Krewski et al.*:*

Development of an Evidence-Based Risk Assessment Framework

Supplementary Data

Supplementary Material I

Best Practices in Weight of Evidence Analysis: A Review

Patrick Saunders-Hastings[1], Lorenz Rhomberg[2] and Daniel Krewski[1]
[1]McLaughlin Centre for Population Health Risk Assessment, University of Ottawa, Ottawa, Canada; [2]Gradient, Cambridge, USA

Correspondence: Patrick Saunders-Hastings, PhD
McLaughlin Centre for Population Health Risk Assessment, University of Ottawa
600 Peter Morand Crescent, Room 216B, Ottawa, ON Canada K1G 5Z3
(patrick.saundershastings@gmail.com)

1 Introduction

The US Environmental Protection Agency (US EPA) first introduced the concept of weight of evidence (WoE) in 1986 as a component of risk assessment for carcinogenic effects (EPA, 1986). Since then, regulatory agencies have incorporated a WoE (also known as 'evidence integration') approach in evaluating and quantifying chemical risks on human health. Though it is currently adopted in many disciplines, there is no clear definition as to what constitutes a WoE analysis (Linkovet al. 2009; Weed, 2005). Further, WoE approaches have been found to differ significantly, with frequent reliance on subjective expert judgement (Linkov, 2015; Lutter et al., 2015; Suter et al., 2017).

WoE can be defined as "a framework for synthesizing individual lines of evidence, using methods that are either qualitative (examining distinguishing attributes) or quantitative (measuring aspects in terms of magnitude) to develop conclusions regarding questions concerned with the degree of impairment or risk" (Lutter et al., 2015, p. 189). This supports integration of insights from human, animal, and mechanistic data, including data generated from new approach methodologies (Andersen et al., 2019), broadening the types of data that can be leveraged to inform risk decision-making: Making use of all available evidence from multiple evidence streams will support efforts to reduce the number of animals used in individual experiments, replace animal testing with other approaches, and refine animal testing procedures.

In a report from the National Research Council (NRC, 2014), the reviewing committee concluded that the current use of the phrase *weight of evidence* is too vague and provides limited scientific value. It has also been reported that the terms *weight of evidence* and *systematic review* are sometimes used interchangeably, despite different intended meanings (Buist et al., 2013; NRC, 2014).

Reviews of the literature on WoE methodologies recommend a structured and well-defined approach. In a systematic review that included 92 papers on "weight of evidence" to characterize the concept, Weed (2005) found that the phrase had multiple definitions and applications with a lack of consensus about the associated methods. In particular, the author noted that the concept was used in three ways: (1) metaphorically, with no description of methods; (2) methodologically, based on familiar methods such as meta-analysis or causal criteria; and (3) theoretically, as a label for a conceptual framework.

While the specific approaches to WoE evaluations tend to vary, methodologies generally consist of summarizing, synthesizing, and interpreting a body of evidence to make conclusions. The basic steps can be summarized as (Suter et al., 2017):

(1) Assemble of evidence: relevant information is systematically identified, screened, evaluated, and summarized
(2) Assign weight to the evidence: the relevance, reliability and strength is evaluated, and a score is assigned to each type of evidence
(3) Weighing of the body of evidence: the weighted evidence is integrated and then interpreted with respect to the hypothesis

doi:10.14573/altex.2004041s

The number of stages in a WoE evaluation differs between frameworks, with varying levels of detail. For example, Rhomberg et al. (2013) define four phases with specific features for each phase in a review evaluating 50 existing WoE frameworks from regulatory agencies and other sources between 2010 and 2012. The authors identified the key characteristics of frameworks used in assessing chemical risks on human health, dividing the WoE analysis into four phases: 1) define a causal question and develop criteria for study selection, 2) develop and apply criteria for review of individual studies, 3) integrate and evaluate evidence, and 4) draw conclusions based on inferences. The processes outlined both by Suter and colleagues (2017) and Rhomberg and colleagues (2013) demonstrate the overlap between systematic review and WoE processes, as both outline steps for the identification or acquisition of evidence that could be considered as part of a systematic review.

In addition to the absence of a well-defined framework for WoE, there is also insufficient guidance on how best to conduct each stage of the process. In a recent review of nine regulatory frameworks in chemical risk assessment in the EU, none of the frameworks were found to provide sufficient guidance to carry out the evaluation (Agerstrand and Beronius, 2016). The authors reported that there was a lack of guidance on how to carry out WoE evaluations, highlighting the need for a more structured approach. Moreover, Buist et al. (2013) note that the lack of guidance may explain the lack of consensus regarding the many approaches used in WoE evaluations. They note that to improve the robustness, reproducibility and transparency of WoE evaluations, clear guidance is needed.

Herein, the authors seek to address this knowledge gap. While intended as an independent publication to provide a scoping review of existing WoE frameworks, this study is part of a series of related publications (collected at doi:10.14573/altex.22S2) associated with a workshop of international experts, held in Ottawa, Canada in December 2018 to discuss the theoretical underpinnings, methodological approaches, and applications of evidence integration frameworks. This effort is expected to contribute to the promotion and advancement of the inclusion of non-human, non-animal research findings in scientific assessments. Seeking to avoid duplication while recognizing that this is a rapidly-evolving body of knowledge, authors chose to replicate the approach of Rhomberg and colleagues (2013), updating their review with publications from the past five years. The intent of this article is to establish the most current understanding of WoE approaches, providing a foundation to be built upon over subsequent case studies examining the application of WoE principles and best practices.

2 Methods

2.1 Overview

The aim of this study was to identify the most relevant frameworks and best practices related to WoE, and not to conduct a full systematic review. As such, while the review methodology was developed in keeping with the PRISMA guidelines for systematic reviews (Moher et al., 2009), double-blind reviewer screening and a complete systematic literature search were not conducted. Briefly, authors conducted a survey of the published literature for articles presenting, comparing or assessing WoE frameworks relevant to human health that had been published since the execution of the search strategy in the review by Rhomberg and colleagues (2013).

2.2 Search strategy

A literature search was conducted on March 27, 2018, using PubMed (all dates), with no language restrictions. The strategy searched for the term “weight of evidence”, replicating the database and search terms used in the Rhomberg review. The search excluded articles published before June 1, 2012, as the previous review included all PubMed articles between 2010 and May 2012. A Google Scholar search for articles that cited the Rhomberg review was also conducted, while reference lists of retained articles were searched by hand for additional articles.

The aim of the search strategy mirrors that of Rhomberg and colleagues, wherein the objective is not to obtain every instance of WoE assessment, but rather to compile a representative sample of frameworks in order to build understanding of the diversity, best practices and persisting challenges in WoE assessment.

2.3 Eligibility criteria and study inclusion

Articles were imported into Endnote X7.5^{TM} and subjected to title and abstract review by a single reviewer. Articles were included unless they clearly met one of the exclusion criteria presented in Table S1 (i.e., uncertain or unclear cases were advanced to full review). Full texts were sought for articles that were retained, which were subjected to a round of “full text” assessment by a single reviewer, using the same criteria applied during title and abstract review.

Tab. S1: Exclusion criteria for articles relevant to WoE approaches

Domain	Criterion	Rationale
Population	Human populations are not considered	WoE approaches for environmental or invertebrate approaches may not be representative.
Intervention	None	Any intervention was considered.
Comparator	None	NA
Outcome	Health is not considered	Any health outcome was considered, but non-health contexts may not be representative.
Time	The article only presents frameworks that would have been captured in the Rhomberg et al., 2013 review	Authors sought to avoid duplication of past research.
Setting/Study Design	The article only presents a WoE assessment/application (case study) and no discussion of WoE methodologies or frameworks	Due to issues of feasibility and limited informative value, case studies that did not include methodological or best practice discussions were excluded.

