Scientific Committee on Health, Environmental and Emerging Risks

SCHEER

Memorandum on weight of evidence and uncertainties

Revision 2018

The SCHEER adopted this Opinion by written procedure on 26 June 2018

About the Scientific Committees

Two independent non-food Scientific Committees provide the Commission with the scientific advice it needs when preparing policy and proposals relating to consumer safety, public health and the environment. The Committees also draw the Commission's attention to new or emerging problems that may pose an actual or potential threat.

These committees are the Scientific Committee on Consumer Safety (SCCS) and the Scientific Committee on Health, Environmental and Emerging Risks (SCHEER). The Scientific Committees review and evaluate relevant scientific data and assess potential risks. Each Committee has top independent scientists from all over the world who are committed to working in the public interest.

In addition, the Commission relies upon the work of other Union bodies, such as the European Food Safety Authority (EFSA), the European Medicines Agency (EMA), the European Centre for Disease prevention and Control (ECDC) and the European Chemicals Agency (ECHA).

SCHEER

This Committee, on request of Commission services, provides Opinions on questions concerning health, environmental and emerging risks. The Committee addresses questions on:

- health and environmental risks related to pollutants in the environmental media and other biological and physical factors in relation to air quality, water, waste and soils.

- complex or multidisciplinary issues requiring a comprehensive assessment of risks to consumer safety or public health, for example antimicrobial resistance, nanotechnologies, medical devices and physical hazards such as noise and electromagnetic fields.

Scientific Committee members

Roberto Bertollini, Teresa Borges, Wim de Jong, Pim de Voogt, Raquel Duarte-Davidson, Peter Hoet, Rodica Mariana Ion, Renate Krätke, Demosthenes Panagiotakos, Ana Proykova, Theo Samaras, Marian Scott, Remy Slama, Emanuela Testai, Theo Vermeire, Marco Vighi, Sergej Zacharov

Contact:
European Commission
DG Health and Food Safety
Directorate C: Public Health, Country Knowledge, Crisis management
Unit C2 – Country Knowledge and Scientific Committees
Office: HTC 03/073 L-2920 Luxembourg
SANTE-C2-SCHEER@ec.europa.eu

ISSN 2467-4559 ISBN 978-92-76-00233-8
doi:10.2875/386011 EW-CA-19-005-EN-N

The Opinions of the Scientific Committees present the views of the independent scientists who are members of the committees. They do not necessarily reflect the views of the European Commission. The Opinions are published by the European Commission in their original language only.

http://ec.europa.eu/health/scientific_committees/policy/index_en.htm

ACKNOWLEDGMENTS

Members of the Working Group are acknowledged for their valuable contribution to this Opinion. The members of the Working Group are:

The SCHEER members

Ana Proykova (Chair)
Renate Krätke (Rapporteur)
Roberto Bertollini
Teresa Borges
Raquel Duarte-Davidson
Demosthenes Panagiotakos
Rémy Slama
Marian Scott
Emanuela Testai
Theo Vermeire
Marco Vighi

SCHER and SCENIHR members (2013-2016):

Alena Bartonova
Claire Beausoleil
Teresa Fernandes
Jan Linders
Greta Schoeters
Philippe Hartemann

All Declarations of Working Group members and supporting experts are available at the following webpage:

http://ec.europa.eu/health/scientific_committees/environmental_risks/members_committee/index_en.htm

ABSTRACT

This Memorandum is focussed on how to use the weight of evidence approach (WoE) to conduct a risk assessment for stressors to which humans and/or the environment may be exposed. It is intended to complement the Scientific Committee on Emerging and Newly Identified Health Risks (SCENIHR) report on the identification of emerging issues and the work on the challenges in future risk assessment. The aim of this document is to support the use of the WoE, wherever appropriate, for the risk assessment activities of the Scientific Committee on Health, Environmental and Emerging Risks (SCHEER). In addition, it should support the consistency in the work of different EU bodies performing risk assessments.

Scientific evidence consists of observations, experimental and model results and expert judgements that serve to support, refute, or modify a scientific hypothesis or theory. The search for relevant information and data for the SCHEER comprises of identifying, collecting and selecting possible sources of evidence in order to perform a risk assessment and/or to answer the specific questions being asked. According to the issue being addressed, the SCHEER may utilise data provided by the DG tasking the SCHEER, or provided by a third party (e.g. stakeholder reports, submissions such as confidential data provided by companies or applicants), reports and Opinions of other scientific, governmental or international bodies, scientific (peer-reviewed) publications, meta-analysis and systematic reviews or personal communications.

