, and the left end of a bar represents an RfC. Source: NRC 2010.

Characterizing the empirical-variation range of the overall uncertainty that is due to differences between studies or end points is useful in elucidating the totality of uncertainty (NRC 2010). The limitation of the empirical-variation approach is that it often does not differentiate the relative effect of uncertainties from different sources or stages because the toxicity values arise from different studies or used different end points, dose-response models, exposure metrics, or other factors. Hence, the range of uncertainty estimates is the result of partial horizontal and partial vertical integrations of some elements of uncertainty. That limitation highlights the fact that only a horizontal integration can tell the size of a single source of uncertainty, and only a vertical integration can tell the contribution of a single source uncertainty to the overarching uncertainty relative to others. When data support a fuller assessment of total uncertainty, Bayesian hierarchical modeling (Spiegelhalter et al. 2002) and multilevel probabilistic modeling (Small 2008) are examples of approaches that support vertical integration of multistage uncertainties. An improved uncertainty analysis within an individual IRIS assessment does not necessarily dictate a complex level of sophistication in mathematical, statistical, or computational methods. Simple analyses or qualitative elucidation of various uncertainties—for example, due to plausible mechanisms—can be adequate especially when few data are available or risk-management decisions are robust under competing options (NRC 2009; IOM 2013).

Another short-term strategy that EPA could adopt to improve uncertainty communication is to present clearly two dose-response values in each future toxicity assessment: a central estimate (such as a maximum likelihood estimate or a posterior mean) and a lower-bound estimate

for a POD from which a final toxicity value is derived. The lower bound becomes an upper bound for a cancer slope factor but remains a lower bound for a reference value. That might improve public risk communication for those who are not well versed in the IRIS process and support cost-benefit analysis and other policy evaluations while remaining health-protective. The recommendation is consistent with the EPA *Guidelines for Carcinogen Risk Assessment* (EPA 2005a), which notes that central estimates might be more useful for some purposes, such as uncertainty analysis and ranking of hazardous agents.

Central estimates can be obtained from dose-response modeling software, such as EPA BMDS. For example, EPA (2013c) includes output for a multistage cancer model fitted to hepatocellular-tumor data on female mice exposed to diisononyl phthalate (Example 6). Although the full form of the probability function for this model is P[response] = background + (1-background) × [1-exp(-[$\beta_1 \times dose^1$]-[$\beta_2 \times dose^2$]-[$\beta_3 \times dose^3$]-[$\beta_4 \times dose^4$])], the model is approximately linear at low doses with a slope of β_1. For this example, the maximum likelihood estimate for β_1 is shown as about 0.001155 kg-day/mg in the BMDS output. In contrast, the upper bound cancer slope factor based on EPA's POD method is reported as 0.00206 kg-day/mg. The two values correspond to the low-dose slopes of the red and black curves, respectively, in Figure 6-3 of EPA (2013c). Presenting both toxicity values in IRIS summaries would improve the clarity and utility of IRIS assessments.

Several frameworks—including EPA's methods for considering uncertainty analysis in policy analysis—could be considered by the IRIS program. The RIVM/MNP Guidance for Uncertainty Assessment and Communication developed by the Netherlands Environmental Assessment Agency National Institute for Public Health and the Environment (Janssen et al. 2003, Petersen et al. 2003; van der Sluijs et al. 2003, 2004, 2008) provides one such example. Several elements of that framework could be incorporated into IRIS guidelines for uncertainty analysis and communication, including the following:

- Develop a plan early in the IRIS process (for example, in parallel with the literature-review phase) for the conduct of the uncertainty analysis (Janssen et al. 2003; Petersen et al. 2003; NRC 2009; IOM 2013). The goals of this planning phase include screening of uncertainty sources to identify optional methods, setting priorities for resource allocation for analyses, and finally developing a strategy for uncertainty communication (NRC 2009, pp.120-121). The planning stage should consider the end use in mind (NRC 2009) and can be tailored for each toxicity assessment because the need, scope, and feasibility of uncertainty analyses can vary from one toxicity assessment to another; indeed, in some cases or stages, uncertainties cannot be easily quantified (IPCS 2006; NRC 2007).
- Establish a consistent framework for the application of approaches (for example, from a qualitative discussion to a full probabilistic distribution of uncertainty) and criteria for their conduct (van der Sluijs et al. 2003, 2004; NRC 2009, p. 100). The framework should recognize and permit various degrees of technical sophistication and rigor corresponding to the objectives and feasibility of a particular uncertainty analysis. The guidelines should also establish an inventory of standard methods for uncertainty analysis according to the stage of an assessment and the nature or source of the uncertainty and should offer insights into method choice (van der Sluijs et al. 2004). The framework would help to reveal the distinct sources and nature of uncertainties, including whether they are unquantifiable system-level uncertainties, indeterminacy, or ignorance (van der Sluijs et al. 2003). The level of uncertainty analysis might be tiered according to quantification level, from a single default (no variation); to qualitative and systematic characterization; to quantitative characterization with bounds, ranges, and sensitivity; to a probabilistic distribution (van der Sluijs et al. 2003; EPA 2004; IPCS 2006). The tiered classification matches the degree of sophistication in uncertainty analysis with the level of concern for the problem and feasibility of conducting the analysis. In general, a lower-tier analysis can be used to screen first to determine whether it is adequate or whether there is sufficient concern to warrant an in-depth uncertainty analysis. Uncertainty analyses can also be tied to the range of potential effects on a chosen management decision or policy issue. The uncertainty might be negligible if it has a small

effect on a policy and invites careful examination if the stakes are high. The RIVM/MNP guidance adopted an approach that ranks uncertainties according to their importance for the policy issue and identifies where a more elaborate uncertainty assessment is warranted and its feasibility (van der Sluijs et al. 2003; Walker et al. 2003).

- Develop a template to support unified documentation of uncertainty analysis by stage or source of an IRIS assessment. For example, the template should summarize the options used, the methods used, the rationale of the choice, relevant results (such as range of variation), and open issues. IRIS assessments often use EPA BMDS software for dose-response modeling and declare that a model "fits the underlying dataset" if it meets a conventional goodness-of-fit criterion and then uses the fitted model for toxicity estimation. However, model-fitting has sometimes been achieved by fixing some model parameters rather than estimating them or by deleting some dose groups (NRC 2010; 2011). Failure to document the technical details amounts to omitting a source of uncertainty and potential bias.
- Develop strategies for communicating uncertainties. Engaging stakeholders earlier in the planning process as EPA is now doing can help to formulate critical messages. For example, it helps to ask such questions as, What are the main messages, and how will the stakeholders and general public receive these messages? What are the major assumptions involved in the main messages of policy decision? How robust are the major conclusions in light of the assumptions? Which aspects of uncertainties require additional attention and analysis? How clear should the statements on uncertainty be, and how can uncertainty be reported in a balanced and consistent fashion (van der Sluijs et al. 2004; Kloprogge et al. 2007)? IOM (2013) recommends always clearly acknowledging the existence of uncertainty and describing its source, magnitude, reducibility in the short term, and importance for the policy decision. Good planning and documentation of uncertainty analysis facilitate better communication and will probably improve stakeholders' confidence in the IRIS process.

FINDINGS AND RECOMMENDATIONS

Finding: EPA develops toxicity values for health effects for which there is "credible evidence of hazard" after chemical exposure and of an adverse outcome.

Recommendation: EPA should develop criteria for determining when evidence is sufficient to derive toxicity values. One approach would be to restrict formal dose-response assessments to when a standard descriptor characterizes the level of confidence as medium or high (as in the case of noncancer end points) or as "carcinogenic to humans" or "likely to be carcinogenic to humans" for carcinogenic compounds. Another approach, if EPA adopts probabilistic hazard classification, is to conduct formal dose-response assessments only when the posterior probability that a human hazard exists exceeds a predetermined threshold, such as 50% (more likely than not likely that the hazard exists).

Finding: EPA has made a number of substantive changes in the IRIS program since the publication of the NRC formaldehyde report, including the derivation and graphical presentation of multiple dose-response values and a shift away from choosing a particular study as the "best" study for derivation of dose-response estimates.

Recommendation: EPA should continue its shift toward the use of multiple studies rather than single studies for dose-response assessment but with increased attention to risk of bias, study quality, and relevance in assessing human dose-response relationships. For that purpose, EPA will need to develop a clear set of criteria for judging the relative merits of individual mechanistic, animal, and epidemiologic studies for estimating human dose-response relationships.

Finding: Although subjective judgments (such as identifying which studies should be included and how they should be weighted) remain inherent in formal analyses, calculation of toxicity values needs to be prespecified, transparent, and reproducible once those judgments are made.

Recommendation: EPA should use formal methods for combining multiple studies and the derivation of IRIS toxicity values with an emphasis on a transparent and replicable process.

Finding: EPA could improve documentation and presentation of dose-response information.

Recommendation: EPA should clearly present two dose-response estimates: a central estimate (such as a maximum likelihood estimate or a posterior mean) and a lower-bound estimate for a POD from which a toxicity value is derived. The lower bound becomes an upper bound for a cancer slope factor but remains a lower bound for a reference value.

Finding: Advanced analytic methods, such as Bayesian methods, for integrating data for dose-response assessments and deriving toxicity estimates are underused by the IRIS program.

Recommendation: As the IRIS program evolves, EPA should develop and expand its use of Bayesian or other formal quantitative methods in data integration for dose-response assessment and derivation of toxicity values.

Finding: IRIS-specific guidelines for consistent, coherent, and transparent assessment and communication of uncertainty remain incompletely developed. The inconsistent treatment of uncertainties remains a source of confusion and causes difficulty in characterizing and communicating uncertainty.