2.4 Quality assessment

As no quantitative data synthesis was conducted, and in the absence of a reliable quality assessment tool for the purposes of this study, no assessment of *study* quality was conducted. Rather, the strengths and limitations of the *frameworks* are discussed qualitatively in subsequent sections.

3 Results

Following deduplication, a total of 653 articles were subjected to title and abstract review, with 82 being retained for full review. Of these, 25 articles met all eligibility criteria and were retained for data extraction and analysis. The study selection process is summarized in Figure S1.

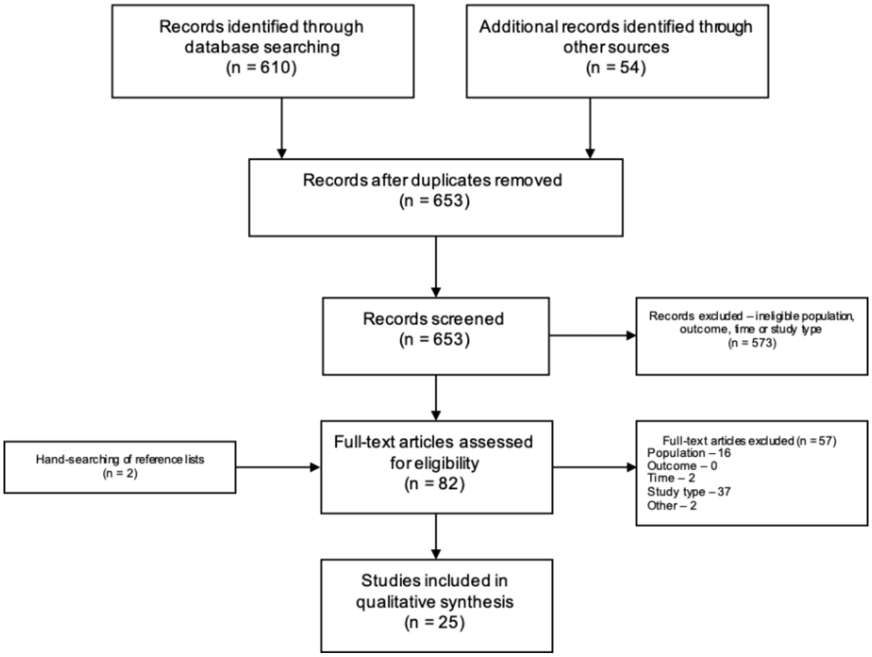
Records identified through database searching (n = 610)
Additional records identified through other sources (n = 54)
Records after duplicates removed (n = 653)
Records screened (n = 653)
Records excluded – ineligible population, outcome, time or study type (n = 573)
Hand-searching of reference lists (n = 2)
Full-text articles assessed for eligibility (n = 82)
Full-text articles excluded (n = 57)
Population – 16
Outcome – 0
Time – 2
Study type – 37
Other – 2
Studies included in qualitative synthesis (n = 25)

Fig. S1: Study selection flow diagram

3.1 Included studies

In total, 25 articles were included for data extraction. Of these, 16 were academic publications (Becker et al., 2015, 2017; Catalan et al., 2017; Collier et al., 2016; Cuddy et al., 2016; Dekant and Bridges, 2016; Dekant et al., 2017; Gross et al., 2017; Gross and Fedak, 2015; Hristozov et al., 2014a,b; Kaltenhauser et al., 2017; Money et al., 2013; Rhomberg, 2015; Sheehan et al., 2018; Vandenberg et al., 2016), eight publications described frameworks affiliated with governments or international bodies (Bridges et al., 2017; Buist et al., 2013; ECHA, 2015a,b; Hardy et al., 2017; Rooney et al., 2014; Tluckiewicz et al., 2013; Vermeire et al., 2013) and one publication was from a non-profit organization (Meek et al., 2013).

Of these 25 publications, 20 WoE frameworks were discussed, as two publications discussed the same quantitative approach (Becker et al., 2015, 2017), three publications examined another quantitative approach (Hristozov et al., 2014a,b; Sheehan et al., 2018) and three publications examined the OSIRIS framework (Buist et al., 2013; Tluckiewicz et al., 2013; Vermeire et al., 2013). These 20 frameworks are categorized into qualitative and quantitative methodologies, and are presented in Section 3.3 and 3.4, respectively. First, however, Section 3.2 provides a summary of the WoE definitions used across the included studies.

3.2 Definition of WoE

A challenge that has been identified in past publications (Buist et al., 2013; NRC, 2014; Rhomberg et al., 2013) is the vague or inconsistent conceptualization of what WoE is meant to entail. In an effort to assess the current state of WoE definition, authors extracted the quoted definition of WoE for each framework (Tab. S2).

A survey of included studies found that definitions were generally consistent across studies, with common elements of the WoE conceptualization including an assessment of "all available" information or data (comprehensiveness), synthesis of different lines of evidence (integration), and an assessment of confidence in the collective body of evidence (weighting).

However, definitions continue to take a vague and general approach to WoE, which may limit their value in informing or comparing approaches. Most problematically, seven of the 20 frameworks (35%) provided no explicit definition of WoE; this risks confusion over the process and value of WoE and can obstruct progress towards shared understanding. For the present paper, we have included within WoE all the components of identifying studies, evaluating studies and their quality, and integrating their results into arguments that gauge the degree of scientific support of an articulated judgment. This recognizes that all the components are essential, even though the last stages of integration and support of judgments are the ones specifically entailing "weighing".

Tab. S2: Summary of WoE definitions provided in included publications

Citation	Definition of WoE
ECHA, 2015a	"A weight of evidence determination means that all available information bearing on the determination of hazard is considered together."
ECHA, 2015b	"A Weight of Evidence assessment involves the consideration of all data that is available and may be relevant to reproductive toxicity."
Cuddy et al., 2016	"In the WOE approach, alternative competing sources of data are compared and integrated to assess the probability of a specific conclusion."
Gross et al., 2017	No explicit definition given.
Money et al., 2013	No explicit definition given.
Rooney et al., 2014	No explicit definition given.

Buist et al., 2013;
Tluckiewicz et al., 2013;
Vermeire et al., 2013
	No explicit definition given.

Bridges et al., 2017	"The identification and objective analysis (using pre-defined, scientifically justified criteria) of all potentially relevant studies, for their quality and relevance in critically testing a hypothesis."
Becker et al., 2015, 2017	"While approaches for conducting WoE evaluations may differ, the essence of all approaches requires considering the collective body of evidence to address the specific questions at hand. The purpose of a WoE evaluation is to document certainty in inferring responses beyond interpolation within the range of empirical observations in a transparent manner"
Catalan et al., 2017	No definition provided.
Collier et al., 2016	"Weight of evidence (WoE) is a term used in multiple disciplines to generally mean a family of approaches to assess multiple lines of evidence in support of (or against) a particular hypothesis."
Dekant and Bridges, 2016	"The identification and objective analysis (using predefined, scientifically justified criteria) of all potentially relevant studies, for their quality and relevance in testing a hypothesis."
Dekant et al., 2017	"A weight of evidence analysis includes definition of the causal question (termed problem formulation by the US EPA), development and application of criteria for review, evaluation and integration of evidence, and conclusions based on inference."
Hristozov et al., 2014a,b; Sheehan et al., 2018	"WoE represents a diverse collection of methods used to synthesise and evaluate individual LOE to form a conclusion."
Gross and Fedak, 2015	"WoE refers to the interpretive methods commonly applied to bodies of literature when conducting hazard and risk assessments."
Kaltenhauser et al., 2017	No explicit definition given.
Meek et al., 2013	No explicit definition given.
Rhomberg et al., 2015	"The application of professional judgment to consider the strengths and weaknesses of individual studies, to compare and contrast their findings, and to try and reconcile or explain inconsistencies so as to arrive at a characterization of what potential toxicological properties are sufficiently supportable to justify the regulatory decisions that will be made."
Vandenberg et al., 2016	No explicit definition given.
Hardy et al., 2017	"Weight of evidence assessment is a process in which evidence is integrated to determine the relative support for possible answers to a scientific question."

