Organisation for Economic Co-operation and Development

ENV/JM/MONO(2019)31

Unclassified

English - Or. English

22 October 2019

ENVIRONMENT DIRECTORATE
JOINT MEETING OF THE CHEMICALS COMMITTEE AND THE WORKING PARTY ON CHEMICALS, PESTICIDES AND BIOTECHNOLOGY

GUIDING PRINCIPLES AND KEY ELEMENTS FOR ESTABLISHING A WEIGHT OF EVIDENCE FOR CHEMICAL ASSESSMENT

Series on Testing and Assessment
No. 311

JT03453231

SERIES ON TESTING AND ASSESSMENT

NO. 311

GUIDING PRINCIPLES AND KEY ELEMENTS FOR ESTABLISHING A WEIGHT OF EVIDENCE FOR CHEMICAL ASSESSMENT

IOMC

INTER-ORGANIZATION PROGRAMME FOR THE SOUND MANAGEMENT OF CHEMICALS

A cooperative agreement among FAO, ILO, UNDP, UNEP, UNIDO, UNITAR, WHO, World Bank and OECD

Environment Directorate
ORGANISATION FOR ECONOMIC COOPERATION AND DEVELOPMENT
Paris 2019

About the OECD

The Organisation for Economic Co-operation and Development (OECD) is an intergovernmental organisation in which representatives of 36 industrialised countries in North and South America, Europe and the Asia and Pacific region, as well as the European Commission, meet to co-ordinate and harmonise policies, discuss issues of mutual concern, and work together to respond to international problems. Most of the OECD's work is carried out by more than 200 specialised committees and working groups composed of member country delegates. Observers from several countries with special status at the OECD, and from interested international organisations, attend many of the OECD's workshops and other meetings. Committees and working groups are served by the OECD Secretariat, located in Paris, France, which is organised into directorates and divisions.

The Environment, Health and Safety Division publishes free-of-charge documents in eleven different series: **Testing and Assessment**; **Good Laboratory Practice and Compliance Monitoring**; **Pesticides**; **Biocides**; **Risk Management**; **Harmonisation of Regulatory Oversight in Biotechnology**; **Safety of Novel Foods and Feeds; Chemical Accidents; Pollutant Release and Transfer Registers; Emission Scenario Documents;** and **Safety of Manufactured Nanomaterials.** More information about the Environment, Health and Safety Programme and EHS publications is available on the OECD's World Wide Web site (www.oecd.org/chemicalsafety/).

This publication was developed in the IOMC context. The contents do not necessarily reflect the views or stated policies of individual IOMC Participating Organizations.

The Inter-Organisation Programme for the Sound Management of Chemicals (IOMC) was established in 1995 following recommendations made by the 1992 UN Conference on Environment and Development to strengthen co-operation and increase international co-ordination in the field of chemical safety. The Participating Organisations are FAO, ILO, UNDP, UNEP, UNIDO, UNITAR, WHO, World Bank and OECD. The purpose of the IOMC is to promote co-ordination of the policies and activities pursued by the Participating Organisations, jointly or separately, to achieve the sound management of chemicals in relation to human health and the environment.

This publication is available electronically, at no charge.

Also published in the Testing and Assessment link

For this and many other Environment, Health and Safety publications, consult the OECD's World Wide Web site (www.oecd.org/chemicalsafety/)

or contact:

OECD Environment Directorate,
Environment, Health and Safety Division
2 rue André-Pascal
75775 Paris Cedex 16
France

Fax: (33-1) 44 30 61 80

E-mail: ehscont@oecd.org

FOREWORD

OECD member countries collaborate in developing and harmonising methods for assessing risk to human health and the environment, including methodologies for hazard and exposure assessment. This document is intended to provide universal Guiding Principles that should be considered when developing or augmenting systematic approaches to Weight of Evidence (WoE) for chemical evaluation and Key Elements to formulating a systematic approach to WoE. The ultimate goal is to facilitate that regulators follow a consistent, clear and transparent delivery of evidence using the Principles and Elements described in this document. This can be especially helpful for countries with no existing WoE frameworks or those looking to augment their approaches. It also allows for stakeholders to understand a WoE decision-making process, including potential for unreasonable bias. These Guiding Principles and Key Elements can be employed to develop frameworks that range from simple and pragmatic approaches to more elaborate systems, depending on the context.

The development of this document was led by Mark Bonnell (Environment Canada) and George Fotakis (European Chemicals Agency). Initial drafts were reviewed by a sub group of the OECD Working Party on Hazard Assessment (WPHA). The WPHA helps member countries to harmonise the methods and approaches used to assess chemicals by integrating various information sources by applying Integrated Approaches for Testing and Assessment (IATA), where establishment of WoE is highly important.

The draft guidance document was endorsed by WPHA during its 3rd meeting on 17-19 June 2019.

This document is published under the responsibility of the Joint Meeting of the Chemicals Committee and the Working Party on Chemicals, Pesticides and Biotechnology.

Table of contents

EXECUTIVE SUMMARY

Weight of evidence (WoE) is not a new concept with respect to decision-making. It has been used in the practice of law for several centuries. With respect to the prioritisation and risk assessment of chemicals, however, WoE definitions and practices vary in complexity. WoE can be generally understood to mean a method for decision-making that involves consideration of known lines of evidence where a "weight" is assigned to each line of evidence based on the confidence associated with the evidence. Evidence is combined and the overall strength of evidence determined to support or refute a hypothesis question posed during a problem formulation stage. The ultimate goal of WoE is to provide a transparent means for communicating decision-making such that decisions can be clearly understood and questioned by all stakeholders.

There are many approaches to WoE in the open literature given that WoE is often context dependent. Several of these approaches, both for human and ecological health have been examined for the preparation of this document. While differences in terminology exist, universal principles and elements have been identified and simplified to provide practical guidance to increase the formal implementation of WoE in chemical management programs, particularly where none yet exist and to help integrate different types of emerging data.

This document first describes universal Guiding Principles that should be considered when developing or augmenting systematic approaches to WoE for chemical evaluation. The Principles can be viewed as principles for "good WoE practice" and are intended to be endpoint and receptor agnostic. They can therefore be applied for both ecological or human health purposes. The Principles include:

- A ***Hypothesis*** which involves a clear formulation and statement of the problem for which evidence is needed and possible alternative hypotheses
- Be ***Systematic and Comprehensive*** in design by documenting a step-wise procedure integrating all evidence and indicating how evidence was collected, evaluated and weighed
- Include a ***Treatment of Uncertainty*** arising from available data (knowns) and data and/or knowledge gaps (unknowns)
- Consider the ***Potential for Bias*** during collection, evaluation and weighing of evidence
- Be ***Transparent*** by including clear documentation to assist the communication of WoE decisions so that they can be understood, reproduced, supported or questioned by all interested parties.

Following from the Principles, Key Elements to formulating a systematic approach to WoE are described. The Elements contain the necessary steps that should be taken to determine the overall strength of evidence to answer a hypothesis question. The Elements include:

- Problem formulation (hypothesis development)
- Evidence collection (establish lines of evidence and knowledge gaps)
- Evidence evaluation (determine data reliability, uncertainty and relevance)
- Evidence weighing (assign weight to evidence)
- Evidence integration and reporting (examine evidence coherence and impact of uncertainty).

Central to the above Elements is the communication and treatment of uncertainty. The level of acceptance and impact of uncertainty affects all elements of WoE and ultimately the strength behind decision-making. It is thus a common thread throughout the Elements described in this document and should be clearly understood from the beginning of problem formulation.

Finally, WoE is based on human judgement and therefore subject to human bias. If the Principles and Elements described in this document are considered, a consistent, clear and transparent delivery of evidence can follow allowing all stakeholders to understand decision-making including potential for unreasonable bias.

1. INTRODUCTION

The guiding principles and key elements presented in this document take into consideration that more than one method to establishing weight of evidence (WoE) is possible (very simple to very complex). To achieve the goal of having widely applicable guiding principles and a systematic approach, a relatively simple and practical approach is taken. It is also important to acknowledge that forming a WoE is judgement or value based. Values are subject to societal and context dependent-bias. As such, the important goal of any WoE evaluation is to provide a clear and transparent process that can be easily followed, reproduced and rationally discussed by all stakeholders.

1.1. Objective and Approach of Document

WoE remains a highly quoted and generally understood concept in chemical evaluation, including risk assessment, at least from first principles, however, its definition remains unclear and approaches can range significantly in complexity (Weed 2005). Weed (2005) also noted that "metaphorical" approaches, where no systematic method is described, are the most common approaches among regulatory agencies partly due to their simplistic nature and lack of operational guidance for conducting WoE within an agency.

Appendix 1, based on EFSA (2017), provides several conceptual descriptions of WoE used by various authors or agencies. Collectively, the descriptions consider WoE as a process for collecting, assessing and integrating evidence to reach a conclusion posed by a hypothesis question (which can be endpoint specific or overall risk or hazard determination). The European Commission's Scientific Committee on Health, Environmental and Emerging Risks (SCHEER) also provide a short compilation of WoE definitions used by various agencies (SCHEER 2018).