Recommendation: Uncertainty analysis should be conducted systematically and coherently in IRIS assessments. To that end, EPA should develop IRIS-specific guidelines to frame uncertainty analysis and uncertainty communication. Moreover, uncertainty analysis should become an integral component of the IRIS process.

REFERENCES

Akaike, H. 1974. A new look at the statistical model identification. IEEE T. Automat. Contr. 19(6):716-723.

Altmann, L., A. Bottger, and H. Wiegand. 1990. Neurophysiological and psychophysical measurements reveal effects of acute low-level organic solvent exposure in humans. Int. Arch. Occup. Environ. Health 62(7):493-499.

Altmann, L., H.F. Neuhann, U. Kramer, J. Witten, and E. Jermann. 1995. Neurobehavioral and neurophysiological outcome of chronic low-level tetrachloroethene exposure measured in neighborhoods of dry cleaning shops. Environ. Res. 69(2):83-89.

Blettner, M., W. Sauerbrei, B. Schlehofer, T. Scheuchenpflug, and C. Friedenreich. 1999. Traditional reviews, meta-analyses and pooled analyses in epidemiology. Int. J. Epidemiol. 28(1):1-9.

Boyes, W.K., M. Bercegeay, W.M. Oshiro, Q.T. Krntz, E.M. Kenyon, P.J. Bushnell, and V.A. Benignus. 2009. Acute perchloroethylene exposure alters rat visual-evoked potentials in relation to brain concentrations. Toxicol. Sci. 108(1):159-172.

Brickman, R., S. Jasanoff, and T. Ilgen. 1985. Controlling Chemicals: The Politics of Regulation in Europe and the United States. Ithaca: Cornell University Press.

Cavalleri, A., F. Gobba, M. Paltrinieri, G. Fantuzzi, E. Righi, and G. Aggazzotti. 1994. Perchloroethylene exposure can induce colour vision loss. Neurosci. Lett. 179(1-2):162-166.

DuMouchel, W.H., and J.E. Harris. 1983. Bayes methods for combining the results of cancer studies in humans and other species: Rejoinder. J. Am. Stat. Assoc. 78(382):313-315.

Echeverria, D., R.F. White, and C. Sampaio. 1995. A behavioral evaluation of PCE exposure in patients and dry cleaners: A possible relationship between clinical and preclinical effects. J. Occup. Environ. Med. 37(6):667-680.

EPA (U.S. Environmental Protection Agency). 1994. Methods for Derivation of Inhalation Reference Concentrations and Application of Inhalation Dosimetry. EPA/600/8-90/066F. Office of Research and Development, U.S. Environmental Protection Agency, Research Triangle Park, NC [online]. Available: http://www.epa.gov/raf/publications/pdfs/RFCMETHODOLOGY.PDF [accessed December 20, 2013].

EPA (U.S. Environmental Protection Agency). 2002. A Review of the Reference Dose and Reference Concentration Processes. Final report. EPA/630/P-02/002F. Risk Assessment Forum, U.S. Environmental Protection Agency, Washington, DC [online]. Available: http://www.epa.gov/raf/publications/pdfs/rfd-final.pdf [accessed December 18, 2013].

EPA (U.S. Environmental Protection Agency). 2004. An Examination of EPA Risk Assessment Principles and Practices. EPA/100/B-04/001. Risk Assessment Task Force, Office of Science Advisor, U.S. Environmental Protection Agency, Washington, DC [online]. Available: http://www.epa.gov/osa/pdfs/ratf-final.pdf [accessed December 19, 2013].

EPA (U.S. Environmental Protection Agency). 2005a. Guidelines for Carcinogen Risk Assessment. EPA/630/P-03/001F. Risk Assessment Forum, U.S. Environmental Protection Agency, Washington, DC. March 2005 [online]. Available: http://www.epa.gov/raf/publications/pdfs/CANCER_GUIDELINES_FINAL_3-25-05.PDF [accessed October 3, 2013].

EPA (U.S. Environmental Protection Agency). 2005b. Supplemental Guidance for Assessing Susceptibility from Early-Life Exposure to Carcinogens. EPA/630/R-03/003F. Risk Assessment Forum, U.S. Environmental Protection Agency, Washington, DC [online]. Available: http://www.epa.gov/ttn/atw/childrens_supplement_final.pdf [accessed December 20, 2013].

EPA (U.S. Environmental Protection Agency). 2012. Benchmark Dose Technical Guidance. EPA/100/R-12/001. Risk Assessment Forum, U.S. Environmental Protection Agency, Washington, DC [online]. Available: http://www.epa.gov/raf/publications/pdfs/benchmark_dose_guidance.pdf [accessed December 18, 2013].

EPA (U.S. Environmental Protection Agency). 2013a. EPA Risk Assessment Glossary [online]. Available: http://www.epa.gov/risk/glossary.htm [accessed December 19, 2013].

EPA (U.S. Environmental Protection Agency). 2013b. Part 1. Status of Implementation of Recommendations. Materials Submitted to the National Research Council, by Integrated Risk Information System Program, U.S. Environmental Protection Agency, January 30, 2013 [online]. Available: http://www.epa.gov/iris/pdfs/IRIS%20Program%20Materials%20to%20NRC_Part%201.pdf [accessed October 22, 2013].

EPA (U.S. Environmental Protection Agency). 2013c. Part 2. Chemical-Specific Examples. Materials Submitted to the National Research Council, by Integrated Risk Information System Program, U.S. Environmental Protection Agency, January 30, 2013 [online]. Available: http://www.epa.gov/iris/pdfs/IRIS%20Program%20Materials%20to%20NRC_Part%202.pdf [accessed December 19, 2013].

EPA (U.S. Environmental Protection Agency). 2013d. Toxicological Review of Methanol (Noncancer) (CAS No. 67-56-1) in Support of Summary Information on the Integrated Risk Information System (IRIS). EPA/635/R-11/001Fa. U.S. Environmental Protection Agency, Washington, DC. September 2013 [online]. Available: http://www.epa.gov/iris/toxreviews/0305tr.pdf [accessed October 22, 2013].

EPA (U.S. Environmental Protection Agency). 2013e. Toxicological Review of Benzo[a]pyrene (CAS No. 50-32-8) in Support of Summary Information on the Integrated Risk Information System (IRIS), Public Comment Draft. EPA/635/R13/138a. National Center for Environmental Assessment, Office of Research and Development, U.S. Environmental Protection Agency, Washington, DC. August 2013 [online]. Available: http://cfpub.epa.gov/ncea/iris_drafts/recordisplay.cfm?deid=66193 [accessed October 22, 2013].

Ferroni, C., L. Selis, A. Mutti, D. Folli, E. Bergamaschi, and I. Franchini. 1992. Neurobehavioral and neuroendocrine effects of occupational exposure to perchloroethylene. Neurotoxicology 13(1):243-247.

Freudenburg, W.R., R. Gramling, and D.J. Davidson. 2008. Scientific Certainty Argumentation Methods (SCAMs): Science and the politics of doubt. Sociol. Inq. 78(1):2-38 http://sciencepolicy.colorado.edu/students/envs_4800/freudenberg_2008.pdf.

Funtowicz, S.O., and J.R. Ravetz. 1992. Three types of risk assessment and the emergence of post-normal science. Pp. 251-274 in Social Theories of Risk, S. Krimsky, and D. Golding, eds. Westport, CT: Praeger.

Gobba, F., E. Righi, G. Fantuzzi, G. Predieri, L. Cavazzuti, and G. Aggazzotti. 1998. Two-year evolution of perchloroethylene-induced color-vision loss. Arch. Environ. Health 53(3):196-198.

Gustafson, P. 2004. Measurement Error and Misclassification in Statistics and Epidemiology: Impacts and Bayesian Adjustments. Boca Raton: Chapman and Hall/CRC Press.

Hake, C.L., and R.D. Stewart. 1977. Human exposure to tetrachloroethylene: Inhalation and skin contact. Environ. Health Perspect. 21:231-238.

Hoeting, J.A., D. Madigan, A.E. Raftery, and C.T. Volinsky. 1999. Bayesian model averaging: A tutorial with discussion. Stat. Sci. 14(4):382-417.

IOM (Institute of Medicine). 2013. Environmental Decision-making in the Face of Uncertainty. Washington, DC: National Academies Press.

IPCS (International Programme on Chemical Safety). 2006. Draft Guidance Document on Characterizing and Communicating Uncertainty of Exposure Assessment, Draft for Public Review. IPCS Project on the Harmonization of Approaches to the Assessment of Risk from Exposure to Chemicals. Geneva: World Health Organization [online]. Available: http://www.who.int/ipcs/methods/harmonization/areas/draftundertainty.pdf [accessed December 19, 2013].

Janssen, P.H.M., A.C. Petersen, J.P. van der Sluijs, J.S. Risbey, and J.R. Ravetz. 2003. RIVM/MNP Guidance for Uncertainty Assessment and Communication, Vol. 2. Quickscan Hints & Actions List. Netherlands Environmental Assessment Agency, National Institute for Public Health and the Environment (RIVM), Bilthoven, The Netherlands [online]. Available: http://www.rivm.nl/bibliotheek/digitaaldepot/Guidance_QS-HA.pdf [accessed December 20, 2013].

Johnson, B.B., and P. Slovic. 1995. Presenting uncertainty in health risk assessment: Initial studies of its effects on risk perception and trust. Risk Anal. 15(4):485-494.

Jones, D.R., J. Peters., J.L. Rushton, A.J. Sutton, and K.R. Abrams. 2009. Interspecies extrapolation in environmental exposure standard setting: A Bayesian synthesis approach. Regul. Toxicol. Pharmacol. 53(3):217-225.