3.3 Qualitative WoE

There were 13 qualitative WoE frameworks or processes identified across 13 publications (Catalan et al., 2017; Cuddy et al., 2016; ECHA, 2015a,b; Gross et al., 2017; Gross and Fedak, 2015; Hardy et al., 2017; Kaltenhauser et al., 2017; Meek et al., 2013; Money et al., 2013; Rhomberg, 2015; Rooney et al., 2014; Vandenberg et al., 2016); these are summarized in Table S3.

Across these frameworks, there was a consistent general approach to WoE assessment; while specific steps and approaches varied, the frameworks could generally be organized into five steps: formulate the problem, assemble the evidence, assess individual studies, weigh the body of evidence, and characterize the hazard.

A common set of best practices also began to emerge. These included assembling all available evidence (ECHA, 2015a,b; Gross and Fedak, 2015; Meek et al., 2013); assessing evidence within each line of evidence before integrating findings across lines of evidence (Cuddy et al., 2016; Hardy et al., 2017; Rhomberg, 2015; Rooney et al., 2014; Vandenberg et al., 2016); and weighing evidence based upon reliability (quality), consistency of findings and relevance to human populations (ECHA, 2015a,b; Hardy et al., 2017; Kaltenhauser et al., 2017). Principles of flexibility and transparency were also valued in WoE approaches, and calls for transparency point to the value of a research protocol developed *a priori*, with any subsequent changes documented and justified in final reports (Gross and Fedak, 2015; Hardy et al., 2017; Meek et al., 2013; Rooney et al., 2014; Vandenberg et al., 2016).

The most common limitations were a lack of stepwise guidance to direct an individual in conducting a WoE assessment (especially with respect to integrating different lines of evidence) (ECHA, 2015a,b; Kaltenhauser et al., 2017) and a reliance on subjective guidance (Money et al., 2013; Rhomberg, 2015). Even frameworks that prioritize transparency and objective scientific review note limitations in the reliance on “inherently subjective” expert judgements in the assessment of confidence in a body of evidence (Rooney et al., 2004, p.713*)*. Together, these limitations can impede the reproducibility of WoE assessments and lead to an erosion of public trust from suspicions of arbitrary decision-making. A lack of a clear WoE definition (Gross and Fedak, 2015) and of empirical support for risk categorization (Catalan et al., 2017) may further contribute to such an issue.

3.4 Quantitative WoE

There were 7 quantitative WoE frameworks or processes identified across 12 publications (Becker et al., 2015, 2017; Bridges et al., 2017; Buist et al., 2013; Collier et al., 2016; Dekant and Bridges, 2016; Dekant et al., 2017; Hristozov et al., 2014a,b; Sheehan et al., 2018; Tluckiewicz et al., 2013; Vermeire et al., 2013). These are summarized in Table S4.

While the general approach was similar to that of qualitative frameworks, a notable addition was the articulation of mechanisms of action (MoA) or adverse outcome pathways (AOP) during the formulation of hypotheses (Becker et al., 2015, 2017; Collier et al., 2016; Dekant et al., 2017); this was less clearly expressed in the qualitative frameworks. The other notable addition was the application of a diverse range of statistical methods to arrive at a quantitative estimate of weight of evidence, though it should be noted that in most cases this amounted to the assignment of a quantitative value to a qualitative assessment.

The principles of WoE were similar to those reported in qualitative frameworks, and included a transparent (Becker et al., 2015, 2017; Buist et al., 2013; Dekant and Bridges, 2016; Tluckiewicz et al., 2013; Vermeire et al., 2013) and objective, consistent and reproducible approach (Bridges et al., 2017; Buist et al., 2013; Dekant and Bridges, 2016; Tluckiewicz et al., 2013; Vermeire et al., 2013). However, these frameworks posited that quantitative approaches could more reliably achieve these goals than could qualitative ones. Again, evidence tended to be assessed on the basis of reliability, relevance and validity (Buist et al., 2013; Tluckiewicz et al., 2013; Vermeire et al., 2013), though other frameworks used a similar paradigm targeting biological plausibility, empirical evidence and essentiality (for human studies) or human relevance (for non-human studies) (Becker et al., 2015, 2017; Dekant et al., 2017)

The most common limitations were inadequate documentation or guidance to support execution of a WoE assessment (Becker et al., 2015, 2017; Bridges et al., 2017; Buist et al., 2013; Collier et al., 2016; Hristozov et al., 2014a,b; Sheehan et al., 2018; Tluckiewicz et al., 2013) and that quantitative approaches were time-consuming, complex and challenging; this was especially true for approaches where customized score cards had to be developed following problem formulation (Bridges et al., 2017; Dekant and Bridges, 2016; Dekant et al., 2017).

Tab. S3: Summary of qualitative WoE frameworks

Citation	Scope	Process	Key principles/ Best practices	Challenges/ limitations
Catalan et al., 2017	Mutagenic potential of nanomaterials	1. Determine whether toxicity studies are available (similar, quality, relevance) 1.2. If yes, conduct a mutagenic assessment 2. Classify relevant studies (setting, genotoxic event, outcome direction) 3. Assess consistency of findings across study types 4. Summarize mutagenic potential	• Uses thresholds for categorical assignment • In vivo studies are weighted more heavily than in vitro (and considered separately) • Irreversible damage weighted more heavily than reversible DNA damage	There is a lack of empirical support for the thresholds for categorical assignment
Cuddy et al., 2016	Nanomaterials in consumer products (e.g., sunscreen/ personal care products)	Use three lines of evidence, each comprising multiple analytical techniques: particle size, particle composition and product composition	The use of multiple complimentary techniques and different lines of evidence increases confidence in results	Standard practices for characterizing the physical state of particles in viscous, complex matrices (such as personal care products) are lacking
ECHA, 2015a	All industrial sectors in the EU	Stepwise instructions are lacking.	• Use all available data. • Consider the quality and consistency of data. • Integrate positive and negative results • High-quality data from the same substance should have more weight than data extrapolated from other substances.	• Guidance on development of a protocol is lacking. • Little guidance on how to integrate complex/contradictory data.
ECHA, 2015b	Biocides in the EU	Stepwise instructions are lacking. 1. Clarify objective of the review and establish study selection criteria. 2. Search literature. 3. Selection relevant studies for inclusion	• Evaluate the relevance and reliability of all available evidence. • Consider the severity, adversity and reversibility of effects. • Give more weight to high-quality studies. • Give more weight to studies enabling the identification of a NOAEL. • Weigh effects that are consistent between humans and animals more heavily.	No guidance on how to integrate complex/contradictory data.
Gross and Fedak, 2015	Molecular events related to disease onset associated with exposure to an environmental carcinogen	1. Collect baseline information on the outcome of interest. 2. Conduct literature search to establish molecular landscape 3. Compare molecular landscapes for overlapping exposure/outcome pieces. 4. Rank overlapping pieces of molecular landscape by value of functional protein in disease process.	Transparency in decision-making for expert judgements	Definition of WoE and its practical applications are either unclear or poorly defined
Gross et al., 2017	Endocrine disruptors	1a. Define hypothesis 1b. Select and screen studies 2. Review individual studies (assess, document and justify reliability) 3. Integrate and assess data (examine relationship between mechanistic and adverse effect data)	• Use EATS pathways to define hypotheses • Obtain all available data from open literature and regulatory studies • Use ToxRTool to assess reliability of individual studies • Arrange data according to OECD CF	Systematic review and WoE assessment are conflated.