The WoE is an iterative process involving:

- Problem formulation
- Identification, collection and selection of the possible sources of evidence
- Assessment and weighing of individual lines of evidence
- Integration of lines of evidence
- Description of uncertainties
- Conclusion and reporting

For each line of evidence, the criteria of validity, reliability and relevance need to be applied and the overall quality has to be assessed. Several tools for the analysis and description of uncertainties are presented. In the integration of the different lines of evidence, the strength of the overall evidence depends on the consistency and the quality of the results. The weighing of the total evidence should be presented in a standard format. A system is proposed that classifies results of analysis for human and environmental risks in terms of:

- Strong weight of evidence: Coherent evidence from a primary line of evidence (human, animal, environment) and one or more other lines of evidence (in particular mode/mechanistic studies) in the absence of conflicting evidence from one of the other lines of evidence (no important data gaps)
- Moderate weight of evidence: good evidence from a primary line of evidence but evidence from several other lines is missing (important data gaps)
- Weak weight of evidence: weak evidence from the primary lines of evidence (severe data gaps)
- Uncertain weight of evidence: due to conflicting information from different lines of evidence that cannot be explained in scientific terms
- Weighing of evidence not possible: No suitable evidence available

Keywords (for literature search): human health risk assessment, environmental risk assessment, scientific literature, risk analysis, uncertainty and variability analysis, weight-of-evidence/weighing of evidence, data integration, lines of evidence/line of evidence, levels of evidence/level of evidence, strength of evidence/strengths of evidence, quality of evidence, quality criteria, evidence integration/integration of

evidence

Opinion to be cited as:
SCHEER (Scientific Committee on Health, Environmental and Emerging Risks), Memorandum on weight of evidence and uncertainties, Date of adoption:

TABLE OF CONTENTS

1. BACKGROUND

According to the Commission Decision C(2015) 53831, the mission of the Scientific Committees is to provide the Commission services with scientific advice and risk assessments in the areas of public health, consumer safety and environmental risks, including, when relevant, identification of research needs to address critical information gaps and the assessment of proposed future research actions and research results.

The scientific assessments carried out by the Scientific Committees should always be based on scientifically accepted standards of best practice, and be transparent with regard to the data, methods and assumptions that are used in the risk assessment process. They should identify uncertainties and use harmonised terminology, where possible, based on internationally accepted terms.

The 'Memorandum on weight of evidence and uncertainties' was adopted by SCENIHR in 2012 to provide greater transparency in the risk assessments carried out by this Scientific Committee, to provide greater consistency between Opinions and to be helpful to stakeholders.

In light of the reorganisation of the Scientific Committees, namely the merger of two committees to form the Scientific Committee on Health, Environmental and Emerging Risks (SCHEER), it is necessary to review this Memorandum intended for use by the SCHEER in formulating their future scientific Opinions.

2. TERMS OF REFERENCE

The SCENIHR and SCHER were requested by the Secretariat to prepare a revised version of the 'Memorandum on weight of evidence and uncertainties', adopted originally by SCENIHR in 2012, making explicit the approach to be used by the SCHEER for determining the weight of evidence and the uncertainties involved in the development of their Opinions. The approach should take into consideration the newest available, evidence-based methodology that has been developed by various national and international risk assessment bodies, including Union bodies. The Revised Memorandum should be applicable for human health, environmental and ecological risk assessments.

3. DEFINITIONS

Term	Definition
Weight of evidence	WHO, (2009) "A process in which all of the evidence considered relevant for a risk assessment is evaluated and weighted" ECHA, (2010) "Weight-of-Evidence can be defined as 'the process of considering the strengths and weaknesses of various pieces of information in reaching and supporting a conclusion concerning a property of the substance" Also in ECHA, (2010) "An evidence based approach involves an assessment of the relative values/weights of different pieces of the available information that have been retrieved and gathered in previous steps. To this end, a value needs to be assigned to each piece of information. These weights/values can be assigned either in an objective way by using a formalised procedure or by using expert judgement. The weight given to the available evidence will be influenced by factors such as the quality of the data, consistency of results, nature and severity of effects, relevance of the information for the given regulatory endpoint" EFSA, (2017) "A process in which evidence is integrated to determine the relative support for possible answers to a question" SCHEER, (2018) "A process of weighted integration of lines of evidence to determine the relative support for hypotheses or answers to a question"
Line of evidence	EFSA, (2017) "Set of evidence of similar type"
Quality	Quality is the combined result of the judgement on relevance, reliability and validity.
Reliability	Klimisch *et al.,* (1997); Nendza *et al.,* (2010); ECHA, (2010) "Evaluating an individual result with regard to the inherent quality of a test report or publication relating to a, preferably standardised, methodology and the way that the experimental procedure and results are described to give evidence of the clarity and plausibility of the findings"

