t⁴ Workshop Report*

Development of an Evidence-Based Risk Assessment Framework

Daniel Krewski[1,2,3], Patrick Saunders-Hastings[2], Robert A. Baan[4], Tara S. Barton-Maclaren[5], Patience Browne[6], Weihsueh A. Chiu[7], Maureen Gwinn[8], Thomas Hartung[9,10], Andrew D. Kraft[11], Juleen Lam[12], R. Jeffrey Lewis[13], Moez Sanaa[14], Rebecca L. Morgan[15], Greg Paoli[3], Lorenz Rhomberg[16], Andrew Rooney[17], Salomon Sand[18], Holger J. Schünemann[15], Kurt Straif[4,19,20], Kristina A. Thayer[21] and Katya Tsaioun[22]

[1]School of Epidemiology and Public Health, University of Ottawa, Ottawa, ON, Canada; [2]McLaughlin Centre for Population Health Risk Assessment, University of Ottawa, Ottawa, ON, Canada; [3]Risk Sciences International, Ottawa, ON, Canada; [4]The IARC Monographs Programme, International Agency for Research on Cancer, Lyon, France (retired); [5]Healthy Environments and Consumer Safety Branch, Health Canada, Ottawa, ON, Canada; [6]Organization for Economic Cooperation and Development, Paris, France; [7]Department of Veterinary Physiology and Pharmacology, Texas A&M University, College Station, TX, USA; [8]Office of Research and Development, US Environmental Protection Agency, Research Triangle Park, NC, USA; [9]Chair for Evidence-based Toxicology and Center for Alternatives to Animal Testing (CAAT), Johns Hopkins University, Baltimore, MD, USA; [10]CAAT-Europe, University of Konstanz, Konstanz, Germany; [11]Center for Public Health and Environmental Assessment, Chemical & Pollutant Assessment Division, US EPA, Washington, DC, USA; [12]Department of Public Health at California State University, East Bay, CA, USA; [13]ExxonMobil Biomedical Sciences, Annandale, NJ, USA (retired); [14]Agence Nationale Sécurité Sanitaire Alimentaire Nationale, Paris, France; [15]Faculty of Health Sciences, McMaster University, Hamilton, ON, Canada; [16]Gradient, Boston, MA, USA; [17]Integrative Health Assessments Branch, National Toxicology Program, US National Institute of Environmental Health Sciences, Research Triangle Park, NC, USA; [18]Department of Risk and Benefit Assessment, Swedish Food Agency, Uppsala, Sweden; [19]Boston College, Chestnut Hill, MA, USA; [20]ISGlobal, Barcelona, Spain; [21]Center for Public Health and Environmental Assessment, Chemical & Pollutant Assessment Division, US EPA, NC, USA; [22]Evidence-Based Toxicology Collaboration, Johns Hopkins University, Baltimore, MD, USA

Abstract

Assessment of potential human health risks associated with environmental and other agents requires careful evaluation of all available and relevant evidence for the agent of interest, including both data-rich and data-poor agents. With the advent of new approach methodologies in toxicological risk assessment, guidance on integrating evidence from multiple evidence streams is needed to ensure that all available data is given due consideration in both qualitative and quantitative risk assessment. The present report summarizes the discussions among academic, government, and private sector participants from North America and Europe in an international workshop convened to explore the development of an evidence-based risk assessment framework, taking into account all available evidence in an appropriate manner in order to arrive at the best possible characterization of potential human health risks and associated uncertainty. Although consensus among workshop participants was not a specific goal, there was general agreement on the key considerations involved in evidence-based risk assessment incorporating 21st century science into human health risk assessment. These considerations have been embodied into an overarching prototype framework for evidence integration that will be explored in more depth in a follow-up meeting.

1 Introduction and background

Risk science has evolved into a well-established interdisciplinary practice incorporating diverse data and methods in order to characterize population health risks and inform decision-making.

Risk science has benefited from advances in biology and toxicology over the last decade, providing powerful new tools and technologies, including high-throughput *in vitro* screening and computational toxicology, that can be used to better assess risks to population health. Risk science has also benefitted from ad-

vances in molecular and genetic epidemiology, which, combined with concomitant advances in exposure science, permit direct estimation of risk in human populations. These and other advances have been incorporated into a framework for the next generation of risk science proposed by Krewski et al. (2014), which was based on work completed under the US Environmental Protection Agency (EPA) NexGen program with input from a large number of stakeholders from North America and Europe.

An important aspect of the evolution of risk science is the desire to ensure that risk decisions are based on the best available scientific evidence, with this evidence identified and evaluated in accordance with appropriate processes and criteria. This trend is consistent with the evolution of evidence-based medicine, which makes use of current best evidence in making clinical decisions about the care of individual patients (Masic et al., 2008; Sackett, 1997). More recently, the concept of evidence-based toxicology has emerged under the leadership of investigators at the Johns Hopkins Bloomberg School of Public Health (Stephens et al., 2013). Like evidence-based medicine, evidence-based toxicology seeks to ensure that the best available data is used in toxicological risk assessment.

The present initiative seeks to build on the scientific advances covered above and the trends towards evidence-based decision-making in multiple disciplines to derive a framework for evidence-based risk assessment that incorporates all relevant data needed to support risk decision-making in a transparent and objective manner. The specific objectives of this project are:

1. to develop a framework for evidence-based risk assessment describing how all relevant evidence relating to a specific risk decision should be assembled and evaluated;
2. to conduct case study prototypes to evaluate the utility of the framework and demonstrate its application in practice; and
3. to lay out a knowledge translation action plan to support the adoption and use of the framework for evidence-based risk assessment in decision-making.

2 Evidence for causation

Establishing causality requires a careful evaluation of the available evidence for and against a causal association between exposure and outcome. Evaluating evidence for causality can be a complex undertaking, particularly in the presence of diverse sources of information, which may report inconsistent findings and which may be of unequal relevance or reliability. A systematic review can be used to summarize the available evidence in a comprehensive and reproducible manner (Wang et al., 2014). Although not all systematic reviews are designed to evaluate causality, there has been a trend towards including causality evaluation as a component of systematic review in recent years. Historically, the Hill criteria (strength, consistency, specificity, temporality, biological gradient, plausibility, coherence, experiment, and analogy) have provided useful general guidance on weighing the evidence for causality (Lucas and McMichael, 2005), though they were originally designed with only observational (epidemiologic) data in mind and do not address other

aspects of causality determination such as consideration of experimental data and integrating different sources of evidence. The grading of recommendations, assessment development, and evaluation (GRADE) approach incorporates aspects of the considerations for causality identified by Hill as well as other considerations, providing an approach to evaluate the certainty of the body of evidence across the following domains: risk of bias, inconsistency, indirectness, imprecision, publication bias, magnitude of effect, dose-response gradient, and opposing residual confounding (Schünemann et al., 2008). On the other hand, the International Agency for Research on Cancer (IARC), for example, has developed a detailed approach for identifying agents that can cause cancer in humans based on a careful evaluation of the available human, animal and mechanistic data (IARC, 2019; Baan and Straif, 2022).

Rhomberg and colleagues (2013) recently reviewed 50 different frameworks that have been proposed in different contexts in the interests of developing a "transparent and defensible" methodology for evaluating the evidence for causation. This review identified four key phases for such assessments: (1) defining the causal question and developing criteria for study selection, (2) developing and applying criteria for review of individual studies, (3) evaluating and integrating evidence, and (4) drawing conclusions based on inferences. Although a specific framework that would be widely applicable in different contexts was not proposed, this work serves to define important attributes of what a broadly applicable framework might include. Five years later, another review of the body of knowledge presented a framework with a similar four-step approach: (1) plan and scope the weight of evidence (WoE) assessment, (2) establish lines of evidence, (3) integrate lines to assess WoE, and (4) summarize conclusions (Martin et al., 2018). While the specific principles, practices, and approaches proposed by these two reviews may differ, together they offer a general approach to evaluating evidence for causation that can be refined and adapted as needed.

Several organizations have provided more detailed guidance for evaluating evidence of causation in various circumstances, depending in part on the nature of the available data (predominantly epidemiological or toxicological) and the risk decision context. Following a review of risk assessment approaches used by the EPA's Integrated Risk Information System (IRIS), the US National Research Council (NRC, 2014a) identified systematic review and evidence integration as key components of a qualitative and quantitative risk assessment paradigm for environmental chemicals. More broadly, the National Toxicology Program Office of Health Assessment and Translation developed a *Handbook for Conducting a Literature-Based Health Assessment Using OHAT Approach for Systematic Review and Evidence Integration*. In the context of establishing dietary reference intakes (DRIs) taking into account chronic disease outcomes, a committee of the National Academies of Sciences, Engineering, and Medicine (NASEM, 2017) developed *Guiding Principles for Developing Dietary Reference Intakes Based on Chronic Disease adopted GRADE* (Guyatt et al., 2008) as the preferred approach to both establishing evidence of causation as well as for intake-response assessment.

3 Sources of evidence

In conducting evidence-based risk assessment, it is important that all available and relevant sources of information be considered. Risk assessment may be informed by toxicological, epidemiological, clinical, surveillance, mechanistic, and other data, all of which need to be considered collectively in order to ensure that the evidence base assembled to potentially support the assessment conclusions is appropriately comprehensive.

An important aspect of evidence-based risk assessment is the amount of data that may or may not be available to support the assessment. Cote and colleagues (2016) note that data-rich and data-poor risk decisions necessarily require different approaches and offer advice on what might be done in a data-poor situation where a risk decision must be made without the luxury of filling key data gaps.

In elaborating the proposed framework for evidence-based risk assessment, the strengths and limitations of different sources of information will be identified and their complementary role in informing the overall assessment outlined. The framework will address both data-rich and data-poor risk decision contexts and establish minimum data requirements to support evidence-based risk decision-making.

3.1 Current approaches to evidence integration

3.1.1 EFSA
With a mandate to provide scientific expertise related to food and feed products in the European Union, the European Food Safety Authority (EFSA) is among the global leaders in hazard identification and risk assessment. Established in 2002 with funding from the European Union, EFSA has developed a variety of frameworks and approaches to support transparency, rigor, and quality in evidence-based risk assessment for products under their remit (EFSA, 2010, 2014, 2015a). While additional information on some of these is provided in a separate publication (Aiassa et al., 2022), a brief summary of the 4-step framework for conducting a scientific risk assessment is included herein.

The 4-step process comprises the following stages:
1. Plan: Formulate the key research question and (as relevant) associated sub-questions, outlining the methodology for answering the question(s) in a protocol developed *a priori*
2. Do: Execute the methodology outlined in the protocol to collect, analyze and leverage data to inform conclusions (specifics will depend on study design and type(s) of data being collected)
3. Verify: Compare the methodology taken with that outlined in the protocol, making note of any deviations from the original plan
4. Report: Promote transparency through the publication of relevant methodologies, assumptions, results and uncertainties

This approach has been piloted across EFSA with some success. For example, such an approach promotes the impartiality, rigor and overall scientific value of the assessment process – as well as the resulting conclusions – by reducing the risk of bias from decisions made in light of the data collected (Munafo et al., 2017;

Shamseer et al., 2015). Conversely, it was found that the implementation of the proposed approach among new or novice users was a resource-, effort- and time-intensive process, and EFSA continues to work towards building capacity and expertise to deliver scientific assessments that are both efficient and in line with current best practices in risk assessment. Nevertheless, EFSA's prioritization of the principles of impartiality, methodological rigor, transparency, and public engagement point to the value of the promotion and integration of such a framework within everyday practice, recognizing that continued improvement should further advance EFSA's ability to deliver high-quality scientific assessments of relevance to public health promotion across the European Union.

3.1.2 EPA IRIS
Created in 1985 and located within the EPA Center for Public Health and Environmental Assessment, the IRIS Program conducts chemical hazard assessments (EPA, 2018). These assessments examine the health consequences of lifetime exposure to environmental chemicals and are both a primary source of certain chemical toxicity information used in support of regulatory and non-regulatory decisions within EPA program offices and regions and important sources of information for other state, federal, and international organizations.

The IRIS approach for assessment (illustrated as interpreted by the National Academy of Sciences (NAS) in Fig. 1) has increasingly been framed through the lens of rigorous and transparent systematic review processes. Assessment development is part of a larger seven-step process for assessment review, which can be summarized as follows (NRC, 2011):
1. Complete draft IRIS assessment
2. Internal agency review
3. Science consultation on the draft assessment with other federal agencies and White House offices
4. Independent expert peer review, public review and comment, and public listening session
5. Revise assessment
6. A. Internal agency review and EPA clearance of final assessment
   B. EPA-led interagency science discussion
7. Post final assessment on IRIS

Assessments are intended to inform decisions related to hazard identification and dose-response assessments, which can be integrated with exposure assessments by EPA programs and regional offices to characterize potential public health risks associated with exposure to an environmental chemical or group of chemicals.

IRIS supports the principle of transparency in their decision processes, with publicly available summaries and databases of chemical-specific evidence and assessment judgments provided since 1988 (EPA, 2018). Progress towards the application of best practices in systematic review and risk assessment has accelerated since 2011, when recognition of challenges in previous assessments motivated a commentary on the IRIS assessment development process by the NAS that was outside of the scope of the chemical-specific review (NRC, 2011). The resulting recommendations outlined a roadmap for a more systematic review
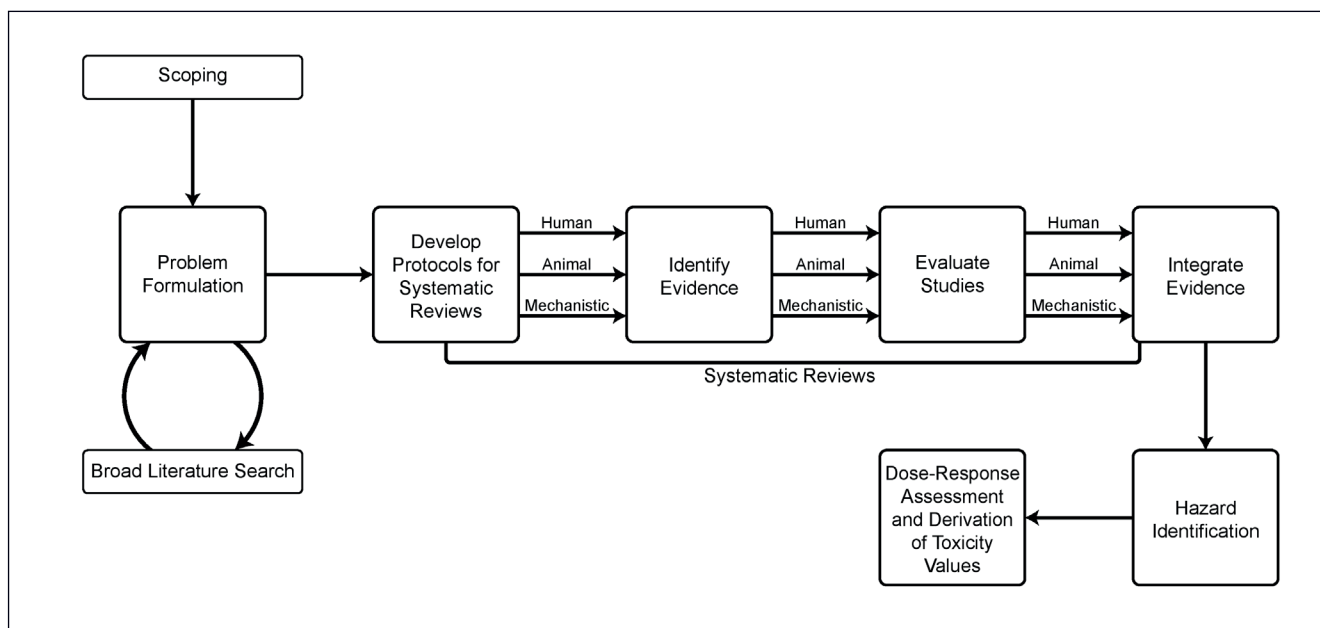
Fig. 1: IRIS approach for assessment development, as interpreted by the NAS
NRC (2014a)

process that triggered numerous planned enhancements to the assessment development process, including on-boarding of systematic review methodologies, adoption of the Health and Environmental Research Online (HERO) tool, and increased public engagement (EPA, 2018; NASEM, 2018).

In order to review changes and progress in the years following the 2011 NAS review, another committee was convened in 2014 (NRC, 2014a); systematic review and application of best practices in evidence integration were again identified as essential elements of environmental chemical human health assessment. Further improvements to the rigor of the IRIS process are being implemented, as reflected in a third assessment conducted in 2018 (NASEM, 2018), which concluded that the IRIS process – while still evolving to adapt to new scientific methodologies and data sources – had successfully undertaken reforms to improve the application and transparency of systematic review methodologies in chemical assessments. Moving forward, it was noted that new tools and approaches would be required to meet some of the outstanding recommendations from the 2014 assessment, *"especially for incorporating mechanistic information and for integrating evidence across studies"* (NASEM, 2018, p. 12).

3.1.3 Health Canada

The Canadian Environmental Protection Act (CEPA), 1999, serves as the main federal policy under which potentially hazardous environmental substances are assessed and regulated. Health Canada and Environment and Climate Change Canada work to-

gether to assess the potential for risk to the environment and the general Canadian population associated with these substances and, as necessary, develop policies and risk management measures for their control. Since its ratification twenty years ago, over 23,000 environmental substances have been registered on the Canadian Domestic Substances List (DSL) (Krewski et al., 2020). The Canadian Chemicals Management Plan (CMP), launched in 2006 based on results of categorization of the DSL and new substances notifications, further sought to evaluate the risk associated with 4,300 prioritized chemicals prior to 2020. Knowledge of chemicals and emerging technologies continues to evolve. Therefore, moving forward, it is important to continue to screen, integrate and consider new information and the increasing complexities of chemicals that may have the potential to cause harm to the environment or human health. Under the CMP, the identification of risk assessment priorities (IRAP)[1] approach is the ongoing prioritization activity for systematically collecting, consolidating, and analyzing information for chemicals and polymers.