Citation	Scope	Process	Key principles/ Best practices	Challenges/ limitations
		4. Draw conclusions based on inferences (comment on data sufficiency)	levels. • Use modified Bradford-Hill considerations to assess evidence of causality	
Hardy et al., 2017	Food safety in the EU (scientific assessments for use in all areas under the remit of the European Food Safety Authority)	1. Assemble the evidence into lines of evidence of similar type 2. Weigh the evidence 3. Integrate the evidence	• Three basic considerations: reliability, relevance and consistency • Flexibility and transparency in WoE and harmonization with existing methodologies	• A single WoE framework that is broadly application, user-friendly, transparent and scientifically sound may not be available. • Trade-offs between feasibility and scientific rigor.
Kaltenhauser et al., 2017	Pesticides	1. Define/refine criteria 2. Search bibliographic databases 3. Inclusion/exclusion of studies based on relevance (checklists) 4. Reliability of relevant studies 5-8. Collect, synthesize, present and interpret data from relevant studies	• Data must be relevant and reliable. • Weighting is influenced by factors including the test organism, study design and statistical methods, as well as test item identification, documentation and reporting of results.	No discussion/guidance on integrating lines of evidence
Meek et al., 2013	Human relevance of hypothesized mechanisms of action	1. Consider evidence for causality for a hypothesized MOA in animals using the evolved Bradford Hill criteria (evidence sufficiency). 2. Assess species concordance (evidence relevance) 3. Use kinetic and dynamic data to estimate dose-response.	Transparency and use of all available data	The framework focuses on comparing WoE for two substances, rather than assessing whether evidence is sufficient to inform a particular decision.
Money et al., 2013	Registered substance under the EU REACH Regulation	Assess the reliability of evidence and identify studies for stronger weighting using the adapted Klimish categories (adapted for use in human data evaluation).	Adapted Klimisch categories for human data evaluation	Reliance on subjective judgements
Rhomberg et al., 2015	Regulatory toxicology	1. Systematic review of individual studies. 2. Assess consistency, specificity and reproducibility of outcomes for each endpoint. 3. Identify and articulate hypotheses and explain relevance of available studies 4. Assess the logic and evidentiary support of each hypothesis based upon each line of evidence. 5. Assess the logic and evidentiary support of each hypothesis across lines of evidence. 6. Describe and compare support for hypotheses, including uncertainties, inconsistencies and assumptions 7. Characterize the WoE and propose next steps.	• Rigorous systematic review processes • Apply set of rules for evaluating evidence • Rely on personal judgements of a panel of experts	Reliance on subjective judgements can lead to conflict, suspicion of arbitrary decision-making (erosion of trust), and contention in selection of judges.

Citation	Scope	Process	Key principles/ Best practices	Challenges/ limitations
Rooney et al., 2014	Environmental hazard identification in the United States	1. Formulate research question and develop protocol. 2. Search for and screen studies 3. Data extraction 4. Quality assessment of individual studies 5. Assessment of confidence in the body of evidence, 6. Translate the confidence ratings into levels of evidence. 7. Integrate lines of evidence to characterize hazard.	• Take an objective and transparent approach (*a priori* criteria, with changes catalogued and justified). • Keep lines of evidence separate until Step 7. • Incorporates adapted Bradford-Hill considerations and GRADE approach.	Reliance on subjective judgements for assessment of confidence Process focuses on systematic review methodology rather than WoE approaches.
Vandenberg et al., 2016	Endocrine- disrupting chemicals	1. Formulate the problem 2. Develop a review protocol 3. Identify relevant evidence 4. Evaluate evidence from individual studies 5. Summarize and evaluate each line of evidence 6. Integrate lines of evidence 7. Interpret findings, implications and uncertainties	• Transparency in expert judgement. • GRADE approach to assessing confidence in the body of evidence	Reliance on subjective judgements can reduce reproducibility and transparency

Tab. S4: Summary of quantitative WoE frameworks

Citation	Scope	Process	Key principles / Best practices	Challenges / limitations
Becker et al., 2015, 2017	Risk evaluation for regulation of commercial products	1. Identify postulated MOAs. 2. Qualitatively evaluate the evidence for each key event (KE) and key event relationship (KER). 3. Quantitatively rate each KE/KER using the evolved BH causal considerations. 4. Derive a composite score for each KE/KER. 5. Integrate the evidence of causality for the MOA. 6. Compare the quantitative confidence scores for the hypothesized MOA. 7. As relevant, consider species concordance or human relevance. 8. Characterize the hazard.	• Use the three Bradford-Hill considerations of biological plausibility, essentiality and empirical evidence. • Score quality of both supportive and counter evidence.	• A transparent approach to integrating sources of varying quality, purpose and biological organization into lines of evidence remains lacking. • Insufficient guidance on assigning empirical scores to the multi-criterion decision analysis.
Bridges et al., 2017	Regulatory toxicity testing in the EU	1. Formulate hypothesis, identify lines of evidence an develop score sheet templates. 2. Search literature and categorize included studies into the lines of evidence. 3. Assess study quality. 4 Assess data relevance. 5. Evaluate the strength of evidence for individual lines of evidence. 6. Evaluate the strength of evidence across lines of evidence. 7. Characterize hazard and report uncertainties.	Quantitative approach to WoE is a movement towards a consistent and reproducible methodologies, which has to date been unavailable.	An absence of formal WoE procedures has led to differences in scope and detail for WoE, which can make the process and findings unclear (in particular as "strength of evidence" may be viewed differently)
Buist et al., 2013; Tluckiewicz et al., 2013; Vermeire et al., 2013	Chemical risk assessment in the EU	1. Gather all substance-specific information. 2. Weigh each type of information using statistical methods and/or expert knowledge. 3. Decide whether available information is sufficient for decision-making (if not: identify data gaps). 4. Gather information on structurally related chemicals. 5. Decide whether exposure-based waiving and Thresholds of Toxicological Concern can be used. 6. Perform animal testing as a last resort.	• Weigh based on validity and adequacy (reliability and relevance). • Formal, transparent, and statistical approach to decision-making. This approach should increase transparency, reproducibility and objectivity in WoE.	There is often inadequate documentation of WoE methods and the influence of expert judgement presents risks to transparency.
Collier et al., 2016	Chemical hazard and risk assessment	1. Prepare the AOP (assemble the evidence. 2. Weigh and score the evidence (weighing the line of evidence) 3. Aggregate the lines of evidence (weighing the body of evidence)	• The EPA General Assessment Factors (soundness; applicability and utility; clarity and completeness; uncertainty and variability; and evaluation and review) are used to assess the quality of individual studies. • The modified Bradford-Hill criteria are used to assess the body of evidence.	Judgement-based data analysis is subject to bias (and it is recommended that a group of subject matter experts are engaged to improve consistency).