Joslyn, S.L., and J.E. LeClerc. 2012. Uncertainty forecasts improve weather related decisions and attenuate the effects of forecast error. J. Exp. Psychol. Appl. 18(1):126-140.

Joslyn, S., L. Nemec, and S. Savelli. 2013. The benefits and challenges of predictive interval forecasts and verification graphics for end users. Weather Climate Soc. 5(2):133-147.

Kjellstrand, P., B. Holmquist, I. Jonsson, S. Romare, and L. Mansson. 1985. Effects of organic solvents on motor activity in mice. Toxicology 35(1):35-46.

Kloprogge, P., J. van der Sluijs, and J. Wardekker. 2007. Uncertainty Communication: Issues and Good Practice. Copernicus Institute for Sustainable Development and Innovation, Utrecht University, The Netherlands [online]. Available: http://www.nusap.net/downloads/reports/uncertainty_communication.pdf [accessed December 19, 2013].

Loccisano, A., S. Peddada, M. Andersen, H. Clewell, and M. Longnecker. 2012. Use of physiologically-based pharmacokinetic models to evaluate epidemiologic associations that may be due to reverse causality. Epidemiology 23(5S):S-331 [Abstract No. S-078].

Lyman, G.H., and N.M. Kuderer. 2005. The strengths and limitations of meta-analyses based on aggregate data. Med. Res. Method. 5(1):14.

Mattsson, J.L., R.R. Albee, B.L. Yano, G. Bradley, and P.J. Spencer. 1998. Neurotoxicologic examination of rats exposed to 1,1,2, 2-tetrachloroethylene (perchloroethylene) vapor for 13 weeks. Neurotoxicol. Teratol. 20(1):83-98.

NRC (National Research Council). 2006. Health Risks from Dioxin and Related Compounds: Evaluation of the EPA Reassessment. Washington, DC: National Academies Press.

NRC (National Research Council). 2007. Models in Environmental Regulatory Decision Making. Washington, DC: The National Academies Press.

NRC (National Research Council). 2009. Science and Decisions: Advancing Risk Assessment. Washington, DC: National Academies Press.

NRC (National Research Council). 2010. Review of the Environmental Protection Agency's Draft IRIS Assessment of Tetrachloroethylene. Washington, DC: National Academies Press.

NRC (National Research Council). 2011. Review of the Environmental Protection Agency's Draft IRIS Assessment of Formaldehyde. Washington, DC: National Academies Press.

NRC (National Research Council). 2012. Science for Environmental Protection: The Road Ahead. Washington, DC: National Academies Press.

Orsini, N., R. Li, A. Wolk, P. Khudyakov, and D. Spiegelman. 2012. Meta-analysis for linear and nonlinear dose-response relations: Examples, an evaluation of approximations, and software. Am. J. Epidemiol. 175(1):66-73.

Oshiro, W.M., Q.T. Krantz, and P.J. Bushnell. 2008. Characterization of the effects of inhaled perchloroethylene on sustained attention in rats performing a visual signal detection task. Neurotoxicol. Teratol. 30(3):167-174.

Peters, J.L., L. Rushton, A.J. Sutton, D.R. Jones, K.R. Abrams, and M.A. Mugglestone. 2005. Bayesian methods for the cross-design synthesis of epidemiological and toxicological evidence. J. R. Stat. Soc. C-Appl. Stat. 54(1):159-172.

Petersen, A.C., P.H.M. Janssen, J.P. van der Sluijs, J.S. Risbey, and J.R. Ravetz. 2003. RIVM/MNP Guidance for Uncertainty Assessment and Communication, Vol. 1. Mini-Checklist and Quickscan Questionnaire. Netherlands Environmental Assessment Agency, National Institute for Public Health and the Environment (RIVM), Bilthoven, The Netherlands [online]. Available: http://www.rivm.nl/bibliotheek/digitaaldepot/Guidance_MC_QS-Q.pdf [accessed December 19, 2013].

Rosengren, L.E., P. Kjellstrand, and K.G. Haglid. 1986. Tetrachloroethylene: Levels of DNA and S-100 in the gerbil CNS after chronic exposure. Neurobehav. Toxicol. Teratol. 8(2):201-206.

Schmid, C.H., M. Landa, T.H. Jafar, I. Giatras, T. Karim, M. Reddy, P.C. Stark, and A.S. Levey. 2003. Constructing a database of individual clinical trials for longitudinal analysis. Control. Clin. Trial. 24(3):324-340.

Schreiber, J.S., H.K. Hudnell, A.M. Geller, D.E. House, K.M. Aldous, M.S. Force, K. Langguth, E.J. Prohonic, and J.C. Parker. 2002. Apartment residents' and day care workers' exposures to tetrachloroethylene and deficits in visual contrast sensitivity. Environ. Health Perspect. 110(7):655-664.

Seeber, A. 1989. Neurobehavioral toxicity of long-term exposure to tetrachloroethylene. Neurotoxicol. Teratol. 11(6):570-583.

Small, M.J. 2008. Methods for assessing uncertainty in fundamental assumptions and associated models for cancer risk assessment. Risk Anal. 28(5):1289-1308.

Smithson, M. 2008. The many faces and masks of uncertainty. Pp. 13-25 in Uncertainty and Risk: Multidisciplinary Perspectives, G. Bammer, and M. Smithson, eds. London: Earthscan.

Spiegelhalter, D.J., N.G. Best, B.P. Carlin, and A. van der Linde. 2002. Bayesian measures of model complexity and fit. J. R. Stat. Soc. B 64(4):583-639.

Spinatonda, G., R. Colombo, E.M. Capodaglio, M. Imbriani, C. Pasetti, G. Minuco, and P. Pinelli. 1997. Process of speech production: Application in a group of subjects chronically exposed to organic solvents (II) [in Italian]. G. Ital. Med. Lav. Ergon. 19(3):85-88.

Stedeford, T., Q.J. Zhao, M.L. Dourson, M. Banasik, and C.H. Hsu. 2007. The application of non-default uncertainty factors in the U.S. EPA's Integrated Risk Information System (IRIS). Part I: UF(L), UF(S), and "other uncertainty factors". J. Environ. Sci. Health C Environ. Carcinog. Ecotoxicol. Rev. 25(3):245-279.

Steenland, K., Mannetje, A., Bofetta, P., Stayner, L., Attfield, M., Chen, J., Dosemeci, M., DeKlerk, N., Hnizdo, E, Koskela, R., and H. Checkoway. 2001. Pooled exposure-response analyses and risk assessment for lung cancer in 10 cohorts of silica-exposed workers: an IARC multicentre study. Can. Caus. Control. 12(9):773-784.

Stroup, D.F., J.A. Berlin, S.C. Morton, I. Olkin, G.D. Williamson, D. Rennie, D. Moher, B.J. Becker, T.A. Sipe, and S.B. Thacker. 2000. Meta-analysis of observational studies in epidemiology: A proposal for reporting. JAMA 283(15):2008-2012.

Sutton, A.J., and K.R. Abrams. 2001. Bayesian methods in meta-analysis and evidence synthesis. Stat. Methods Med. Res. 10(4):277-303.

van der Sluijs, J.P., J.S. Risbey, P. Kloprogge, J.R. Ravetz, S.O. Funtowicz, S.C. Quintana, A.G. Pereira, B. De Marchi, A.C. Petersen, P.H.M. Janssen, R. Hoppe, and S.W.F. Huijs. 2003. RIVM/MNP Guidance for Uncertainty Assessment and Communication, Vol. 3. Detailed Guidance. Copernicus Institute for Sustainable Development and Innovation, Utrecht University, Utrecht, The Netherlands [online]. Available: http://www.nusap.net/downloads/detailedguidance.pdf [accessed December 20, 2013].

van der Sluijs, J.P., P.H.M. Janssen, A.C. Petersen, P. Kloprogge, J.S. Risbey, W. Tuinstra, and J.R. Ravetz. 2004. RIVM/MNP Guidance for Uncertainty Assessment and Communication, Vol. 4. Tool Catalogue for Uncertainty Assessment. Copernicus Institute for Sustainable Development and Innovation, Utrecht University, Utrecht, The Netherlands [online]. Available: http://www.nusap.net/downloads/toolcatalogue.pdf [accessed December 20, 2013].

van der Sluijs, J.P., A.C. Petersen, P.H.M. Janssen, J.S. Risbey, and J.R. Ravetz. 2008. Exploring the quality of evidence for complex and contested policy decisions. Environ. Res. Lett. 3(2):024008.

Walker, W.E., P. Harremoes, J. Rotmans, J.P. van der Sluijs, M.B.A. van Asselt, P. Janssen, and M.P. Krayer von Krauss. 2003. Defining uncertainty: A conceptual basis for uncertainty management in model-based decision support. Integ. Assess. 4(1):5-17.

Wang, S., J.E. Karlsson, T. Kyrklund, and K. Haglid. 1993. Perchloroethylene-induced reduction in glial and neuronal cell marker proteins in rat brain. Pharmacol. Toxicol. 72(4-5):273-278.

Zhu, Y. 2005. Dose-time-response modeling of longitudinal measurements for neurotoxicity risk assessment. Environmetrics 16(6):603-617.

Zhu, Y., M.R. Wessel, T. Liu, and V.C. Moser. 2005a. Analyses of neurobehavioral screening data: Dose-time-response modeling of continuous outcomes. Regul. Toxicol. Pharmacol. 41(3):240-255.