Conceptually WoE can be seen as a method for decision-making that involves consideration of known lines of evidence (LoEs) where a "weight" is assigned to each LoE, according to its relevance and reliability. A conclusion can be reached by combining the various LoEs to determine if sufficient strength of evidence is available to address the question posed under the hypothesis (e.g., a molecular initiating event will lead to an adverse outcome). Useful definitions in this regard are provided by the USEPA (2016):

> *"**Weight of Evidence:** (1) A process of making inferences from multiple pieces of evidence, adapted from the legal metaphor of the scales of justice. (2) The relative degree of support for a conclusion provided by evidence. The result of weighing the body of evidence."*

and by SCHEER (2018)

> *"**Weight of Evidence:** A process of weighted integration of lines of evidence to determine the relative support for hypotheses or answers to a question"*

and by OECD (2017)

> "***Weight of Evidence*** *refers to a positive expert opinion that considers available evidence from different independent sources and scientific viewpoints on a particular issue, coming to a considered view of the available, oftentimes conflicting data. It is preferred when every source does not provide sufficient information individually"*

Defining a prescriptive and systematic universal method for WoE is difficult given the context dependency of WoE (e.g., site-specific risk assessment versus chemical prioritisation versus risk assessment) (Bonnell 2011; Hall 2017). For the purposes of this document, however, it is possible to document Guiding Principles and Key Elements that a WoE framework should include when making decisions on the impact of chemicals to human health and the environment.

1.2. Terminology

Many investigators of WoE approaches have noted the confusion resulting from common terms cited in the scientific literature for describing how evidence can be weighed for risk assessment. Subjective terms such as "relevance", "confidence" or "adequacy", for example, can have different meanings depending on context. A glossary of terms is therefore provided in Appendix 2 to ensure clarity of key terminology used in this document. The glossary is not intended to be an exhaustive. Other WoE approaches (e.g., SCHEER 2018, EFSA 2017) can be consulted for additional terms.

1.3. Historical Background and Overview of Approaches

Some of the earliest guidance for the application of WoE by a regulatory body dates back to the mid-1980s for human health risk assessment when the USEPA published their *Guidelines for Carcinogen Risk Assessment* (USEPA 2005 updated) and in the 1990s, for example, by Menzie et al. (1996) for site-specific risk assessment in the United States. More recently, several regulatory bodies in Europe and North America [e.g., European Commission (SCHEER 2018), European Food Safety Authority (EFSA 2017), European Chemicals Agency (ECHA 2017), Government of Canada (GoC 2017), US Environmental Protection Agency (USEPA 2016), and USEPA Endocrine Disruptor program (USEPA 2011)] have been active in establishing evidence-based frameworks (see Appendix 1). More recently, OECD initiatives such as its Integrated Approaches to Testing and Assessment (IATA)[1] and Adverse Outcome Pathways (AOP)[2] have also used WoE approaches for integrating traditional and alternative data for hazard assessment. Appendix 1 contains a list of WoE related literature sources consulted for this document. This list is also not exhaustive, but includes some primary sources of WoE literature to augment references cited throughout this document.

[1] http://www.oecd.org/chemicalsafety/risk-assessment/iata-integrated-approaches-to-testing-and-assessment.htm
[2] http://www.oecd.org/chemicalsafety/testing/adverse-outcome-pathways-molecular-screening-and-toxicogenomics.htm

2. GUIDING PRINCIPLES FOR ESTABLISHING WEIGHT OF EVIDENCE

Examination of WoE approaches reveals that there are common or universal principles that underpin most approaches. It follows that development of WoE for chemical assessment should include *Guiding Principles* in order to provide an effective means of communicating decisions. The Guiding Principles are as follows:

- Include a ***Hypothesis*** which involves a clear formulation and statement of the problem for which evidence is needed and possible alternative hypotheses
- Be ***Systematic and Comprehensive*** in design by documenting a step-wise procedure integrating all evidence and indicating how evidence was collected, evaluated and weighed
- Include a ***Treatment of Uncertainty*** arising from available data (knowns) and data and/or knowledge gaps (unknowns)
- Consider the ***Potential for Bias*** during collection, evaluation and weighing of evidence
- Be ***Transparent*** by including clear documentation to assist the communication of WoE decisions so that they can be understood, reproduced, supported or questioned by all interested parties.

The above principles are intended to guide the development and application of evidence-based investigations, but do not themselves describe a systematic approach to WoE (i.e., elements and steps contained in WoE). Chapter three of this document will therefore describe Key Elements to WoE in a step-wise approach. The key elements help to ensure that the Guiding Principles above can be applied systematically when establishing and conducting evidence-based evaluation.

3. KEY ELEMENTS FOR ESTABLISHING A WEIGHT OF EVIDENCE FOR CHEMICAL EVALUATION

3.1. Overview

Figure 1 illustrates the WoE elements following a step-wise approach. The elements are organised to systematically allow for the incorporation of the Guiding Principles for establishing a WoE within the context of chemical evaluation and are conceptually consistent with several approaches reviewed (e.g., SCHEER 2018, Martin et al. 2018, GoC 2017, ECHA 2017, etc. - see Appendix 1). Terminology among existing WoE approaches varies, but conceptually, they almost always contain the concepts presented in Figure 1. The following chapters will discuss these elements in relation to the Guiding Principles outlined in Chapter 2 in more detail.

Figure 1: Key Elements for Weight of Evidence for Chemical Assessment

3.2. Problem Formulation

The first step to establishing a WoE is asking the right question in order to set the hypothesis to be tested against the available evidence. In the analysis of WoE approaches provided in Lutter et al. (2015), based on work by Rhomberg et al. (2013), the authors strongly supported a hypothesis-based approach for chemical evaluation. This is a critical step in the WoE process because if the wrong question is asked, the wrong evidence will be

gathered and the results will likely be meaningless (i.e., lack of plausibility of the hypothesis). Hypothesis-based questions are prepared during the problem formulation stage of chemical evaluation. Problem formulation should define the scope and goals of the assessment, the level of uncertainty that is acceptable as well as the urgency of the assessment. Here, hypothesis questions can be phrased either as general text in relation to a specific process or can have additional sub-questions that would assist information gathering. Evidence is gathered in subsequent WoE steps to accept or reject the hypothesis question(s). Brunk (2007) has noted, in scientific enquiry the "*standard of proof*" required to reject a null hypothesis is typically high (i.e., "greater than 95% confidence" or equivalently "beyond reasonable doubt"). However in risk assessment, as in some legal contexts, the "standard of proof" may be lower – e.g., "a preponderance of evidence" or "greater than 50% confidence" (Krimsky 2005). The level of confidence required to accept or reject a hypothesis formed during problem formulation can be associated with the acceptable level of uncertainty given the context in which the question is being asked. Acceptance of the level of uncertainty is directly linked to the protection goal(s) outlined during problem formulation and may differ between the human receptor (single species, higher specificity often required) and ecological receptors (multiple species, lower specificity is inherent). It may also vary according to the level or tier of evaluation undertaken depending on the decision context. It should also be noted that in chemical assessment, uncertainty is often bidirectional, that is, a decision may be overly conservative or not conservative enough as compared to an optimally informed decision. For example, during chemical prioritisation, a higher level of uncertainty can be acceptable if further risk evaluation is planned for prioritised chemicals. However, when "deprioritising chemicals" (which may not receive further risk evaluation), it is ideal to achieve a balanced uncertainty, with a low chance of both false positives and negatives (i.e., good specificity and good sensitivity) to ensure confidence in prioritisation results and maximize efficiencies. Ultimately, the acceptance level for uncertainty is related to the decision context and the protection goal and thus needs to be defined upfront at the problem formulation stage.

Examples of problem formulation questions within regulatory processes include:

- Is there sufficient evidence to conclude on the presence or absence of hazardous properties such as endocrine disrupting properties, aquatic toxicity, carcinogenicity, mutagenicity or reproductive and developmental toxicity?
- Does the substance fulfill relevant criteria for globally harmonized system (GHS) classification?
- Is there sufficient evidence on physical-chemical properties such as persistence and bioaccumulation to fulfill relevant criteria?
- Is there sufficient evidence to conclude that a mechanism of action observed in experimental animals is (not) relevant for humans?
- Is there sufficient evidence to accept the read-across of hazardous properties for similar chemicals for a specific endpoint or property?
- Is there sufficient evidence of potential exposure to indicate potential risk to human health or the environment?

Problem formulation becomes a key element when considering the regulatory use of emerging or alternative data. A roadmap for Risk Assessment in the 21st Century (Embry et al. 2014) available at https://risk21.org/webtool-user-guide/ provides guidance for establishing a problem formulation and is useful when integrating alternative data with traditional data sources. The National Academy of Sciences, Engineering, and Medicine also (NAS) provide useful problem formulation guidance including on substances with the

potential for endocrine disruption (NAS 2017). Balls et al. (2006) include WoE guidance for test methods and testing strategies based on output from a workshop sponsored by the European Centre for the Validation of Alternative Methods (ECVAM).