Given the ambitious timelines and number of chemicals to be assessed and addressed, an important element in Health Canada's success to date has been the development and application of the CMP Risk Assessment Toolbox (Fig. 2). This Toolbox was developed to delineate the various types of approaches used to address the remaining substances or groups of substances prioritized under the CMP. To make best use of available information, gain efficiencies, and ensure the ability to focus on substances of highest concern, the Risk Assessment Toolbox outlines three

---

[1] https://www.canada.ca/en/health-canada/services/chemical-substances/fact-sheets/identification-risk-assessment-priorities.html
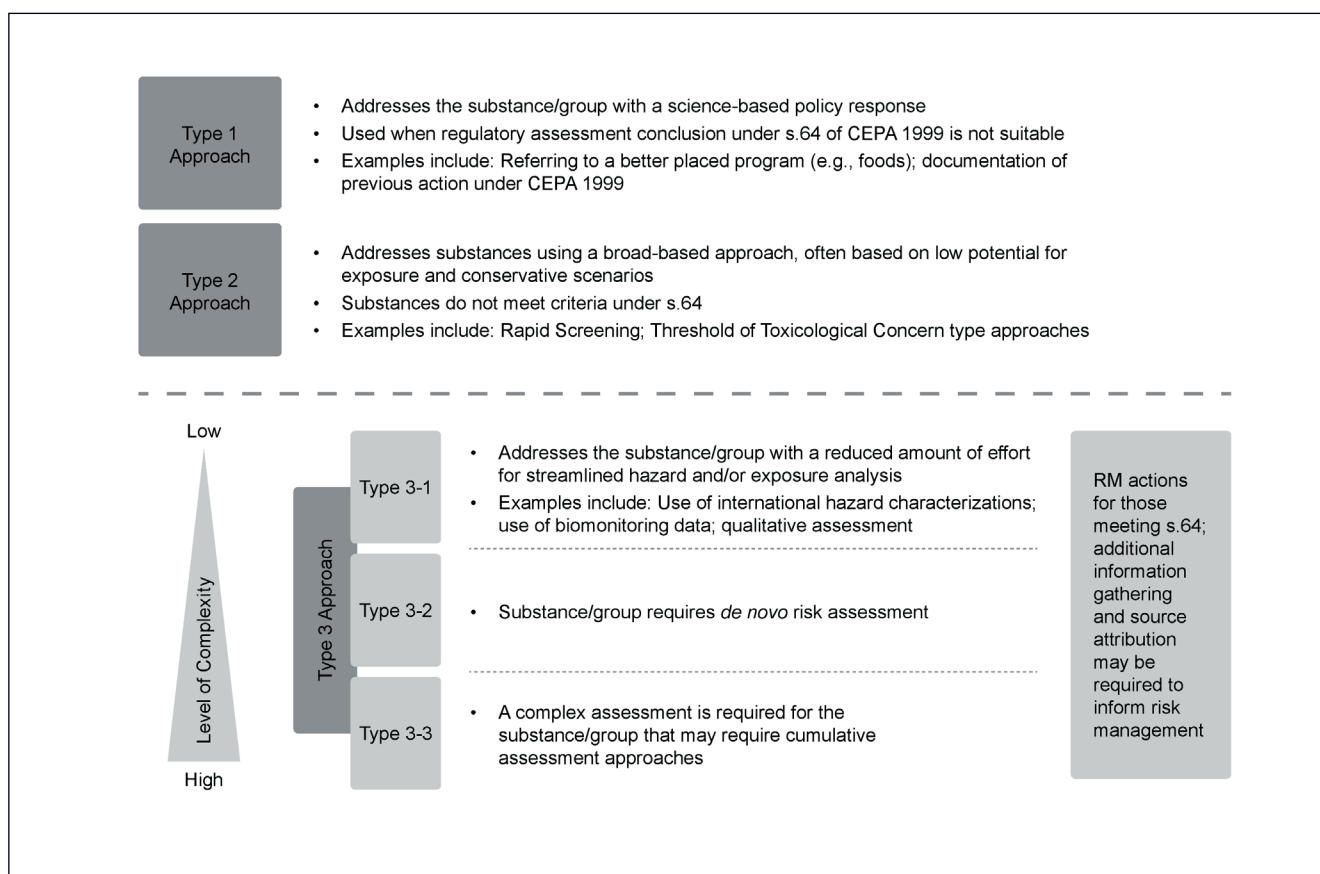
Fig. 2: Chemical Management Plan Risk Assessment Toolbox[2]

types of approaches that can be selected as appropriate and used in a fit-for-purpose manner based on the complexity of the assessment required[2]:

− *Type 1 approaches* use science-based policy responses such as referral of the assessment to a better-placed federal risk assessment program or documentation that the substance has already been addressed by an existing action or previous initiative under CEPA.

− *Type 2 approaches* address substances using broad-based quantitative or qualitative approaches and apply conservative (protective) assumptions. Formal CEPA conclusions may or may not be made under section 64.

− *Type 3 approaches* are applied for substances requiring a standard risk assessment approach including both hazard and exposure and may consider qualitative and quantitative lines of evidence in the determination of whether the substances or group of substances meet the criteria under section 64 of CEPA 1999. Further, the Toolbox proposes three approach subtypes spanning a continuum of complexity and methodology considerations in order to focus the risk assessment efforts.

Select examples of the types of approaches are noted in Figure 2, and the scientific details are described in the published CMP Science Approach Documents (SciADs). More information on the approaches, application, and results, including the threshold of toxicological concern, ecological risk classification, and biomonitoring-based approaches can be found on the Chemical Substances webpage[3].

As the Government of Canada embarks on the next phase of its chemicals assessment and management program, new approach methodologies (NAMs) are being considered and developed for inclusion in the CMP Risk Assessment Toolbox, particularly for Type 2 and 3 approaches, to rapidly and effectively identify their potential for risk assessment in support of the 21st century paradigm shift in risk science. To date, there has been a high degree of success in advancing the use of new technologies and analytical tools through several case studies that have illustrated the practical and positive impacts of integrating multiple lines of evidence, including emerging science. A solid foundation and proof-of-concept have been illustrated for the application of several important NAMs including the example presented at the workshop

---

[2] https://www.canada.ca/en/health-canada/services/chemical-substances/fact-sheets/chemicals-management-plan-risk-assessment-toolbox.html

[3] https://www.canada.ca/en/health-canada/services/chemical-substances/science-approach-documents.html

on the use of integrated approaches for testing and assessment (IATA) for screening-level risk assessment (Webster et al., 2019).

As Health Canada continues to advance chemicals assessment and management, NAMs will be considered in the evolving risk assessment toolbox and incorporated into decision-making through the application of robust methodologies that are context-specific and fit-for-purpose.

3.1.4 ANSES

Over the period 2015-2016, the French Agency for Food, Environmental and Occupational Health and Safety, ANSES (*l'agence national de sécurité sanitaire alimentation, environment, travail*), convened an expert panel to provide a critical review and advice on best practices in the evaluation of the WoE (*"le poids des preuves"*) in the hazard identification step of the risk assessment process. ANSES is somewhat unique in this regard due to the breadth of their mandate. They sought to harmonize the application of WoE concepts across multiple hazardous domains, including public and occupational exposures to chemical hazards, radiation hazards, nutrients and microbial hazards, but all within the hazard identification stage. In this way, questions such as "Does exposure to this chemical cause cancer?" and "Are these particular prions transmissible to humans?" would be answered in a rigorous and harmonized approach. The result of the process was a report with several findings and recommendations (Makowski et al., 2016).

Following the literature review, the panel described a four-step process that is similar in general structure to other frameworks in the literature (planning, evaluation of lines of evidence, integration of lines of evidence, and reporting on the overall weight-of-evidence). The panel's recommendations to ANSES were described in line with this framework. Further work by ANSES has included, among other activities, consideration of the role of quantitative approaches to WoE.

3.1.5 IARC

Several national and international health agencies have established programs with the aim of identifying agents and exposures that cause cancer in humans. The *IARC Monographs on the Identification of Carcinogenic Hazards to Humans* are published by the IARC, the cancer research arm of the World Health Organization (WHO). Each *IARC Monograph* represents the consensus of an international working group of expert scientists based on a critical review of the pertinent peer-reviewed scientific literature as the basis for an evaluation of the weight of the evidence that an agent may be carcinogenic to humans. Published continuously since 1972, the scope of the *IARC Monographs* has expanded beyond chemicals to include complex mixtures, occupational exposures, lifestyle factors, physical and biologic agents, and other potentially carcinogenic exposures. To date, 120 volumes are available on-line[4] and four more are in preparation. More than 1000 agents, mixtures, and exposures have been evaluated. Among these, 120 have been characterized as carcinogenic to humans, 82 as probably carcinogenic to humans, and 311 as possibly carcinogenic to humans.

From the very beginning of the monographs program, there have been two criteria for consideration of an agent for evaluation: (a) There is evidence of human exposure and (b) there are published scientific data suggestive of carcinogenicity. For each agent considered, systematic reviews of the available scientific evidence on its carcinogenicity in humans and in experimental animals are conducted by an international working group of independent experts. Data on human exposure to the agent and toxicological data on pertinent mechanisms of carcinogenesis are also reviewed. An overall evaluation that integrates epidemiological and experimental cancer data as well as mechanistic evidence is reached according to a structured process. Agents with "sufficient evidence of carcinogenicity" in humans are assigned by default to the highest category, "carcinogenic to humans" (IARC Group 1), whereas the categories of "probably" (Group 2A) or "possibly" (Group 2B) carcinogenic to humans, or "not classifiable as to its carcinogenicity to humans" (Group 3) are assigned according to the combined strength of the human, animal, and mechanistic evidence. Agents may be placed in a higher category when the evidence for a relevant mechanism of carcinogenesis is sufficiently strong. The *IARC Monograph* classifications refer to the strength of the evidence for a cancer hazard, rather than to the quantitative level of cancer risk. The monographs integrate the three streams of evidence on cancer in humans, cancer in experimental animals, and mechanistic data into an overall evaluation of the strength of evidence in terms of a hazard identification on the carcinogenicity of an agent. Identification and critical appraisal of the published literature includes both exposure and (when possible) exposure-response characterization, supporting both hazard identification and characterization. In addition to identifying cancer hazards, the *IARC monographs* may include quantitative statements about the level of cancer risk, as with Monograph Volume 114 on the consumption of red and processed meat and Monograph Volume 120 on benzene.

The *Preamble* to the *IARC Monographs* describes the objective and scope of the program, the scientific principles and procedures used in developing a monograph, the types of evidence considered, and the scientific criteria that guide the evaluations. The *IARC Monographs* are prepared according to principles of scientific rigor, impartial evaluation, transparency, and consistency. The criteria defining those principles evolved during the early years of the program and were outlined in the first *Preamble* (IARC, 1978), which has been refined and updated a dozen times since. In the recently revised *Preamble* (IARC, 2019), mechanistic evidence has gained in prominence, and relevance to cancer hazard evaluation. One of the important changes in the new *Preamble* is the introduction of systematic review of mechanistic data facilitated by organization into key characteristics (Smith et al., 2016), which is now common practice since Monograph Volume 112. Further discussion of the evolution of the *Preamble* and other important changes incorporated into the

---

[4] See https://monographs.iarc.fr/monographs-and-supplements-available-online/

most recent update is given by Samet and colleagues (2019). In their account of the history of the *Preamble to the IARC Monographs*, Baan and Straif (2022) note that NAMs have the potential of reducing or avoiding the use of experimental animals in future evaluations conducted by the IARC.

3.2 NAMs

Following publication of the US National Research Council report, *Toxicity Testing in the 21st Century: A Vision and A Strategy* (NRC, 2007), there has been increasing emphasis on alternatives to animal testing in toxicological risk assessment (Bolt, 2019; Perkins et al., 2019a,b; Price et al., 2020). A mid-term update on progress made towards the realization of this vision over the original 20-year planning horizon has recently been prepared by Krewski and colleagues (2020). The broad suite of tools and strategies offering viable alternatives to animal testing is now referred to as NAMs.

The EPA (2018, p. 6) describes the term "NAM" as a "*descriptive reference to any technology, methodology, approach, or combination thereof that can be used to provide information on chemical hazard and risk assessment that avoids the use of intact animals*". EPA (2019a) has developed a list of NAMs considered potentially relevant for evaluating chemical toxicity under the Toxic Substances Control Act. These new approaches can be broadly classified as including:

− computational toxicology and bioinformatics;
− high-throughput screening methods;
− testing of categories of chemical substances;
− tiered testing strategies;
− *in vitro* studies;
− systems biology; and
− new or revised methods from validation bodies such as ICCVAM, ECCVAM, NICEATM, and OECD[5].

The EPA National Center for Computational Toxicology (NCCT) has recently developed a roadmap outlining an approach for making greater use of NAMs (Thomas et al., 2019). Key elements of the EPA CompTox Blueprint include an emphasis on computational modeling and high-throughput approaches to supplement traditional approaches in chemical assessments for regulatory decision-making. On December 17, 2019, EPA hosted its First Annual Conference on the State of the Science on Development and Use of NAMs for Chemical Safety Testing (EPA, 2019b). The US National Toxicology Program has developed a listing of alternative test methods currently accepted by US regulatory agencies[6].

In administering the European Union's REACH (Registration, Evaluation, Authorization and Restriction of Chemicals) regulation (Armstrong et al., 2020), the European Chemicals Agency (ECHA, 2020) relies heavily on test guidelines (TG) that have been validated in accordance with guidance provided by the OECD[7], including alternative test methods. The OECD Toolbox provides a rich suite of alternative computational tools for evaluating quantitative structure-activity relationships[8]. Alternative test methods are an important component of REACH dossiers submitted by registrants: of 6,290 substances evaluated within REACH, 89% were found to include at least one data endpoint based on an alternative test method (ECHA, 2017). Alternative test methods considered by EPA include alternatives that have been validated by OECD, such as the *Bacterial Reverse Mutation Test* (OECD TG 471) and the *Performance-Based Test Guideline for Human Recombinant Estrogen Receptor (hrER) In Vitro Assays* (OECD TG 493).

Andersen et al. (2019) have suggested a multi-level strategy for incorporating NAMs into toxicological risk assessment practice, seeking to deploy new methods in a context-specific manner. Level 1 in the proposed strategy focuses on computational screening, with quantitative structure-activity relationships (QSAR)/read-across, cheminformatics, and threshold of toxicological concern approaches used to assess bioactivity and high-throughput exposure modeling approaches used to evaluate potential human exposure. Level 2 relies on high-throughput *in vitro* screening to assess bioactivity through transcriptomics, high-content imaging, and bioinformatics, along with judiciously chosen test batteries, to evaluate bioactivity. Refined exposure models may be used, along with high-throughput *in vitro* to *in vivo* extrapolation (HT-IVIVE), to estimate human doses. Level 3 invokes fit-for-purpose assays for bioactivity, including lower throughput cell-based assays and consideration of metabolism. More specific exposure models are employed, along with quantitative *in vitro* to *in vivo* extrapolation (q-IVIVE). Level 4 employs more complex *in vitro* assays for bioactivity, advanced systems such as organ chips, and tailored *in vivo* studies to confirm *in vitro* results. Tailored exposure models may also be employed at this level, including physiologically based pharmacokinetic (PBPK) models for *in vivo* species extrapolation.

As envisaged by Andersen et al. (2019), level 1 approaches may be most useful in the context of priority setting, with level 2 approaches more suited to screening level assessments. Level 3 approaches afford greater insight into toxicity pathways and are consistent with the vision put forward by the US National Research Council (NRC, 2007) for toxicity testing in the 21st century (TT21C). When required, level 4 approaches may provide additional data using more integrated assays at the biological system level. At each level, margins of exposure (MOEs) based on a comparison of predicted human doses to doses at which bioactivity is expected to provide valuable information in support of context-specific risk decisions.

The OECD IATA Case Studies Project includes a successful application of NAMs in risk assessment in which performance-based approaches were used to assess the reliability and accuracy of *in vitro* predictions (OECD 2019a,b). Specifically, using estrogen receptor bioactivity models based on 18 HTS as-

---

[5] EURL ECVAM: European Center for the Validation of Alternative Methods; ICCVAM: Interagency Coordinating Committee on the Validation of Alternative Methods; NICEATM: NTP Interagency Center for the Evaluation of Alternative Toxicological Methods; OECD: Organization for Economic Cooperation and Development.

[6] https://ntp.niehs.nih.gov/whatwestudy/niceatm/accept-methods/index.html (accessed 15.06.2020)

[7] http://www.oecd.org/env/ehs/testing/seriesontestingandassessmenttestingforhumanhealth.htm

[8] https://www.oecd.org/chemicalsafety/risk-assessment/oecd-qsar-toolbox.htm

says, 43 reference chemicals achieved a balanced accuracy of 86-95%; similar validation exercises for androgen receptor activity and anti-androgen activity also produced strong measures of validation performance. Encouraged by the success of these *in vitro* approaches, the OECD has published several guidance documents on incorporating NAMs into IATA.