Zhu, Y., Z. Jia, W. Wang, J. Gift, V.C. Moser, and B.J. Pierre-Louis. 2005b. Analyses of neurobehavioral screening data: Benchmark dose estimation. Regul. Toxicol. Pharmacol. 42(2):190-201.

8

Future Directions

The US Environmental Protection Agency (EPA) clearly has embraced the need for revision of its Integrated Risk Information System (IRIS). It has begun to implement changes that follow the general guidance given in Chapter 7 of the report *Review of the Environmental Protection Agency's Draft IRIS Assessment of Formaldehyde* (NRC 2011). The present report has reviewed and evaluated those changes and has provided additional suggestions for further improving the process. This chapter highlights several recommendations that should receive high priority, considers issues that extend over the full process, and presents suggestions directed at ensuring that the IRIS program provides the best possible assessments in the future.

OVERALL EVALUATION

The IRIS program provided substantial materials to the committee that documented the general strategy for the continuing revisions of its process (see Table 1-1) and various materials that constitute the building blocks of the revision: the draft preamble (EPA 2013a, Appendix B) and the draft handbook (EPA 2013a, Appendix F). EPA also provided examples of assessments in which elements of the revisions have been implemented (EPA 2013b,c). Although the committee recognizes that EPA has not yet completed its revisions of the methods used for IRIS assessments, it is able to assess the general approach taken by the agency and the trajectory of change in the assessment approach. The committee commends EPA for its substantive new approaches, continuing commitment to improving the process, and successes to date. Overall, the committee expects that EPA will complete its planned revisions in a timely way and that the revisions will transform the IRIS program.

EPA has responded to many of the suggestions made in the 2011 NRC formaldehyde report at a particularly challenging time for the IRIS program. Its response to the formaldehyde report acknowledges the current context and makes changes, not only in scientific methods but in the underlying principles and in public processes. Critical changes in leadership positions have occurred since the release of the NRC formaldehyde report. Kenneth Olden, the newly named director of the EPA National Center for Environmental Assessment, has made a far-reaching effort to engage the full array of stakeholders, including the general public, in providing input into the changes being made. The revisions embrace stakeholder engagement in all relevant phases of the process. Under its acting director, Vincent Cogliano, the IRIS program has moved forward steadily in planning for and implementing changes in each element of the assessment process. The committee is confident that there is an institutional commitment to completing the revisions of the process even as the program continues through the current transition phase and is closely watched by both stakeholders and Congress.

As reviewed in the preceding chapters, the committee found that appropriate revisions of all elements of the IRIS assessment process (see Figure 1-2) were underway or planned. The preamble represents a useful general framework for the assessments, and the committee recommends that this document and the draft handbook be completed and reviewed. The committee

also found that the proposed format for the assessments should enhance "user friendliness" and transparency. The evidence tables and data displays in the new documents are moving to the standard practice for systematic reviews.

SPECIFIC RECOMMENDATIONS

The present committee offers a long series of findings and recommendations in this report. The recommendations can be categorized as those directly relevant to current revisions of the process, those related to future refinements of the methods, or those calling for research related to the assessment process that will provide results for guiding the further evolution of the methods used in the process. The committee urges that high priority be assigned to those in the first category and has listed them in Box 8-1. A timetable should be developed for the other recommendations. For the longer term, as discussed below, EPA will need to establish procedures to ensure that there is continuing refinement of the assessment methods. In addition, the IRIS program will need resources to conduct or commission focused methodologic research.

LESSONS LEARNED

The present committee has looked retrospectively at the methods and performance of the IRIS program and evaluated the changes that have been implemented. The 2011 NRC formaldehyde report cited a number of "lessons learned." Here, the committee offers several that are deemed critical for ensuring that the IRIS program provides the best possible assessments.

- *Assessment methods should be updated in a continuing, strategic fashion*. The 2011 NRC formaldehyde report found that state-of-the-art approaches that had widespread application in other fields, such as systematic review, were not being used in the IRIS assessment process. Even as the IRIS program undergoes revision, consideration needs to be given to how methods relevant to all elements of the process will evolve continuously. The revisions in progress should include consideration of how relevant progress in risk assessment and other domains will be tracked and incorporated into the IRIS assessment approach.
- *Inefficiencies in the IRIS program need to be systematically identified and addressed.* Although the present committee recognizes that factors beyond the program itself create delay, it urges the IRIS program to consider systematically how delay occurs so that it can be anticipated and addressed. Some of the most controversial assessments have had long histories with multiple cycles of revision and review. Some assessments that have been delayed have involved review of substantial bodies of evidence and continuing publication of relevant evidence. EPA has examined the timing of its process and has proposed principles for stopping rules related to such issues as literature identification and the addition of important new documents to assessments. Although principles have been offered, their application in practice could prove challenging, and monitoring of adherence to stopping rules will be needed. Collaborating with other agencies, such as the National Toxicology Program, to avoid duplication of effort is another important efficiency-promoting activity.
- *Evolving competences that reflect new scientific directions are needed.* The conduct of an IRIS assessment necessarily involves multiple scientific disciplines, and as research methods and data streams change, EPA management will need to ensure that the chemical-assessment teams and chemical-assessment support teams have appropriate expertise and training. The IRIS program needs continuing evaluation of its expertise in relation to changing scientific contexts.

BOX 8-1 Recommendations Directly Relevant to Current Revisions

Chapter 2

- EPA needs to complete the changes in the IRIS process that are in response to the recommendations in the NRC formaldehyde report and specifically complete documents, such as the draft handbook, that provide detailed guidance for developing IRIS assessments. When those changes and the detailed guidance, such as the draft handbook, have been completed, there should be an independent and comprehensive review that evaluates how well EPA has implemented all the new guidance. The present committee is completing its report while those revisions are still in progress.
- EPA should provide a quality-management plan that includes clear methods for continuing assessments of the quality of the process. The roles of the various internal entities involved in the process, such as the chemical-assessment support teams, should be described. The assessments should be used to improve the overall process and the performance of EPA staff and contractors.

Chapter 3

- EPA should establish a transparent process for initially identifying all putative adverse outcomes through a broad search of the literature. The agency should then develop a process that uses guided expert judgment to identify the specific adverse outcomes to be investigated, each of which would then be subjected to systematic review of human, animal, and in vitro or mechanistic data.
- EPA should include protocols for all systematic reviews conducted for a specific IRIS assessment as appendixes to the assessment.

Chapter 4

- The current process can be enhanced with more explicit documentation of methods. Protocols for IRIS assessments should include a section on evidence identification that is written in collaboration with information specialists trained in systematic reviews and that includes a search strategy for each systematic-review question being addressed in the assessment. Specifically, the protocols should provide a line-by-line description of the search strategy, the date of the search, and publication dates searched and explicitly state the inclusion and exclusion criteria for studies.

Chapter 5

- To advance the development of tools for assessing risk of bias in different types of studies (human, animal, and mechanistic) used in IRIS assessments, EPA should explicitly identify factors that can lead to bias in animal studies—such as control for litter effects, dosing, and methods for exposure assessment—so that these factors are consistently evaluated for experimental studies. Likewise, EPA should consider a tool for assessing risk of bias in in vitro studies.
- When considering any method for evaluating individual studies, EPA should select a method that is transparent, reproducible, and scientifically defensible. Whenever possible, there should be empirical evidence that the methodologic characteristics that are being assessed in the IRIS protocol have systematic effects on the direction or magnitude of the outcome. The methodologic characteristics that are known to be associated with a risk of bias should be included in the assessment tool. Additional quality-assessment items relevant to a particular systematic-review question could also be included in the EPA assessment tool.

(Continued)

BOX 8-1 Continued

- Although additional methodologic work might be needed to establish empirically supported criteria for animal or mechanistic studies, an IRIS assessment needs to include a transparent evaluation of the risk of bias of studies used by EPA as a primary source of data for the hazard assessment. EPA should specify the empirically based criteria it will use to assess risk of bias for each type of study design in each type of data stream.
- To maintain transparency, EPA should publish its risk-of-bias assessments as part of its IRIS assessments. It could add tables that describe the assessment of each risk-of-bias criterion for each study and provide a summary of the extent of the risk of bias in the descriptions of each study in the evidence tables.
- The risk-of-bias assessment of individual studies should be carried forward and incorporated into the evaluation of evidence among data streams.

Chapter 6

- EPA should continue to improve its evidence-integration process incrementally and enhance the transparency of its process. It should either maintain its current guided-expert-judgment process but make its application more transparent or adopt a structured (or GRADE-like) process for evaluating evidence and rating recommendations along the lines that NTP has taken. If EPA does move to a structured evidence-integration process, it should combine resources with NTP to leverage the intellectual resources and scientific experience in both organizations. The committee does not offer a preference but suggests that EPA consider which approach best fits its plans for the IRIS process.
- EPA should expand its ability to perform quantitative modeling of evidence integration; in particular, it should develop the capacity to do Bayesian modeling of chemical hazards. That technique could be helpful in modeling assumptions about the relevance of a variety of animal models to each other and to humans, in incorporating mechanistic knowledge to model the relevance of animal models to humans and the relevance of human data for similar but distinct chemicals, and in providing a general framework within which to update scientific knowledge rationally as new data become available. The committee emphasizes that the capacity for quantitative modeling should be developed in parallel with improvements to existing IRIS evidence-integration procedures and that IRIS assessments should not be delayed while this capacity is being developed.