3.3. Evidence Collection

Evidence is gathered and assembled into "lines of evidence" (LoE) relevant for addressing the hypotheses under the problem formulation. As noted by Linkov et al. (2009), assembling the lines of evidence is not a common feature of older approaches to WoE, but is more common in recent approaches (e.g., Hull and Swanson 2006; Suter and Cromier 2011; Hoke and Clarkson 2014, Hall 2017, EFSA 2017, ECHA 2017, GoC 2017, Martin et al. 2018, Balls et al. 2018). Assembling the lines of evidence is a key transparency element of WoE that allows stakeholders to ensure all relevant lines of evidence have been considered and helps to identify relevant information gaps. Different types of evidence from multiple sources may be gathered or submitted and considered in context of "all" available evidence to date. The full spectrum of sources and types of evidence may include: company and/or third party generated studies of a proprietary nature, peer-reviewed published scientific literature, expert opinion reports, decisions and analysis reports from regulatory authorities, incident reports, randomised controlled clinical trials, adverse reactions submitted to regulatory authorities, and unpublished data. The application of a systematic review process (e.g., Beronius and Vandenberg 2016) is a useful standardised approach to document sources of information collected and cited as evidence for evaluation.

In principle, all direct lines of evidence (e.g., measured endpoints or properties) and indirect lines of evidence (e.g., regulatory decisions in other jurisdictions) can initially be considered as lines of evidence. As noted by Suter and Cormier (2011), evidence can be assembled into categories that share a common line of inquiry (e.g., by endpoint or property), share common qualities in a weighting scheme (e.g., common mechanism of action) or come from common sources (e.g., laboratory data). For example, toxicological evidence can be collected or generated for various levels of biological organisation such as:

- *In situ* (field or epidemiological population-based) evidence
- *In vivo* (living organism) evidence
- *In vitro* (e.g. cell and tissue based) evidence
- *In chemico* (chemical reactivity) evidence
- *In silico* (computer modeling) evidence

In practice, only some of the data collected will be considered relevant to the hypothesis formed under problem formulation and will depend on the context of the question being asked. For example, following a problem formulation for hazard assessment, data may be collected for specific endpoints. If substance Y is suspected of being endocrine active, lines of evidence can be assembled to establish that an adverse outcome is the result of cascading biological effects consistent with a plausible molecular initiating event. When the scope is larger, for example, for an overall conclusion of risk, more lines of evidence are needed to be assembled and organised in a manner such that they tell "the story" of the risk outcome (e.g., from physico-chemical properties to exposure and effects). Indicating the criteria for

inclusion or exclusion of the information/evidence when assembling lines of evidence will also improve transparency and reproducibility of the WoE.

The collection of information is typically comprised of the following:

a. **Collection Strategy (data generation or identification)**
b. **Documentation of the collection strategy**
c. **Selection and grouping of the evidence for use in the WoE**

To increase transparency in the decision making process it is recommended that the search strategy be reported and include the key words and sources searched. For example, Table 1 below summarises limited examples of typical sources of information that are relevant for hazard assessment in a regulatory setting under REACH (Registration, Evaluation, Authorisation and Restriction of Chemicals in European Union). Additional lists of sources for information collection are available within ECHA Guidance (IR/CSA R2-R4), and as an example of data collection using an IATA approach for endocrine disrupting substances, by ECHA and EFSA (2018).

Table 1. Examples of accessible sources of information for identification of relevant information for hazard assessment

Sources	Type of information	Link
eChemPortal	Experimental studies (in vivo, in vitro) and regulatory decisions	https://www.echemportal.org/
QSAR Toolbox	Experimental studies, QSAR models, read-across generator	https://qsartoolbox.org/
Science Direct	Scientific literature	https://www.sciencedirect.com/
Pubmed	Scientific literature	https://www.ncbi.nlm.nih.gov/pubmed/

It is often the case that evidence for a line of investigation important to the problem formulation has not been measured and cannot be estimated. When quantifying human exposure via the dermal route, for example, if no measured data are available for a substance's rate of absorption across human skin, the dermal uptake rate remains unknown. Dermal absorption, therefore, cannot be considered a measured line of evidence for the WoE analysis and it remains a critical unknown affecting the strength of the overall evidence for human dermal exposure. Unmeasured, but highly relevant lines of evidence to the hypothesis are therefore considered remaining uncertainties, which should be made known when reporting the outcomes of the WoE. Remaining uncertainty has a direct impact on the "confidence" underpinning chemical assessment conclusions. It can provide reasoning for further data requisition (see section 3.6).

3.4. Evidence Evaluation

Perhaps the most critical of the key elements in a WoE approach is evaluation of the available evidence because it provides the outcomes to determine the weight given to evidence. Regulatory and non-regulatory agencies often use different terms for describing evidence evaluation. However, in all cases the aim of evidence evaluation is to determine the inherent quality, usefulness and completeness of the available data. For the purposes of this document, this is described using two key terms: reliability and relevance. These terms are defined in the sections below. When combined, these elements provide the strength of inference, that is, the ability to infer the likelihood of an outcome based on the available evidence. This also directly affects the degree of extrapolations or assumptions used in chemical assessments. The application of a systematic review approach to evidence evaluation (Beronius and Vandenberg 2016; NAS 2017) may be beneficial at this point depending on the regulatory context and practicality. A systematic review approach also offers additional benefit in evaluating variability from biological differences and reduces bias in weighing evidence consideration.

3.4.1. Determining Data Reliability

There are many definitions of reliability as it pertains to WoE, as outlined in Moermond et al. (2017), as well as criteria for assessing the quality of individual test data (e.g., Hall et al. 2017, SCHEER 2018). Most definitions and terminology relate to determining the confidence of a study or data set. For example, the definition of reliability in USEPA (2016) for ecological assessment is "a property of evidence determined by the degree to which it has quality or other attributes that inspire confidence". This document adopts the concept of the USEPA definition in that greater reliability in available evidence inspires confidence and in turn provides greater strength of inference. Reliability can refer to the assessment of individual studies as well as entire data sets and can scale from very low to very high. In this document, reliability refers to the quality, sufficiency (quantity) and consistency of the data evaluated under a LoE, and considers the associated contribution of uncertainty of each of these parameters. These key concepts are described in more detail in the following sections.

3.4.1.1 Data Quality

There are numerous approaches to assess data quality and it is not the purpose of this document to recommend one approach over another. A selected approach for data quality assessment should be applied consistently across all lines of evidence. Selected examples of data quality evaluation criteria are given below. Many regulatory agencies internationally use the Klimisch criteria (Klimisch et al. 1997) to assess the reliability of toxicological data, including the European Chemicals Agency (ECHA) (for the assessment of reliability) and the US Environmental Protection Agency (EPA). In this approach, available studies are sorted into four categories: 1) Reliable without restrictions, 2) Reliable with restrictions, 3) Not reliable and 4) Not assignable. Other bodies that support policy-making such as SCHEER have similar criteria for assessing the quality of the data (SCHEER 2018):

a. Studies are considered to be of good scientific quality if they are appropriately designed, conducted, reported, and use a valid (which in many cases means: validated according to internationally accepted guidelines) methodology;

b. Studies of adequate/utilisable scientific quality have some significant limitations, yet are scientifically acceptable. These have some important deficiencies in the design and/or conduct and/or the reporting of the experimental findings;

c. Studies of inadequate scientific quality have serious flaws or concerns with the design or conduct of the study;

d. Studies that are not assignable have insufficient detail to assess their reliability.

However, in many cases, regulatory agencies may be assessing substances with few available data requiring a more qualitative review of data quality (Beronius and Vandenberg, 2016). ECHA (ECHA 2017) also acknowledges this in the ECHA Guidance IR/CSA R.4, "the scoring of information should not exclude all unreliable data from further consideration by expert judgment because of possible pertinence of these data related to the evaluated endpoints". The Hazardous Substances Advisory Committee (HSAC, 2015), in the United Kingdom, considers the quality of the data through examining a number of related concepts that lend themselves to determining the confidence of the evidence. HSAC examines the transparency of the aims of a study, or ability to appropriately test or falsify the hypothesis. The Committee recommends that studies be weighed based on having clear and transparently recorded methodology and that known biases have been addressed and acknowledged to the extent possible.

The process and level of complexity for determining data quality is context dependent and linked to the acceptable level of uncertainty set during the problem formulation. Lack of data of sufficient quality ultimately affects the reliability of the available evidence and results in lower inferential strength. It is therefore necessary to transparently communicate how data were evaluated for quality to allow the process to be reproducible and so that stakeholders may understand the impact of data quality on reliability.