New approaches to risk decision-making are increasingly being used by regulatory authorities worldwide, motivated in large part by the need for increased throughput in risk decision-making. The EPA used Attagene assays, including multiple gene targets such as PPARα and NRF2, to rapidly evaluate the relative toxicity of eight dispersants in response to the Deepwater Horizon oil spill in the Gulf of Mexico in 2009, representing an early practical application of NAMs in emergency risk decision-making by the EPA (Anastas et al., 2010). In 2019, the Methodology Working Group of the Scientific Committee on Consumer Safety organized a working group to discuss the use of NAMs to advance "Next Generation Risk Assessment" in the context of animal-free safety testing of cosmetic ingredients in Europe (Rogiers et al., 2020). With the diverse set of NAMs currently available, there are unprecedented opportunities to apply new high- and medium-throughput assays in support of human health risk assessment.

As an additional example, Gannon et al. (2019) conducted a tiered evaluation of hexabromocyclododecane (HBCD), a widely used flame retardant. This analysis incorporated ToxCast™ HTS data and *in vitro-in vivo* extrapolation, or IVIVE (Tier 1), rat liver transcriptomic data (Tier 2), and conventional rat bioassay data (Tier 3). Toxicity pathway perturbations were closely aligned between Tiers 1 and 2, and consistent with apical effects seen in Tier 3. Bioactivity-exposure ratios (BERs) calculated as the ratio of levels (expressed in human dose equivalents) at which biological activity was seen in these tiered assays to Canadian exposure levels were smallest in Tier 1 and similar in Tiers 2 and 3.

Other comparisons of indicators of toxicity used in regulatory risk assessment between alternative and traditional test methods include benchmark dose estimates reported by Thomas et al. (2012), who showed good agreement between transcriptomic- and bioassay-based BMDs in a pilot study of five chemicals. More recently, Paul Friedman (2020) conducted a comparison of points of departure (PODs) for 448 chemicals based on quantitative high-throughput screening (qHTS) predictions of bioactivity, denoted PODNAM, with results from mammalian toxicity tests, denoted $POD_{traditional}$. The 95% lower confidence limit (LCL) on the PODNAM was less than the $POD_{traditional}$ for the great majority of these chemicals, suggesting that qHTS-based exposure guidelines are likely to afford at least as much public health protection as exposure guidelines based on traditional animal tests.

The elaboration of strategies for deploying NAMs in a systematic manner, such as that suggested by Andersen et al. (2019), will be of great value in choosing the most appropriate approaches to employ within specific risk decision contexts. While such efforts are both needed and welcome, the increased use of NAMs by regulatory authorities will also require new thinking on how to incorporate new types of data into WoE evaluations.

3.3 Trade-offs between cost, timeliness, and uncertainty reduction in toxicity testing

A fundamental tenant of vision for the future of toxicity testing advanced by the NRC (2007) was the need for more rapid toxicity testing strategies in order to expand the coverage of the large number of chemicals for which toxicological data is lacking. This need was underscored by the NRC (2009), which recommended the use of formal value-of-information (VOI) techniques for the adoption of methods to evaluate alternative testing strategies. Price et al. (2021) and Hagiwara et al. (2022) have recently developed informal and formal VOI methods to evaluate the trade-offs between the cost, timeliness, and uncertainty reduction. These contributions suggest that while rapid alternative test methods may be subject to greater uncertainty than traditional, more expensive animal tests of longer duration, having sufficient information to support risk decision-making in a timely manner can result in a lower total social cost by avoiding public health impacts that can accrue when decisions are delayed pending the results of toxicity tests of longer duration. Although further VOI analyses are needed to reaffirm these findings under broader real-world toxicity testing scenarios, initial results appear to be supportive of alternative test methods that can provide cost-effective toxicity data in a timely manner.

4 Defining the research question

Formulating a clear and actionable research question creates structure in the approach to conducting systematic reviews and developing health guidance (Guyatt et al., 2011a). Within the field of risk assessment, this question may be tailored for studies of exposure as a PECO question, which is used to outline the population, exposure, comparator, and outcomes (Morgan et al., 2018). Guidance has been provided to help users operationalize the PECO question, as this informs many stages of the evidence review and quality assessment of the findings (Section 7). The population may be defined based upon characteristics relevant to the exposure or outcome of interest, including geographic, demographic, socioeconomic, or genetic and biological factors. Approaches for identifying the exposure and comparator are discussed by Morgan and colleagues (2018). Research question formulation may also benefit from consideration of the FINER criteria (feasible, interesting, novel, ethical and relevant) (Farrugia et al., 2010).

In designing a systematic review, it is important to avoid the risk of formulation bias. Researchers should clearly identify the objectives of the review, which in turn should clearly reflect the ultimate use of the review in risk decision-making (Farhat et al., 2022). Phrasing of the problem and definition of the search parameters and selection criteria *a priori* should help limit formulation bias. Many risk assessment frameworks (e.g., Krewski et al., 2014; NRC, 2009; Paoli et al., 2022) emphasize the importance of problem formulation, including consideration of VOI, to ensure that the information collected is informative with respect to the objective of the assessments.

Tab. 1: Five case studies of the Navigation Guide

Citation	Case study	Evidence streams	Findings
Johnson et al., 2014; Koustas et al., 2014; Lam et al., 2014	Developmental exposure to perfluorooctanoic acid (PFOA) and fetal growth outcomes	Human and animal	Rated 18 epidemiology studies and 21 animal toxicology studies. Both evidence streams were rated as "moderate" quality and "sufficient" strength, leading to a final conclusion that PFOA was "known to be toxic" to human reproduction and development.
Vesterinen et al., 2015	Association between fetal growth and maternal glomerular filtration rates	Human and animal	Rated 31 human and non-human observational studies as "low" quality and two experimental non-human studies as "very low" quality. All three evidence streams were rated as "inadequate." There was insufficient evidence to support the plausibility of a reverse causality hypothesis for associations between environmental exposures during pregnancy and fetal growth.
Johnson et al., 2016	Exposure to triclosan and human development or reproduction	Human and animal	Rated three human studies and eight experimental animal studies in rats reporting hormone concentration outcomes (thyroxine levels). Human studies were rated as "moderate/ low" and animal studies were rated as "moderate." There was "sufficient" non-human evidence and "inadequate" human evidence, leading to the conclusion that triclosan was "possibly toxic" to reproductive and developmental health.
Lam et al., 2016	Exposure to air pollution and autism spectrum disorder (ASD)	Human only	Rated 23 epidemiology studies. Evidence was rated as "moderate" quality, leading to the conclusion that there was "limited evidence of toxicity" between exposure to air pollution and ASD diagnosis.
Lam et al., 2017	Developmental exposure to polybrominated diphenyl ethers (PBDEs) and IQ/ ADHD outcomes	Human only	Rated 10 epidemiology studies for intelligence outcomes and 9 studies for ADHD outcomes. Evidence was rated as "moderate" quality with "sufficient" evidence for IQ outcomes and as "moderate" quality with "limited" evidence for ADHD outcomes.

5 Assembling the evidence

Because of the diversity of evidence that could be considered in risk decision-making, it is essential to have structured approaches for identifying and summarizing all relevant information. Systematic review provides a powerful approach to meet this need. Guidelines for systematic review of clinical data have been established by the Cochrane Collaboration, including the preferred reporting items for systematic reviews and meta-analyses (PRISMA) guidelines (Moher et al., 2009). Although guidelines for best practices in systematic review were developed first for summarizing clinical evidence, similar guidelines are now being developed for other sources of evidence, including toxicology and epidemiology. An overview of best practices in systematic review is provided in a separate article (Farhat et al., 2022).

Within the field of environmental health, there is a clear need for rigorous systematic review methodologies. Although outside the scope of the present report, less intensive expedited or rapid reviews (see Farhat et al., 2022) may be conducted when time and resources do not permit the completion of a comprehensive systematic review. Review of the scientific evidence plays a critical role in decision-making about exposures to environmental chemicals by local and federal government agencies. However, challenges exist where there are large data sets, variable study quality, conflicting evidence, or limited information, which impedes integration and final conclusions. Approaches used to assemble and synthesize evidence have evolved over the last three decades, with expert judgement increasingly supported by guidance developed by national and international organizations. Improved methods of chemical assessment that better reflect scientific knowledge have been articulated by the National Research Council in several reports (NRC, 2008, 2009, 2014a,b), in particular identifying systematic review as an approach that could substantially improve the processes used to inform policy- and decision-making regarding environmental chemicals.

Several systematic approaches have been evolving and undergoing applications to chemical assessment at the National Toxicology Program (Rooney et al., 2014) and U.S. EPA (NRC, 2014a,b). One novel approach to systematic review in environmental health is the Navigation Guide, developed in 2009 through a collaboration between academic scientists and clinicians with the goal of expediting the development of evidence-based recommendations for preventing harmful environmental exposures (Woodruff and Sutton, 2011). The Navigation Guide was developed by drawing from the rigor of systematic review methods used in the clinical sciences with modifications allowing for the unique challenges faced with evidence streams specific to environmental health (i.e., animal toxicology and human epidemiology data).
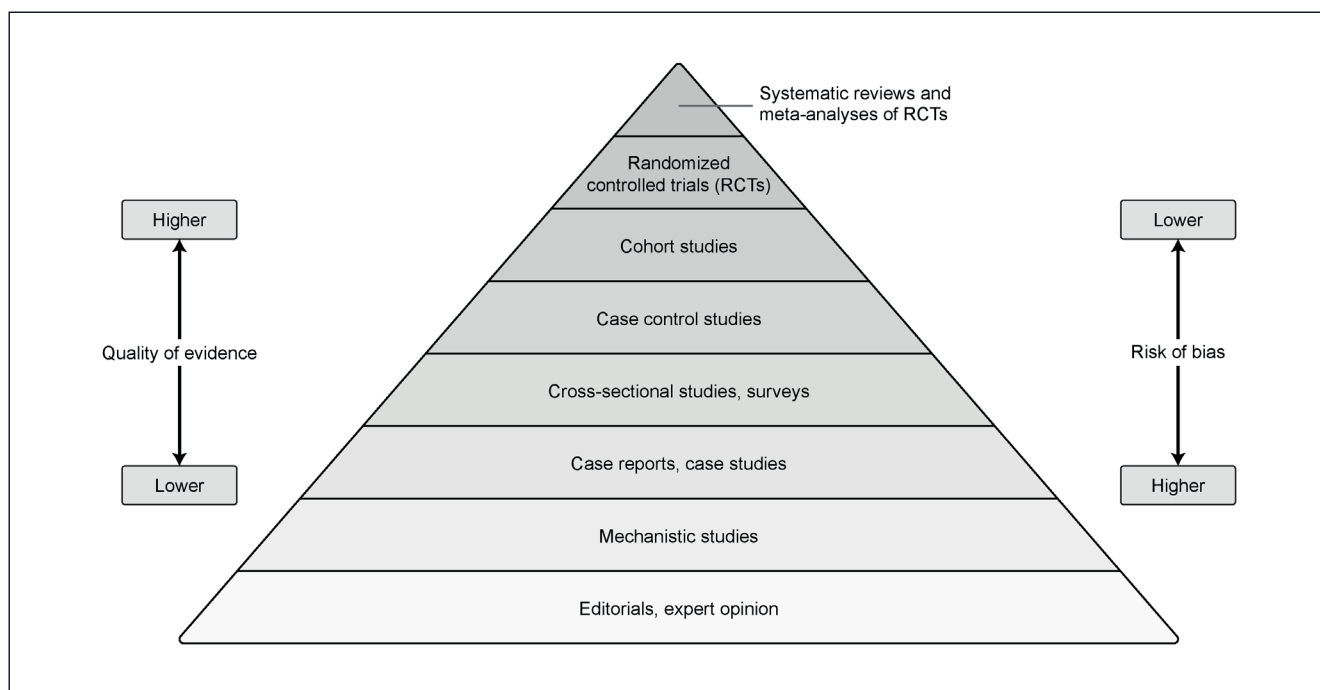
Fig. 3: Hierarchy of evidence pyramid
adapted from Yetley et al. (2017)

The Navigation Guide has been applied in five published case studies as proof-of-concept (Johnson et al., 2014, 2016; Koustas et al., 2014; Lam et al., 2014, 2016, 2017; Vesterinen et al., 2015) (Tab. 1). These case studies were some of the first to demonstrate that systematic and transparent review approaches in environmental health were not only achievable but also advantageous over existing methodologies such as narrative reviews.

The Navigation Guide systematic review methodology involves three main steps:
1. Specify the study question: Frame a specific research question relevant to decision-makers about whether human exposure to a chemical or other environmental exposure is a health risk.
2. Select the evidence: Conduct and transparently document a systematic search for published and unpublished evidence.
3. Rate the quality and strength of the evidence: Rate the potential risk of bias (i.e., internal validity) of individual studies and the quality/strength of the overall body of evidence based on prespecified and transparent criteria (typically outlined in a pre-published, publicly available protocol). The Navigation Guide methodology conducts this process separately by evidence stream (i.e., human and animal evidence). Ultimately, evidence is combined by integrating the quality ratings of each of these two evidence streams. The end result is one of five possible statements about the overall strength of the evidence: "known to be toxic," "probably toxic," "possibly toxic," "not classifiable," or "probably not toxic."

To date, the traditional application of the Navigational Guide has been restricted to human and animal data (Woodruff and Sutton, 2014). Future iterations of the guide could benefit from extensions to include consideration of alternative test data derived from NAMs.

Although various authors have suggested different data hierarchies for use in evidence integration (Burns et al., 2011; Petrisor and Bhandari, 2007), consensus on a single hierarchy for application across diverse risk assessment contexts is lacking. As an example, Yetley et al. (2017) identified hierarchies of evidence considering sources of information to support the establishment of dietary reference intakes (DRIs) of nutrients present in the food supply. In this paradigm, well-conducted randomized clinical trials (RCTs) represent the 'gold standard' in terms of obtaining unbiased information directly in human populations. RCTs will not be available for most hazards of concern, though the hierarchy of evidence pyramid shown in Figure 3 also identifies other valuable sources of information that are frequently used in risk assessment applications. Other variations of this hierarchy have also been suggested, such as that discussed by Murad et al. (2016), which includes consideration of both study design and quality to allow for departures from a strict *a priori* hierarchy of evidence.

The emphasis on the application of systematic review to assemble the data needed to support evidence-based risk assessment is consistent with recommendations made by the US National Research Council (NRC, 2014a) as part of its review of the EPA's IRIS program. As indicated in Figure 1, adapted from the NRC review, systematic review represents a critical first step in assembling all human, animal, and mechanistic data relevant to the assessment of potential risks associated with environmental health hazards.

Additional guidance on the use of systematic review in risk assessment is available from numerous sources. Farhat and colleagues (2022) trace the evolution of incorporating systematic review into evidence-based risk assessment, give contemporary examples of the application of current methods in systematic review across different domains – including clinical, epidemiological, and toxicological applications – and provide a summary of available tools to support best practices in systematic review.

6 Synthesizing the evidence

Once all relevant information has been assembled in a systematic review, this information needs to be synthesized, qualitatively and sometimes quantitatively. Qualitative synthesis involves a determination as to whether the exposure of interest constitutes a human health hazard. Within the context of evidence-based risk assessment, this is done using an appropriate evidence integration framework. Should a potential human health hazard be identified through qualitative synthesis, the next step is to determine whether the available data are sufficient to support an evidence-based quantitative estimate of population health risk.

6.1 Qualitative synthesis

A number of frameworks for specific types of risk have been developed by different authorities. IARC, for example, has elaborated and refined a well-known scheme for evaluating human carcinogenicity, which, over the last 50 years, has led to the identification of 120[9] agents as known causes of human cancer. It is important to recognize that while the IARC framework for evaluating potential cancer risk to humans identifies cancer hazards, it generally does not result in a quantitative estimate of human cancer risk.

Other schema for evaluating non-cancer hazards have also been proposed by other authorities. Rhomberg and colleagues (2013) recently reviewed 50 frameworks in the literature. An update to this search from 2013 to 2018 is provided in the supplementary material[10]: It is noted that a number of these frameworks specifically address the incorporation of data derived from alternative test methods. Meanwhile, Martin and colleagues (2018) conducted a critical assessment of 24 approaches in an effort to develop a generalized approach. A major conclusion of this assessment was that incorporation of factors such as extent of prescription, relevance to the question of interest, and ease of implementation increases the transparency of the approach taken.

6.2 Quantitative synthesis

In many cases, a quantitative estimate of risk will be needed to complete the risk assessment. In the past, quantitative estimates of risk have often been based on identifying a key study (or studies) that is amenable to fitting an appropriate dose-response or exposure-response model. [Although outside the scope of this report, sophisticated analytical techniques such as Bayesian mod-

el averaging (Thomas et al., 2007) can be used to incorporate results from multiple exposure-response models that are compatible with the data.] These models can then be used to develop projections of potential population health risk and associated uncertainty under specified exposure scenarios. These models can also be used to identify a point of departure (PoD) on the dose-response curve that can be used to establish a reference value (RfV) or other toxicity benchmarks to serve as a guideline for human exposure (EPA, 2012). The PoD can also be used to establish a margin of exposure (MoE) reflecting the ratio between the toxicity benchmark and estimated or predicted human exposure levels (Thomas et al., 2013).

When multiple studies with quantitative information on dose-response are available, it may be possible to combine the results of these studies. As discussed below, combined analysis of the primary raw data may be possible when the study designs are compatible. When the primary raw data are not accessible for analysis, meta-analysis of summary risk estimates from the individual studies is often done. Another potentially useful approach to combining data involving different toxicological endpoints is categorical regression. Each of these approaches is described briefly below.