Chapter 7

- EPA should develop criteria for determining when evidence is sufficient to derive toxicity values. One approach would be to restrict formal dose-response assessments to when a standard descriptor characterizes the level of confidence as medium or high (as in the case of noncancer end points) or as "carcinogenic to humans" or "likely to be carcinogenic to humans" for carcinogenic compounds. Another approach, if EPA adopts probabilistic hazard classification, is to conduct formal dose-response assessments only when the posterior probability that a human hazard exists exceeds a predetermined threshold, such as 50% (more likely than not likely that the hazard exists).
- EPA should clearly present two dose-response estimates: a central estimate (such as a maximum likelihood estimate or a posterior mean) and a lower-bound estimate for a POD from which a toxicity value is derived. The lower bound becomes an upper bound for a cancer slope factor but remains a lower bound for a reference value.

LOOKING FORWARD

The committee has looked beyond the approach for revising the process and the timespan of its recommendations. As EPA completes the current revisions, it needs to consider developing a strategic plan for continuous updating of the IRIS methodology. The strategic plan should be sufficiently flexible to consider a variety of approaches that incorporate advances in fields relevant to the IRIS program. For example, such a strategic plan should address

- Using data from emerging technologies of molecular toxicology.
- Incorporating new statistical methods.
- Applying advances in data retrieval and text-mining.

The committee urges that the strategic plan consider the human and technologic resources that are needed to carry out the IRIS assessments and to support methodologic research and the implementation of new approaches. The program is already challenged by the number of assessments that it conducts and by the need to increase the pace at which assessments are completed. It is now faced with the additional burden of fundamental revisions of its methods. EPA needs to evaluate carefully the demands on its staff and the load carried by its contractors and consultants. For now, sufficient financial and staff resources need to be available to complete the revisions of the process; for the future, increased capacity is needed for methodologic work and incorporation of modifications into the assessment approach.

Consideration will also need to be given to how changes in the IRIS assessment process will be reviewed before implementation. The committee suggests that EPA consider whether its Chemical Assessment Advisory Committee or Science Advisory Board could be used for this purpose. Peer review should be an integral part of the revision process, and the IRIS system should be sufficiently dynamic to reflect relevant advances.

REFERENCES

EPA (U.S. Environmental Protection Agency). 2013a. Part 1. Status of Implementation of Recommendations. Materials Submitted to the National Research Council, by Integrated Risk Information System Program, U.S. Environmental Protection Agency, January 30, 2013 [online]. Available: http://www.epa.gov/iris/pdfs/IRIS%20Program%20Materials%20to%20NRC_Part%201.pdf [accessed November 13, 2013].

EPA (U.S. Environmental Protection Agency). 2013b. Part 2. Chemical-Specific Examples. Materials Submitted to the National Research Council, by Integrated Risk Information System Program, U.S. Environmental Protection Agency, January 30, 2013 [online]. Available: http://www.epa.gov/iris/pdfs/IRIS%20Program%20Materials%20to%20NRC_Part%202.pdf [accessed December 19, 2013].

EPA (U.S. Environmental Protection Agency). 2013c. Toxicological Review of Benzo[a]pyrene (CAS No. 50-32-8) in Support of Summary Information on the Integrated Risk Information System (IRIS), Public Comment Draft. EPA/635/R13/138a. National Center for Environmental Assessment, Office of Research and Development, U.S. Environmental Protection Agency, Washington, DC. August 2013 [online]. Available: http://cfpub.epa.gov/ncea/iris_drafts/recordisplay.cfm?deid=66193 [accessed Nov. 13, 2013].

NRC (National Research Council). 2011. Review of the Environmental Protection Agency's Draft IRIS Assessment of Formaldehyde. Washington, DC: National Academies Press.

Appendix A

Biographic Information on the Committee to Review the IRIS Process

Jonathan M. Samet (*Chair*) is a pulmonary physician and epidemiologist. He is a professor and the Flora L. Thornton Chair of the Department of Preventive Medicine of the Keck School of Medicine of the University of Southern California (USC) and director of the USC Institute for Global Health. Dr. Samet's research has focused on the health risks posed by inhaled pollutants. He has served on numerous committees concerned with public health: the US Environmental Protection Agency Science Advisory Board; committees of the National Research Council (NRC), including chairing the Biological Effects of Ionizing Radiation VI Committee, the Committee on Research Priorities for Airborne Particulate Matter, the Committee to Review EPA's Draft IRIS Assessment of Formaldehyde, the Committee to Develop a Research Strategy for Environmental, Health, and Safety Aspects of Engineered Nanomaterials, and the Board on Environmental Studies and Toxicology; the National Cancer Advisory Board; and committees of the Institute of Medicine (IOM). He is a member of IOM. Dr. Samet received his MD from the University of Rochester, School of Medicine and Dentistry.

Scott Bartell is associate professor in public health, statistics, and epidemiology at the University of California, Irvine. His research interest is in environmental-health methodology with applications in environmental epidemiology, exposure science, and risk assessment. His recent projects include epidemiologic analysis of particulate-matter exposure and arrhythmia in the Cardiovascular Health and Air Pollution Study, linkage of fate and transport models and a pharmacokinetic model for perfluorooctanoic acid with data from the C8 Health Project, and development of statistical methods for biomarker-based exposure estimation and for epidemiologic analysis of aggregated data. He has served on a variety of scientific advisory committees for the National Research Council, the Environmental Protection Agency, the Centers for Disease Control and Prevention, the National Institute of Environmental Health Sciences, and the Department of Energy. Dr. Bartell earned a PhD in epidemiology and an MS in statistics from the University of California, Davis and an MS in environmental health from the University of Washington.

Lisa Bero is a professor in the Department of Clinical Pharmacy and Institute for Health Policy Studies of the University of California, San Francisco. She is also the director of the San Francisco Branch of the United States Cochrane Center. Her research interests include methods for meta-analysis and critical appraisal of research, academic–industry relations, pharmaceutical outcomes assessment, pharmacology, tobacco-control policy, and translation of research into policy. Dr. Bero is a member of the World Health Organization Guideline Review Committee and the Advisory Committee on Health Research of the Pan American Health Organization. In addition, she is a member of the Institute of Medicine Board on Health Care Services. Dr. Bero received a PhD in pharmacology and toxicology from Duke University.

Ann Bostrom is the Weyerhaeuser endowed professor of environmental policy at the Daniel J. Evans School of Public Affairs of the University of Washington. Her research focuses on risk perception, communication, and management and on environmental policy and decision-making under uncertainty. She serves as an associate editor or a risk-communication editor for *Journal of Risk Research* and the journal *Human and Ecological Risk Assessment* and is on the editorial board of *Risk Analysis*. Dr. Bostrom is an elected fellow of the American Association for the Advancement of Science and past president and an elected fellow of the Society for Risk Analysis. She has served on several National Research Council committees, including the Committee on Use of Emerging Science for Environmental Health Decisions. Dr. Bostrom received a PhD in public-policy analysis from Carnegie Mellon University.

Kay Dickersin is a professor and director of clinical trials at the Johns Hopkins Bloomberg School of Public Health. Her major research interests are related to randomized clinical trials, trial registers, systematic reviews and meta-analysis, publication bias, peer review, evidence-based health care, and comparative-effectiveness research. Dr. Dickersin has also conducted studies in such fields as women's health, eyes and vision, and surgery. She is director of the US Cochrane Center, one of 14 centers worldwide participating in the Cochrane Collaboration, which aims to help people to make well-informed decisions about health by preparing, maintaining, and promoting the accessibility of systematic reviews of available evidence on the benefits and risks associated with health care. She has served as a member of several National Research Council committees and is a member of the Institute of Medicine. Dr. Dickersin received a PhD in epidemiology from the Johns Hopkins University School of Hygiene and Public Health.

David C. Dorman is a professor of toxicology in the Department of Molecular Biosciences of North Carolina State University. The primary objective of his research is to provide a refined understanding of chemically induced neurotoxicity in laboratory animals that will lead to improved assessment of potential neurotoxicity in humans. Dr. Dorman's research interests include neurotoxicology, nasal toxicology, pharmacokinetics, and cognition and olfaction in military working dogs. He served as a member of the National Research Council Committee on Animal Models for Testing Interventions Against Aerosolized Bioterrorism Agents, as chair of the Committee on Emergency and Continuous Exposure Guidance Levels for Selected Submarine Contaminants and of the Committee to Evaluate Potential Health Risks from Recurrent Lead Exposure of DOD Firing Range Personnel, and as a member of the Committee to Review EPA's Draft IRIS Assessment of Formaldehyde. He received his DVM from Colorado State University. He completed a combined PhD and residency program in toxicology at the University of Illinois at Urbana-Champaign and is a diplomate of the American Board of Veterinary Toxicology and the American Board of Toxicology.

David L. Eaton is a professor of environmental and occupational health sciences and dean and vice provost of the graduate school at the University of Washington (UW). He also serves as the director of the National Institute of Environmental Health Sciences Center for Ecogenetics and Environmental Health at UW. He has held several other UW positions, including toxicology program director in and associate chairman of the Department of Environmental Health and associate dean for research in the School of Public Health. Dr. Eaton maintains an active research and teaching program that is focused on the molecular basis of environmental causes of cancer and how human genetic differences in biotransformation enzymes may increase or decrease individual susceptibility to chemicals in the environment. He has published over 150 scientific articles and book chapters in toxicology and risk assessment. Nationally, he has served on the Board of Directors and as treasurer of the American Board of Toxicology, as secretary and later as president of the Society of Toxicology, as a member of the Board of Directors and as vice-president

of the Toxicology Education Foundation, and as a member of the Board of Trustees of the Academy of Toxicological Sciences. Dr. Eaton is a member of the Institute of Medicine and has served on several National Research Council committees. He is an elected fellow of the American Association for the Advancement of Science and the Academy of Toxicological Sciences. Dr. Eaton earned a PhD in pharmacology from the University of Kansas Medical Center.