3.4.1.2 Data Sufficiency and Consistency

Data sufficiency, for the purposes of this document, refers to the concept of enough data to address the question from the problem formulation. When there is sufficient data, there is an acceptable number of data points or studies to address the hypothesis question set within the context of the problem formulation. The degree of data sufficiency is context dependent, and is based on the accepted level of uncertainty set during problem formulation (section 3.2). The term "adequacy" is also used by some agencies when referring to data quantity, but it can also refer to aspects of data relevance as well. In this document, the term adequacy is discussed further under evidence relevance (section 3.4.3).

The term "completeness" can also be associated with sufficiency when referring to available data. However, completeness more often refers to a prescribed regulatory requirement for data submission (e.g., new substances regulations in Canada, REACH dossier requirements), which may not provide sufficient data for a WoE.

Data consistency refers to the degree of consensus or corroboration among available data within a line of evidence (e.g., developmental effects). Data sufficiency and consistency are interrelated. A higher number of acceptable studies for a line of evidence will allow for a greater understanding of natural variation. For example, if only one data point is available for an endpoint or property, there is no knowledge of the variability, sensitivity or specificity of the endpoint or property (i.e., the direction of uncertainty is unknown). Thus, higher consistency may occur as a result of lower variability and/or uncertainty. Conversely, data may be inconsistent due to high variability or because of a plausible

methodological or mechanistic reason (e.g., different species sensitivity). Reasons for consistency or lack thereof should be investigated and reported.

In summary, the highest level of data reliability is achieved when, given the context of the question, there is a sufficient amount of quality data that are internally consistent and can be shown to support the hypothesis with a plausible explanation (see also evidence integration in section 3.6).

3.4.2. Integrating Uncertainty

Before discussing the integration of uncertainty into WoE, it is important to understand what uncertainty encompasses in the context of this document. There are varying definitions and viewpoints of what uncertainty involves depending on the agency, program and the assessment of human versus ecological receptors. For example, the subject of uncertainty characterisation for hazard assessment is discussed in detail for human health in the guidance document from the WHO-IPCS (WHO 2018) and for human health exposure assessment (WHO 2008). According to WHO/IPCS (WHO 2018) uncertainty is defined as

> *"imperfect knowledge concerning the present or future state of an organism, system, or (sub)population under consideration". In relation to the specific topic of this monograph, it can be further defined as lack of knowledge regarding the "true" value of a quantity, lack of knowledge regarding which of several alternative model representations best describes a system of interest, or lack of knowledge regarding which probability distribution function and its specification should represent a quantity of interest."*

The above WHO/IPCS definition generally describes "true" uncertainty from lack of knowledge such as that generated from significant data gaps or suitable model. However, uncertainty has been used to refer to other sources of potential error such as those described by the USEPA (1998) below:

> *"Sources of uncertainty that are encountered when evaluating information include unclear communication of the data or its manipulation and errors in the information itself (descriptive errors)."*

Quantitative uncertainty analysis is conducted to address potential model bias and address variability in data sets, typically using statistical techniques such as probabilistic analysis (e.g., Monte Carlo analysis) (e.g., USEPA 1994). However, the majority of chemical risk assessments conducted for commercial chemicals typically involve a qualitative approach to uncertainty analysis (e.g., safety factors, qualitative confidence statements) simply because of the lack of adequate data to support statistical methods.

Therefore, for the purposes of this document, uncertainty shall be generalised to include data variability and that created from lack or limitations of knowledge (unknowns) from all sources. A useful generic definition in this regard is provided by EFSA (2016) which states that

> *"uncertainty refers to all types of limitations in the knowledge available to assessors at the time an assessment is conducted and within the time and resources agreed for the assessment."*

It is likely that the majority of uncertainty assessment will be qualitative given the high degree of data paucity for the majority of commercial chemicals (the exception is for regulatory programs with prescribed data requirements, such as some pesticides or new

industrial chemicals, that may provide sufficient data to conduct quantitative uncertainty analysis). In situations where quantitative uncertainty analysis is conducted using quality data and a point of departure (PoD) is selected from the distribution, the PoD is inherently more reliable (e.g., vs single values) because it encompasses data variability, regardless of where on the distribution it is taken from (e.g., 50th or 95th percentile). It does not, however, account for potential unknown methodological issues in studies. The choice of PoD selection is context specific and can range for less conservative to highly conservative. Figure 2, for example, illustrates the impact of variability on the point of departure. At the extremes of a data distribution, there is less chance the "truth" is represented (denoted in Figure 2 as the average). When values from the extremes of distributions are consistently used as "conservative estimates" in an assessment (e.g., in exposure modeling), in order to have a higher certainty in 'capturing' the "truth" within the PoD, the error is exponentially compounded such that unrealistic values with high uncertainty may result.

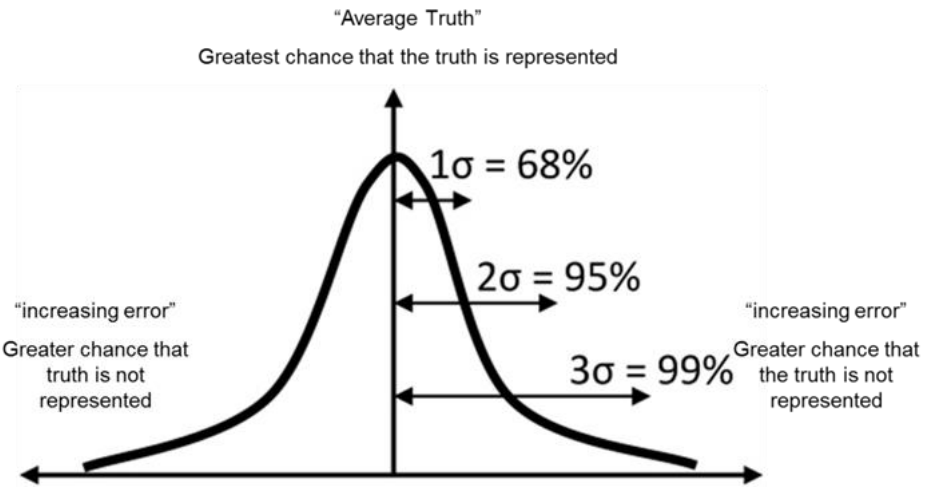

Figure 2: Impact of data variability on the point of departure

The assessment of uncertainty is essential for determining the weight that evidence should receive and has a direct impact on the reliability or "overall confidence" in the data used for a LoE. Many regulatory and non-regulatory bodies include a separate discussion of uncertainty outside of, but not always linked to, a WoE analysis. SCHEER (2018), however, provide a useful description of how uncertainty can be considered within weight of evidence:

> *"In SCHEER Opinions, the uncertainty should be expressed in relation to the question asked by risk managers and decision makers and should be appropriate regarding the quality and quantity of information or data available to the Committee."*

As proposed in this document, the above statement from SCHEER suggests that uncertainty can be integrated into decision making for individual lines of evidence based on an evaluation of available data, but importantly related to the hypothesis question. When evidence is gathered with higher levels of uncertainty, confidence that the data describe an outcome is lower and ideally there should be less reliance on these data when weighing the

evidence (see section 3.5)[3]. Conversely, when uncertainty is determined to be low, there is a higher level of confidence that the data describe an outcome and more reliance on these data should be given when weighing evidence.

During the determination of data reliability outlined in previous sections, uncertainty is implicitly considered when evaluating the available data (e.g., under data quality, sufficiency and consistency) and it is possible to integrate this when evaluating data reliability such that the "score" given to reliability reflects both data variability and, if possible, lack of knowledge (data gaps). How uncertainty analysis is conducted and the degree to which it is conducted will be case dependent and can vary in terms of complexity depending on the goals and the acceptable level of uncertainty from the problem formulation. SCHEER (2018) and read-across case studies under the OECD IATA initiative (OECD 2018b) provide examples of qualitative and quantitative approaches for expressing the results of uncertainty analysis as well as the impact of uncertainty on a line of evidence.

Finally, it is often not possible to undertake the evaluation of "unknowns" or true uncertainties generated as a result of data gaps or other limitations. These are remaining uncertainties that should be communicated along with assessment outcomes (see section 3.6) noting that the impact of remaining uncertainties on the hypothesis question posed in the problem formulation will vary from not significant to very significant. For example, lack of soil toxicity data may have no impact on an assessment outcome in circumstances where a highly water-soluble chemical is only released to water. Whereas new aquatic toxicity data generated to fill data gaps for key trophic levels may impact and alter an assessment outcome.