6.2.1 Combined analysis
When access to the primary raw data from a series of related studies is available, a combined analysis of the raw data can be conducted. Having access to the raw data affords maximum flexibility in modelling exposure-response relationships across the studies being combined, as well as an opportunity to evaluate the effects of potential modifying factors included in the original studies. For example, Krewski and colleagues (2006) conducted a combined analysis of the primary raw data from a series of case-control studies on residential radon and lung cancer risk, demonstrating for the first time a strong association between radon and lung cancer in residential settings.

6.2.2 Pooling epidemiological data
There are various methods for pooling data from epidemiological studies, each with its own strengths and limitations. Combining primary data from individual studies to yield a large dataset – referred to as *pooled analysis* when used in epidemiological studies – has many advantages. In particular, the increase in sample size allows for more precise calculations of risk estimates (Tobias et al., 2004). The large sample size can also improve the statistical power to allow the assessment of risks in specific subgroups or restricted subsets of data that would not be possible in smaller data sets of individual studies. These strengths of pooling data make it appealing when investigating effects of rare exposures or risk factors of diseases that have long induction periods, such as cancer (Cardis et al., 2011; Fehringer et al., 2017; Felix et al., 2015; Gaudet et al., 2010; Kheifets et al., 2010; Peres et al., 2018; Wyss et al., 2013).

Although pooling primary sources of data can be expensive and time-consuming and requires agreement on data-sharing and coop-

---

[9] https://monographs.iarc.fr/agents-classified-by-the-iarc/
[10] doi:10.14573/altex.2004041s

eration of investigators from multiple study centers, pooling primary data can allow investigators to make a broader range of conclusions compared to meta-analyses that are based on published study findings (Checkoway, 1991). This approach also has other advantages; it allows having unified inclusion criteria and definitions of variables across the centers. It also allows the use of the same statistical model on all the combined data. This is particularly important since individual studies commonly adjust for different confounders in their analysis (Friedenreich, 1993). Standardizing the methods reduces potential sources of heterogeneity across the studies.

This type of analysis can be done retrospectively or prospectively. Prospective planning has the added advantage that it allows co-investigators to plan ahead and ensure uniform methods are used for data collection and reporting (Blettner et al., 1999). Nevertheless, many pooled analyses have been conducted retrospectively after individual studies had reported their findings.

For example, as mentioned above, Krewski and colleagues (2006) retrospectively combined primary data from seven case-control studies in North America. Their pooled findings, based on 4,081 cancer cases, indicate an association between residential radon and lung cancer, although findings from the individual case control studies had provided inconsistent evidence on the risks of lung cancer. The pooling of data further allowed analysis on subsets of the data with more complete radon dosimetry in the most critical exposure time windows. The investigators also performed dose-response analyses based on the histological type of lung cancer (Field et al., 2006). These analyses had not been possible in the prior analyses of each study data individually. Similar radon risk analyses were conducted by Darby and colleagues (2005) from 13 European case-control studies on 7,148 lung cancer cases. The large sample size achieved by pooling provided sufficient statistical power to detect moderate risks that could not be detected in individual studies.

To perform a pooled analysis of primary data from multiple studies, Friedenreich (1993) suggested the need for a strict protocol and detailed eight steps to follow for pooling data and analyzing the combined dataset: 1) identify relevant studies; 2) select (sufficiently similar) studies from which to pool data; 3) combine the data after obtaining each study data from original investigators; 4) estimate study-specific effects using logistic regression; 5) examine the homogeneity of study-specific effects; 6) estimate pooled effects (if study-specific effects are homogenous); 7) explain heterogeneity between studies (if study-specific effects are not homogenous); and 8) perform sensitivity analyses to examine the robustness of the pooled effects.

While pooling of data can serve to both increase precision and support subgroup analyses that would otherwise not be possible in individual studies, inferences regarding causality (see Section 2) will typically still require establishment of a plausible biological hypotheses, with consideration of mechanistic data, including data derived using NAMs.

6.2.3 Meta-analysis
Meta-analysis has become a popular and useful technique for developing a more pragmatic estimate of risk by quantitatively combining compatible study-specific risk estimates. Meta-analysis requires that the designs for the studies being combined are reasonably compatible and that the study results do not demonstrate a high degree of heterogeneity. While meta-analysis has been applied in toxicological risk scenarios, it is predominantly applied in cases where human data is available.

When pooling data, primary data from individual studies are combined to provide a much larger dataset that is then analyzed to obtain an overall effect estimate. This is the main difference compared to meta-analyses, where effect estimates reported from individual studies are combined into one overall effect estimate. Meta-analyses are very common in epidemiology and have many advantages including low associated costs, time efficiency, and the ability to provide an overall quantitative assessment of risk and uncertainty (Friedenreich, 1993). However, limitations do arise in instances where significant heterogeneity between studies is present due to variations in study design, eligibility criteria, exposure and outcome ascertainment, and statistical analyses (Blettner et al., 1999). Meta-analyses are also limited by the information provided in the original publications and may preclude dose-response analysis and specific subgroup analyses (Friedenreich, 1993; Tobias et al., 2004).

Recent examples of informative meta-analyses of epidemiological data include analyses of the association between exposure to diesel exhaust and lung cancer (Vermeulen et al., 2014) and analyses of the association between talc and ovarian cancer (Taher et al., 2019). Vermeulen and colleagues (2014) conducted a meta-analysis of three epidemiological studies on the association between occupational exposure to diesel exhaust emissions in the mining and trucking industries and lung cancer, using elemental carbon as an indicator of exposure to diesel exhaust. This analysis provides a possible approach to characterizing the exposure-response relationship between diesel exhaust and lung cancer risk (HEI Diesel Epidemiology Panel, 2015). Taher et al. (2019) conducted a meta-analysis of 27 case-control and cohort studies with limited evidence of study heterogeneity to estimate the odds for ever vs never use of talc to be 1.28 (95% CI: 1.20-1.37). An important component of this work was the conduct of a series of subgroup analyses, focusing on the nature of talc use, tumor characteristics, and the possible effect of menopausal state, hormone use, and pelvic surgery.

6.2.4 Categorical regression
Categorical regression can be used to combine data from diverse sources, including different (both toxicological and epidemiological) types of studies and studies focusing on diverse health endpoints. This is done by developing a severity scoring system to place different adverse health outcomes on a common severity scale, following which categorical regression modelling of the severity scores can be done. In analyses involving both animal and human data, adjustments for inter-species differences in sensitivity can be included in the model.

The EPA has invested considerable effort in developing a software package called CatReg to perform categorical regression (EPA, 2017). More recently, Milton and colleagues (2017a,b) have extended the EPA CatReg approach to permit modelling of

U-shaped dose-response curves for essential elements that demonstrate toxicity due to both excess and deficiency. This modelling approach facilitates minimization of the joint risks of excess and deficiency after incorporating all relevant dose-response data within a single categorical regression model. Yetley and colleagues (2017) have identified categorical regression as a potentially useful tool for combining data from multiple sources in establishing DRIs for nutrients. To illustrate the use of categorical regression in practice, Farrell and colleagues (2022) provide a description of the application of this technique to rich datasets on two essential elements – copper and manganese – that include extensive human and animal data. Farrell and colleagues indicate how animal and human evidence on diverse endpoints associated with excess and deficiency of essential elements could be supplemented with alternative test data on the severity of biological changes occurring as upstream effects in adverse outcome pathways.

6.2.5 Combining outcomes with different severities

Assessment of chemical hazards is based on specific critical health effect(s). As an extension, Sand and colleagues (2018) introduced a method for characterizing the dose-related sequence of the development of multiple (lower- to higher-order) toxicological health effects caused by a chemical. Although novel, the method proposed by Sand for combining data on different endpoints, like categorical regression, is not restricted to specific types of data and can accommodate data from NAMs as well as traditional toxicological and epidemiological data. A "reference point profile" was defined as the relation between benchmark doses (BMDs) for selected health effects, and a standardized severity score determined for these effects (Fig. 4). For a given dose of a chemical or mixture, the probability of exceeding the reference point profile can be assessed. Following severity weighing, an overall toxicological response (expressed in terms of the most severe outcomes) at the same dose can then be derived by integrating contributions across all health effects. Conversely, dose equivalents corresponding to specified levels of the new response metric can also be estimated. The reference point profile is a cross-section of the dose-severity-response volume, and in its generalized form the method accounts for all three dimensions (dose, severity, response). In this case, the new response metric becomes a proxy for the probability of response for the most severe health effects rather than the probability for exceeding the BMD for such effects.

Conceptually, there are similarities between this method and categorical regression (e.g., Hertzberg and Miller, 1985; Dourson et al., 1997; Milton et al., 2017b). The latter methods have for example been used to calculate the probability for a given severity category. The method introduced by Sand and colleagues (2018) provides this type of output simultaneously across all categories. In addition, as indicated earlier, probabilities are integrated over the entire severity domain to produce an overall response expressed in terms of the most severe health effects. Integration across different severities requires weighting, and a developed system with nine severity categories (C1 to C9) is therefore mapped to a quantitative severity scale ($S = 0$ to $S = 1$) (see Fig. 4).

For BMD analysis, Sand and colleagues (2018) used data on dioxin-like chemicals from U.S. National Toxicology Program long-term toxicology and carcinogenesis studies in rats describing various health effects in the liver (Fig. 4). They demonstrated that results derived by the method are largely insensitive to the choice of model used to describe the reference point profile. The proposed method also appears to be robust with respect to minor and moderate changes in severity classification of BMDs. Further analyses indicate that the interpretation of effective doses or PODs, based on individual health effects, may change when considering health effects jointly along the lines proposed (Sand, 2022). This influences the consideration of equipotent doses for different chemicals and the concept of acceptable response levels for individual effects. In addition, results suggest that estimation of exposure guidelines, or similar, by the proposed method may be sufficiently accurate and precise even if data for the most severe health effects, associated with the highest severity categories, are omitted (Sand, 2022). The method may therefore enable derivation of a surrogate for the probability of severe health effects, and/or the probability for exceeding corresponding BMDs, also in the case of using data on comparatively "mild" effects only.

Sand (2022) further extend the proposed concept to incorporate genomic dose-response information. A "genomic reference point profile", describing variability within and across affected gene sets, is characterized by using an iterative method for combining all unique BMDs from a transcriptomic study. Similar to the implementation for traditional toxicity data, results suggest that joint consideration of different effects and associated BMDs differentiates the consequence of chemical exposure to a greater extent compared to using a specific/lowest BMD only. This can help to refine establishment of POD for various levels of health concern. Analysis and comparison of apical and genomic reference point profiles, as well as consideration of functional relations between gene sets within such analyses, may aid in the transition towards a NAM-based risk assessment paradigm.

6.2.6 Structured expert elicitation

Structured expert elicitation (SEE) is a well-established approach for gauging expert opinion, of particular value in contexts characterized by limited data, low risk, and substantial uncertainty (Aspinall, 2008, 2010; Cooke, 2013, 2015). One challenge associated with expert elicitation is the risk or perception of overly subjective expert input. For example, Schünemann and colleagues (2019b) point to the need to distinguish expert-elicited evidence from expert opinion, relying on evidence to inform decision-making. Concerns about expert bias are reduced through an anonymized elicitation procedure with a formal, transparent, and auditable processing of responses and a performance-based weighting scheme for pooling judgements. This encourages experts to be open-minded in responding with their estimates and uncertainties, based on their own personal knowledge, expertise, and experience.

There are number of specific SEE methodologies, which may concern a generic scenario or circumstances related to the assess-
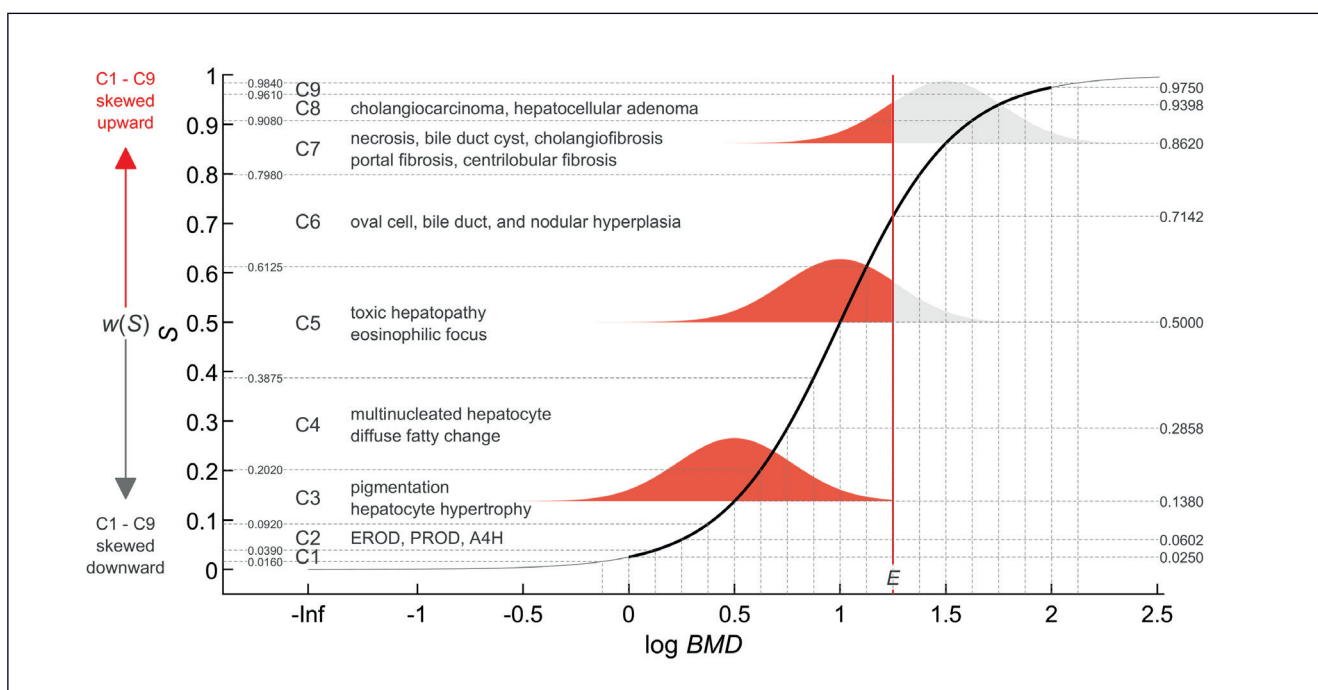
Fig. 4: Technical illustration of the reference point profile (RPP) reproduced from Sand (2018)

The solid S-shaped curve describes the relation between benchmark doses (corresponding to a normalized benchmark response, BMR) for selected apical health effects and the relative severity of toxicity (S) determined for these effects. The RPP applies to the specified value of the normalized BMR and thus represents a cross-section of the dose-S-BMR volume. The relative severity for individual health effects is first determined categorically according to a hierarchical classification scheme: The classification performed in Sand et al. (2018) of considered health effects in the liver is illustrated. The nine-graded categorical scale, C1-C9, is then mapped to a quantitative scale that ranges from $S = 0$ to $S = 1$. The default mapping distributes severity categories symmetrically across S, e.g., with C1, C5, and C9 centered at $S = 0.025$, $S = 0.5$, and $S = 0.975$, respectively (see Sand et al., 2018 for details). The variability is assumed to be normally distributed on the log-scale with constant variance. Red areas describe probabilities, p, for exceeding the RPP at exposure level, E, corresponding to the vertical (red) line. Here, E corresponds to a response (RTR, Equation 3) of 50%, and E intersects the solid curve at $S \approx 0.71$. The midpoint of C6 thus represents the center of the effect/category sequence in terms of RTR. This point of calibration is approximately independent of the RPP model parameters (Sand et al., 2018). A non-linear severity-weight, $w(S) \neq S$, can indirectly modify the symmetrical mapping of C1-C9 to S. This allows the midpoint of the system, corresponding to RTR = 50%, to be lower (C1-C9 skewed upward) or higher (C1-C9 skewed downward) than the midpoint of C6, which would also increase or decrease the RTR associated with E, respectively.

ment or management for a specific project scenario (Colson and Cooke, 2017; Cooke and Goossens, 2008). Generally, the method quantifies subjective judgements through the weighting of expert responses to generate a collective view represented as a median value and accompanying uncertainty distribution. An expert panel between four and twenty members is considered adequate to obtain meaningful results (Colson and Cooke, 2017).

The elicitation may be convened either in person or through video conferencing (thus also reducing the carbon footprint of the event), with experts being offered the opportunity to comment on the process, decline to answer specific questions, or withdraw from the exercise entirely. Per Cooke's Classical Model, an SEE should include the following steps (Cooke, 2013, 2015):

− A draft version of the elicitation instrument is reviewed by independent experts (modifications as necessary).
− The elicitation instrument is introduced to the expert participants, with a thorough review of the relevant terms and con-

ditions. Questions are permitted throughout the process, such that the problem, context, definitions, and question content are understood. Refinements may be made, with the goal of ensuring the same understanding among experts.

− Experts complete each question individually using pre-formatted response tables, which are submitted by the experts to the investigators upon completion. Experts first complete a series of "calibration" questions on technical issues for the topic of interest. This calibration exercise enables distinct performance weights to be given to individual experts based on their accuracy and ability to judge uncertainties.
− Experts then respond to numerical uncertainty distribution "target" questions of the same format, with a central value (median), best judgement (50th percentile), and the 90% credible range (lower limit 5th percentile and upper limit 95th percentile).
− Following the analysis, another facilitated meeting or video conference is arranged, providing the expert panel with an

opportunity to review preliminary findings. Another round of modification and elicitation may be conducted if necessary.