Citation	Scope	Process	Key principles / Best practices	Challenges / limitations
Dekant and Bridges, 2016	Chemical hazard classification (focus on classification and labelling) in the EU	1. Define hypothesis 2. Search literature 3. Develop scoring categories for quality/strength of evidence 4. Score each study for quality/strength of evidence 5. Graph/tabulate scores for all studies. 6. Calculate overall score (multiply scores for quality by scores for strength). 7. Compare overall score with predefined thresholds for classification.	Consider all available data sources in an objective, transparent and reproducible manner.	• Vague definition of WoE can lead to variation in assessment findings. • It is time-consuming and challenging to develop a scoring system for the quality/strength of each line of evidence.
Dekant et al., 2017	Adverse effects by chemicals (extrapolating to humans from animal studies	1. Formulate problem (identify adverse effects in the appropriate literature) 2. Literature search 3. Define hypothesis (MOAs for adverse effect) 4. Define molecular initiating event (MIE) and key event (KE) in MoA 5. Develop scoring categories for quality/strength 6. Score each report on MIEs/KEs for quality and strength of effects 7. Tabulate scores for all MIEs/KEs and calculate summary score for support of MoA in animals for biologically plausible MoAs. 8. Assess human relevance of all steps for best supported MoA in animals	• Score individual studies on the relevance of the model system to the MIE/KE, relevance of exposure (e.g., concentration) conditions and strength/consistency of effects • Score body of evidence based upon biological plausibility experimental support and human relevance.	• Development of scoring system is challenging and time-consuming. • Basic principles for development of regulatory guidelines tend not to reflect the best practices employed to investigate MoA.
Hristozov et al., 2014a,b Sheehan et al., 2018	Nanomaterial hazard identification/screening	1. Calculate hazard scores based on physico-chemical properties and toxicological effects. 2. Indices for each are aggregated into a hazard index using a weighted sum operator for each line of evidence.	• Characterize hazard is based upon sets of criteria related to material properties, toxicity and data quality. • Data quality is assessed using Klimisch scores for adequacy, reliability and relevance	Reliance on expert elicitation for Klimish scores, development of risk classes and data quality evaluation. There is a lack of stepwise guidance in the WoE process.

3.5. Alternative test methods in WoE

A expanding array of alternative test methods, also known as NAMs, are available as a source of evidence on potential human health risks of environmental agents (Andersen et al., 2019), the present authors sought to understand their relevance and potential application in the context of WoE frameworks.

Some have been explicitly mentioned in the qualitative (Gross et al., 2017) and quantitative (Dekant et al*.,* 2017) frameworks mentioned above. However, while frameworks did not restrict or preclude the incorporation of NAMs, there was little guidance regarding how alternative testing procedures could be incorporated in WoE approaches. Most of the discussion in this regard was focused on AOPs that can be used to consider mechanistic data and link molecular initiating events to biological outcomes.

In their quantitative WoE framework, Becker and colleagues (2015) describe a method for assessing WoE of an AOP using guidance provided by the Organization for Economic Cooperation and Development (OECD). While authors note that further refinement is needed, they point to the potential value of incorporating AOP information in WoE assessments. Similarly, Collier and colleagues (2016) note the lack of guidance on WoE determinations for AOPs, advocating an approach based on expert judgement, and illustrate how this approach could be applied in two exemplars. Rocca and colleagues (2018) suggest the use of target biology and molecule-specific pharmacokinetics for biopharmaceutical risk assessments, turning to animal studies only in cases where an unacceptable level of uncertainty persists. Although these suggestions on how to incorporate NAM data into WoE evaluations are welcome, more detailed guidance on the broader use of NAMs in support of human health risk assessment is needed.

4 Discussion

This review updates a publication by Rhomberg and colleagues (2013), providing the most current understanding of the body of literature on WoE approaches. In the main, the present update is consistent with the findings of the earlier survey in that the array of approaches covers the same span and new developments have not obviated any findings. The Rhomberg *et al*. survey included an extensive discussion and evaluation of the understanding of issues and challenges as revealed by the surveyed WoE approaches. We will not repeat that discussion here, but instead focus on what the update has shown about trends and developments in the ongoing evolution of WoE processes.

In the past five years, it appears that there has been a movement towards quantitative WoE approaches, with seven of the 20 included frameworks (35%) dealing with quantitative methodologies. However, it is important to note that quantitative approaches will only improve consistency and reliability if they are paired with transparent and rigorous approaches to assess and quantify the body of evidence; methodologies based on assigning numerical values to qualitative assessments may do little more than obscure the subjective judgements that are informing the assessment.

A similar consideration relates to the relative value of ranking (Gross and Fedak, 2015) or categorizing (Catalan et al*.,* 2017) weight of evidence, whether qualitatively or quantitatively. It may be the case that a ranking of different weights of evidence is more valid and reliable, as it allows the user to draw comparisons across bodies of evidence. However, this approach may not be as well-suited to informing decision-making, where qualitative categorizations may be most appropriate if hazard characterization thresholds are informed by evidence. It is likely that the most appropriate approach will vary with the research question and decision-making context, though this distinction is still poorly understood.

A common set of principles for WoE assessment began to emerge from the body of literature; these most commonly referred to the reliability (or quality), consistency and relevance of evidence. Hardy and colleagues (2017) defined reliability as the extent to which evidence (or a line of evidence) was correct; relevance related to the extent to which evidence (or a line of evidence) would help answer the research question if correct (including whether non-human studies are relevant to human populations); consistency was understood as the degree to which different lines of evidence were compatible. Other relevant principles include biological plausibility (assessment of the biological evidence of a mechanistic link between an upstream and downstream event), essentiality (assessment of whether downstream events are prevented by blocking upstream events) and empirical evidence (consistency of support for the hypothesized exposure-outcome relationship) (Becker et al., 2015, 2017). Interestingly, authors found no discussion of the application of principles of risk-based decision making; the implications of principles such as the precautionary principle, risk acceptability and cost-effectiveness are likely of direct relevance to WoE assessment and hazard characterization for decision-making and warrant more explicit discussion.

A variety of tools, scales and scoring systems were proposed for various elements of individual study and collective evidence assessment. The most commonly proposed was an adaptation of the Bradford-Hill considerations, summarized in Table S5; these assess the epidemiological evidence for causality, and were commonly applied in both qualitative and quantitative approaches (Becker et al., 2015, 2017; Collier et al., 2016; Gross et al., 2017; Meek et al., 2013; Rooney et al., 2014).

Tab. S5: Modified Bradford-Hill considerations (adapted from Meek et al., 2014)

Consideration	Definition
Concordance of dose-response relationships between key and end events	Dose-response relationships for key events are compared with one another and with those for endpoints of concern. (Are the key events always observed at doses below or similar to those associated with toxic outcome?)
Temporal association	Key events and adverse outcomes are evaluated to determine if they occur in expected order
Consistency and specificity (essentiality)	Is the incidence of the toxic effect consistent with that for the key events? Is the sequence of events reversible if dosing is stopped or a key event prevented?
Biological Plausibility	Is the pattern of effects across species/strains/systems consistent with the hypothesized MoA? Does the hypothesized MoA make sense based on broader knowledge?

Another notable assessment tool was the Klimisch scores, which are used to assess the reliability of data based upon the previously discussed principles of adequacy, reliability and relevance (Hristozov et al., 2014a,b; Sheehan et al., 2018). This scale forms the foundation of the European regulation on Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH) (ECHA, 2008). Lastly, two studies (Rooney et al., 2014; Vandenberg et al., 2016) made reference to using the GRADE guidelines to assess the overall quality of evidence. Covering subjective assessments of risk of bias, imprecision, inconsistency, indirectness and publication bias, the GRADE framework overlaps with some of the modified Bradford-Hill considerations, though both publications focused more generally on systematic reviews rather than WoE. Uncertainty remains regarding the contexts in which a particular individual or collective assessment tool might be more appropriate than others.

A notable result of the present survey is that the methods for assembly and evaluation of validity of relevant studies are found to be extensive and, at least in broad terms, similar across existing WoE-evaluation methods. That is, they all proceed on the premise that an objective, systematic, and openly-documented process for identifying evidence – and for noting the strengths and potential shortcomings of individual studies in providing reliable results – is key to a sound evaluation approach. The aims are to guard against even unintentional tendencies to select or emphasize studies that support a particular favored conclusion while overlooking or downplaying studies with contradictory implications, justifying a prior impression rather than objectively assembling evidence. The same principle applies to evaluations of study quality, discouraging the selective attention to shortcomings of studies at odds with one’s favored conclusion.