Validity	Klimisch *et al.,* (1997); Nendza *et al.,* (2010); ECHA, (2010) "Evaluating the method used for the generation of data for a specific endpoint relative to accepted guidelines. Or: Evaluating the model used for the generation of data against validation principles such as the OECD validation principles"
Relevance	Klimisch *et al.,* (1997); Nendza *et al.,* (2010); Relevance/potential importance. This defines whether a set of data (e.g. from a publication) is appropriate for a particular hazard identification or risk characterisation.

4. METHODOLOGY (principles of WoE)

4.1 Introduction

This memorandum is intended to make explicit the approach used by the SCHEER for determining the weight of evidence (WoE) and the uncertainties involved in the development of its Opinions. The Memorandum draws on the methodology sections of previous Opinions of the SCENIHR and the SCHER, identifying the best common practices in the different domains. It involves a staged approach and a number of additional elements that are considered to improve the transparency and consistency of human health and environmental risk assessments carried out by the Scientific Committee. The approach draws on a number of schemes that have been developed by various national and international bodies. Particular attention has been paid to ensuring that the format of the schemes can be applied to a wide range of lines of evidence and types of publication.

This chapter provides a brief description of the framework and steps required to complete a WoE to be used for risk assessments by the SCHEER. It updates the approach developed previously in 2012 (SCENIHR, 2012). A number of organisations have established their own frameworks for assessing/evaluating evidence (e.g. EFSA 2017; ECHA 2017). These have been drawn upon wherever appropriate in the development of this memorandum.

There are a number of definitions for the WoE including those from the WHO (2009), "a process in which all of the evidence considered relevant for a risk assessment is evaluated and weighed" and from EFSA (2017), "a process in which evidence is integrated to determine the relative support for possible answers to a question". The SCHEER approach is consistent with both of these definitions, viewing WoE as a process of weighted integration of lines of evidence to determine the relative support for hypotheses or answers to a question.

4.2 General SCHEER approach

The SCHEER has identified 6 key steps in performing a WoE, which will be addressed in detail in the subsequent chapters (Figure 1):

1. Problem formulation of the risk assessment requested (see section 4.2.1)
2. Identification, collection and selection of the possible sources of evidence and gaps in relation to the aim of the assessment, including initial screening of these evidence sources to identify those that are relevant to address the question(s) posed by the Commission Services (see section 5)
3. Assessment and weighing of individual lines of evidence (see section 6)
4. Integration of different lines of evidence (see section 7)
5. Uncertainty assessment (see section 8)
6. Conclusions/reporting (see section 9)