7 Evidence assessment and presentation of findings

7.1 Determining certainty in the evidence based on the GRADE approach

The evidence identified by systematic reviews can be assessed collectively in a framework such as that described for GRADE (Guyatt et al., 2011a). This moves forward the individual study assessment so that end-users understand the strengths and limitations (i.e., certainty) across the body of evidence. Informed by Bradford Hill criteria and the iterative development of evidence-based medicine, the GRADE approach for evidence assessment evaluates the certainty in the evidence based on the following domains that may decrease the certainty in the body of evidence: risk of bias (i.e., study limitations), inconsistency (i.e., heterogeneity), indirectness, imprecision, and publication bias (Balshem et al., 2011; Guyatt et al., 2011b). In addition, for nonrandomized studies, the certainty of the body of evidence may be increased by the following domains: magnitude of effect (e.g., large or very large effect size), dose-response gradient, or opposing residual confounding (an effect seen in the opposite direction expected from confounders).

These eight domains can help assessors understand the body of the evidence across outcomes as it relates to the research question of interest (i.e., our PECO question, Section 3), even when the evidence comes from non-human studies. As mentioned previously, five domains relate to lowering the certainty in the body of evidence. Risk of bias is informed by the individual study assessments performed as part of the systematic review. While GRADE was originally developed in the context of RCTs, its application has expanded to include risk of bias related to randomized and nonrandomized intervention and exposure studies (Guyatt et al., 2011a; Morgan et al., 2019; Schünemann et al., 2019a). When considering inconsistency across the pooled evidence, the distinction is made between explained or unexplained inconsistency (Guyatt et al., 2011a). Indirectness, based on how directly the identified evidence answers the research question, is a key element for evidence integration (Guyatt et al., 2011c). Information used to inform indirectness would be, when the population of interest is humans, how directly can evidence from animal experiments or other types of research (*in vitro* or *in vivo*) be extrapolated or help inform the association between an exposure and outcome. Some research has explored how the domain of indirectness relates to evidence from pre-clinical studies (e.g., research from animals) (Hooijmans et al., 2018). Imprecision considers whether the overall estimate of effect is precise or due to random error. Lastly, publication bias summarizes whether all the studies that have been conducted were captured in the review.

When considering the three domains that allow for increased certainty across the body of evidence, the magnitude of effect captures the extent of the observed effect, dose-response considers the exposure-effect relationship, and opposing residual con-

founding captures whether or not the worst-case scenario still allows for drawing strong conclusions (Guyatt et al., 2011b).

Operationalizing these domains relates to the understanding of the relationship between the research question and the evidence extrapolated to inform the findings to that question. In most instances, this can be informed by exploring the various sources of indirectness. For example, one may be interested in humans as a population. For example, humans who are exposed to a carcinogen with an interest in exploring how this exposure would relate to an adverse outcome of interest. Within GRADE, the best available evidence is understood to come from human studies; however, indirect evidence from other sources (animal, mechanistic) is also considered. The exposure or comparator could also introduce an element of indirectness. To apply this to a review, five paradigmatic scenarios exist (Morgan et al., 2018). In one particular scenario, little is known about the association between the exposure and outcome, therefore the assessment seeks to define that relationship. In this situation, mechanistic data or modelling may be utilized to see whether we have some confidence in a statement between the exposure and outcome. Indirectness may also be identified within the outcome, as to whether our evidence is extrapolated from a surrogate.

The starting point when determining the certainty across the body of evidence for an outcome typically starts at high certainty for RCTs and low certainty for nonrandomized studies. However, with the development of risk of bias instruments applied to nonrandomized studies that use a standardized scale from RCTs, the level of certainty could be increased (Morgan et al., 2019; Schünemann et al., 2019a).

7.2 Relevance to risk assessment

Characterizing the relationship between exposure levels and the health impacts they exert has been the focus of much research conducted by risk assessors, health practitioners, and regulatory experts in developing health protection programs to establish safe intake levels for humans (Krewski et al., 2010; Stern et al., 2007). Any substance, including but not limited to chemicals, nutrients, vitamins, or pharmaceuticals, has the potential to be harmful to humans if they are exposed to too much or too little. To establish a range of allowable intake for a substance that may be harmful to humans in excess or deficient amounts, it is necessary to strike a balance between the health impacts exerted by exposures across the excess-deficiency spectrum. The challenge of identifying an acceptable medium between excess and deficiency motivates the development of exposure-response models; they provide the foundation for identifying recommended levels of exposure to essential and nonessential substances (Krewski et al., 2010).

The EPA developed the concept of a RfV for toxic substances, which has been widely accepted and used in practice to prescribe safe intake levels for humans. The RfV can be derived from a single key study that considers one critical health effect and is defined by applying uncertainty or adjustment factors to the no-observed-adverse-effects level (NOAEL), which corresponds to the level of exposure that does not result in a significant increase in the risk of adverse effects in the exposed group when compared

with controls (Barnes and Dourson, 1988). Currently, the benchmark dose (BMD) is more often used as the basis for determining the RfV. The BMD was introduced by Crump (1984) and uses a mathematical model to identify the dose corresponding to a specified increase in response. It has also been used as a point of departure on the dose-response curve for establishing human exposure guidelines (ECHA, 2013). More recently, the signal-to-noise crossover dose (SNCD), defined as the dose at which the uncertainty in the biological signal is indistinguishable from the background noise, has been introduced as an alternative to the BMD (Sand et al., 2011). Categorical regression has also been applied in dose-response modelling for health risk assessment (Allen et al., 2005; Chambers et al., 2010; Gift et al., 2008; Haber et al., 2001; Milton et al., 2017a,b). Categorical regression facilitates the inclusion of multiple studies in an exposure-response model by applying a severity scoring scheme to standardize different outcomes reported in each experiment. Further discussion of these techniques may be found in previous publications (Milton et al., 2017a; Yetley et al., 2017).

8 Uncertainty analysis

8.1 Assessing uncertainty in risk assessment
In a great majority of situations of interest, risk assessment is confronted with a wide range of imperfections in the evidence base that collectively constrain the ability to provide a certain answer with respect to causality, and even if causality is assumed, a certain estimate of the level of risk to be expected in the exposed population of interest. The level of uncertainty in the level of risk can span multiple orders of magnitude, even when excluding uncertainty in causality.

In general, guidance documents that prescribe best practices in risk assessment suggest that a formal treatment of uncertainty is a fundamental component of risk assessment processes (NRC, 2009). The sources of uncertainty (i.e., the limitations of the evidence base) can be from a wide spectrum of sources, including:

− inability to infer a causal relationship due to conflicting evidence within the same line of evidence
− inability to infer a causal relationship due to conflicting evidence from different lines of evidence
− lack of testing data for numerous types of possible health outcomes
− testing data which is available but of questionable relevance to human health
− uncertainty in what to predict at low doses given data restricted to much higher doses
− uncertainty in the fate and transport of substances and the resulting environmental concentrations in near- and far-field exposure situations
− the levels of exposure in the population of interest to the environmental media or product in question; and
− uncertainty stemming from imprecision of knowledge of mechanisms of action

Each of these forms of uncertainty can be dealt with in a piece-wise fashion with the uncertainty characterization limited to an interim conclusion within different stages of the risk assessment process. In addition, the means of characterizing uncertainties can vary from a narrative approach (explaining the limitations or extent of doubt of any conclusions and the basis for that doubt), to qualitative (labelling uncertainty as high, medium, low) to categorical (e.g., causal, likely to be causal, suggestive evidence of causal, etc.) through to quantitative characterizations (providing a probability estimate for causality, providing a confidence interval on an estimated value, or explicitly defining or computing a probability distribution representing uncertainty in an estimated or computed value).

The various practices employed determine, partly, to what extent a final and overall characterization of uncertainty can be rendered, even if the uncertainties in each part of the risk assessment are relatively complete. Fully quantitative estimates of uncertainty in each part of a risk assessment can be propagated in a reliable way to capture the overall uncertainty. However, it is currently rare that practitioners capture uncertainty in a quantitative way in all the ways that the inadequacies of the evidence base might contribute to the overall uncertainty in a risk assessment.

For the three main risk assessment steps preceding risk characterization, recent methods have emerged, some have long existed, and other methods need to be developed to allow for the careful and comprehensive treatment of uncertainty in risk assessment.

In hazard identification, progress is being made in minimizing some sources of uncertainty due to the increasing use of formal systematic methods in the gathering and treatment of evidence (e.g., explicit criteria for inclusion and quality scoring of studies) within the principal lines of evidence. However, methods to formally capture the uncertainty and implications of imperfections in the evidence base when integrating evidence across evidence lines appears to have limited formal methodological support, often relying on expert judgement and consensus-based processes in the ultimate weighing of evidence.

One potential method of formally weighing and combining evidence across evidence lines was described during the workshop in a proof-of-concept application. This method captures the imperfections of each type of evidence (both in a general sense as well as in a study-by-study sense) in the form of conditional probability tables linking *evidence* of various types and qualities to *hypotheses* in a Bayesian network. The Bayesian network is a means of computationally combining all the evidence, through application of Bayes' rule within the software tool, to yield intermediate and overall statements of the uncertainty in a hypothesis linked to the various sources of evidence in the network. The explicit linkages captured in the network allow for detailed sensitivity analysis both in a general sense (How strongly is an animal-based test of type X linked to evidence of health outcomes Y) and in a specific sense (How strongly does the overall conclusion of carcinogenicity for substance A depend on the quality score applied to study Z?). Jaworska et al. (2010) described an example of this approach applied to mutagenicity testing. Another decision-analytic method of combining evidence, the Dempster-Shafer theory of evidence, has also been applied in the context of computational toxicology (Rathman et al., 2018).

For exposure assessment, with some effort, quantitative estimates of uncertainty can be readily combined (e.g., through

Monte Carlo simulation or other methods of propagating uncertainty) from the uncertainties in individual quantitative components (contaminant levels, food intakes, product use, occupational working conditions, inhalation rates, body weights) that are combined to generate estimates of dose. While non-trivial, conducting an uncertainty assessment within the step of exposure assessment is supported by existing methods.

Dose-response assessment is an area where recent advances may be considered the most dramatic. The availability of software tools to enable dose-response assessment, including capturing uncertainty in the process, has been an important contribution. Efforts to harmonize non-cancer and cancer dose-response assessment into a formal and quantitative process have made significant progress and are now widely applicable with supporting publicly available tools (see 8.2 below; Chiu et al., 2018). The ability to consider varying levels of severity in a quantitative way is another important consideration in removing uncertainty that stems from the relative vagueness of the treatment of severity in the traditional minimally quantitative dose-response approaches (see 6.2.5 above; Sand et al., 2018).

The risk characterization step of risk assessment integrates insights from the hazard identification, exposure assessment, and dose-response steps to generate overall estimates of risk and uncertainty surrounding those estimates. With recent advances in the treatment of dose-response assessment, the ability to render a more complete characterization of uncertainty in risk estimates should be enabled. The inclusion of uncertainty in causality aspects of hazard identification in an integrated way may be the most challenging to render and to communicate.

8.2 The IPCS approach to quantitative uncertainty analysis

As noted above, the RfV and similar approaches to quantitative synthesis suffer from some key limitations, many of which have been identified by several National Academies reports (NRC, 1994, 2009). For instance, the way in which RfV have traditionally been derived characterize neither the degree of residual risk that may be present nor the shape of the dose-response curve for adverse effects. This is because the RfV is usually obtained by identifying a NOAEL of intake from an experimental study (predominantly animal studies) and then dividing this intake level by a number of "uncertainty factors" to account for limitations in the data. The most commonly applied uncertainty factors are a factor of 10 to address differences between experimental animals (UF_A) and a second factor of 10 to address variability among humans (UF_H). This "NOAEL divided by 100" concept dates back to the 1950s in the context of FDA regulation of food additives (Lehman and Fitzhugh, 1954). Each of these components – the NOAEL, UF_A, and UF_H – is assumed to be "conservative" in the sense of erring on the side of protecting public health, but without much specificity as to "how conservative" they actually are (WHO/IPCS, 2014). For instance, with respect to the NOAEL, it is assumed that the severity of effects at this exposure level is negligible, but the extent to which this is true depends on the endpoint examined and the statistical power of the study (Crump, 1984; EPA, 2012). For UF_A, it is assumed that humans are gen-

erally no more than 10-fold more sensitive than the experimental animal species, but it is unclear at what confidence level this 10-fold factor is supposed to be (90%, 95%, 99%?). Similarly, for UF_H, it is assumed that individuals more susceptible to toxicity are no more than 10-fold more sensitive than more typical individuals. Here, there are two ambiguities: first, like UF_A, the confidence level of this 10-fold factor is unclear; 90%, 95%, 99%? Second, it is unclear what "susceptible" means in terms of the more sensitive tail of the population distribution; 5%, 1%, 1 in a million?

More recently, the World Health Organization/International Program on Chemical Safety (WHO/IPCS) developed a guidance document describing a "probabilistic" framework that results in substantially better characterization of the intake-response for adverse effects (Chiu and Slob, 2015). The key concept underlying the WHO/IPCS approach is that the goal of deriving quantities like the reference dose (RfD) is a "target human dose" HD_M^I, defined as the estimated human dose (or intake) at which effects with magnitude M occur in the population with an incidence I, along with an associated confidence interval (Fig. 5A).

The derivation of the HD_M^I can be summarized into the following steps (illustrated in Fig. 5B):

− *Replace the NOAEL with so-called "benchmark dose (BMD) modeling."* The limitations of the NOAEL as a starting point for toxicological risk assessment have been recognized for decades (EFSA, 2015b; NAS, 2001; EPA, 1995, 2012; WHO/IPCS, 2009), and it is generally accepted that the BMD, introduced by Crump (1984), is more scientifically appropriate. The BMD is the dose associated with a specific size of effect, the benchmark response (BMR) [we use M, for magnitude of effect, to denote this value]. The BMD_M is estimated, with associated confidence interval/statistical distribution, by statistical model fitting to dose-response data. In this way, it provides information that the NOAEL does not regarding the nature and severity of the adverse effects under consideration as well as the precision with which the associated intake level can be estimated.

− *Replace interspecies UF_A with results from mathematical/computational modeling.* The approaches for replacing UF_A all involve disaggregating it into two components. The first component is related to dosimetry, denoted as the dosimetric adjustment factor (DAF), to convert experimental animal exposures to "human equivalent" exposures. If chemical-specific toxicokinetic (TK) data are available, PBPK models have been used to derive the DAF (Corley et al., 2012; Dorman et al., 2008; Schroeter et al., 2008; Teeguarden et al., 2008). Otherwise, generic approaches have been developed based on physiological differences, such as allometric scaling by body mass, between experimental animals and humans (EPA, 1994, 2011; West, 1999). A second component, which is denoted UF_A, accounts for any remaining (i.e., unknown) chemical-specific interspecies differences, such as due to toxicodynamics (TD). WHO/IPSC reviewed analyses of historical data on toxicity thresholds across species and chemicals, deriving confidence intervals for the DAF and remaining interspecies differences for oral intakes (WHO/IPCS, 2014).
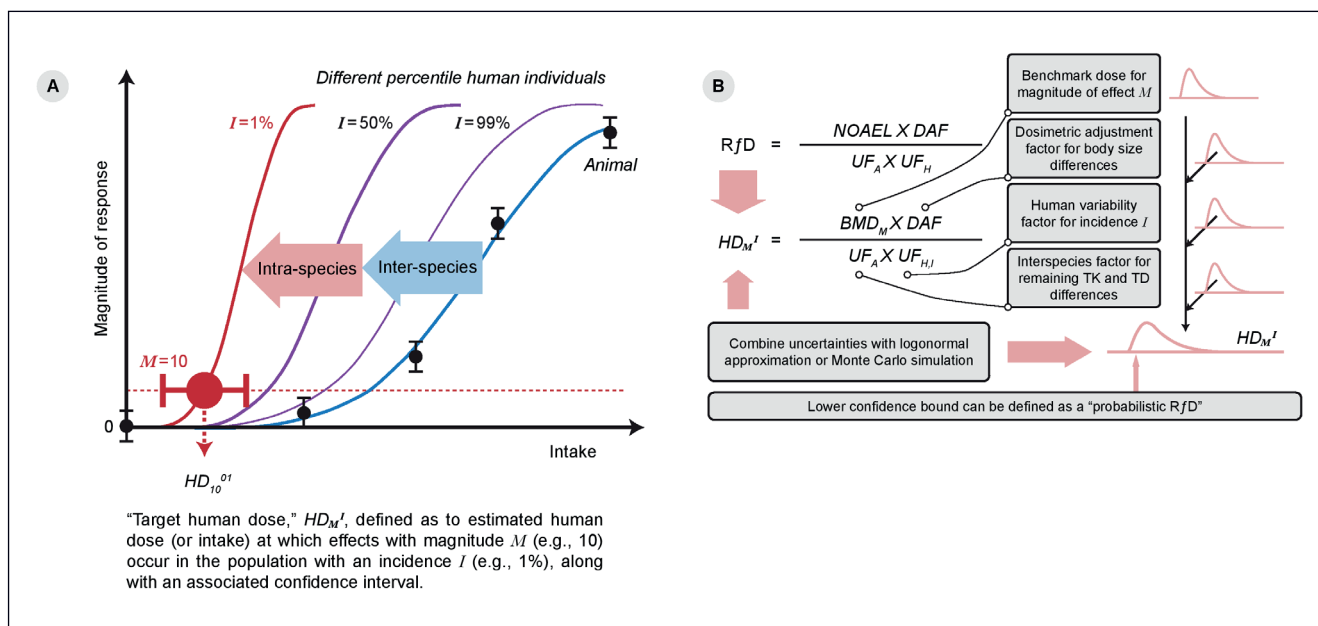
Fig. 5: Summary of probabilistic approach to characterize quantitative uncertainty and variability in dose-response assessment
(A) Illustration of the concept of the "human target dose," HD_M^I, which replaces traditional toxicity values such as the reference dose (RfD).
(B) General approach to deriving HD_M^I probabilistically in comparison to deriving a traditional deterministic RfD. For additional details, see WHO/IPCS (2014) and Chiu and Slob (2015).