Joe G. Garcia is a professor of medicine and the senior vice president for health sciences at the Arizona Health Sciences Center of the University of Arizona. He is internationally recognized for his expertise in the genetic basis of lung disease and the prevention of and treatment for inflammatory lung injury. Dr. Garcia's research focuses on understanding the biochemical and molecular basis of lung inflammation, especially vascular leak, in which blood cells and fluid escape from small vessels and cause edema in the surrounding tissues, especially the lungs. He is a past president of the Central Society for Clinical Research and a member of the Board of Directors of the American Thoracic Society and has been a member or chairman of several committees of the National Institutes of Health. In addition, Dr. Garcia is a member of the Institute of Medicine, the Association of American Physicians, and the American Society of Clinical Investigation. He received an MD from the University of Texas Southwestern Medical Center.

Miguel Hernán is a professor of epidemiology and biostatistics at the Harvard University School of Public Health and affiliated faculty at the Harvard-MIT Division of Health Sciences and Technology. His research is focused on methods for causal inference, including comparative effectiveness of policy and clinical interventions. Dr. Hernán and his collaborators combine observational data, mostly untestable assumptions, and statistical methods to emulate hypothetical randomized experiments. His research group emphasizes the need to formulate well-defined causal questions and the use of analytic approaches whose validity does not require assumptions that conflict with current subject-matter knowledge. Dr. Hernán is an editor of the journal *Epidemiology* and has served on the Institute of Medicine Committee on Ethical and Scientific Issues in Studying the Safety of Approved Drugs. He received an MD from the Universidad Autónoma de Madrid in Spain.

James S. House is Angus Campbell Distinguished University Professor of Survey Research, Public Policy, and Sociology at the University of Michigan. His research interests include social psychology, political sociology, social structure and personality, psychosocial and socioeconomic factors in health, and survey research methods. Dr. House has worked in sociology and social epidemiology to understand the effects of broader social structures and processes on people's attitudes, behavior, well-being, and especially health. His and his colleagues' research has helped to demonstrate the adverse effects of occupational and other forms of stress on health and how social relationships and supports can buffer or mitigate the deleterious health effects of stress and promote health more generally. Over the last 2 decades he has focused on describing and understanding social disparities in health over time and the life course, especially as related to socioeconomic position. Dr. House is a member of the American Academy of Arts and Sciences, the National Academy of Sciences, and the Institute of Medicine. He has served on the National Research Council Panel on Race, Ethnicity, and Health in Later Life. Dr. House received a PhD in social psychology from the University of Michigan.

Margaret M. MacDonell is a program manager in the Environmental Science Division of Argonne National Laboratory. She conducts integrated environmental health analyses, primarily for federal agencies. She has professional interests in cumulative impact and risk; integrated environmental fate, exposure, and health-effects analyses on multiple stressors, including chemical mixtures, nanomaterials, and other hazards, such as ones related to energy development; integrated impact analyses of sustainability; and community involvement in environmental health protection. Dr. MacDonell developed risk training workshops for environmental managers and

practitioners, including people in state agencies and tribal nations. She collaborated with the Environmental Protection Agency National Homeland Security Research Center to develop acute and short-term exposure advisories for chemical, radiologic, and biologic contaminants released into drinking water and buildings. She serves on two National Research Council committees: the Committee on Toxicology and the Committee on Acute Exposure Guideline Levels. Dr. MacDonell received a PhD in environmental health engineering from Northwestern University.

Richard P. Scheines is a professor and the head of philosophy at Carnegie Mellon University. His research focuses on causal discovery, specifically the problem of learning about causation from statistical evidence. Dr. Scheines also works in building and researching the effectiveness of educational software, ranging from intelligent-proof tutors to virtual-causality laboratories to a full-semester course on causal and statistical reasoning. Because of that work, he has a courtesy appointment in the Human-Computer Interaction Institute of Carnegie Mellon. He served as a member of the National Research Council Committee on Evaluation of the Presumptive Disability Decision-Making Process for Veterans and the Committee on Food Marketing and the Diets of Children and Youth. Dr. Scheines received a PhD in the history and philosophy of science from the University of Pittsburgh.

Leonard M. Siegel is director of the Center for Public Environmental Oversight, a project of the Pacific Studies Center that facilitates public participation in the oversight of military environmental programs, federal facilities cleanup, and brownfield revitalization. He is one of the environmental movement's leading experts in military-facility contamination, community oversight of cleanup, and the vapor-intrusion pathway. For his organization, he runs two Internet newsgroups: the Military Environmental Forum and the Brownfields Internet Forum. He is a member of the Interstate Technology and Regulatory Council Munitions Response Work Team, the California Department of Toxic Substances Control External Advisory Group, and the Moffett Field (formerly Moffett Naval Air Station) Restoration Advisory Board. He has served on several committees of the National Research Council, currently as a member of the Committee on the Future Options for Management in the Nation's Subsurface Remediation Effort. Mr. Siegel studied physics at Stanford University.

Robert B. Wallace is a professor in and director of the Center on Aging in the Departments of Epidemiology and Internal Medicine of the University of Iowa. His research interests include the epidemiology and prevention of aging-related chronic conditions, such as disabling illnesses of older persons, including arthritis, cancer, cardiovascular diseases, and dementia; clinical trials; disease prevention; epidemiology; health promotion; preventive medicine; and public health. Dr. Wallace is a member of the Institute of Medicine (IOM), chairs the IOM Board on the Health of Select Populations, and has been a member or chair of numerous IOM committees. Dr. Wallace received an MD from Northwestern University.

Yiliang Zhu is a professor in the Department of Epidemiology and Biostatistics of the University of South Florida College of Public Health. He is also director of the college's Center for Collaborative Research. His current research is focused on quantitative methods in health risk assessment, including modeling of biologic systems via pharmacokinetics and pharmacodynamics, dose-response modeling, benchmark-dose methods, and uncertainty quantification. He also conducts research in disease surveillance, health-outcome evaluation, and impact assessment of health-care systems and policies in rural China. Dr. Zhu was a member of the National Research Council Committee on EPA's Exposure and Human Health Assessment of Dioxin and Related Compounds, Committee on Tetrachloroethylene, and Committee to Review EPA's Draft IRIS Assessment of Formaldehyde. He received a PhD in statistics from the University of Toronto.

Appendix B

Workshop Agenda on Weight of Evidence

March 27-29, 2013
National Academy of Sciences
2101 Constitution Ave., N.W.
Washington, DC 20418

WEDNESDAY, MARCH 27, 2013

8:00 **Welcome to Workshop**
Jonathan Samet
Chair, Committee to Review the IRIS Process
Professor and Flora L. Thornton Chair, Department of Preventive Medicine
Keck School of Medicine, University of Southern California

ASSEMBLING THE EVIDENCE

This session will address approaches to identifying evidence on agents being considered in IRIS assessments. It will cover methods for searching literature and other data bases. The session will also consider the complicating issues of publication bias, "the grey literature," selective publication of model results, and access to primary data. A further major set of topics include the use of systematic approaches for characterizing study quality, methods for qualitatively and quantitatively assessing heterogeneity across studies, and use of quantitative synthesis (meta-analysis). An additional topic, potentially relevant to some assessments, is whether all assessments need a comprehensive review of the literature.

8:15 **Introduction and Overview of Session**
Lisa Bero
Member, Committee to Review the IRIS Process
Professor, Department of Clinical Pharmacy
University of California, San Francisco

8:25 **Systematic Review of Animal Studies and Approaches for Characterizing Study Quality**
Malcolm MacLeod
Professor of Neurology and Translational Neuroscience
University of Edinburgh

8:40 **Systematic Review of Human Studies and Approaches for Characterizing Study Quality**
Karen Robinson
Associate Professor
Medicine, Epidemiology, and Health Policy and Management
Johns Hopkins Medical Institutions

8:55 **Development, Maintenance, and Use of an Air Pollution Data Base**
Richard Atkinson
Senior Lecturer in Epidemiology
St. George's University of London

9:10 **Panel Discussion with Speakers on Assembling the Evidence**

Key Questions

(1) Do IRIS assessments necessarily require full systematic reviews? (2) How might assessment of risk of bias differ between studies of chemicals and studies of other interventions, such as drugs? (3) What are the implications of heterogeneity of findings for risk relationships? (4) What approaches should be used for assembling different types of evidence, such as epidemiological and toxicological? (5) How can mechanistic information be systematically identified?

MECHANISM AND MODE OF ACTION

There is a pressing need to improve efficiency in the risk-assessment process and incorporate high-throughput technology in evaluating the potential health effects of chemicals. Several efforts are underway by EPA to improve chemical risk assessment. For example, EPA's high-throughput testing program (ToxCast) is designed to identify chemicals with the greatest potential risk to human health. EPA's IRIS program is charged with evaluating and integrating these and other multiple types of evidence regarding potential adverse effects of environmental contaminants on human health: mechanistic studies, animal bioassays, and human studies. This panel will discuss current and future use of data on mechanism and mode of action in weight-of-evidence considerations. Specific topics of interest are (a) evaluation of strength-of-evidence related to mechanisms, (b) the use and interpretation of high-throughput toxicity screening data, and (c) application of genomic dose-response data to chemical risk assessment. Consideration of application of mechanistic data to cancer and noncancer chemical risk assessment within IRIS assessments is of overarching interest.