The prior specification of an evidence-assembly and -review process guards against such biases. It also makes the process transparent (by publicly reporting on how the rules have been followed and applied), and it makes for consistency in the process across applications. Further, it communicates to stakeholders and the affected public how the decision process for characterizing toxicity proceeds.

As we have noted, however, the surveyed methods are quite unspecific about how the “integration” or “synthesis” of these assembled evidence elements is to be conducted so as to gauge the evidentiary support for an overall conclusion regarding the motivating fundamental question of the evaluated agent’s potential (human) toxicity. It is difficult to articulate a pre-specified, rules-based process for how integration of evidence across studies is to proceed, beyond the recognition that decisions should be made on the consideration of inferences across available sources of evidence rather than being keyed on single studies that are somehow identified as dispositive.

The amount of information required to reach a clear conclusion using weight of evidence evaluation remains somewhat decision-specific. While standard processes and best practices for conducting WoE assessment can be identified, it is likely inappropriate to recommend a single, one-size-fits all approach for when and how WoE can be viewed as sufficient to reach a conclusion, as this will be dependent on context-specific factors that include the strength of different lines of evidence, qualitative or quantitative approaches to synthesis of the lines of evidence, and the risk context itself. Weighting and prioritization of evidence is likely to vary both within and across lines of evidence, recognizing that some studies may be more powerful, informative or rigourous, deserving more weight than others. Similarly, certain lines of evidence (e.g., direct evidence of harm or risk in human populations) may be more pertinent that others. As such, conclusive evidence may not be required across lines of evidence in order to arrive at a conclusion or regulatory decision.

Historically, the Bradford-Hill criteria provided guidance on evaluating the weight of evidence for concluding causality; these often-invoked criteria formed the foundation for more structured GRADE evaluations of the degree of confidence afforded by the available data in concluding causality. Examples of context-specific weight of evidence evaluation schemes include the recently revised Preamble to the IARC Monographs (Baan and Straif, 2022), which integrate human, animal, and mechanistic evidence streams to reach graded decisions of the likelihood that an agent poses a cancer risk to humans. The European Union’s REACH Regulation – which was established to evaluate potential health and environmental risks of commercial chemicals – includes similarly elaborate guidance on identifying both cancer and non-cancer hazards (Armstrong et al., 2020), considering all relevant and reliable key, supporting, and ‘weight of evidence’ studies in an organized fashion (see Willhite et al., 2021, Figure 1). These two examples both involve expert scientific judgment in reaching weight of evidence conclusions, as does the European Food Safety Authority framework for scientific assessments (Aiassa et al., 2022).

Attention to strengths, shortcomings, or ambiguities of individual studies is usually asked for, but the further questions as to how to judge applicability of each different type of study result, how to deal with apparent disagreements or inconsistencies among studies, and how to weigh the influence among studies of differing strength but also of potentially differing relevance are usually not spelled out with any specificity. Importantly, the ways to resolve apparent inconsistencies among studies and their differing implications is rarely set out in any prescribed method.

This reflects the fact that most of the study results in question are not simple repeated instances of direct observations of the causal effect in question (where the main question for integration would be an evaluation of their consistency). Rather, they are attempts (which may be more or less successful) to examine potential for toxicity in a controlled (and therefore limited or even rather artificial) setting such that any effects can clearly be attributable to the treatment applied, combined with the further inference that any such effects are generalizable from the constrained tested setting to the setting of ultimate interest (usually, the ability to cause toxicity in humans at the levels of exposure they actually experience). A rodent bioassay result, for instance, needs to be judged not only as to whether the results are reliably attributable to the tested agent ("internal validity") but also as to whether a finding in the particular bioassay system should be taken as evidence that the target human population should be expected to have a similar reaction. This inference must be made in view of our wider experience with the degree of consistency of concordance of effects across bioassay systems, the agreeing and disagreeing results for the particular agent and the particular toxicity in question, knowledge (or hypothesis) about similarities or differences in apparent modes of action among species, and so on.

Simple and consistently applicable rules (to which adherence can be systematically documented) are challenging to formulate for such complex inferences. Sound inferences depend not only on the specific results at hand but also on wider understanding of the biological basis for invoking relevance of results and the history of and nature of exceptions or limitations to tenable extrapolation of effects seen in test systems to the target human population. The surveyed methods tend to invoke more general principles to be borne in mind by those conducting expert judgment, rather than specifying rules by which those judgments are to be carried out. They usually acknowledge that for now and for the foreseeable future, these integration processes must be matters of professional judgment rather than the application of an algorithmic system of ex-ante decision rules. They encourage the articulation of the basis of judgments and its tying to the objectively and systematically assembled base of evidence, so that the reasoning is public and openly debatable.

In such a process, it needs to be presumed that competent professional judgment will be similar among practitioners faced with a given objectively developed array of evidence (i.e., that the judgment is driven mainly by the results alone and is largely independent of the specific judges). That is, it is generally presumed in the surveyed systems that competent scientists would read the evidence similarly if it has objectively been set out to them, and so those designated to conduct the evaluation can be taken as representatives of scientific opinion in general.

Given the challenges already noted in formulating an ex-ante set of interpretation and evaluation rules to achieve the integration of inferences across all the available evidence, however, it would appear difficult to achieve the desired independence of decisions from the choice of judges. But it is hard to avoid the concerns that stakeholders whose own judgments are contradicted might challenge the objectivity of the designated evidence-interpreters. It is also difficult to document the consistency of application of judgments among cases, since whether pre-stated overarching principles of proper interpretation have been adhered to is itself a matter of scientific judgment. The methods surveyed here have not, in general, addressed this issue. Further work seems warranted on how to develop consistent, pre-specified, and repeatable processes for evidence integration, such that adherence to good practices can be documented and questions about the soundness of judgments can be avoided.

This review was not without limitations. First, a decision was made to conduct a comprehensive and structured review as opposed to a systematic one, as it was felt that the full body of knowledge was not necessary to elicit the insights to inform further discussion and application of WoE approaches; as a result, only a single database was searched, and articles were subjected to screening and extraction by a single reviewer. This methodology mirrors that of a previous publication (Rhomberg et al., 2013) intended for a similar purpose. Further, a lack of clarity in the distinction between systematic review and WoE – the boundaries between which vary and overlap across publications and organizations – presented challenges in conducting a review of WoE approaches without considering the broader scope of systematic reviews (such as *how* to access all available data).

Put briefly, the diversity of WoE approaches and vagueness over best practices has obstructed progress towards a formal, consistent, and universal procedure that reflects the WoE principles – transparency, flexibility, reproducibility, objectivity, quality, consistency, relevance – about which there is some degree of consensus. Particular areas where further guidance could be of value would include the integration of lines of evidence and assessing the sufficiency of evidence to inform decision-making. While it is unlikely – and perhaps undesirable – that a reliance on expert judgement can be eliminated, more formal guidance can help experts speak the same language while making WoE approaches and findings more transparent and accessible to the public. In particular, additional guidance on how alternative test methods and mechanistic data can be better incorporated in WoE assessments will help in further reducing reliance on animal testing in risk assessment.

5 Conclusion

With this review, the authors have established the most current understanding of WoE approaches. Across a diverse range of qualitative and quantitative frameworks, a consistent set of principles was reflected across varying methodologies. This review was intended as a foundation for further research that built upon best practices and addressed persisting knowledge gaps. As informed by the challenges identified above, future research will focus on understanding the role of risk-based decision-making in WoE, developing case studies to understand the role of context in determining best practices, and generating formal and stepwise guidance on best practices to improve transparency, consistency, and reliability in WoE.