– *Replace human variability UF_H with results from mathematical/computational modeling*. In the WHO/IPCS framework, UF_H is replaced with a value that depends on the Ith population percentile of susceptibility, denoted $UF_{H,I}$ (Chiu and Slob, 2015; WHO/IPCS, 2014). This quantity reflects TK and TD differences between individuals at the median and the Ith percentile of the population distribution. For TK, PBPK models have been used for specific substances to estimate the degree of human variability. For TD, the most promising approach for substance-specific information has been based on population-based *in vitro* studies (e.g., Chiu et al., 2017); quantitative AOP models may have potential in the future (e.g., Perkins et al., 2019a), but their expansion to include human variability has not been pursued. WHO reviewed previous analyses of historical data on TK and/or TD variability across chemicals (WHO/IPCS, 2014), based on work by Hattis and colleagues (Hattis et al., 2002), and recommend "default" factors (as probability distributions) that can be applied.
– *Combine the newly defined BMD, UF_A, and $UF_{H,I}$ in a probabilistic manner to derive an intake-response function and its uncertainty*. The integration of BMD modeling, allometric scaling, and historical TK/TD variability data across chemicals leads to the HD_M^I (Chiu and Slob, 2015; WHO/IPCS, 2014). The HD_M^I can then replace the UL and disaggregates "risk" into the distinct concepts of magnitude of effect (M), incidence of effect (I), and uncertainty (reflected in the confidence interval). The HD_M^I can furthermore be mathematically "inverted" to derive an intake-response function for a specified fraction I of the population (Chiu and Slob, 2015).

The resulting output is a two-dimensional distribution of intake-response functions: one reflecting human variability in terms of Ith percentiles (as shown by the intake-response curves in Fig. 5A), and the other reflecting statistical uncertainty (as shown by the confidence interval on the HD_M^I in Fig. 5A). By providing intake-response functions rather than "bright lines," changes in risk of adverse effects from changes in dose can be quantified. This type of "risk-benefit" comparison would be infeasible under the traditional "NOAEL divided by 100" approach, because there is no characterization of the gradient of the dose-response over a wide enough range of doses. However, the approach to derive an HD_M^I would enable such comparison to be made much more easily. Chiu and colleagues (2018) recently applied this approach to over 600 chemicals and 1,500 endpoints, demonstrating the feasibility of broadly implementing this approach in chemical risk assessments.

9 Preliminary evidence-based risk assessment framework

The overarching goal of this project was to develop an evidence-based risk assessment framework to guide the conduct of evidence-based risk assessment, including assembling and synthesizing all relevant data as discussed above. This document provides detailed guidance on each of the key steps involved in conducting an evidence-based risk assessment and is of value to practitioners seeking to ensure that risk assessments are completed according to the highest possible scientific standards and are
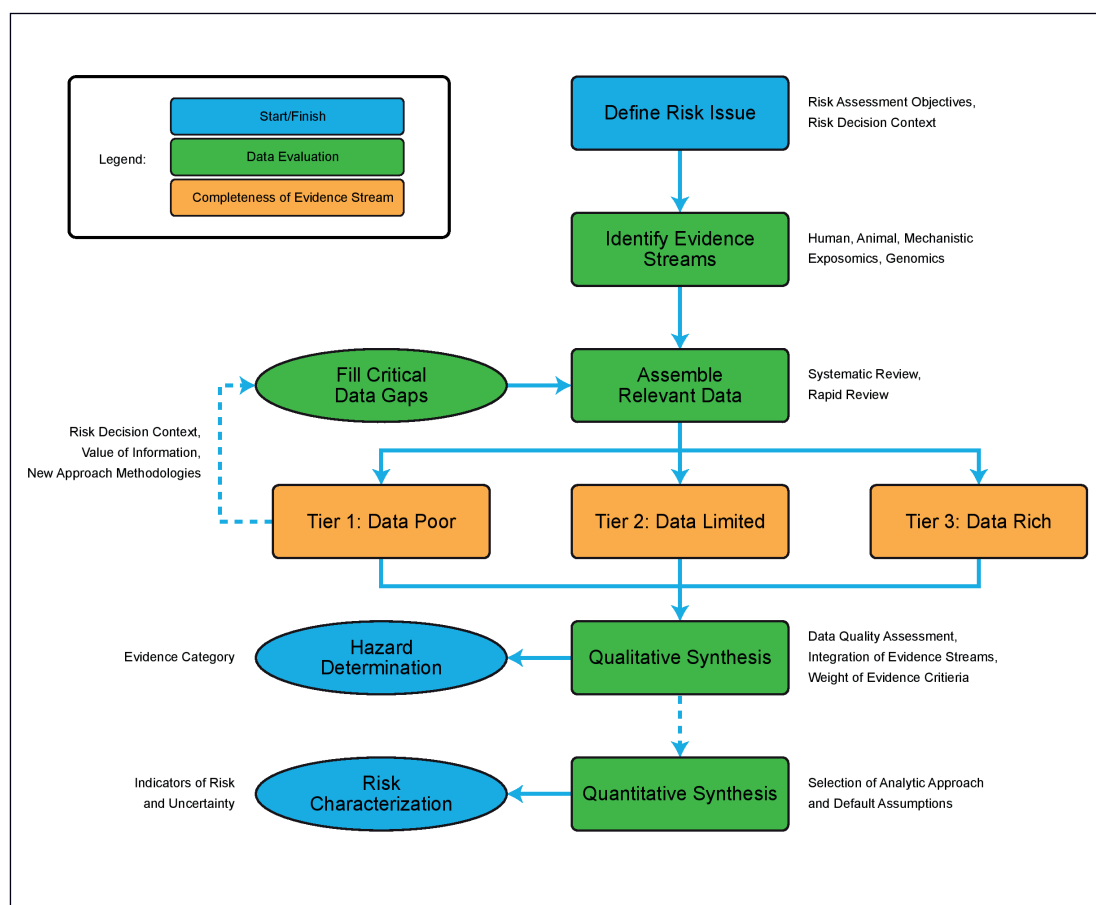
Fig. 6: Preliminary evidence-based risk assessment framework

conducted in an open, transparent, and reproducible manner. This guidance aligns with and builds upon foundations of good practice established in prior publications by individual experts and assessment agencies, such as that recently provided by OECD (2019a).

An important aspect of the framework is the distinction between the related steps of assembling and synthesizing the evidence. Systematic review offers a powerful approach to assembling all relevant data in support of the assessment, with objective inclusion/exclusion criteria and study quality assessment. In the past, the selection of studies to be included in risk assessments has sometimes been a source of controversy. By invoking current best practices in systematic review, it is expected that much of this controversy can be circumvented.

Having agreed on the evidence base to support risk assessment, attention can then focus on qualitative and possibly quantitative syntheses of the available information. At this stage, clear criteria for evaluating the available data will serve to support data-driven determinations regarding the existence or otherwise of a human health hazard. Should a human health hazard be identified using the criteria embodied in the framework, methods for quantitative syntheses of the available data can then be applied in cases where the available data are sufficient to support an evidence-based estimate of potential population human health risk and associated uncertainty.

An initial framework for the evaluation of all available evidence on the association between a particular exposure and adverse outcome is presented in Figure 6. Although practitioners will be familiar with each of the components of this framework, it may serve as a useful paradigm for ensuring consistency in evidence-based risk assessment. Following problem formulation (an important starting point, but outside the scope of the present framework), the first step is to assemble all relevant evidence relating to the specific risk issue under consideration. As discussed, systematic review provides a powerful set of methodologies for accomplishing this in a comprehensive and reproducible manner.

In the framework laid out in Figure 6, once all relevant evidence has been summarized, the volume of evidence is assigned to one of three tiers: data-poor (Tier 1), limited data (Tier 2), or data-rich (Tier 3). In the data-poor context, a decision is required as to whether the assessment should proceed to a formal evaluation: should the data be judged inadequate to support a meaningful evaluation, key data gaps should be identified and filled before proceeding.

Within the context of the present framework, the term "limited data" (Tier 2) is used to represent at least a minimal amount of data that would support a credible risk assessment. The term "data-rich" (Tier 3) represents the case in which considerable data is available from multiple sources (including human, animal, and

other experimental sources) to support a credible evaluation.

Once it has been determined that the data are adequate to support assessment development (regardless of whether the available evidence falls into Tier 1, 2 or 3), the next step is to conduct a qualitative synthesis of the available data, resulting in a determination as to whether or not a human health hazard exists. This is a non-trivial undertaking and will involve the application of explicit (likely context-specific) criteria for hazard determination.

Although such criteria have been elaborated for certain outcomes such as cancer (specifically those elaborated upon by IARC in the 2019 update to the IARC monographs) (Samet et al., 2019; Baan and Straif, 2022), other criteria will need to be developed for other adverse outcomes. Lessons may be learned from current REACH guidance, which focuses on 11 broad adverse outcomes (Armstrong et al., 2020). The end result of the qualitative synthesis is a statement about the evidence for a causal association between the exposure and outcome(s) of interest. Should an inconclusive outcome be reached, outstanding data gaps should be noted and addressed for use in a future re-evaluation. In its review of the U.S. EPA IRIS program, the NRC recommended categories of evidence that included "sufficient to infer a causal relationship", "suggestive but not sufficient to infer a causal relationship", "inadequate to infer the presence of a causal relationship", and "suggestive of no causal relationship" (NRC, 2014a, p.94). Other investigators have also suggested simplified evidence categorization schemes, including Wigle and colleagues (2008) and Krewski and colleagues (2017), who proposed categories for classifying evidence as "sufficient", "limited" or "inadequate".

Should the qualitative synthesis conclude that a human health hazard exists, a quantitative synthesis of the available data can be attempted, with the goal of characterizing the level of risk, and attendant uncertainty, in quantitative terms. Recent trends in data aggregation have provided powerful new approaches to quantitative data synthesis, including techniques such as categorical regression, that permit the inclusion of quantitative data from multiple sources and on multiple endpoints into a single dose-response analysis. It is important to note that the successful completion of a qualitative synthesis of the available data does not guarantee that the data can support a meaningful quantitative synthesis; a quantitative synthesis will only be possible when there is reliable data on the dose-response relationship between the agent and outcome of interest from one or more sources.

10 Summary of workshop deliberations

Experts from the field of evidence integration met at a workshop hosted by the University of Ottawa on December 17-18, 2018, to discuss the development of an evidence-based risk assessment framework. The workshop benefitted from strong attendance from government, academia, and industry.

The workshop agenda (provided in full in the supplementary material[10]) covered a range of topics relevant to risk science and evidence integration, from emerging trends and new methodologies to approaches for putting evidence integration into practice.

During the second day of the workshop, participants were divided into breakout groups and tasked with addressing four key questions relating to the development of a framework for evidence-based risk assessment. The four discussion areas, described in more detail below, were:
1. Lessons learned from previous experience
2. Benchmarks of good practice
3. Problem formulation and data requirements
4. Potential challenges

10.1 Breakout group 1:
Lessons learned from previous experience
This breakout group explored what can be learned from previous experience in the development of an organizing framework for evidence integration. Participants explored past successes and challenges, as well as opportunities for future improvement.

The iterative development and refinement of systematic and transparent approaches for data collection, screening, and abstraction were identified as key areas of strength where insights could be drawn from past experience. Participants noted, however, that guidance tends to be less clear for mechanistic data. Also, there was a discussion of the need to preserve evidence and decision context throughout the risk assessment process.

It was also suggested that the ability to both summarize individual lines of evidence and use insights from one line to inform judgements for another had improved over time. This was driven in part by best practices in the clinical and pharmaceutical industry, and participants suggested that similar efforts for evidence-based risk assessment could benefit from building upon this foundation.

Participants noted that "getting the data right does not mean getting the answer right". In other words, while a systematic approach is essential, both subject matter expertise and public review should be included as necessary components of the risk assessment process. Similarly, transparency should not be conflated with objectivity, and both decisions and their impacts on the assessment process should be made explicit.

Moving forward, participants recommended a focus on developing discrete steps for evidence integration that record and summarily present choices, assumptions, and justifications. They also recommended the development of reporting guidelines for author publication of individual studies to facilitate their incorporation in risk assessments. Lastly, participants advocated for continued knowledge-sharing and collaboration to foster best practices and reduce duplication of work.

10.2 Breakout group 2:
Benchmarks of good practice
This breakout group explored examples of previous evidence integration efforts in an effort to identify examples of best practices. Participants sought to address this topic by identifying both useful frameworks and relevant case studies to examine their application in practice.

With regard to applicable frameworks, the participants made note of several of the frameworks discussed in earlier sections of this workshop report. One framework that received particular at-

tention was the updated Whaley literature review appraisal tool-kit[11], which participants suggested could be useful as a mechanism for refining best practices and evaluation criteria for systematic reviews.

The case studies discussed were intended to explore various systematic review frameworks, highlighting strengths and areas that could be improved moving forward. Participants compared numerous case studies, such as OHAT immunotoxicity (involving human, animal, and mechanistic data), Health Canada data-poor case studies, and EPA IRIS chemical assessments.

From this discussion, participants developed three key recommendations for future consideration.

1. Prospective case studies could be considered as an avenue for advancing discussion of best practices.
2. Empirical studies of how to reduce risk of bias could further increase the confidence with which risk assessment findings could be interpreted.
3. Efforts to evaluate the impact of different levels of literature review comprehensiveness could be informative in guiding best practices and pragmatism in review efforts.

10.3 Breakout group 3:
Problem formulation and data requirements

This breakout group explored how problem formulation can be used to define data requirements for evidence-based risk assessment as well as how contextual factors may impact these requirements. Participants strongly agreed that risk assessment data requirements are context-dependent and focused their discussion on the use of GRADE to respond to research questions with varying degrees of urgency.

Participants examined how GRADE can be applied to assess research questions across different timescales, including emergency response (within hours), urgent response (1-2 weeks), rapid response (1-3 months), and routine response (more than three months). The types of evidence incorporated and risk of bias associated with findings would vary across these categories, demonstrating the context-specific nature of data requirements (and flexibility of the GRADE process). While GRADE principles can be applied to results of both systematic and expedited reviews, non-systematic reviews may often lead to greater uncertainty in the interpretation of review findings. In short, the overarching recommendation of the group was that contextual factors – including the urgency of characterizing a risk – must be considered in problem formulation efforts and data requirement discussions.

10.4 Breakout group 4:
Potential challenges

This breakout group discussed what challenges may arise in efforts to strengthen evidence-based risk assessment. Participants considered risk assessment within the broad categories of evidence retrieval (data collection), integration within a stream of evidence, and integration across evidence streams.

Within the context of evidence retrieval, participants highlighted the need to be comprehensive without diluting the findings of the review. They identified the need to improve the quality of evidence that informs risk assessment and identified knowledge translation, data sharing, and methodological standardization as possible avenues for overcoming this challenge.

In assessing individual lines of evidence, participants discussed challenges relating to the weighting of difference sources (such as among several animal models) and reaching agreement on the "quality" of information available. Specifically, they discussed what limitations should be viewed as more problematic than others.

When integrating evidence across evidence streams, participants discussed challenges driven by uncertainty about the informative value of different evidence. Integration of qualitative and quantitative data can be particularly challenging in this regard, and participants proposed standardized metrics and probabilistic approaches as potential avenues for overcoming these difficulties.

At a general level, authors suggested that key challenges included building confidence in the risk assessment process, synthesizing different types of evidence, and navigating inter-reviewer disagreements and judgements. They suggested that complementary evidence could be used to support generalizations of risk assessment findings and that evidence integration at different levels (e.g., to understand the biological plausibility of mechanistic data) could be of value.

11 Conclusion

The overarching objective of this initial workshop was to explore the development of an evidence-based framework for risk assessment. Participants at the workshop discussed issues relating to recent advances in risk science, new methodologies in evidence evaluation, approaches for qualitative and quantitative evidence synthesis, and putting evidence integration into practice.

A preliminary framework was distributed in advance of the workshop and refined based on discussion and debate among the workshop participants. The framework presents a practical approach to evidence evaluation and synthesis designed to ensure that the relevant evidence for human health risk assessment is considered in a comprehensive and objective manner. The framework relies on current best practices in systematic review to summarize human, animal, and mechanistic evidence relevant to the risk issue under consideration. With the recent advances in systematic review methodology and powerful software to support the conduct of systematic review, the relevant evidence can be readily summarized in a comprehensive and reproducible manner.

Once the available evidence has been summarized, evidence can be used to conduct a qualitative synthesis of the available data. Application of appropriate integration approaches can be used for hazard determination. If a human health hazard is identified, a quantitative synthesis of the data can then be undertaken to

---

[11] Further details on this and other appraisal tools are provided by Whaley (2015) and Menon et al. (2022).

characterize population health risks and uncertainties in quantitative terms. Both the qualitative and quantitative syntheses will require adequate data to support these syntheses, with key data gaps being identified and filled as relevant to the context of the risk assessment.

Elaborating on the evidence-based risk assessment framework proposed here will require more in-depth consideration of the criteria to be applied in conducting both qualitative and quantitative synthesis of the data. However, expert judgement is essential to informing the proper application of such criteria, which can limit the reproducibility and certainty of the overall findings; rigorous, stepwise, and transparent processes are needed to calibrate and ensure reproducibility of specific applications of the framework. Such criteria will be proposed in a subsequent phase of this work. Upon completion of the evidence-based risk assessment framework, case study prototypes could be conducted to evaluate its use in practice. A follow-up workshop is currently being planned to flesh out the preliminary framework for evidence-based risk assessment in more detail.