10:30 **Introduction and Overview of Session**
David Dorman
Member, Committee to Review the IRIS Process
Professor of Toxicology, College of Veterinary Medicine
North Carolina State University

10:40 **Use of High-Throughput and High-Data-Content Technologies in Chemical Risk Assessment**
Rusty Thomas
Director, Institute for Chemical Safety Sciences
The Hamner Institutes for Health Sciences

11:00 **Panel Discussion of High-Throughput Data for Determining Mechanism or Mode of Action**

Panelists: ***David Schwartz***, Chair of Medicine, Professor of Medicine and Immunology, University of Colorado; ***George Leikauf***, Professor of Environmental and Occupational Health, Graduate School of Public Health, University of Pittsburgh; ***Rusty Thomas***, Director, Institute for Chemical Safety Sciences, The Hamner Institutes for Health Sciences; ***Joe Rodricks***, Principal, ENVIRON; and ***Thomas Hartung***, Professor and Doerenkamp-Zbinden Chair for Evidence-based Toxicology, Director Center for Alternatives to Animal Testing, Johns Hopkins Bloomberg School of Public Health

Key Questions

Topic 1: How will findings from new high-throughput assays be used? Can data from high-throughput assays replace more traditional apical end points that are examined in animal toxicity studies? How can dose-dependent changes in mechanisms identified from high-throughput assays be incorporated into chemical risk assessments? How can pharmacokinetic and similar data inform the interpretation of high-throughput screening assays?
Topic 2: How should mechanistic information be incorporated into IRIS assessments? How can the science be advanced to improve qualitative and quantitative application of mechanistic information? What are the evidence criteria for concluding that a mechanism is established as relevant to an agent and outcome?

INTEGRATION OF DATA

EPA's IRIS program is charged with evaluating and integrating multiple types of evidence regarding potential effects of environmental contaminants on human health: mechanistic studies, animal bioassays, and human studies. Assessments are often challenging due to sparse evidence, the use of relatively high doses in experimental bioassays, unclear toxicological mechanisms of action, and unmeasured co-exposures and other threats to validity in observational designs. This session will address qualitative and quantitative strategies for integrating evidence of different types in human health risk assessments.

1:00 **Introduction and Overview of Session**
Scott Bartell
Member, Committee to Review the IRIS Process
Associate Professor, Program in Public Health
University of California, Irvine

1:10 **Qualitative and Quantitative Methods for Integrating Evidence**
Duncan Thomas
Professor and Verna Richter Chair in Cancer Research, Keck School of Medicine
University of Southern California

Panel Discussion on Integrating Various Data

Panelists: ***Steve Goodman***, Professor of Medicine and Epidemiology, Stanford University; ***Kristina Thayer***, Director, Office of Health Assessment and Translation, National Toxicology Program; ***Duncan Thomas***, Professor and Verna Richter Chair in Cancer Research, Keck School of Medicine, University of Southern California; ***Tracey Woodruff***, Professor and Director, Program on Reproductive Health and the Environment, University of California, San Francisco; and ***Lauren Zeise***, Deputy Director for Scientific Affairs, Office of Environmental Health Hazard Assessment, California EPA

Key Questions

Topic 1: Hypothetical mechanisms or modes of action have been proposed for some toxicants, largely based on research in animal models. Consequently, it might be difficult to identify or exclude additional mechanisms for toxic effects in humans. Should mechanistic information be used in a qualitative manner, such as in Hill's biological "plausibility" criterion? Can information from observational or clinical studies on intermediate end points related to mechanisms be helpful? How can mechanistic understanding best be reflected in dose-response model selection or parameter estimation?

Topic 2: How should evidence of toxicity from high-dose animal studies be weighed against null findings from one or more epidemiologic studies at lower exposures? What level of epidemiologic evidence would be sufficient to dismiss a toxic effect in animals as

irrelevant to humans? How can dose-response relationships be combined from different types of research, for example, animal bioassay and epidemiological?

Topic 3: Should positive epidemiologic studies with weaker designs (for example, ecological studies, or studies with unmeasured known confounders) or with positive but non-significant associations contribute to the weight of evidence, or should they be considered only as hypothesis generating?

CAUSALITY

The IRIS assessments evaluate hazard, specifically whether the chemical of concern is a cause of one or more adverse outcomes. The goal of the causal criteria session is to consider the best methods available for systematically evaluating the evidence from individual studies with respect to whether, and to what degree, a chemical causes a particular health outcome, and for combining the evidence in individual studies into an overall judgment as to the likelihood of a causal relationship. Specific goals of the session include (1) considering the utility of existing causal criteria outlined in the most recent IRIS documents; (2) comparing causal assessment methods used by other national and international organizations, with the potential goals of elaborating new guidelines for assessing strength of evidence for causation and of achieving some harmonization across agencies; and (3) considering whether the Hill "criteria" are still useful as guides to synthesizing the overall evidence for causation, or whether alternative criteria or guidelines might be an improvement on approaches developed almost half a century ago.

3:00 **Introduction and Overview of Session**
Richard Scheines
Member, Committee to Review the IRIS Process
Professor and Head of Philosophy Department
Carnegie Mellon University

3:10 **The Role of Mechanism in Causal Assessments and the State of Bradford-Hill**
Steve Goodman
Professor of Medicine and Epidemiology
Stanford University

3:25 **Application of Causal Methods to Assess Effects of Chemical Exposures in Practice**
Lauren Zeise
Deputy Director for Scientific Affairs
Office of Environmental Health Hazard Assessment
California EPA

3:40 **Comparing Weight-of-Evidence Frameworks for Causation**
Lorenz Rhomberg
Principal
Gradient

3:55 **Panel Discussion with Speakers on Causal Methods**

Key Questions

Should the approach to causal inference within EPA guidelines be revised? Are the long-standing causal criteria still useful, given the range of evidence considered in IRIS assessments? How should causal judgments be made in practice? How can they be most useful for practitioners?

4:55 **Opportunity for Public Comment**

THURSDAY, MARCH 28, 2013

8:00 **Welcome to Concluding Session of Workshop**
Jonathan Samet
Chair, Committee to Review the IRIS Process
Professor and Flora L. Thornton Chair, Department of Preventive Medicine
Keck School of Medicine, University of Southern California

8:15 **Putting the Pieces Together: A Case Study**
Tracey Woodruff
Professor and Director
Program on Reproductive Health and the Environment
University of California, San Francisco

8:45 **Workshop Discussion: From Start to Finish – Systematic Review and Evidence Integration**
Speakers, Panelists, and Committee Members

METHODS FOR CHARACTERIZING AND COMMUNICATING UNCERTAINTY

One of the primary aims of systematic reviews is to characterize and communicate the state-of-evidence on a specific topic. Absence of evidence and uncertainties may be characterized using different approaches that range from implicit characterization (qualitative discussion, unexplained variance) to explicit and quantitative characterization. In most cases, communicating uncertainty qualitatively or quantitatively should be an intrinsic element of such efforts. Numerical, verbal, and graphical tools are all widely used to characterize and communicate uncertainty, but with varying success. In this session, methods for characterizing and communicating uncertainties in the IRIS assessment will be considered.

9:15 **Introduction and Overview of Session**
Ann Bostrom
Member, Committee to Review the IRIS Process
Professor, Daniel J. Evans School of Public Affairs
University of Washington

9:25 **Characterizing Uncertainty**
Jay Kadane
Leonard J. Savage University Professor of Statistics, Emeritus
Carnegie Mellon University

9:45 **How the Public Interprets Uncertainty Communication: Some Lessons from the IPCC**
David Budescu
Anne Anastasi Professor of Psychometrics and Quantitative Psychology
Fordham University

10:00 **Panel Discussion on Uncertainty**

Panelists: Tim Lash, Professor, Rollins School of Public Health, Emory; ***Chris Frey***, Distinguished University Professor, North Carolina State University; ***David Budescu***, The Anne Anastasi Professor of Psychometrics and Quantitative Psychology, Fordham University; ***Jay Kadane***, Leonard J. Savage University Professor of Statistics, Emeritus, Carnegie Mellon University; and ***Thomas Wallsten***, Professor, Department of Psychology, University of Maryland

Key Questions

What approaches would enhance the consideration and presentation of uncertainty in IRIS assessment? What attributes of users and uses of IRIS should guide methods for characterizing uncertainties in IRIS assessments? What do we know about tools that are readily available for use in quantifying uncertainty in IRIS?

USE OF EXPERT JUDGMENT

Expert judgment is used in systematic review processes and throughout IRIS assessments, as discussed in the earlier sessions at this workshop. Expert judgment is also used in risk analysis to fill gaps when data are unavailable. Although it is an inherent component of IRIS assessments, this has not been explicitly acknowledged. In this session, the use of expert judgment in the IRIS assessment will be considered, identifying those points in the review and assessment process where expert judgment is important. The session will consider processes for using expert judgment as discussed throughout the workshop in previous sessions and in risk assessment, including elicitation and Delphi approaches.

11:00 **Introduction and Overview of Session**
Ann Bostrom
Member, Committee to Review the IRIS Process
Professor, Daniel J. Evans School of Public Affairs
University of Washington

11:15 **Panel Discussion on Expert Judgment**

Panelists: Tim Lash, Professor, Rollins School of Public Health, Emory; ***Chris Frey***, Distinguished University Professor, North Carolina State University; ***David Budescu***, The Anne Anastasi Professor of Psychometrics and Quantitative Psychology, Fordham University; ***Jay Kadane***, Leonard J. Savage University Professor of Statistics, Emeritus, Carnegie Mellon University; and ***Thomas Wallsten***, Professor, Department of Psychology, University of Maryland

[NOTE: All invited workshop participants are urged to participate in this particular discussion.]

Suggested topics to address by the panel: (a) elicitation techniques (b) understanding the specificity of expertise and to what extent interdisciplinary expertise is required or possible, (c) opportunities (when and where) for the value of expert judgments in IRIS and (d) limitations (including expert bias) on the value of expert judgments in IRIS.