3.4.3. Determining Relevance

Much like reliability, the term relevance has various meanings in the WoE literature, but more commonly refers to the appropriateness or degree of correspondence of evidence for the hypothesis. When referring to ecotoxicological effects in decision-making, Hall et al. (2017), for example, note that

> *"Relevance assessment can be divided into 3 categories: regulatory relevance (fit for purpose to the regulatory framework, protection goal, and assessment endpoints), biological relevance (e.g., related to the test species, life stage, endpoints, and response function), and exposure relevance (e.g., related to test substance, exposure route and exposure dynamics)."*

ECHA (2017) defines relevance as covering the extent to which data and tests are *appropriate* for a particular hazard identification or risk characterisation, which is conceptually similar to that provided in USEPA (2016):

> **"***A property of a piece or type of evidence that expresses the degree of correspondence between the evidence and the assessment endpoint to which it is applied."*

According to OECD (2018a) the term *relevance describes whether a procedure is meaningful and useful for a particular purpose.*

[3] Noting that in many cases uncertain data may be the only data available for a LoE and judgement must be used to accept or reject the data depending on the goals of the problem formulation

However, the term "adequacy" is also used when referring to the appropriateness or relevance of data, but this term can also refer to the sufficiency of data in some literature and by some agencies. Consequently, for the purposes of this document, the term relevance is preferred and refers to the appropriateness or degree of correspondence of evidence to the hypothesis question outlined in the problem formulation. The definition provided in Hall et al. (2017) is also critical for chemical assessment because relevance can refer to both the relevance of regulatory and scientific evidence (biological and exposure) for the hypothesis question posed during the problem formulation. For example, the octanol-water partition coefficient ($logK_{ow}$) is relevant for describing bioaccumulation in aquatic receptors. However, in most jurisdictions it has lower regulatory and scientific relevance compared with laboratory bioconcentration factor (BCF) data.

Human and ecological chemical assessment often requires that data be extrapolated beyond their designed intent (e.g., laboratory to field, *in vitro* to *in vivo*). In such cases, the degree of extrapolation directly affects the relevance of the data. For example, application of aquatic toxicity data to assess the hazards for soil dwelling organisms is often performed for ecological risk assessment, but the scientific relevance of aquatic data for soil organisms might be questionable on a case by case basis as the data are being applied outside of their designed intent (aquatic laboratory studies with no soil). Similarly, from a human health perspective, consideration of data for dermal toxicity may not be appropriate when exposures are expected to occur via an oral or inhalation route or at least not without significant extrapolation.

Even when data are applied within the designed intent (appropriate), relevance may also be called into question. For example, laboratory water-based test data for superhydrophobic substances requiring the use of solubilising agents to achieve exposures may not be fit for purpose when considering actual exposures in the environment or to humans. In this case, considerable uncertainty arises when extrapolating to real world exposures, both internal and external to the organism. This issue is mentioned by the HSAC in the UK (HSAC 2015):

> *"one of the caveats to the use of any study is the relationship of the broader environment or human condition to the experimental conditions described, so it is important that the conditions are relevant to the problem under investigation."*

The determination of relevant versus non-relevant data will be context specific and care should be taken when determining non-relevant data. Beronius and Vandeberg (2016), for example, point out that much of the research on endocrine disruption would be considered of lower quality according to the Klimisch et al. (1997) criteria, but that these criteria would miss key information that might be relevant to a health endpoint. SCENIHR (2012) suggests a qualitative assessment of the relevance of data in which relevance is assessed as

> *"direct relevance, addressing the agent (stressor), model and outcome of interest; indirect relevance, addressing a related agent (stressor), model or outcome of interest or insufficient relevance, not useful for the specific risk assessment being conducted*."

In summary, highly relevant lines of evidence are those that correspond closely to the hypothesis generated in the problem formulation (i.e., have high regulatory and scientific impact and require lower extrapolation).

3.5. Evidence Weighing for Lines of Evidence

In the comprehensive review of WoE approaches by Linkov et al. (2009), the authors concluded that a quantitative approach to WoE was preferred. However, others such as Weed (2005), Suter and Cormier (2011) or Rhomberg et al. (2013), note that from their review of approaches, more often a qualitative approach is used to weigh evidence. Suter and Cormier (2011) highlight that numerical scores can be used to weigh lines of evidence, but "*those scores are numerical but not quantitative*". Thus, there is no advantage to a numerical system; in fact it may appear to be more "*rigorous*" than is possible (Suter and Cormier, 2011). Regardless of the approach used, the above authors all conclude that one of the most important aspects of weighing evidence is transparency of the approach. It follows that the goal of this element is to develop a transparent approach to assigning a weight to each line of evidence assembled based on the combined influences of reliability and relevance. Scoring can involve symbols or numerical values that are related to a qualitative description or rule, but these are not essential if a fully qualitative approach is preferred (e.g., low, moderate, high). Hall et al. (2017) provide a good example of this idea for scoring ecotoxicological data. Determining a score is judgement based and context dependent and thus absolute rules or criteria to judge the level of reliability and relevance are not provided here and should be developed by individual agencies. The number of scoring descriptors is also subjective and can be selected according to the context of the problem. Table 2 below provides a simplistic qualitative example of a scoring scheme for reliability and relevance using three descriptors (low, moderate, high) where weight assigned is a combination of the score for reliability and relevance for each line of evidence (based on Bonnell 2011; ECCC 2016; GoC 2017).

Table 2. Example Qualitative Scoring Scheme for Determining Weight for Each Line of Evidence

Line of evidence	Reliability	Relevance	Weight assigned
LoE 1	[low, moderate; high]	[low, moderate, high]	[low; low to moderate; moderate; moderate to high; high]
LoE 2			
LoE 3			
etc.			

In Table 2, there are a total of nine possible weight outcomes (i.e., 3x3 matrix). This can be simplified to five outcomes because some outcomes are repetitive. To provide the best transparency, it may be important to document all lines of evidence considered, not just those considered more relevant. In such cases, a scoring scheme can include a "null" weight, that is a descriptor to indicate that a line of evidence is not of sufficient reliability and/or relevance to be used for the weight of evidence (i.e., is less than low in Table 2 example), but was examined nonetheless. This is explained by several agencies and organisations (ECHA, 2017; HSAC, 2015; NAS 2017; SCHEER 2018), who recommend documenting all evidence viewed, whether used or not. However, from a practical standpoint, there are often inadequate resources to formally tally all of the evidence reviewed, but not used. At minimum, it is recommended that agencies be able to justify why certain evidence has or has not been incorporated into the WoE.

Regardless of the simplicity or complexity of the approach used for scoring and WoE, transparency and tracking of outcomes should be communicated, noting that a more complex approach will make this more difficult.

3.6. Evidence Integration and Reporting

The last step of the Weight of Evidence is the integration and final reporting of the process. According to SCHEER (2018):

"*The main objectives of the integration procedure are:*

- *To check the consistency of different lines of evidence, that is the extent to which the contributions of different lines of evidence drawing a specific conclusion are compatible (EFSA, 2017)*
- *In case of inconsistencies, to try to understand and explain the reasons for them, possibly deciding if more than one answer to the formulated problem is plausible*
- *To reject cases of unacceptable or inconsistent outliers to conclude on the WoE based on consistency and quality"*

Additional guidance on performing the integration of evidence is provided in SCHEER (2018).

The two key factors that need to be considered during the integration of evidence are the coherence (or in other words consistency) of the lines of evidence for a property/endpoint and the relationship of the collected evidence with the WoE outcome(s).

This can be done by examining the plausibility and causality of collected evidence as it relates to WoE conclusions. This process is inherent during the evaluation of evidence for hazard assessment, but at this stage all evidence for a hypothesis should be examined in a coherent fashion to ensure outcomes are mechanistically supported.

Plausibility is one of the classic Bradford Hill (Hill 1965) considerations for determining the confidence in the evidence. It is defined as "the likelihood a known mechanism explains causality". As Hill (1965) noted, however, this explanation will depend on the current level of understanding of the phenomenon. Causality is the relationship between cause and effect. It is also one of the key drivers of the Bradford Hill criteria and is used in the WHO/ IPCS Mode of Action Framework (Meek et al. 2013). Causality is integrally related to temporality – determining that the exposure to the substance occurred prior to the effect. In the absence of causality, the WoE to support or refute the hypothesis becomes much more difficult, but not impossible. In chemical assessment, for example, if there is no plausible mechanistic explanation for the hypothesis that observed reproductive effects are the result of disruption of the endocrine system via substance X binding to the estrogen receptor (ER), other competing hypotheses become possible. Effects could be the result of another receptor mediated process such as interactions with thyroid or androgen receptors or of a non-endocrine mechanism. Likewise, should there be multiple mechanisms of action that could explain an outcome, more confidence should be put to those that are more plausible.

Evidence can be combined in a tabular format for reporting purposes to add clarity when many lines of evidence are assembled. The weighting approach and results should also be included for transparency. A narrative should follow the table to provide a verbal summary of:

- Each line of evidence and the rationale for its weight or null weight (outcomes of the evaluation of the level of reliability and relevance).
- The coherence of the evidence. That is, how well individual lines of evidence corroborate each other and with the hypothesis (e.g., Figure 3). Greater coherence among lines of evidence with higher weighting yields greater strength of evidence and should form the basis for accepting or rejecting the hypothesis.
- The conclusion - acceptance or rejection of the hypothesis according to the strength of evidence and where relevant, possible alternative explanations of the results.
- Identification of remaining uncertainty generated from gaps in knowledge and the sensitivity of the hypothesis to remaining uncertainty. Understanding the sensitivity of the hypothesis to remaining uncertainty is particularly useful when determining if new data will be of benefit to the problem formulation as well as the overall uncertainty of a LoE. That is, the hypothesis or LoE may be sensitive, not sensitive or sensitivity is neutral or unknown. SCHEER (2018) provide guidance that can be consulted for expressing uncertainty at this stage of the WoE particularly when indicating the direction of the uncertainty. Acquisition of further data, if necessary, could therefore proceed for LoE(s) in which uncertainty has the highest impact on the hypothesis under the problem formulation.