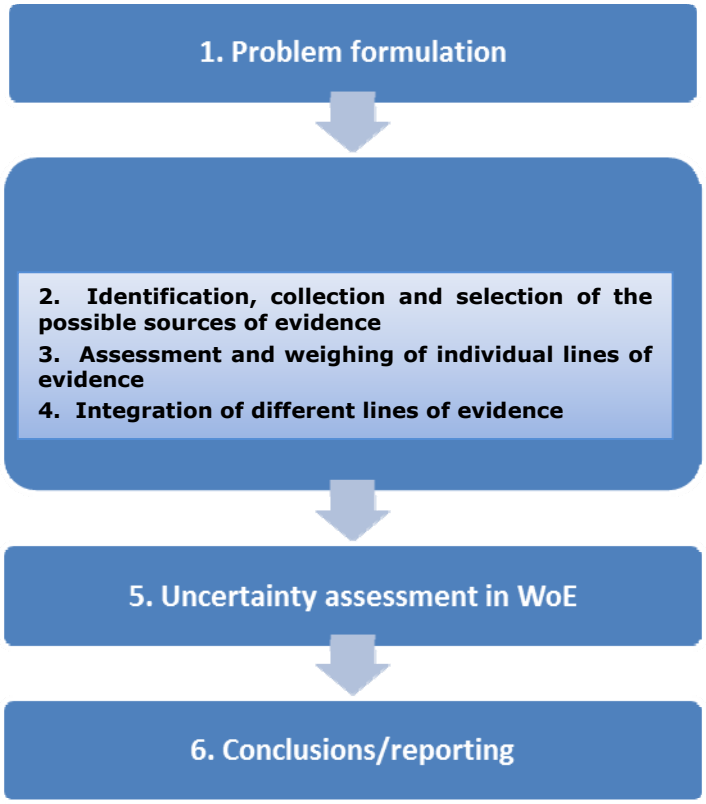

Fig. 1: The Weight of Evidence process in risk assessment

Steps 2, 3 and 4 of the weight of evidence assessment process, described in Figure 1, may be taken at one or more points in the course of the hazard and exposure assessment in response to the questions or in the terms of reference as elaborated in the problem formulation (step 1). The output of WoE and the determination of uncertainties in the WoE (step 5) feed into the overall conclusion of the scientific assessment. In a risk assessment this will be at the stage of risk characterisation. Although Figure 1 shows a linear process, iterations may occur.

4.2.1 Problem formulation

A critical aspect of the risk assessment process is the framing of the questions asked of the committee. Problem formulation should address the risk management needs and clearly state the purpose of the risk assessment and should include for example the relevant population exposed, the level of uncertainty that is acceptable and the urgency of the assessment. The problem formulation can also include questions on the characterisation of experimental data sets, the analysis plan and remaining uncertainty (IPCS, 2014). The issue is addressed in an earlier SCENIHR/SCCS/SCHER Opinion (Addressing the New Challenges for Risk Assessment, 2012) and is not further considered here, other than to emphasise that this is a critical first step.

4.2.2 Identification collection and selection of the possible sources of evidence

Scientific evidence includes observations, experimental and model results and expert judgements that serve to support, refute, or modify a scientific hypothesis or theory. There may be sufficient evidence for one line of evidence from several independent sources of information leading e.g. to the assumption/conclusion that a substance or an issue (e.g. radiation) has or does not have a particular property, while the information from each single source alone is regarded insufficient to support this conclusion. According to the issue being addressed, the SCHEER uses different sources for information. For an Opinion based on publically available scientific information, the

SCHEER primarily relies on original peer reviewed publications, though this is obviously not always possible. In general non peer-reviewed reports should be considered and weighted on case-by-case basis based on expert judgement of the SCHEER.

4.2.3 Assessment and weighing of individual lines of evidence

A major task of the SCHEER in conducting a risk assessment is to evaluate and assess the lines of evidence and to judge their validity, reliability and relevance (Klimisch *et al.,* 1997; Nendza *et al.,* 2010; ECHA, 2010):

- Relevance: This defines whether a set of data (e.g. from a publication) is appropriate for a particular hazard identification or risk characterisation and therefore has the potential to contribute to answering the questions asked by the Commission Services
- Validity: Evaluating the method used for the generation of data for a specific endpoint relative to accepted guidelines. Or: Evaluating the model used for the generation of data against validation principles such as the OECD validation principles
- Reliability: Evaluating an individual result with regard to the inherent quality of a test report or publication relating to a, preferably standardised, methodology and the way that the experimental procedure and results are described to give evidence of the clarity and plausibility of the findings

Key issues to be evaluated are:

- Characterisation of the stressor
- Soundness and appropriateness of the methodology and models
- Extent to which the full details of methodology are provided
- Reproducibility of findings between experiments/observations
- Relevance of a set of data for a particular endpoint

4.2.4 Integration of individual lines of evidence

Integrative assessment means that the results from all relevant individual lines of evidence are compiled into an overall assessment, taking into account their reliability, validity and relevance. The integration of the different lines of evidence may demand an element of expert judgement. The WoE depends on the consistency and the quality of the results. Consistency is defined as the agreement in the results of the analysis between all the lines of evidence (SCENIHR, 2012); but also as the extent to which contributions of different pieces or lines of evidence to answering the specified question are compatible (EFSA, 2017). Quality is defined as the combined result of the judgement on relevance, reliability and validity.

In the final weight of evidence assignment, the basis for the judgement should be outlined as far as practicable. Information gaps should also be clearly identified.

4.2.5 Uncertainty assessment in WoE

The strength of evidence is inversely related to the degree of uncertainty. Characterisation of the uncertainties in WoE is important for transparency and should also be a valuable aid to help risk managers determine how to respond to risk management advice. In addition, it is a useful way of indicating priorities for further work to improve the robustness of risk assessments. However, if not clearly and suitably described, expressing uncertainty may raise unwarranted concerns and/or provoke unwarranted actions. The degree to which characterisation of uncertainty is needed will depend on the risk assessment and risk management contexts as determined by the questions asked, i.e. the problem formulation. Uncertainty analysis should be incorporated during the weighing of evidence rather than added after this process is completed.