Key Questions

What are best practices for identifying appropriate expertise and eliciting expert judgments, what is the evidence for their effectiveness, and how could they inform the IRIS process? What types of biases in expert judgments might affect IRIS assessments, and how could these be mitigated?

12:15 **Opportunity for Public Comment**

12:30 **Adjourn Workshop**

Appendix C

Primer on Bayesian Method

The Bayesian approach to statistical inference allows scientists to use prior information about the probability of a given hypothesis or other pieces of a model and to combine it with observed data to arrive at a "posterior"—the probability of the hypothesis *given* the observed data and our prior information. A simple illustration is HIV testing. Suppose that the hypothesis is about whether John Smith is infected with HIV. And suppose that the evidence is whether a new blood test comes out positive or negative. The abbreviations are as follows:

H = John Smith has HIV
~H = John Smith does not have HIV
E = blood test for John Smith is positive
~E = blood test for John Smith is negative

One wants to determine the probability that John Smith has HIV *after* receiving the results of a blood test. Suppose that the test is 95% reliable; this means that among those who have HIV, the test will read positive 95% of the time, which can be represented as P(E|H) = 0.95. And suppose that the false-positive rate is tiny (only 1%). That is, among those who do not have HIV, the test will read positive 1% of the time, which can be represented as P(E|~H) = 0.01.

Now suppose that John Smith is routinely screened for HIV with the new blood test and that the test comes back positive. After being informed of the result, he panics because he imagines that he has a 95% or 99% chance of having HIV. That conclusion is not correct. The fundamental theorem, attributed to the Reverend Bayes in the 18th century, is simple in this case to state as follows:

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)}$$

The equation states that the probability of the hypothesis H *given* the evidence E (the posterior) is equal to the product of the probability of the evidence E given that H is true (the likelihood) and the probability of H before any evidence (the prior) is provided divided by the probability of E. To avoid computing P(E), scientists sometimes consider the ratio of the posterior of H and ~H, after E is seen, which in this case is as follows:

$$\frac{P(H|E)}{P(\sim H|E)} = \frac{\frac{P(E|H)P(H)}{P(E)}}{\frac{P(E|\sim H)P(\sim H)}{P(E)}} = \frac{P(E|H)P(H)}{P(E|\sim H)P(\sim H)} = \frac{.95\,P(H)}{.01\,P(\sim H)}$$

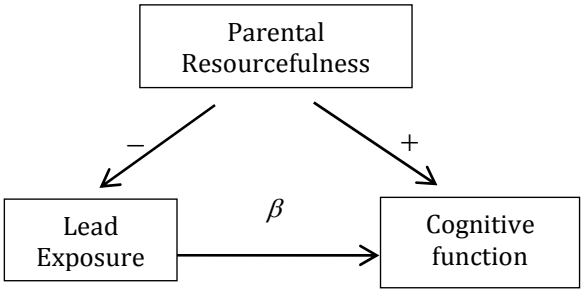
Suppose that, before seeing a blood test, one had no idea whether John Smith had HIV and translated one's ignorance into a 50-50 probability by saying P(H) = P(~H) = 0.5. Then the ratio

above would equal 95, so P(H|E) = 0.9896, which means that John Smith probably has HIV.[1] But suppose that instead of saying that John Smith has a 50% chance of having HIV before one sees a test, one assesses his prior probability of having HIV as the frequency of HIV in people of his age, sex, sexual habits, and drug habits. If John Smith is 30 years old, a middle-class American, heterosexual, and monogamous and does not use any illicit drugs that require needles, his prior might be the frequency of HIV in that group, which might be as low as 1 in 10,000. In this problem, that frequency is referred to as the base rate. If we use P(H) = 0.0001 , the posterior looks much different:

$$\frac{P(H|E)}{P(\sim H|E)} = \frac{0.95 \times 0.0001}{0.01 \times 0.9999} = 0.0095$$

in which case P(H | E) = 0.0094. Thus, with a base rate of 1 in 10,000, John Smith has less than a 1% chance of having HIV, even though his blood test was positive and the test is a highly reliable one. In that case, the Bayesian approach allows one to incorporate base rates easily and test reliability into a calculation of what one actually cares about: the probability of having HIV after getting a test result.

In more general settings, the Bayesian approach can be used to transfer prior knowledge in one part of a model effectively into posterior knowledge in another part of the model of interest. For example, suppose that the basic causal model of the effect of exposure to lead on a child's developing brain is as follows:

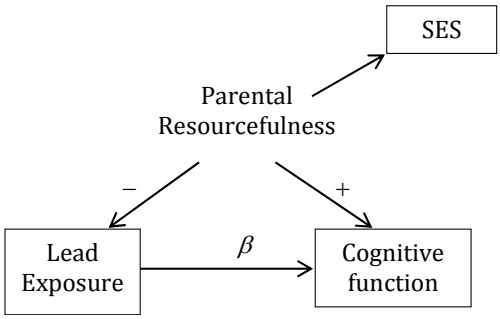


In this model, β, the parameter of interest, represents the size of the effect of lead on cognitive function.[2] β can be estimated from the observed association between lead exposure and cognitive function after adjusting for parental resourcefulness. One problem, especially if one needs to be able to detect statistically even a fairly small β, is that one must be able to measure parental resourcefulness precisely and reliably.

Suppose that socioeconomic status (SES), such as the mother's education and income, is used to measure parental resourcefulness.

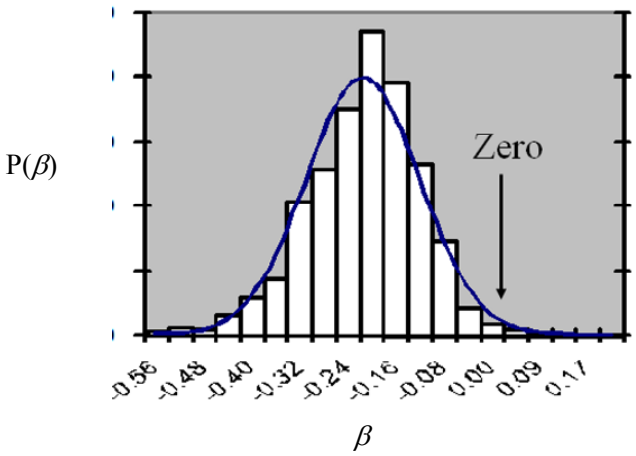
[1]This is because P(H|E) + P(~H|E) = 1, and P(H|E)/P(~H|E) = 0.95, which entails that P(H|E) = 0.9896 and P(H|~E)= 0.0104

[2]Prior beliefs about β can be incorporated directly into a Bayesian model that is used to compute one's degree of belief about β after seeing data. Prior beliefs about other parts of the model will influence the posterior degree of belief about β indirectly.



Then the estimate of β will be biased in proportion to how poorly SES measures parental resourcefulness relevant to preventing a child from being exposed to lead and relevant to stimulating the child's developing brain. The worse SES is as a measure of parental resourcefulness, the more biased the estimate of β. On a scale of 0-100, where 0 means that SES is just random noise and 100 means that it is a perfect measure of parental resourcefulness, is SES a 95? 55? A sensitivity analysis would build a table in which the estimate of β is displayed for each possible level of the quality-of-measure scale of SES, making no judgment about which level is more likely. That can be extremely useful because it might reveal, for example, that as long as one assumes that SES is above a 30 on the quality-of-measure scale, the bias in the estimate of β is below 50%.

Perhaps it is not known where SES sits on a quality-of-measure scale precisely, but one's best guess is that it is 70, and one is pretty sure that it is between 50 and 90. Then, a Bayesian analysis can incorporate this prior information into a posterior over β. For example, after eliciting information on the amount of measurement error in SES, one can conduct a Bayesian analysis of the size of β that might produce the plot below. The X-axis shows the size of β (which in a simple linear model is the size of the IQ drop that one would expect in a 6-year-old after an exposure to enough lead to increase blood lead by 1 µg/dL), and the Y-axis is the posterior probability of β, given our prior knowledge and the data that have been measured in 6-year-olds.



As can be seen, the modal value for β in the posterior is somewhere around -0.2. The spread of the distribution expresses uncertainty about β. Roughly, it shows that the bulk of the posterior distribution over β is roughly between about -0.4 and -0.04. If β is in fact -0.2,

increasing a child's exposure to lead by an amount that would produce a 20-μg/dL increase in its blood concentration would cause an expected drop in IQ of 4 points.[3]

In an IRIS assessment, the analogue of β is any parameter that expresses something about the dose-response relationship in humans. Prior knowledge that a Bayesian analysis might incorporate includes

- The degree to which animal data on rodents are relevant to humans.
- The degree to which mechanistic information informs the dose-response relationship in humans.
- The amount of confounding that might still be unmeasured in epidemiologic studies.
- The quality of the measures of exposure in epidemiologic studies.

Although prior elicitation is important for choosing good informative priors, in some situations, particularly if data are sufficient, moderately informative or even noninformative priors might be sufficient. The major danger with Bayesian models for meta-analysis comes with specifying prior distributions for the between-study variance because information for this parameter is limited by the number of studies available and not by the size of each study. Typical noninformative priors do not work well, and some care must be taken to choose one that is sufficiently informative. Enough is often known to establish reasonably loose bounds that enable estimation, although sensitivity analyses that check how much the final answer is affected by prior choice are still necessary.

[3]Children often test now around 3-5 μg/dL, but children in the 1970s, who were often exposed to lead paint and air with a lot of lead from leaded gasoline, often tested at 20-30 μg/dL.