In addition, WoE reporting templates have been published by certain agencies that can be considered or modified to accommodate this concept (e.g., ECHA 2017, EFSA 2017, ECCC 2016).

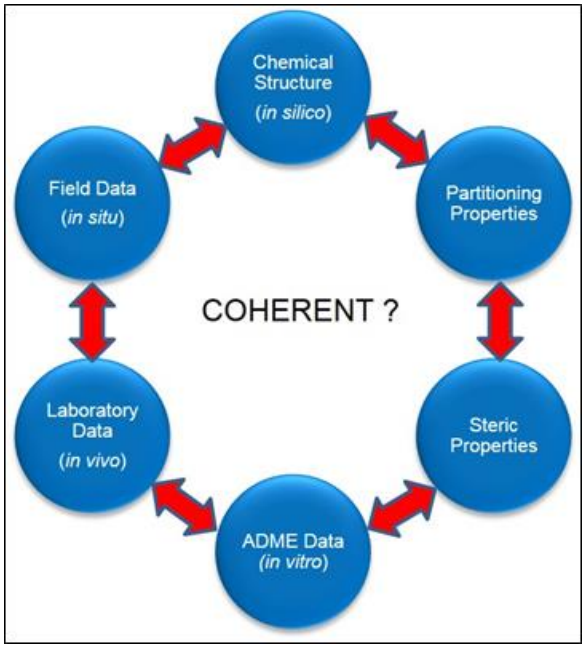

Figure 3: Possible lines of evidence for bioaccumulation or aquatic toxicity; (ADME: Absorption, Distribution, Metabolism, Elimination).

4. CONCLUSIONS

This document presents guiding principles that can be used to construct approaches to WoE for chemical evaluation. The principles are intended to be endpoint and receptor agnostic and therefore can be used for both human and ecological purposes. They can be viewed as principles for "good WoE practice". Based on a review of approaches from the literature, this document also describes key elements for constructing or revising approaches to WoE for chemical evaluation. They are described in a practical manner to facilitate the implementation of WoE, particularly in cases where a WoE approach does not yet exist. The data landscape for chemical prioritisation and risk assessment continues to change and now includes multiple sources of non-traditional data. WoE approaches will necessarily need to continue to evolve and adapt to meet this data context, particularly as it relates to 21st century risk science (including exposure science), IATA and development and support for AOPs. A case study on the integration and application of alternative data for decision-making could provide an ideal example to demonstrate the principles and elements described in this document.

Most literature on WoE agrees that having a clear and transparent process for all stakeholders to track decision-making is a critical, and necessary, component of a WoE approach. Central to this communication is the treatment of uncertainty. The acceptance and impact of uncertainty is a primary factor that influences all elements of WoE. It is an essential consideration at the very beginning of WoE formulation (before and during problem formulation), during evidence evaluation and is accounted for when weighing evidence. Its treatment should therefore be clearly documented, particularly when assessing data reliability. Finally, most aspects of WoE are based on judgement which means that the potential for bias is inherent in the process. The degree of bias will again depend on context and the protection goals sought for the chemical evaluation. However, if both the guiding principles and key elements captured in this document are adhered to, a consistent, clear and transparent delivery of evidence can follow. All stakeholders can then trace decision-making to determine reasons for bias and determine when bias becomes unreasonable.

5. REFERENCES

Ågerstrand M and Beronius A. 2016. Weight of evidence evaluation and systematic review in EU chemical risk assessment: Foundation is laid but guidance is needed. Environment International, 92-93: 590-596. https://doi.org/10.1016/j.envint.2015.10.008

ANSES, 2016. Evaluation du poids des preuves a l'Anses: revue critique de la litterature et recommandations a l'etape d'identification des dangers. Rapport d'expertise collective. Saisine 2015-SA-0089, 116 pp.

Balls M, Amcoff P, Bremer S, Casati S, Coecke S, Clothier R, Combes R, Corvi R, Curren R, Eskes C, Fentem J, Gribaldo L, Halder M, Hartung T, Hoffmann S, Schectman L, Scott L, Spielmann H, Stokes W, Tice R, Wagner D, Zuang V. 2006. The principles of weight of evidence validation of test methods and testing strategies. The report and recommendations of ECVAM workshop 58. Altern Lab Anim. 34(6): 603-620.

Beronius A, Molander L, Rudén C and Hanberg A. 2014. Facilitating the use of non-standard in vivo studies in health risk assessment of chemicals: a proposal to improve evaluation criteria and reporting. Journal of Applied Toxicology, 34, 607-617. https://doi.org/10.1002/jat.2991

Beronius, A. and Vandenberg L.N. 2016. Using systematic reviews for hazard and risk assessment of endocrine disrupting chemicals. Rev Endocr Metab Disord, 16(4): 273-287. doi: 10.1007/s11154-016-9334-7

Bonnell, M. 2011. Weighing Evidence for Bioaccumulation. SETAC North America 32nd Annual Meeting. Society of Environmental Toxicology and Chemistry, Pensacola, Florida.

Borgert CJ, Mihaich EM, Ortego LS, Bentley KS, Holmes CM, Levine SL, Becker RA. 2011. Hypothesis-driven weight of evidence framework for evaluating data within the US EPA's Endocrine Disruptor Screening Program. Regul Toxicol Pharmacol. 61(2):185-91. https://doi.org/10.1016/j.yrtph.2011.07.007

Brunk, CG. 2007. Science and the precautionary principle. PowerPoint presentation for Workshop on Risk Analysis, Ottawa, Ontario, June 6-8, 2007.

Collier ZA, Gust KA, Gonzalez-Morales B, Gong P, Wilbanks MS, Linkov I and Perkins EJ. 2016. A weight of evidence assessment approach for adverse outcome pathways. Regulatory Toxicology and Pharmacology, 75, 46–57. https://doi.org/10.1016/j.yrtph.2015.12.014

ECCC (Environment and Climate Change Canada). 2016. Draft State of the Science Report Certain Organic Flame Retardants Substance Grouping benzoic acid, 2,3,4,5-tetrabromo-, 2-ethylhexyl ester (TBB) and 1,2 benzenedicarboxylic acid, 3,4,5,6-tetrabromo-, bis(2-ethylhexyl) ester (TBPH). http://www.ec.gc.ca/ese-ees/default.asp?lang=En&n=844D1EBA-1#toc93 (last accessed Feb 2018).

ECHA and EFSA. 2018. (European Chemicals Agency) and EFSA (European Food Safety Authority) with the technical support of the Joint Research Centre (JRC), Andersson N, Arena M, Auteri D,Barmaz S, Grignard E, Kienzler A, Lepper P, Lostia AM, Munn S, Parra Morte JM, Pellizzato F, Tarazona J,Terron A and Van der Linden S. Guidance for the identification of endocrine disruptors in the context of Regulations (EU) No 528/2012

and (EC) No 1107/2009. EFSA Journal 2018;16(6):5311,135 pp.https://doi.org/10.2903/j.efsa.2018.5311. ECHA-18-G-01-EN.

ECHA (European Chemicals Agency) 2017. Weight of Evidence/ Uncertainty in Hazard Assessment, Helsinki, Finland, https://echa.europa.eu/support/guidance-on-reach-and-clp-implementation/formats

ECHA (European Chemicals Agency) 2013 Guidance (IR/CSA R2-R4): IR/CSR Guidance on Information Requirements and Chemical Safety Assessment, Chapter R.2 Overall framework for meeting the information requirements on intrinsic properties of substances, Guidance on collection of available information (Chapter R.3), Evaluation of information (Chapter R.4)

EFSA (European Food Safety Authority). 2016. European Food Safety Authority Guidance on Uncertainty in EFSA Scientific Assessment, Parma, Italy, https://www.efsa.europa.eu/sites/default/files/160321DraftGDUncertaintyInScientificAssessment.pdf

EFSA (European Food Safety Authority). 2017. Guidance on the assessment of the biological relevance of data in scientific assessments. Scientific Committee. European Food Safety Authority Journal, Place, Country, Vol. 15/8 pp. 4970.