4.2.6 Conclusions/reporting

Clear and transparent documentation and argumentation is essential for allowing stakeholders and policy-makers to understand how the lines of evidence were selected, assessed and integrated in the WoE used by the SCHEER for the development of the Scientific Opinion.

More specifically, what is needed is explicit and transparent documentation of the assumptions, defaults, data sources, decision criteria, applications of expert judgment and other descriptive information used to reach the conclusions of the assessment. The rationale should include any uncertainties and gaps.

The following chapters provide details on each of these areas.

5. IDENTIFICATION, COLLECTION AND SELECTION OF THE POSSIBLE SOURCES OF EVIDENCE

The search for relevant information and data is restricted to identifying, collecting and selecting possible sources of evidence to answer the specific questions being asked. According to the issue being addressed, the SCHEER may utilise one or more of the following:

- Data provided by the DG tasking the SCHEER
- Data provided by a third party (e.g. stakeholder submissions such as confidential data provided by companies or applicants)
- Reports and Opinions of other scientific bodies
- Reports of various governmental and international bodies (e.g. WHO, FAO, JECFA, IARC, OECD, WMO, NIEHS)
- Reports of stakeholder bodies (e.g. ILSI, ECETOC, WWF)
- Scientific (peer reviewed) publications
- Meta-analysis and systematic reviews
- Modelled data, such as read across and exposure estimations
- Personal communications

5.1 Use of confidential data

For the purposes of the work of SCHEER, it is unavoidable that confidential data are sometimes used. For example, when the data will only be made publically available in the near future or when data and/or information needed are not publically available but provided by e.g. an applicant and/or are provided by different stakeholders after a call for information is published by the Commission.

It should be made clear in any request for access to confidential reports that the data can only be considered by the Committee if the provider agrees that the summary of the evaluation of the data may be incorporated in the text of the Opinion and thereby made publicly available.

The Commission Services retain the confidential files if they have been used to generate an Opinion.

5.2 Initial screening of data sources

Initial electronic searches may be a starting point for data gathering. Appropriate data bases and search engines need to be used, for example: PubMed, Scopus, Toxline, US-EPA-ECOTOX, Chemical and Biological Abstracts, or Google Scholar. In each Opinion, the search engines used and the period covered in searching relevant documents should be identified, along with the search terms used. Both keywords and index/subject search terms (e.g., MeSH, Medical Subjects Headings, the NLM-controlled vocabulary thesaurus used for indexing articles for PubMed) are useful and should be used in the search procedure. Keywords should be tagged to search all of the texts in the documents. Index/subject terms help to focus your search appropriately, looking for items that have had a specific term applied by an indexer. A typical database search filter may be applied to improve, and in some cases narrow, the results, so that the retrieved articles are the most relevant to the mandate. Filter types may include: article/publication type, publication dates, species, intervention types, endpoints, etc. An additional filter for ecotoxicological data is the type of environment (aquatic/terrestrial). The search strategy (including inclusion and exclusion criteria) should be clearly stated in the Opinion.

As the issues that need to be addressed differ significantly between Opinions, data sources will also differ. As far as possible, all relevant data sources should be identified to address the questions being asked. Inevitably, this is subject to practical constraints of accessibility, the time available to complete the Opinion and the language used to publish this information.

For an Opinion based on publically available scientific information, the SCHEER primarily relies on original peer-reviewed publications, though this is obviously not always possible.

In general, secondary sources, i.e. reports of the work of others, should only be considered if there is insufficient peer-reviewed, published scientific data to provide an Opinion. However, this is not intended to include meta-analysis.

In preparing an Opinion, the SCHEER should cross check all references that are intended to be cited from reports and Opinions of other scientific bodies included in the reference list. Personal communications can only be used if supported by raw data and details of the methodology used.

Prior to the screening of data sources, it is important to consider all aspects of the risk(s) under consideration since incomplete identification of the risk(s) and/or risk factors may lead to an inappropriate literature search.

The process of incorporating systematic review methodology into literature-based evaluations has been published by OHAT (2015), though this approach is very resource-intensive which often limits its use by the SCHEER.

Typically, a substantial number of publications will be identified that are of possible interest and an initial screening process is undertaken to identify the references that are suitable for the purposes of answering the questions.