References

Aiassa, E., Merten, C. and Martino, L. (2022). EFSA's framework for evidence-based scientific assessments: A case study on uncertainty analysis. *ALTEX 39*, 451-462. doi:10.14573/altex.2004211

Allen, B., Zeiger, E., Lawrence, G. et al. (2005). Dose-response modeling of in vivo genotoxicity data for use in risk assessment: Some approaches illustrated by an analysis of acrylamide. *Regul Toxicol Pharmacol 41*, 6-27. doi:10.1016/j.yrtph.2004.09.006

Anastas, P. T., Sonich-Mullin, C. and Fried, B. (2010). Designing science in a crisis: The Deepwater Horizon oil spill. *Environ Sci Technol 44*, 9250-9251. doi:10.1021/es103700x

Andersen, M. E., McMullen, P. D., Phillips, M. B. et al. (2019). Developing context appropriate toxicity testing approaches using new alternative methods (NAMs). *ALTEX 36*, 523-534. doi:10.14573/altex.1906261

Armstrong, V., Karyakina, N., Nordheim, E. et al. (2020). Overview of REACH: Issues involved in the registration of metals. *Neurotoxicology 83*, 186-198. doi:10.1016/j.neuro.2020.01.010

Aspinall, W. (2008). Expert judgment elicitation using the Classical Model and EXCALIBUR. In *Briefing notes for Seventh Session of the Statistics and Risk Assessment Section's International Expert Advisory Group on Risk Modeling: Iterative Risk Assessment Processes for Policy Development Under Conditions of Uncertainty/Emerging Infectious Diseases: Round IV*. http://dutiosc.twi.tudelft.nl/~risk/extrafiles/EJcourse/Sheets/Aspinall%20Briefing%20Notes.pdf

Aspinall, W. (2010). A route to more tractable expert advice. *Nature 463*, 294-295. doi:10.1038/463294a

Baan, R. A. and Straif, K. (2022). The Monographs programme of the international agency for research on cancer: A brief history of its Preamble. *ALTEX 39*, 443-450. doi:10.14573/altex.2004081

Balshem, H., Helfand, M., Schünemann, H. J. et al. (2011). GRADE guidelines: 3. Rating the quality of evidence. *J Clin Epidemiol 64*, 401-406. doi:10.1016/j.jclinepi.2010.07.015

Barnes, D. G. and Dourson, M. (1988). Reference dose (RfD): Description and use in health risk assessments. *Regul Toxicol Pharmacol 8*, 471-486. doi:10.1016/0273-2300(88)90047-5

Blettner, M., Sauerbrei, W., Schlehofer, B. et al. (1999). Traditional reviews, meta-analyses and pooled analyses in epidemiology. *Int J Epidemiol 28*, 1-9. doi:10.1093/ije/28.1.1

Bolt, H. M. (2019). Highlight report: Adverse outcome pathways: The need of research on mechanisms of toxicity. *Arch Toxicol 93*, 3385-3386. doi:10.1007/s00204-019-02596-2

Burns, P. B., Rohrich, R. J. and Chung, K. C. (2011). The levels of evidence and their role in evidence-based medicine. *Plast Reconstr Surg 128*, 305-310. doi:10.1097/PRS.0b013e318219c171

Cardis, E., Armstrong, B. K., Bowman, J. D. et al. (2011). Risk of brain tumours in relation to estimated RF dose from mobile phones: Results from five Interphone countries. *Occup Environ Med 68*, 631-640. doi:10.1136/oemed-2011-100155

Chambers, A., Krewski, D., Birkett, N. et al. (2010). An exposure-response curve for copper excess and deficiency. *J Toxicol Environ Health B Crit Rev 13*, 546-578. doi:10.1080/10937404.2010.538657

Checkoway, H. (1991). Data pooling in occupational studies. *J Occup Med 33*, 1257-1260.

Chiu, W. A. and Slob, W. (2015). A unified probabilistic framework for dose-response assessment of human health effects. *Environ Health Perspect 123*, 1241-1254. doi:10.1289/ehp.1409385

Chiu, W. A., Wright, F. A. and Rusyn, I. (2017). A tiered, Bayesian approach to estimating of population variability for regulatory decision-making. *ALTEX 34*, 377-388. doi:10.14573/altex.1608251

Chiu, W. A., Axelrad, D. A., Dalaijamts, C. et al. (2018). Beyond the RfD: Broad application of a probabilistic approach to improve chemical dose-response assessments for noncancer effects. *Environ Health Perspect 126*, 067009-067009. doi:10.1289/EHP3368

Colson, A. R. and Cooke, R. M. (2017). Cross validation for the classical model of structured expert judgment. *Reliab Eng Syst Saf 163*, 109-120. doi:10.1016/j.ress.2017.02.003

Cooke, R. M. and Goossens, L. L. H. J. (2008). TU Delft expert judgment database. *Reliab Eng Syst Saf 93*, 657-674. doi:10.1016/j.ress.2007.03.005

Cooke, R. M. (2013). Validating Expert Judgment with the Classical Model. http://www.expertsinuncertainty.net/LinkClick.aspx?fileticket=HlcTmEoDunY%3D&tabid=4385&mid=8296

Cooke, R. M. (2015). The aggregation of expert judgment: Do good things come to those who weight? *Risk Anal 35*, 12-15. doi:10.1111/risa.12353

Corley, R. A., Kabilan, S., Kuprat, A. P. et al. (2012). Comparative computational modeling of airflows and vapor dosimetry in the respiratory tracts of rat, monkey, and human. *Toxicol Sci 128*, 500-516. doi:10.1093/toxsci/kfs168

Cote, I., Andersen, M. E., Ankley, G. T. et al. (2016). The next

generation of risk assessment multi-year study-highlights of findings, applications to risk assessment, and future directions. *Environ Health Perspect 124*, 1671-1682. doi:10.1289/ehp233

Crump, K. S. (1984). A new method for determining allowable daily intakes. *Fundam Appl Toxicol 4*, 854-871. doi:10.1016/0272-0590(84)90107-6

Darby, S., Hill, D., Auvinen, A. et al. (2005). Radon in homes and risk of lung cancer: Collaborative analysis of individual data from 13 European case-control studies. *BMJ 330*, 223. doi:10.1136/bmj.38308.477650.63

Dorman, D. C., Struve, M. F., Wong, B. A. et al. (2008). Derivation of an inhalation reference concentration based upon olfactory neuronal loss in male rats following subchronic acetaldehyde inhalation. *Inhal Toxicol 20*, 245-256. doi:10.1080/08958370701864250

Dourson, M. L., Teuschler, L. K., Durkin, P. R. et al. (1997). Categorical regression of toxicity data: A case study using aldicarb. *Regul Toxicol Pharmacol 25*, 121-129. doi:10.1006/rtph.1996.1079

ECHA (2013). *Guidance for Human Health Risk Assessment* (Vol. 3, Part B). https://echa.europa.eu/documents/10162/23492134/biocides_guidance_vol_iii_part_b_v10_superseded_en.pdf/8ce06b02-2a0b-a348-7a44-162a8c83e633

ECHA (2017). The use of alternatives to testing on animals for the REACH Regulation: Third report under Article 117(3) of the REACH Regulation. https://echa.europa.eu/documents/10162/13639/alternatives_test_animals_2017_en.pdf/075c690d-054c-693a-c921-f8cd8acbe9c3 (accessed 15.06.2020).

ECHA – European Chemicals Agency (2020). OECD and EU test guidelines. https://echa.europa.eu/support/oecd-eu-test-guidelines (accessed 15.06.2020).

EFSA – European Food Safety Authority (2010). Application of systematic review methodology to food and feed safety assessments to support decision making. *EFSA J 8*, 1637. doi:10.2903/j.efsa.2010.1637

EFSA (2014). Guidance on expert knowledge elicitation in food and feed safety risk assessment. *EFSA J 12*, 1831-4732. doi:10.2903/j.efsa.2014.3734

EFSA (2015a). Principles and process for dealing with data and evidence in scientific assessments. *EFSA J 13*, 4121-4136. doi:10.2903/j.efsa.2015.4121

EFSA (2015b). Conclusion on the peer review of the pesticide risk assessment of the active substance glyphosate. *EFSA J 13*, 4302. doi:10.2903/j.efsa.2015.4302

EPA – US Environmental Protection Agency (1994). *Methods for Derivation of Inhalation Reference Concentrations (RfCs) and Application of Inhalation Dosimetry*. Washington, DC, USA: US EPA. https://www.epa.gov/risk/methods-derivation-inhalation-reference-concentrations-and-application-inhalation-dosimetry

EPA (1995). *The Use of the Benchmark Dose Approach in Health Risk Assessment*. Washington, DC, USA: US EPA. http://hero.epa.gov/index.cfm/reference/download/reference_id/5992

EPA (2011). *Recommended Use of Body Weight 3/4 as the Default Method in Derivation of the Oral Reference Dose*. Washington, DC, USA: US EPA. https://www.epa.gov/risk/recommended-use-body-weight-34-default-method-derivation-oral-reference-dose

EPA (2012). Benchmark Dose Technical Guidance. https://www.epa.gov/sites/production/files/2015-01/documents/benchmark_dose_guidance.pdf

EPA (2017). Categorical Regression (CATREG) User Guide Version 3.1.0.7. https://www.epa.gov/sites/default/files/2016-03/documents/catreg_user_guide.pdf

EPA (2018). Basic Information about the Integrated Risk Information System. https://www.epa.gov/iris/basic-information-about-integrated-risk-information-system

EPA (2019a). List of Alternative Test Methods and Strategies (or New Approach Methodologies [NAMs]). https://www.epa.gov/sites/production/files/2019-12/documents/alternative_testing_nams_list_first_update_final.pdf

EPA (2019b). First Annual Conference on the State of the Science on Development and Use of New Approach Methods (NAMs) for Chemical Safety Testing. https://www.epa.gov/chemical-research/first-annual-conference-state-science-development-and-use-new-approach-methods-0 (accessed 15.06.2020)

Farhat, N., Saunders-Hastings, P., Morgan, R. et al. (2022). Best practices in systematic review. *ALTEX 39*, 463-479. doi:10.14573/altex.2004111

Farrell, P. J., Milton, B., Ramoju, S. et al. (2022). The use of categorical regression in evidence integration. *ALTEX 39*, 656-666. doi:10.14573/altex.2012022

Farrugia, P., Petrisor, B. A., Farrokhyar, F. et al. (2010). Practical tips for surgical research: Research questions, hypotheses and objectives. *Can J Surg 53*, 278-281.

Fehringer, G., Brenner, D. R., Zhang, Z. F. et al. (2017). Alcohol and lung cancer risk among never smokers: A pooled analysis from the international lung cancer consortium and the SYNERGY study. *Int J Cancer 140*, 1976-1984. doi:10.1002/ijc.30618

Felix, A. S., Gaudet, M. M., La Vecchia, C. et al. (2015). Intrauterine devices and endometrial cancer risk: A pooled analysis of the epidemiology of endometrial cancer consortium. *Int J Cancer 136*, E410-422. doi:10.1002/ijc.29229

Field, R. W., Krewski, D., Lubin, J. H. et al. (2006). An overview of the North American residential radon and lung cancer case-control studies. *J Toxicol Environ Health A 69*, 599-631. doi:10.1080/15287390500260960

Friedenreich, C. M. (1993). Methods for pooled analyses of epidemiologic studies. *Epidemiology 4*, 295-302. doi:10.1097/00001648-199307000-00004

Gannon A. M., Moreau, M., Farmahin, R. et al. (2019). Hexabromocyclododecane (HBCD): A case study applying tiered testing for human health risk assessment. *Food Chem Toxicol 131*, 110581. doi:10.1016/j.fct.2019.110581

Gaudet, M. M., Olshan, A. F., Chuang, S. C. et al. (2010). Body mass index and risk of head and neck cancer in a pooled analysis of case-control studies in the international head and neck cancer epidemiology (INHANCE) consortium. *Int J Epidemiol 39*, 1091-1102. doi:10.1093/ije/dyp380

Gift, J. S., McGaughy, R., Singh, D. V. et al. (2008). Health as-

sessment of phosgene: Approaches for derivation of reference concentration. *Regul Toxicol Pharmacol 51*, 98-107. doi:10.1016/j.yrtph.2008.03.004

Guyatt, G. H., Oxman, A. D., Kunz, R. et al. (2008). Going from evidence to recommendations. *BMJ 336*, 1049-1051. doi:10.1136/bmj.39493.646875.AE

Guyatt, G. H., Oxman, A. D., Kunz, R. et al. (2011a). GRADE guidelines: 2. Framing the question and deciding on important outcomes. *J Clin Epidemiol 64*, 395-400. doi:10.1016/j.jclinepi.2010.09.012

Guyatt, G. H., Oxman, A. D., Sultan, S. et al. (2011b). GRADE guidelines: 9. Rating up the quality of evidence. *J Clin Epidemiol 64*, 1311-1316. doi:10.1016/j.jclinepi.2011.06.004

Guyatt, G. H., Oxman, A. D., Kunz, R. et al. (2011c). GRADE guidelines: 8. Rating the quality of evidence – Indirectness. *J Clin Epidemiol 64*, 1303-1310. doi:10.1016/j.jclinepi.2011.04.014

Haber, L., Strickland, J. and Guth, D. J. (2001). *Categorical Regression Analysis of Toxicity Data* (Vol. 7). https://www.tera.org/Publications/catreg2001.pdf

Hagiwara, S., Paoli, G., Price, P. S. et al. (2022). A value of information framework for assessing the trade-offs associated with uncertainty, duration, and cost of chemical toxicity testing. *Risk Anal*, online ahead of print. doi:10.1111/risa.13931

Hattis, D., Baird, S. and Goble, R. (2002). A straw man proposal for a quantitative definition of the RfD. *Drug Chem Toxicol 25*, 403-436. doi:10.1081/dct-120014793

HEI Diesel Epidemiology Panel (2015). *Diesel Emissions and Lung Cancer: An Evaluation of Recent Epidemiological Evidence for Quantitative Risk Assessment*. Boston, MA, USA: Health Effects Institute. https://www.healtheffects.org/publication/diesel-emissions-and-lung-cancer-evaluation-recent-epidemiological-evidence-quantitative

Hertzberg, R. C. and Miller, M. (1985). A statistical model for species extrapolation using categorical response data. *Toxicol Ind Health 1*, 43-57. doi:10.1177/074823378500100405

Hooijmans, C. R., de Vries, R. B. M., Ritskes-Hoitinga, M. et al. (2018). Facilitating healthcare decisions by assessing the certainty in the evidence from preclinical animal studies. *PLoS One 13*, e0187271. doi:10.1371/journal.pone.0187271

IARC (1978). IARC monographs on the evaluation of the carcinogenic risk of chemicals to humans: Some N-nitroso compounds. *IARC Monogr Eval Carcinog Risk Chem Man 17*, 1-349.

IARC (2019). *Preamble to the IARC Monographs on the Identification of Carcinogenic Hazards to Humans*. Lyon, France: International Agency for Research on Cancer.

Jaworska, J., Gabbert, S. and Aldenberg, T. (2010). Towards optimization of chemical testing under REACH: A Bayesian network approach to integrated testing strategies. *Regul Toxicol Pharmacol 57*, 157-167. doi:10.1016/j.yrtph.2010.02.003

Johnson, P. I., Sutton, P., Atchley, D. S. et al. (2014). The navigation guide – Evidence-based medicine meets environmental health: Systematic review of human evidence for PFOA effects on fetal growth. *Environ Health Perspect 122*, 1028-

1039. doi:10.1289/ehp.1307893

Johnson, P. I., Koustas, E., Vesterinen, H. M. et al. (2016). Application of the navigation guide systematic review methodology to the evidence for developmental and reproductive toxicity of triclosan. *Environ Int 92-93*, 716-728. doi:10.1016/j.envint.2016.03.009

Kheifets, L., Ahlbom, A., Crespi, C. M. et al. (2010). Pooled analysis of recent studies on magnetic fields and childhood leukaemia. *Br J Cancer 103*, 1128-1135. doi:10.1038/sj.bjc.6605838

Koustas, E., Lam, J., Sutton, P. et al. (2014). The navigation guide – Evidence-based medicine meets environmental health: Systematic review of nonhuman evidence for PFOA effects on fetal growth. *Environ Health Perspect 122*, 1015-1027. doi:10.1289/ehp.1307177

Krewski, D., Lubin, J. H., Zielinski, J. M. et al. (2006). A combined analysis of North American case-control studies of residential radon and lung cancer. *J Toxicol Environ Health A 69*, 533-597. doi:10.1080/15287390500260945

Krewski, D., Chambers, A., Stern, B. R. et al. (2010). Development of a copper database for exposure-response analysis. *J Toxicol Environ Health A 73*, 208-216. doi:10.1080/15287390903340815

Krewski, D., Westphal, M., Andersen, M. E. et al. (2014). A framework for the next generation of risk science. *Environ Health Perspect 122*, 796-805. doi:10.1289/ehp.1307260

Krewski, D., Barakat-Haddad, C., Donnan, J. et al. (2017). Determinants of neurological disease: Synthesis of systematic reviews. *Neurotoxicology 61*, 266-289. doi:10.1016/j.neuro.2017.04.002

Krewski, D., Anderson, M., Tyshenko, M. et al. (2020). Toxicity testing in the 21st century: Progress in the past decade and future perspectives. *Arch Toxicol 94*, 1-58. doi:10.1007/s00204-019-02613-4

Lam, J., Koustas, E., Sutton, P. et al. (2014). The navigation guide – Evidence-based medicine meets environmental health: Integration of animal and human evidence for PFOA effects on fetal growth. *Environ Health Perspect 122*, 1040-1051. doi:10.1289/ehp.1307923

Lam, J., Sutton, P., Kalkbrenner, A. et al. (2016). A systematic review and meta-analysis of multiple airborne pollutants and autism spectrum disorder. *PLoS One 11*, e0161851. doi:10.1371/journal.pone.0161851

Lam, J., Lanphear, B. P., Bellinger, D. et al. (2017). Developmental PBDE exposure and IQ/ADHD in childhood: A systematic review and meta-analysis. *Environ Health Perspect 125*, 086001. doi:10.1289/ehp1632

Lehman, A. J. and Fitzhugh, O. G. (1954). 100-fold margin of safety. *Quarterly Bulletin of Food and Drug Officials 18*, 33-35. https://hero.epa.gov/hero/index.cfm/reference/details/reference_id/3195

Lucas, R. M. and McMichael, A. J. (2005). Association or causation: Evaluating links between "environment and disease". *Bull World Health Organ 83*, 792-795.