EFSA Scientific Committee, Hardy A, Benford D, Halldorsson T, Jeger MJ, Knutsen HK, More S, Naegeli H, Noteborn H, Ockleford C, Ricci A, Rychen G, Schlatter JR, Silano V, Solecki R, Turck D, Benfenati E, Chaudhry QM, Craig P, Frampton G, Greiner M, Hart A, Hogstrand C, Lambre C, Luttik R, Makowski D, Siani A, Wahlstroem H, Aguilera J, Dorne J-L, Fernandez Dumont A, Hempen M, Valtuena Martınez S, Martino L, Smeraldi C, Terron A, Georgiadis N and Younes M, 2017. Scientific Opinion on the guidance on the use of the weight of evidence approach in scientific assessments. EFSA Journal 2017, 15(8): 4971, 69 pp. https://doi.org/10.2903/j.efsa.2017.4971

Embry MR, Bachman AN, Bell DR, Boobis AR, Cohen SM, Dellarco M, Dewhurst IC, Doerrer NG, Hines RN, Moretto A, et al. 2014. Risk assessment in the 21st century: Roadmap and matrix. Critical Reviews in Toxicology, 44: 6-16. https://doi.org/10.3109/10408444.2014.931924

GoC (Government of Canada). 2017. Application of weight of evidence and precaution in risk assessment. https://www.canada.ca/en/health-canada/services/chemical-substances/fact-sheets/application-weight-of-evidence-precaution-risk-assessments.html (last accessed March 2018)

Good IJ, 1979. Studies in the History of Probability and Statistics. XXXVII A. M. Turing's Statistical Work in World War II. Biometrika, 66: 393-396.

Good IJ, 1985. Weight of Evidence: A Brief Survey. Bayesian Statistics, 2: 249-270.

Hall AT, Belanger SE, Guiney PD, Galay-Burgos M, Maack G, Stubblefield W and Martin O. 2017. New Approach to Weight-of-Evidence Assessment of Ecotoxicological Effects in Regulatory Decision-Making. Integrated Environmental Assessment and Management, 13 (4): 573-579. https://doi.org/10.1002/ieam.1936|

Health Canada. 2011. Weight of Evidence: General Principles and Current Applications at Health Canada. Retrieved December 22, 2017

Hill AB. 1965. The environment and disease: Association or causation? Proceedings of the Royal Society of Medicine. 58(5):295-300.

Hope BK and Clarkson JR. 2014. A strategy for using weight-of-evidence methods in ecological risk assessments. Hum and Ecol Risk Assess. 20(2): 1549-7860. https://doi.org/10.1080/10807039.2013.781849

HSAC (Hazardous Substances Advisory Committee). 2015. "Considering evidence: The approach taken by the Hazardous Substances Advisory Committee in the UK", United Kingdom: Hazardous Substances Advisory Committee, Environment International, Elsevier, place, Vol. 92-93, pp. 565-568.

Hull RN and Swanson S. 2006. Sequential analysis of lines of evidence-An advanced weight-of-evidence approach for ecological risk assessment. Integrated Environmental Assessment and Management 2(4): 302-311.

Klimish H-J, Andreae M, Tillmann U. 1997. A systematic approach for evaluating the quality of experimental toxicological and ecotoxicological data. Reg Toxicol Pharmacol 25(1): 1-5. https://doi.org/10.1006/rtph.1996.1076

Krimsky S. 2005. The weight of scientific evidence in policy and law. Amer J Public Health. 95(Supplement 1): S129-S136. doi: 10.2105/AJPH.2004.044727

Linkov I, Loney D, Cormier S, Satterstrom FK, Bridges T. 2009. Weight-of-evidence evaluation in environmental assessment: Review of qualitative and quantitative approaches. Sci Total Environ 407(19): 5199-5205. https://doi.org/10.1016/j.scitotenv.2009.05.004

Lutter R, Abbott L, Becker R, Borgert C, Bradley A, Charnley G, Dudley S, Felsot S, Golden N, Gray G, Juberg D, Mitchell M, Rachman N, Rhomberg L, Solomon K, Sundlof S, Willett K. 2015. Improving weight of evidence approaches to chemical evaluations. Risk Analysis 35(2): 186-192. https://doi.org/10.1111/risa.12277

Martin P., et al., Weight of Evidence for Hazard Identification: A Critical Review of the Literature. Environ Health Perspect. 2018 Jul; 126(7): 076001 https://doi.org/10.1289/EHP3067

Meek, M.E., A. Boobis, I. Cote, V. Dellarco, G. Fotakis, S. Munn, J. Seed and C. Vickers, 2013. New developments in the evolution and application of the WHO/IPCS framework on mode of action/species concordance analysis. Journal of Applied Toxicology 34(1): 1-18. https://doi.org/10.1002/jat.2949

Meek ME, Palermo CM, Bachman AN, North CM and Lewis RJ, 2014. Mode of action human relevance (species concordance) framework: evolution of the Bradford Hill considerations and comparative analysis of weight of evidence. Journal of Applied Toxicology, 34: 595-606. https://doi.org/10.1002/jat.2984

Menzie CA, Henning MH, Cura J, Finkelstein K, Gentile J, Maughan J, Mitchell D, Petron S, Potocki B, Svirsky S, Tyler P. 1996. Special report of the Massachusetts weight-of-evidence workgroup: A weight-of-evidence approach for evaluating ecological risks. Human Ecological Risk Assessment. 2: 277-304. https://doi.org/10.1080/10807039609383609

Moermond C, Beasley A, Breton R, Junghans M, Laskowski R, Solomon K, Zahner H. 2017. Assessing the reliability of ecotoxicological studies: An overview of current needs and approaches. Integr Environ Assess Manag. 13: 640-651. https://doi.org/10.1002/ieam.1870|

Morgan RL, Thayer KA, Bero L, Bruce N, Falck-Ytter Y, Ghersi D, Guyatt G, Hooijmans C, Langendam M, Mandrioli D, Mustafa RA, Rehfuess EA, Rooney AA, Shea B, Silbergeld EK, Sutton P, Wolfe MS, Woodruff TJ, Verbeek JH, Holloway AC, Santesso N and Schünemann HJ, 2016. GRADE: Assessing the quality of evidence in environmental and occupational health. Environment International, 92-93, 611-616. https://doi.org/10.1016/j.envint.2016.01.004.

National Research Council (U.S.) Committee on Improving Risk Analysis Approaches Used by the U.S. EPA, 2009. Science and decisions: Advancing risk assessment/ Committee on Improving Risk Analysis Approaches Used by the U.S. EPA, Board on Environmental Studies and Toxicology, Division on Earth and Life studies.

NAS (National Academies of Sciences, Engineering, and Medicine). 2017. Application of Systematic Review Methods in an Overall Strategy for Evaluating Low-Dose Toxicity from Endocrine Active Chemicals. Washington, DC: The National Academies Press. https://doi.org/10.17226/24758

OECD, 2017, Guidance on Grouping of Chemicals, second edition, Series on Testing and Assessment, No. 194., https://doi.org/10.1787/20777876

OECD, 2018a, Guidance Document on Good In Vitro Method Practices, Series on Testing and Assessment, No. 286., https://doi.org/10.1787/20777876

OECD, 2018b, Integrated Approaches to Testing and Assessment (IATA). http://www.oecd.org/chemicalsafety/risk-assessment/iata-integrated-approaches-to-testing-and-assessment.htm

Rhomberg LR, Goodman JE, Bailey LA, Prueitt RL, Beck NB, Bevan C, Honeycutt M, Kaminski NE, Paoli G, Pottenger LH, Scherer RW, Wise KC, Becker RA. 2013. A survey of frameworks for best practices in weight-of-evidence analyses. Critical Reviews in Toxicology. 43(9): 753-784. https://doi.org/10.3109/10408444.2013.832727

Rooney AA, Boyles AL, Wolfe MS, Bucher JR, and Thayer KA, 2014. "Systematic review and evidence integration for literature-based environmental health science assessments." Environmental Health Perspectives, 122, 711–718. https://doi.org/10.1289/ehp.1307972

SCENIHR (Scientific Committee on Emerging and Newly Identified Health Risks).2012. Memorandum on the use of scientific literature for human health risk assessment purposes - weighing of evidence and expression of uncertainty, European Commission, Brussels, Belgium.

SCHEER (Scientific Committee on Health, Environmental and Emerging Risks). 2018. Memorandum on weight of evidence and uncertainties Revision 2018. European Commission, Brussels, Belgium.

Suter II GW and Cormier SM. 2011. Why and how to combine evidence in environmental assessments: Weighing evidence and building cases. Sci Total Environ. 409: 1406-1417. https://doi.org/10.1016/j.scitotenv.2010.12.029

USEPA (United States Environmental Protection Agency). 1994. Use of Monte Carlo Simulation in Risk Assessments: Region 3 Technical Guidance Manual, Risk Assessment. EPA 100/B-03/001. https://www.epa.gov/risk/use-monte-carlo-simulation-risk-assessments

USEPA (United States Environmental Protection Agency). 2003. A Summary of General Assessment Factors for Evaluating the Quality of Scientific and Technical Information Prepared for the U.S. Environmental Protection Agency by members of the Assessment Factors Workgroup, a group of the EPA's Science Policy Council. EPA 100/B-03/001.