There is no universal, formal and transparent procedure for the evaluation of the acceptability of data for risk assessment purposes. However, the acceptability of a publication for the purposes of its use to answer specific questions can be based on the criteria proposed by Klimisch *et al.,* (1997) and described in the OECD Manual for the investigation of high production volume (HPV) chemicals and further elaborated by Nendza *et al.* 2010 and ECHA (2017) and should include relevance/potential importance and quality of the data published (relevance, validity and reliability).

Publications that are identified initially but do not meet the criteria of relevance, reliability and validity for the development of the Opinion should appear in the reference list or as an additional document for the report on which the Opinion is based as: "Publications noted but not considered suitable for the purposes of developing the Opinion".

Potential conflicts of interest among the authors of a study or the funding sources need to be identified for each source of data.

5.3 Assessment of the quality of individual data

This involves determining the contribution of a publication to the knowledge base, for the purpose of developing an Opinion. This 'level' is likely to vary according to the nature and extent of the evidence available. The implications of the findings may be considered and addressed in the text while taking into account the increased uncertainty caused by, for example, the quality of the data presented.

A number of organisations have established their own frameworks for assessing/ evaluating evidence (e.g. the preamble to the IARC Monograph Series IARC, 2006) where as a result of the assessment a weight is attributed to the study findings. In this process, risk assessors need to assess uncertainties in the underlying data as well as in their own interpretations of these data (Levin *et al.,* 2004). Unfortunately, formal procedures and consistent terminology for weight of evidence processes are lacking. EFSA (2015; 2010) provides guidance on literature searching and systematic review and ECHA provides practical guidelines on how to apply weight of evidence (ECHA, 2010, 2017).

If modelling data are used for exposure assessment (in particular for prospective risk assessment), an issue that must be carefully considered is the relevance and validity of the environmental scenario used. More details on predictive approaches in exposure assessment are reported in section 6.2.2

6. ASSESSMENT AND WEIGHING OF INDIVIDUAL LINES OF EVIDENCE

Each set of information is reviewed to determine its quality based on relevance, validity and reliability to address the problem. Therefore, relevance, validity and reliability are each classified in one of the categories of high, medium or low. Based on expert judgement, an overall category for quality is derived. The overall category for the quality will be used in for the integration of different lines of evidence (see section 7).

In subsequent sections, specific aspects determined by the nature of the line of evidence are discussed in order to enable risk assessors to judge on the quality of the different lines of evidence. Important general aspects are:

- Differentiation between non-adverse and adverse effects
- Ensuring that the adverse effect is related to exposure (e.g. substance or material-related)
- Assessment of biologically relevance, not simply statistical significance. A biologically relevant effect can be defined as an effect considered by expert judgement as important and meaningful for human, animal, plant or environmental health. It therefore implies a change that may alter how decisions for a specific problem are taken (EFSA, 2017)
- Presence of dose/time-effect relationship
- Data on the reversibility of effect
- Information on normal variation in the incidence of the disease/effect of interest (e.g. consideration of historical controls).

When assessing and weighing individual lines of evidence, any potential for bias should be considered accordingly. It is also important to determine whether a study has been conducted in accordance with accepted test guidelines, e.g. OECD guidance documents, and whether a study was performed under a quality system e.g. GLP (Good Laboratory Practice).

When weighing individual lines of evidence, data gaps may be identified if data are inconsistent, uncertain, fail to fulfil requirements or are lacking.

Table 1. Assessment and weighing of overall quality of individual lines of evidence

Quality of a line of evidence/data sources/publication (high medium or low)		
Relevance	Validity	Reliability
high/medium/low	high/medium/low	high/medium/low

Further scoring systems for each line of evidence

Several organisations have proposed grading systems for assessing the quality of evidence. For example, the U.S. Preventive Services Task Force (USPSTF) suggested five different levels, based on type and quality of the study (Level I/Level IIa, IIb, IIc/Level III). The Oxford (UK) CEBM provided, similar to the aforementioned, levels of evidence

for claims about the prognosis, diagnosis, treatment benefits or harms and screening of a treatment. In 2000, a systematic approach was developed by the GRADE (Grading of Recommendations Assessment, Development and Evaluation) working group that takes into account more dimensions than just the quality of research, including the assessment of: risk of bias (on the basis of the chance that bias has influenced the estimate of effect), imprecision and indirectness (on the basis of how the study was conducted), how the results are actually going to be applied and any inconsistency and publication bias. Under these principles, the GRADE classifies the levels of evidence in 4 categories related to the quality of the evidence (Balshem *et al*., 2011).