Makowski, D., Albert, I., Bonvallot, N. et al. (2016). Opinion of the French agency for food, environmental and occupation-

al health and safety regarding the progress report on the assessment of the weight of evidence at ANSES: Critical literature review and recommendations at the hazard identification stage. *ANSES Opinion Request, No 2015-SA-0089*.

Martin, P., Bladdier, C., Meek, B. et al. (2018). Weight of evidence for hazard identification: A critical review of the literature. *Environ Health Perspect 126*, 1-15. doi:10.1289/EHP3067

Masic, I., Miokovic, M. and Muhamedagic, B. (2008). Evidence based medicine – New approaches and challenges. *Acta Inform Med 16*, 219-225. doi:10.5455/aim.2008.16.219-225

Menon, J., Struijs, F. and Whaley, P. (2022). The methodological rigour of systematic reviews in environmental health. *Cri Rev Toxicol 52*, 167-187. doi:10.1080/10408444.2022.2082917

Milton, B., Farrell, P. J., Birkett, N. et al. (2017a). Modeling U-shaped exposure-response relationships for agents that demonstrate toxicity due to both excess and deficiency. *Risk Anal 37*, 265-279. doi:10.1111/risa.12603

Milton, B., Krewski, D., Mattison, D. R. et al. (2017b). Modeling U-shaped dose-response curves for manganese using categorical regression. *Neurotoxicology 58*, 217-225. doi:10.1016/j.neuro.2016.10.001

Moher, D., Liberati, A., Tetzlaff, J. et al. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Med 6*, e1000097. doi:10.1371/journal.pmed.1000097

Morgan, R. L., Whaley, P., Thayer, K. A. et al. (2018). Identifying the PECO: A framework for formulating good questions to explore the association of environmental and other exposures with health outcomes. *Environ Int 121*, 1027-1031. doi:10.1016/j.envint.2018.07.015

Morgan, R. L., Thayer, K. A., Santesso, N. et al. (2019). A risk of bias instrument for non-randomized studies of exposures: A users' guide to its application in the context of GRADE. *Environ Int 122*, 168-184. doi:10.1016/j.envint.2018.11.004

Munafo, M., Nosek, B., Bishop, D. et al. (2017). A manifesto for reproducible science. *Nat Hum Behav 1*, 0021. doi:10.1038/s41562-016-0021

Murad, M. H., Asi, N., Alsawas, M. et al. (2016). New evidence pyramid. *Evid Based Med 21*, 125-127. doi:10.1136/ebmed-2016-110401

NAS – National Academies of Sciences (2001). *Standard Operating Procedures for Developing Acute Exposure Guideline Levels*. Washington, DC, USA: National Academies Press. https://www.nap.edu/catalog/10122/standing-operating-procedures-for-developing-acute-exposure-guideline-levels-for-hazardous-chemicals

NASEM – National Academies of Sciences Engineering and Medicine (2017). *Guiding Principles for Developing Dietary Reference Intakes Based on Chronic Diseases*. Washington, DC, USA: NASEM. http://www.nationalacademies.org/hmd/Reports/2017/guiding-principles-for-developing-dietary-reference-intakes-based-on-chronic-disease.aspx

NASEM (2018). *Progress Toward Transforming the Integrated Risk Information System (IRIS) Program: A 2018 Evalua-tion*. Washington, DC, USA: The National Academies Press. https://www.nap.edu/catalog/25086/progress-toward-trans-forming-the-integrated-risk-information-system-iris-program

NRC – National Research Council (1994). *Science and Judgement in Risk Assessment*. Washington, DC, USA: National Academies Press. doi:10.17226/2125

NRC (2007). *Toxicity Testing in the 21st Century: A Vision and Strategy*. https://www.nap.edu/catalog/11970/toxicity-testing-in-the-21st-century-a-vision-and-a

NRC (2008). *Phthalates and Cumulative Risk Assessment: The Task Ahead*. Washington, DC, USA: National Academies Press. https://www.ncbi.nlm.nih.gov/pubmed/25009926

NRC (2009). *Science and Decisions: Advancing Risk Assessment*. Washington, DC, USA: National Academies Press. https://www.nap.edu/catalog/12209/science-and-decisions-advancing-risk-assessment

NRC (2011). *Review of the Environmental Protection Agency's Draft IRIS Assessment of Formaldehyde*. Washington, DC, USA: The National Academies Press. https://www.nap.edu/catalog/13142/review-of-the-environmental-protection-agen-cys-draft-iris-assessment-of-formaldehyde

NRC (2014a). *Review of EPA's Integrated Risk Information System (IRIS) Process*. Washington, DC, USA: National Academics Press. https://www.nap.edu/catalog/18764/review-of-epas-integrated-risk-information-system-iris-process

NRC (2014b). *Review of the Environmental Protection Agency's State-of-the-science Evaluation of Nonmonotonic Dose-response Relationships as They Apply to Endocrine Disruptors*. Washington, DC, USA: National Academies Press. https://bit.ly/3LA4GZZ

OECD (2019a). Case Study on the Use of an Integrated Approach to Testing and Assessment for Estrogen Receptor Active Chemicals. *Series on Testing and Assessment, No. 309*. https://www.oecd.org/officialdocuments/publicdisplaydocu-mentpdf/?cote=ENV/JM/MONO(2019)28&docLanguage=en

OECD (2019b). Guiding Principles and Key Elements for Establishing a Weight of Evidence for Chemical Assessment. *Series on Testing and Assessment, No. 311*. Environment, Health and Safety Division, Environment Directorate. https://www.oecd.org/chemicalsafety/risk-assessment/guiding-principles-and-key-elements-for-establishing-a-weight-of-evidence-for-chemical-assessment.pdf

Paoli, G., Momoli, F., Tyshenko, M. G. et al. (2022). Problem formulation for EFSA Scientific Assessments. *EFSA Supporting Publications 19*, 7349E. doi:10.2903/sp.efsa.2022.EN-7349

Paul Friedman K., Gagne, M., Loo, L. H. et al. (2020). Utility of in vitro bioactivity as a lower bound estimate of in vivo adverse effect levels and in risk-based prioritization. *Toxicol Sci 173*, 202-225. doi:10.1093/toxsci/kfz201

Peres, L. C., Risch, H., Terry, K. L. et al. (2018). Racial/ethnic differences in the epidemiology of ovarian cancer: A pooled analysis of 12 case-control studies. *Int J Epidemiol 47*, 1011. doi:10.1093/ije/dyy054

Perkins, E. J., Ashauer, R., Burgoon, L. et al. (2019a). Build-

ing and applying quantitative adverse outcome pathway models for chemical hazard and risk assessment. *Environ Toxicol Chem 38*, 1850-1865. doi:10.1002/etc.4505

Perkins, E. J., Gayen, K., Shoemaker, J. E. et al. (2019b). Chemical hazard prediction and hypothesis testing using quantitative adverse outcome pathways. *ALTEX 36*, 91-102. doi:10.14573/altex.1808241

Petrisor, B. and Bhandari, M. (2007). The hierarchy of evidence: Levels and grades of recommendation. *Indian J Orthop 41*, 11-15. doi:10.4103/0019-5413.30519

Price, P. S., Jarabek, A. M., and Burgoon, L. D. (2020). Organizing mechanism-related information on chemical interactions using a framework based on the aggregate exposure and adverse outcome pathways. *Environ Int 138*, 105673. doi:10.1016/j.envint.2020.105673

Price, P. S., Hubbell, B., Hagiwara, S. et al. (2021). A framework that considers the impacts of time, cost, and uncertainty in the determination of the cost effectiveness of toxicity-testing methodologies. *Risk Anal 42*, 707-729. doi:10.1111/risa.13810

Rathman, J. F., Yang, C. and Zhou, H. (2018). Dempster-Shafer theory for combining in silico evidence and estimating uncertainty in chemical risk assessment. *Comput Toxicol 6*, 16-31. doi:10.1016/j.comtox.2018.03.001

Rhomberg, L. R., Goodman, J. E., Bailey, L. A. et al. (2013). A survey of frameworks for best practices in weight-of-evidence analyses. *Crit Rev Toxicol 43*, 753-784. doi:10.3109/10408444.2013.832727

Rogiers, V., Benfenati, E., Bernauer, U. et al. (2020). The way forward for assessing the human health safety of cosmetics in the EU – Workshop proceedings. *Toxicology 436*, 152421. doi:10.1016/j.tox.2020.152421

Rooney, A. A., Boyles, A. L., Wolfe, M. S. et al. (2014). Systematic review and evidence integration for literature-based environmental health science assessments. *Environ Health Perspect 122*, 711-718. doi:10.1289/ehp.1307972

Sackett, D. L. (1997). Evidence-based medicine. *Semin Perinatol 21*, 3-5. doi:10.1016/s0146-0005(97)80013-4

Samet, J. M., Chiu, W. A., Cogliano, V. et al. (2019). The IARC Monographs: Updated procedures for modern and transparent evidence synthesis in cancer hazard identification. *J Natl Cancer Inst 112*, 30-37. doi:10.1093/jnci/djz169

Sand, S., Portier, C. J. and Krewski, D. (2011). A signal-to-noise crossover dose as the point of departure for health risk assessment. *Environ Health Perspect 119*, 1766-1774. doi:10.1289/ehp.1003327

Sand, S., Lindqvist, R., von Rosen, D. et al. (2018). Dose-related severity sequence, and risk-based integration, of chemically induced health effects. *Toxicol Sci 165*, 74-89. doi:10.1093/toxsci/kfy124

Sand, S. (2022). A novel method for combining outcomes with different severities. *ALTEX 39*, 480-497. doi:10.14573/altex.2004212

Schroeter, J. D., Kimbell, J. S., Gross, E. A. et al. (2008). Application of physiological computational fluid dynamics models to predict interspecies nasal dosimetry of inhaled acrolein. *Inhal Toxicol 20*, 227-243. doi:10.1080/08958370701864235

Schünemann, H. J., Oxman, A. D., Brozek, J. et al. (2008). Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. *BMJ 336*, 1106-1110. doi:10.1136/bmj.39500.677199.AE

Schünemann, H. J., Cuello, C., Akl, E. A. et al. (2019a). GRADE guidelines: 18. How ROBINS-I and other tools to assess risk of bias in nonrandomized studies should be used to rate the certainty of a body of evidence. *J Clin Epidemiol 111*, 105-114. doi:10.1016/j.jclinepi.2018.01.012

Schünemann, H. J., Zhang, Y. and Oxman, A. D. (2019b). Distinguishing opinion from evidence in guidelines. *BMJ 366*, l4606. doi:10.1136/bmj.l4606

Shamseer, L., Moher, D., Clarke, M. et al. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: Elaboration and explanation. *BMJ 350*, g7647. doi:10.1136/bmj.g7647

Smith, M. T., Guyton, K. Z., Gibbons, C. F. et al. (2016). Key characteristics of carcinogens as a basis for organizing data on mechanisms of carcinogenesis. *Environ Health Perspect 124*, 713-721. doi:10.1289/ehp.1509912

Stephens, M. L., Andersen, M., Becker, R. A. et al. (2013). Evidence-based toxicology for the 21st century: Opportunities and challenges. *ALTEX 30*, 74-103. doi:10.14573/altex.2013.1.074

Stern, B. R., Solioz, M., Krewski, D. et al. (2007). Copper and human health: Biochemistry, genetics, and strategies for modeling dose-response relationships. *J Toxicol Environ Health B Crit Rev 10*, 157-222. doi:10.1080/10937400600755911

Taher, M. K., Farhat, N., Karyakina, N. A. et al. (2019). Critical review of the association between perineal use of talc powder and risk of ovarian cancer. *Reprod Toxicol 90*, 88-101. doi:10.1016/j.reprotox.2019.08.015

Teeguarden, J. G., Bogdanffy, M. S., Covington, T. R. et al. (2008). A PBPK model for evaluating the impact of aldehyde dehydrogenase polymorphisms on comparative rat and human nasal tissue acetaldehyde dosimetry. *Inhal Toxicol 20*, 375-390. doi:10.1080/08958370801903750

Thomas, D. C., Jerrett, M., Kuenzli, N. et al. (2007). Bayesian model averaging in time-series studies of air pollution and mortality. *J Toxicol Environ Health A 70*, 311-315. doi:10.1080/15287390600884941

Thomas, R. S., Clewell, 3rd, H. J., Allen, B. C. et al. (2012). Integrating pathway-based transcriptomic data into quantitative chemical risk assessment: A five chemical case study. *Mutat Res 746*, 135-143.

Thomas, R. S., Philbert, M. A., Auerbach, S. S. et al. (2013). Incorporating new technologies into toxicity testing and risk assessment: Moving from 21st century vision to a data-driven framework. *Toxicol Sci 136*, 4-18. doi:10.1093/toxsci/kft178

Thomas, R. S., Bahadori, T., Buckley, T. J. et al. (2019). The next generation blueprint of computational toxicology at the U.S. Environmental Protection Agency. *Toxicol Sci 169*, 317-332. doi:10.1093/toxsci/kfz058

Tobias, A., Saez, M. and Kogevinas, M. (2004). Meta-analysis of results and individual patient data in epidemiological studies. *J Mod Appl Stat Meth 3*, 176-185. https://pdfs.semanticscholar.org/819f/aa4d45cbdacc501d97293e1a4914c3a175c1.pdf

Vermeulen, R., Silverman, D. T., Garshick, E. et al. (2014). Exposure-response estimates for diesel engine exhaust and lung cancer mortality based on data from three occupational cohorts. *Environ Health Perspect 122*, 172-177. doi:10.1289/ehp.1306880

Vesterinen, H. M., Johnson, P. I., Atchley, D. S. et al. (2015). Fetal growth and maternal glomerular filtration rate: A systematic review. *J Matern Fetal Neonatal Med 28*, 2176-2181. doi:10.3109/14767058.2014.980809

Wang, M. D., Gomes, J., Cashman, N. R. et al. (2014). A meta-analysis of observational studies of the association between chronic occupational exposure to lead and amyotrophic lateral sclerosis. *J Occup Environ Med 56*, 1235-1242. doi:10.1097/jom.0000000000000323

Webster, F., Gagné, M., Patlewicz, G. et al. (2019). Predicting estrogen receptor activation by a group of substituted phenols: An integrated approach to testing and assessment case study. *Regul Toxicol Pharmacol 106*, 278-291. doi:10.1016/j.yrtph.2019.05.017

West, G. B. (1999). The fourth dimension of life: Fractal geometry and allometric scaling of organisms. *Science 284*, 1677-1679. doi:10.1126/science.284.5420.1677

Whaley, P. (2015). Improving the evaluation of evidence: A framework for systematic review and integrated assessments. https://www.researchgate.net/publication/279537641 (accessed 31.07.2022).

WHO/IPCS (2009). *Principles for Modelling Dose-Response for the Risk Assessment of Chemicals*. Geneva, Switzerland: World Health Organization International Program on Chemical Safety. https://apps.who.int/iris/handle/10665/43940

WHO/IPCS (2014). *Guidance Document on Evaluating and Expressing Uncertainty in Hazard Characterization*. Geneva, Switzerland: World Health Organization International Program on Chemical Safety. https://apps.who.int/iris/handle/10665/259858

Wigle, D. T., Arbuckle, T. E., Turner, M. C. et al. (2008). Epidemiologic evidence of relationships between reproductive and child health outcomes and environmental chemical contaminants. *J Toxicol Environ Health B Crit Rev 11*, 373-517.

doi:10.1080/10937400801921320

Woodruff, T. J. and Sutton, P. (2011). An evidence-based medicine methodology to bridge the gap between clinical and environmental health sciences. *Health Aff (Millwood) 30*, 931-937. doi:10.1377/hlthaff.2010.1219

Woodruff, T. J. and Sutton, P. (2014). The navigation guide systematic review methodology: A rigorous and transparent method for translating environmental health science into better health outcomes. *Environ Health Perspect 122*, 1007-1014. doi:10.1289/ehp.1307175

Wyss, A., Hashibe, M., Chuang, S. C. et al. (2013). Cigarette, cigar, and pipe smoking and the risk of head and neck cancers: Pooled analysis in the international head and neck cancer epidemiology consortium. *Am J Epidemiol 178*, 679-690. doi:10.1093/aje/kwt029

Yetley, E. A., MacFarlane, A. J., Greene-Finestone, L. S. et al. (2017). Options for basing dietary reference intakes (DRIs) on chronic disease endpoints: Report from a joint US-/Canadian-sponsored working group. *Am J Clin Nutr 105*, 249s-285s. doi:10.3945/ajcn.116.139097