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

Agerstrand, M. and Beronius, A. (2016). Weight of evidence evaluation and systematic review in EU chemical risk assessment: Foundation is laid but guidance is needed. *Environ Int 92-93*, 590-596. doi:10.1016/j.envint.2015.10.008

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., N. Karyakina, E. Nordheim 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

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

Becker, R., Ankley, G., Edwards, S. et al. (2015). Increasing scientific confidence in adverse outcome pathways: Application of tailored Bradford-Hill considerations for evaluating weight of evidence. *Regul Toxicol Pharmacol 72*, 514-537. doi:10.1016/j.yrtph.2015.04.004

Becker, R., Dellarco, V., Seed, J. et al. (2017). Quantitative weight of evidence to assess confidence in potential modes of action. *Regul Toxicol Pharmacol 86*, 205-220. doi:10.1016/j.yrtph.2017.02.017

Bridges, J., Sauer, U. G., Buesen, R. et al. (2017). Framework for the quantitative weight-of-evidence analysis of 'omics data for regulatory purposes. *Regul Toxicol Pharmacol 91, Suppl 1*, S46-S60. doi:10.1016/j.yrtph.2017.10.010

Buist, H., Aldenberg, T., Batke, M. et al. (2013). The OSIRIS weight of evidence approach: ITS mutagenicity and ITS carcinogenicity. *Regul Toxicol Pharmacol 67*, 170-181. doi:10.1016/j.yrtph.2013.01.002

Catalan, J., Stockmann-Juvala, H. and Norppa, H. (2017). A theoretical approach for a weighted assessment of the mutagenic potential of nanomaterials. *Nanotoxicology 11*, 964-977. doi:10.1080/17435390.2017.1382601

Collier, Z. A., Gust, K. A., Gonzalez-Morales, B. et al. (2016). A weight of evidence assessment approach for adverse outcome pathways. *Regul Toxicol Pharmacol 75*, 46-57. doi:10.1016/j.yrtph.2015.12.014

Cuddy, M., Poda, A., Moser, R. et al. (2016). A weight-of-evidence approach to identify nanomaterials in consumer products: A case study of nanoparticles in commercial sunscreens. *J Expo Sci Environ Epidemiol 26*, 26-34. doi:10.1038/jes.2015.51

Dekant, W. and Bridges, J. (2016). A quantitative weight of evidence methodology for the assessment of reproductive and developmental toxicity and its application for classification and labeling of chemicals. *Regul Toxicol Pharmacol 82*, 173-185. doi:10.1016/j.yrtph.2016.09.009

Dekant, W., Bridges, J. and Scialli, A. R. (2017). A quantitative weight of evidence assessment of confidence in modes-of-action and their human relevance. *Regul Toxicol Pharmacol 90*, 51-71. doi:10.1016/j.yrtph.2017.08.012

EPA (1986). Guidelines for Carcinogen Risk Assessment. (51 FR 33992-34003). https://cfpub. epa.gov/ncea/risk/recordisplay.cfm?deid=54933

ECHA (2008). Chapter R.4: Evaluation of Available Information. https://echa.europa.eu/documents/10162/13643/ information_requirements_r4_en.pdf/d6395ad2-1596-4708-ba86-0136686d205e

ECHA (2015a). Guidance on the Application of the CLP Criteria. https://echa.europa.eu/documents/ 10162/23036412/clp_en.pdf

ECHA (2015b). *Guidance on Biocidal Products Regulation*. https://echa.europa.eu/documents/10162/23036412/ biocides_guidance_human_health_ra_iii_part_bc_en.pdf/30d53d7d-9723-7db4-357a-ca68739f5094

Gross, S. A. and Fedak, K. M. (2015). Applying a weight-of-evidence approach to evaluate relevance of molecular landscapes in the exposure-disease paradigm. *Biomed Res Int 2015*, 515798. doi:10.1155/2015/515798

Gross, M., Green, R. M., Weltje, L. and Wheeler, J. R. (2017). Weight of evidence approaches for the identification of endocrine disrupting properties of chemicals: Review and recommendations for EU regulatory application. *Regul Toxicol Pharmacol 91*, 20-28. doi:10.1016/j.yrtph.2017.10.004

Hardy, A., Benford, D., Halldorsson, T. et al. (2017). Guidance on the use of the weight of evidence approach in scientific assessments. *EFSA J 15*, e04971. doi:10.2903/j.efsa.2017.4971

Hristozov, D. R., Gottardo, S., Cinelli, M. et al. (2014a). Application of a quantitative weight of evidence approach for ranking and prioritising occupational exposure scenarios for titanium dioxide and carbon nanomaterials. *Nanotoxicology 8*, 117-131. doi:10.3109/17435390.2012.760013

Hristozov, D. R., Zabeo, A., Foran, C. et al. (2014b). A weight of evidence approach for hazard screening of engineered nanomaterials. *Nanotoxicology 8*, 72-87. doi:10.3109/17435390.2012.750695

Kaltenhauser, J., Kneuer, C., Marx-Stoelting, P. et al. (2017). Relevance and reliability of experimental data in human health risk assessment of pesticides. *Regul Toxicol Pharmacol 88*, 227-237. doi:10.1016/j.yrtph.2017.06.010

Linkov, I., Loney, D., Cormier, S. et al. (2009). Weight-of-evidence evaluation in environmental assessment: Review of qualitative and quantitative approaches. *Sci Total Environ 407*, 5199-5205. doi:10.1016/j.scitotenv.2009.05.004

Linkov, I. (2015). From “weight of evidence” to quantitative data integration using multicriteria decision analysis and Bayesian methods. *ALTEX 32*, 3-8. doi:10.14573/altex.1412231

Lutter, R., Abbott, L., Becker, R. et al. (2015). Improving weight of evidence approaches to chemical evaluations. *Risk Anal 35*, 186-192. doi:10.1111/risa.12277

Meek, B., Palermo, C., Bachman, A. et al. (2013). Mode of action human relevance (species concordance) framework: Evolution of the Bradford Hill considerations and comparative analysis of weight of evidence. *J Appl Toxicol 34*, 595-606. doi:10.1002/jat.2984

Meek, M., Boobis, A., Cote, I. et al. (2014). New developments in the evaluation and application of the WHO/IPCS framework on mode of action/species concordance analysis. *J Appl Toxicol 34*, 1-18. doi:10.1002/jat.2949

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

Money, C. D., Tomenson, J. A., Penman, M. G. et al. (2013). A systematic approach for evaluating and scoring human data. *Regul Toxicol Pharmacol 66*, 241-247. doi:10.1016/j.yrtph.2013.03.011

NRC (2014). *Review of EPA’s Integrated Risk Information System (IRIS) Process.* Washington, DC, USA: National Academies Press. doi:10.17226/18764

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

Rhomberg, L. R. (2015). Hypothesis-based weight of evidence: An approach to assessing causation and its application to regulatory toxicology. *Risk Anal 35*, 1114-1124. doi:10.1111/risa.12206

Rocca, M., Morford, L., Blanset, D. et al. (2018). Applying a weight of evidence approach to the evaluation of developmental toxicity of biopharmaceuticals. *Regul Toxicol Pharmacol 98*, 69-79. doi:10.1016/j.yrtph.2018.07.006

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

Sheehan, B., Murphy, F., Mullins, M. et al. (2018). Hazard screening methods for nanomaterials: A comparative study. *Int J Mol Sci 19*, 649. doi:10.3390/ijms19030649