6.1 Hazard Identification

Hazard identification is carried out to identify the intrinsic toxicological properties of a stressor or a mixture of stressors and to receive information on whether it has the potential to damage human health or the environment. For chemicals, there are different documents available, giving guidance on how to identify hazards of chemicals and mixtures, e.g. the Globally Harmonised System of Classification and Labelling of Chemicals (GHS), which is updated bianually by the United Nations, the ECHA Guidance on the Application of the CLP Criteria for classification and labelling, or the ECHA Guidance on information requirements and chemical safety assessment.

6.1.1. Human Health Hazards

In a tiered approach to identifying human health hazards, emphasis is placed on existing human data such as epidemiological and clinical studies as well as on well-documented case reports and observations followed by animal data, in vitro data and other sources of information. The Scientific Committee on Consumer Safety has given an overview on accepted methods for the different toxicological endpoints to be considered when assessing human health hazards (SCCS, 2016).

Human data

In evidence-based decision making, different study types contribute with different weights. In a tiered approach to identify human health hazards, it is widely accepted that more emphasis is given to results derived from meta-analyses, especially of randomised clinical trials (RCT) and experimental studies, multi- or single-centre RCTs and experimental studies in general, prospective (cohort or case-control studies nested in a cohort) studies and retrospective case-control studies. Robust conclusions can rarely be drawn from cross-sectional studies or case reports. In the context of human studies, ecological studies are defined as relying on an analysis of the relation between exposure and the occurrence of the health outcome at an aggregated scale, i.e. using community-level data instead of individual-level data like in the above-mentioned approaches. Ecological studies generally provide a more limited level of evidence, in particular those based on spatial contrasts (which suffer from limitations in their ability to control for confounding bias), while ecological studies relying on temporal exposure contrasts (before/after studies and time-series studies) can provide a very strong level of evidence regarding short-term effects of exposure.

It should be underlined that there is a considerable range of quality within the variety of different study types. Specifically, clinical trials, with effective random allocation to intervention groups in order to ensure balanced participants' characteristics, and other features to control bias, are usually considered the best methodological approach to test a cause-effect relationship. In epidemiological studies, comprehensive identification of the referent population, sampling procedures (e.g., random selection, stratified and representative, participation rate), validity and accuracy of the exposure assessment and outcome data, sufficient control of potential confounding factors, adequate statistical power and appropriate statistical methods used are among the key elements for a good

study. Prospective studies, although of superior design in order to test for a cause-effect relationship, also have certain limitations, especially when the baseline information is retrieved a long time before the evaluation of the outcome(s).

Animal Data

The advantage of animal studies is that they provide information about effects related to a stressor or a mixture of stressors for a whole living organism that displays the full repertoire of body structures and functions. Beside effects on different organs, or complex systems like the nervous system, the endocrine system or the immune system, aspects of toxicokinetic and metabolism of a chemical or a mixture administered to an experimental animal can be investigated directly. In general, organ- or tissue-specific dosimetry is crucial for associating a stressor or a mixture of stressors with the corresponding health effects. In this respect, animal studies are usually a more powerful experimental tool than cellular studies for assessing health risks to humans. If animal studies are to be used to anticipate potential effects in humans then the extrapolation of the date is needed and appropriate assessment factors have to be applied. The hazard identification should be based on the appropriate species. In case of doubt, the most sensitive species should be chosen.

When evaluating an *in-vivo* study, important aspects like group size, controls, treatment, effect assessment, analytical methods, and statistical analysis should be considered.

In vitro studies

In light of replacing animal experiments, *in vitro* studies, using animal or human tissues, are a further source for toxicological data. *In vitro* studies are available for different toxicological endpoints. Some *in vitro* methods, after having undergone a validation procedure, are accepted as stand-alone methods for hazard assessment and their results have the same regulatory consequences as the corresponding *in vivo* method. Others can be used within a test battery or within a testing strategy combining several *in vitro* /non-testing methods. Being focused on hazard identification, there is a limit in the use of *in vitro* data for a quantitative risk assessment. However, they can support and strengthen lines of evidence in the WoE assessment. The increased use of metabolic systems in *in vitro* data and extrapolation from *in vitro* to *in vivo* by modelling is being explored. This might possibly allow the future use of *in vitro* data for dose-response assessment in humans.

The OECD has published guidance documents for non-animal based integrated testing strategies (Integrated Approach on Testing and Assessment, IATA), e.g. for serious eye damage and eye irritation as well as for skin corrosion and irritation (OECD 2014, OECD 2016). An overview of accepted *in vitro* methods for the different toxicological endpoints is also given in the Notes of Guidance from the SCCS (SCCS, 2016).