USEPA (United States Environmental Protection Agency). 2005. Guidelines for Carcinogen Risk Assessment. EPA/630/P-03/001B. Washington, DC. http://www.epa.gov/raf/publications/pdfs/CANCER GUIDELINES FINAL 3-25-05.pdf

USEPA (United States Environmental Protection Agency) Endocrine Disruptor Screening Programme. 2011. Weight-of-evidence Guidance Document, Available via https://www.epa.gov/endocrine-disruption/endocrine-disruptor-screening-program-documents (direct link to Guidance document: https://www.regulations.gov/document?D=EPA-HQ-OPPT-2010-0877-0021)

USEPA (United States Environmental Protection Agency). 2016. Weight of Evidence in Ecological Assessment. EPA/100/R-16/001. Washington, D.C. December.

Vermeire T, Aldenberg T, Buist H, Escher S, Mangelsdorf I, Paun_e E, Rorije E and Kroese D, 2013. OSIRIS, a quest for proof of principle for integrated testing strategies of chemicals for four human health endpoints. Regulatory Toxicology and Pharmacology 67, 136-145. https://doi.org/10.1016/j.yrtph.2013.01.007

Weed DL. 2005. Weight of evidence: A review of concepts and methods. Risk Analysis. 25(6): 1545-1557.

WHO (World Health Organization), 2008. Guidance Document on Characterizing and Communicating Uncertainty in Exposure Assessment, Harmonization Project Document No. 6

WHO (World Health Organization), 2009. Food Safety. Project to update the principles and methods for the assessment of chemicals in food. Principles and methods for the risk assessment of chemicals in food. EHC 240. ISBN 978 92 4 157240 8.

WHO (World Health Organization). 2018. Guidance document on evaluating and expressing uncertainty in hazard characterization– 2nd edition. Geneva: World Health Organization. https://apps.who.int/iris/bitstream/handle/10665/259858/9789241513548-eng.pdf;jsessionid=91DCD5290EAFFA050F8A2020F81CB9D3?sequence=1

WHO/IPCS 2014, Guidance on Evaluating and Expressing Uncertainty in Hazard Assessment is available in Harmonization Project Document No. 11 http://www.who.int/ipcs/methods/harmonization/areas/hazard_assessment/en/

Appendix 1: Selected Approaches and Methods from EFSA (2017)

Publication	Definitions or descriptions given for weight of evidence
Ågerstrand and Beronius (2016)	'In general terms, weight of evidence and systematic review are processes of summarising, synthesising and interpreting a body of evidence to draw conclusions … these processes differ from the traditional method for risk assessment by promoting the use and integration of information from all the available evidence instead of focusing on a single study'
ANSES (2016)	Defines weight of evidence as 'the structured synthesis of lines of evidence, possibly of varying quality, to determine the extent of support for hypotheses'
Beronius et al. (2014)	States that 'The meaning of weight of evidence intended here is the collective summary and evaluation of all existing evidence after a certain "weight" has been attributed to individual studies, e.g. by evaluating reliability and relevance'
Government of Canada (2017)	It is generally understood as a method for decision-making that involves consideration of multiple sources of information and lines of evidence. Using a WoE approach avoids relying solely on any one piece of information or line of evidence. Risk assessments of substances conducted under the Canadian Environmental Protection Act, 1999 (CEPA 1999) generally consider multiple lines of evidence to support a risk assessment conclusion (as defined in section 64 of CEPA 1999).
Collier et al. (2016)	Describes weight of evidence as '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, although (it) tends to be used inconsistently and vaguely across disciplines'
ECHA 2017; Weight of Evidence/ Uncertainty in Hazard Assessment	Weight of Evidence approach can be generally described as a stepwise process/approach of collecting evidence, assessing, integrating and weighing them to reach a conclusion on a particular problem formulation with (pre)defined degree of confidence.
EFSA (2013) [PPR Aquatic Ecol RA guidance doc]	States that the 'process of combining available lines of evidence to form an integrated conclusion or risk characterisation is frequently referred to as weight-of-evidence assessment. This term reflects the principle that the contribution of each line of evidence should be considered in proportion to its weight'
USEPA (2003)	Describes weight of evidence as an 'approach (which) considers all relevant information in an integrative assessment that takes into account the kinds of evidence available, the quality and quantity of the evidence, the strengths and limitations associated with each type of evidence and explains how the various types of evidence fit together'
Good (1979, 1985)	Defines weight of evidence as the logarithm of the ratio of the likelihood of a given hypothesis to the likelihood of an alternative hypothesis. This expression corresponds to the Bayes factor

Hope and Clarkson (2014)	Refers to Good for quantitative definition Describes weight of evidence as 'basically the process of considering the strengths and weaknesses of various pieces of information in order to inform a decision being made among competing alternatives'
Hull and Swanson (2006)	Describes weight of evidence as 'approaches (that) integrate various types of data (e.g., from chemistry, bioassay, and field studies) to make an overall conclusion of risk'
Linkov et al. (2009)	Defines weight of evidence as 'a framework for synthesising 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'
NRC (2009)	States that 'The phrase weight of evidence is used by EPA and other scientific bodies to describe the strength of the scientific inferences that can be drawn from a given body of evidence'.
Rhomberg et al. (2013)	Defines 'weight of evidence framework' as 'approaches that have been developed for taking the process from scoping an assessment and initial identification of relevant studies through the drawing of appropriate conclusions'
SCHEER 2018	A process of weighted integration of lines of evidence to determine the relative support for hypotheses or answers to a question
Schleier et al. (2015)	Describes weight of evidence as 'approaches in which multiple lines of evidence can be considered when estimating risk'
Suter and Cormier (2011)	'In sum, weighing evidence is a synthetic process that combines the information content of multiple weighted pieces of evidence. The information may be dichotomous (supports or not), quantitative values (e.g., an exposure or risk estimate), qualitative properties (e.g., large, medium or small), or a model. The weights that are applied to the information may express various properties that affect its credibility or importance and the weights themselves may be qualitative or quantitative. The combining of evidence may be a simple quantitative operation (e.g., weighted averages of concentration estimates) but more often involves difficult qualitative judgments'
Vermeire et al. (2013)	Implicit definition: 'The different and possibly contradictory information is weighted and the respective uncertainties taken into account in a weight of evidence approach'
Weed (2005)	Identifies three characteristic uses of the term weight of evidence: metaphorical, methodological and theoretical. Does not propose a definition but recommends that authors using weight of evidence should define the term and describe their methods
WHO (2009)	Defines weight of evidence as 'a process in which all of the evidence considered relevant for a risk assessment is evaluated and weighted'
Meek et al. (2014)	Uses the term weight of evidence but do not include an explicit definition or summary description
Rooney et al. (2014) (OHAT), Morgan et al. (2016) (GRADE)	These publications do not use the term weight of evidence but rather use related terms including 'evidence synthesis' and 'evidence integration'

Appendix 2: Glossary of Selected Terms

Adequacy (adequate): sufficient for a specific need or requirement

Causality (causation): the relationship of cause and effect. The degree to which a cause contributes to an effect, where the cause is partly responsible for the effect, and the effect is partly dependent on the cause[4].

Consistency: the degree of variability and uncertainty within a line of evidence

Coherence: the degree to which multiple lines of evidence are corroborated or are mutually supportive

Direction of Uncertainty: the impact of *uncertainty* on the *hypothesis* (e.g., potential under- or overestimation of risk)

Fit-for-purpose: appropriate and of a sufficient standard for the intended use or purpose

Hypothesis: an initial explanation or supposition to be evaluated based on the available evidence versus alternative explanations

Inference (inferential strength): the degree to which evidence can be reasoned to corroborate or refute a given *hypothesis*. This is different from "fit for purpose" where evidence is deemed suitable for its purpose (e.g., meets acceptance criteria)

Line of Evidence: set of data/evidence with common properties (e.g., same type of test or directed to the same endpoint) of scientific or regulatory *relevance* to the hypothesis

Plausibility: the likelihood a known mechanism explains *causality* (cause-effect)

Relevance: the degree of correspondence of scientific or regulatory *evidence* to the *hypothesis*

Reliability: the confidence assigned to evidence based on the assessment of data quality, sufficiency (quantity), plausibility and uncertainty

Strength of Evidence: the overall level of confidence when various weighted *lines of evidence* are combined

Sufficiency: refers to the concept of enough data to address the question from the problem formulation

Quality: the combined result of the judgement on relevance, reliability and validity

Uncertainty:
the combination of *lack of knowledge* (true uncertainty) and data *variability* OR according to EFSA Guidance on Uncertainty "a general term referring to all types of limitations in available knowledge that affect the range and probability of possible answers to an assessment question "(EFSA, 2016)

[4] Bunge, Mario (1960) [1959]. Causality and Modern Science. Nature. 187 (3, revised ed.) (published 2012). pp. 123–124.

Weighing of Evidence: the process of assigning a weight to assembled *lines of evidence* based on the combined impact of *reliability* and *relevance*

Weight of Evidence: a consideration of the weight assigned to each of the assembled *lines of evidence* to come to an overall decision or conclusion