Suter, G., Cormier, S. and Barron, M. (2017). A weight of evidence framework for environmental assessments: Inferring qualities. *Integr Environ Assess Manag*. doi:10.1002/ieam.1954

Tluckiewicz, I., Batke, M., Kroese, D. et al. (2013). The OSIRIS weight of evidence approach: ITS for the endpoints repeated-dose toxicity (RepDose ITS). *Regul Toxicol Pharmacol 67*, 157-169. doi:10.1016/j.yrtph.2013.02.004

Vandenberg, L. N., Agerstrand, M., Beronius, A. et al. (2016). A proposed framework for the systematic review and integrated assessment (SYRINA) of endocrine disrupting chemicals. *Environ Health 15*, 74. doi:10.1186/s12940-016-0156-6

Vermeire, T., Aldenberg, T., Buist, H. et al. (2013). OSIRIS, a quest for proof of principle for integrated testing strategies of chemicals for four human health endpoints. *Regul Toxicol Pharmacol, 67*, 136-145. doi:10.1016/j.yrtph.2013.01.007

Weed, D. L. (2005). Weight of evidence: A review of concept and methods. *Risk Anal 25*, 1545-1557. doi:10.1111/j.1539-6924.2005.00699.x

Willhite, C. C., N. A. Karyakina, E. Nordheim et al. (2021). The REACH process: A case study of aluminium metal, aluminium oxide and aluminium hydroxide. *Neurotoxicology 83*, 166-178*.* doi:10.1016/j.neuro.2020.12.004

Supplementary Material II

Development of an Evidence Based Risk Assessment Framework
Workshop Agenda

<u>Monday, December 17th, 2018</u>

Welcome and Overview

8:15 am - 8:20 am	Guy Levesque, Associate Vice President, University of Ottawa Welcome Welcome
8:20 am - 8:30 am	Daniel Krewski, University of Ottawa Risk science in the 21st century: Overview

Session 1: Recent Advances in Risk Science: Including New Approach Methodologies in Weight of Evidence Evaluation

This session will take stock of recent scientific developments that will support evidence-based risk assessment, including new approach methodologies (NAMs).

Chair: Thomas Hartung, Johns Hopkins University

8:30 - 8:55 am	Maureen Gwinn, US EPA Current Status of New Approach Methodologies
8:55 am - 9:20 am	Patience Browne. OECD Predictive value of in vitro assays
9:20 am - 9:45 am	Andrew Rooney, NIEHS Incorporating information from new approach methodologies in weight of evidence evaluation)
9:45 am - 10: am	General discussion
10:00 am - 10:30 am	*Break*

Session 2: Summarizing the Evidence

This session will focus on methods for summarizing all relevant data to be included in an evidence-based risk assessment. Methods in systematic review will be examined, along with current approaches to data quality scoring.

Chair: Jeff Lewis, Exxon Mobil Biomedical Research

10:30 am - 10:55 am	Juleen Lam, Cal State East Bay Integrating multiple evidence streams
10:55 am - 11:20 am	Thomas Hartung, Johns Hopkins Systematic review of toxicological data
11:20 am - 11:45 am	Charlotte Bertrand, US EPA Quality scoring of human, animal, and in vitro data
11:45 am - 12:00 pm	General discussion
12:00 pm - 1:00 pm	*Lunch* Demonstration of Bayesian Weight of Evidence Decision-Support Tool Moez Sanaa, ANSES and Greg Paoli, Risk Sciences International

Session 3: Qualitative Data Synthesis

The first step in evidence-based risk assessment is the determination of whether or not a hazard exists. This involves a weight of evidence evaluation of all relevant information in order to reach a decision on whether the available data supports the existence of a human health hazard.

Chair: Kristina Thayer, US EPA

1:00 pm - 1:25 pm	Kurt Straif (confirmed, IARC The IARC Monographs Programme of identification of carcinogenic hazards to humans
1:25 pm - 1:50 pm	Holger Schünemann, McMaster Use of GRADE in evidence integration

1:50 pm - 2:15 pm Andrew Kraft, US EPA
Current and future EPA practices in systematic review
2:15 pm - 2:30 pm General discussion

2:30 pm - 3:00 Break

Session 4: Quantitative Data Synthesis

Once a hazard has been identified on the basis of the available evidence, a quantitative assessment of risk and exposure-response may be undertaken. This session will focus on new methodologies for quantitative synthesis of data from multiple sources, including synthesis of data on diverse toxicological endpoints.

Chair: Greg Paoli

3:00 pm - 3:25 pm Salomon Sand, Swedish National Food Agency
New approaches for quantitative combining of data from multiple sources
3:25 pm - 3:50 pm Don Mattison, Risk Sciences International
Quantitative synthesis of neurotoxicity data on manganese using categorical regression
3:50 pm - 4:15 pm Weihsueh Chiu, Texas AandM University
New approaches to characterizing uncertainty in risk assessment
4:15 pm - 4:40 pm Katya Tsaioun, Johns Hopkins University
In vitro predictions of drug induced liver injury
4:40 pm - 5:00 pm General discussion

5:00 pm Adjourn

Tuesday, December 18th, 2018

8:30 am - 9:00 am Summary of Day 1
Daniel Krewski, University of Ottawa

Session 5: Putting Weight of Evidence into Practice

In order to guide discussions about considerations involved in the practical implementation of weight of evidence, this session will provide an overview of current approaches within EFSA and Health Canada.

Chair: Maureen Gwinn, EPA

9:00 am - 9:25 am Elisa Aiassa, Laura Martino and Caroline Merten, EFSA
Evidence integration: an EU perspective
9:25 am - 9:50 am Tara-Barton Maclaren, Health Canada
Health Canada's evolving framework for evidence synthesis

10:00 am - 10:30 am Break

The remainder of the meeting will be held in closed session.

Session 6: Breakout Groups

Participants at the workshop will be assigned to breakout groups to address a series of key questions relating to the development of an evidence-based framework for risk assessment. (Questions developed by the Steering Committee.)

Moderator: Tara Barton-Maclaren, Health Canada

10:30 am - 12:00 pm Parallel Breakout Group Discussions

Group 1: Lessons learned from previous experience
Chair: Lorenz Rhomberg, Gradient Corporation
Rapporteur: Patrick Saunders-Hastings, Gevity

Group 2: Benchmarks of good practice
Chair: Greg Paoli, Risk Sciences International
Rapporteur: Maureen Gwinn, US EPA

Group 3: Problem formulation and data requirements
Chair: Robert Baan, IARC (retired)
Rapporteur: Kris Thayer, US EPA

Group 4: Potential challenges
Chair: Thomas Hartung, Johns Hopkins
Rapporteur: Rebecca Morgan, McMaster University

12:00 pm - 1:00 pm Lunch

1:00 pm - 2:00 pm Breakout Group Reports
Moderator: Tara Barton-Maclaren, Health Canada
2:00 pm - 2:30 pm Synthesis of Breakout Group Reports
Daniel Krewski, University of Ottawa

2:30 pm - 3:00 pm Break

Session 6: General Discussion and Next Steps

This session will include a general discussion of key themes identified at the workshop and possible components of an evidence-based risk assessment framework. (Steering Committee members will be asked to provide their perspectives on future directions, with input from participants.)

Chair: Thomas Hartung, Johns Hopkins

3:00 pm - 3:30 pm Opening 5-minute presentations by Steering Committee members:
Tara Barton-Maclaren, Health Canada; Thomas Hartung, Johns Hopkins University; Daniel Krewski, University of Ottawa; Kristina Thayer, US EPA; Jeff Lewis, Exxon Mobil Biomedical Research.
3:30 pm - 4:00 pm General discussion
4:00 pm - 4:30 pm Conclusion
Daniel Krewski, University of Ottawa

4:30 pm Adjourn