When evaluating an *in vitro* study, important aspects like cell type, applicability domain, controls, metabolic competence of the cells, type of treatment, biokinetics, effect assessment, analytical methods and statistical analysis should be considered.

Available *in vitro* test data from well-characterised target organ and target system models on, e.g. mode of action(s) (MoA) or mechanism(s) of toxicity, may be useful in the interpretation of observed repeated dose toxicity. Further guidance on mode of action analysis is available from the WHO/IPCS framework on Mode of action and human relevance (Boobis *et al.,* 2006, Boobis *et al.,* 2008, Meek *et al.,* 2016).

The use of a mode (and/or mechanism) of action may be supported by the Adverse Outcome Pathway (AOP) methodology, an approach that provides a framework to collect, organise and evaluate relevant information on chemical, biological and toxicological effects of chemicals and gives a structured representation of biological events leading to adverse effects. Guidance is given in the OECD Guidance Document on Developing and Assessing Adverse Outcome Pathways, Series on Testing and Assessment No. 184 (2013).

When integrating studies on MoA or AOPs in an assessment, important aspects to be considered include the identification of the relevant molecular initiating event, key events, plausibility and concordance between different outcomes.

Omics technologies such as genomics, transcriptomics, proteomics and metabolomics may help in the identification of specific markers of toxicity that occur early in the process of long-term toxic responses and that are mechanistically linked to the underlying pathology. However, omics technologies are not ready for regulatory purposes and their integration in testing strategies is still under development. Detailed information on omics technologies is given by the SCENIHR, SCCS and SCHEER in their Opinion addressing the New Challenges for Risk Assessment (2012).

Non-testing methods: SAR, QSAR, computer expert systems, analogue and category approaches (*in silico* methods)

The predictive computational models are based on either a (quantitative) structure-activity relationship ((Q)SAR), expert systems (rule-based models), or grouping/read-across from experimental data on analogous chemicals. Detailed information of *in-silico* methods can be found in the SCENIHR, SCCS, SCHEER Opinion on Addressing the New Challenges for Risk Assessment (2012).

The models based on (Q)SAR are mathematical descriptions of the biological activity of (a group of) chemical compounds as a function of their structural or physicochemical properties. SARs describe the qualitative relationship between a chemical structure and a property or biological activity. The success of any (Q)SAR model depends on the accuracy of data, the selection of appropriate descriptors and statistical methods, the number of (groups of) compounds for which (experimental) data are available to develop the model and the validation of the developed model (Nendza, *et al.,*2010;). The validity of the computer models used should have been assessed using the OECD principles for the validation of (Q)SARs (OECD, 2007) The OECD QSAR Tool Box is available for a systematic approach to the formation of chemical categories and other chemical analogies and predicting toxicological effects (OECD, 2009). The model validation should have been adequately described, e.g. by using QMRFs ((Q)SAR Model Reporting Formats).[1]

Expert systems guide hazard assessment by predicting toxicity endpoints of certain substance structures based on the available information. They can be based on an automated rule induction system (e.g., TOPKAT, HazardExpert and MultiCASE) or on a knowledge-based system (e.g., DEREK or the BfR-DSS).

Analogue and category approaches require sufficient reliable test data on similar substance(s) and justification of the similarity with the tested substance(s). Guidance on grouping/read-across has been published by the OECD (2014) and by ECHA (2008).

The use of a combination of different approaches in an *in-silico* battery usually increases confidence of the derived predictions. The compounds under consideration should fall within the applicability domain of the respective model.

[1] Mention of any commercial or non-commercial *in silico* system does not constitute a recommendation for its use by the SCHEER

6.1.2 Environmental hazards

Ecotoxicological data

Ecotoxicological data may be produced at different hierarchical levels of the ecological organisation, from individuals to populations, communities and ecosystems. Moreover, effect data may also refer to sub-individual level, cellular or sub-cellular (biochemical and genetic biomarkers, omics, etc.).

One must be aware that the objective of ecological risk assessment is not to protect individuals but to protect the structure and functioning of ecosystems (Hommen *et al.,* 2010). Therefore, an important issue for assessing the relevance and usefulness of ecotoxicological endpoints is their ecological relevance and realism.

Unfortunately, ecological realism is inversely correlated with other characteristics of ecotoxicological testing (technical simplicity, reproducibility, ease of interpretation, etc.) (figure 2).

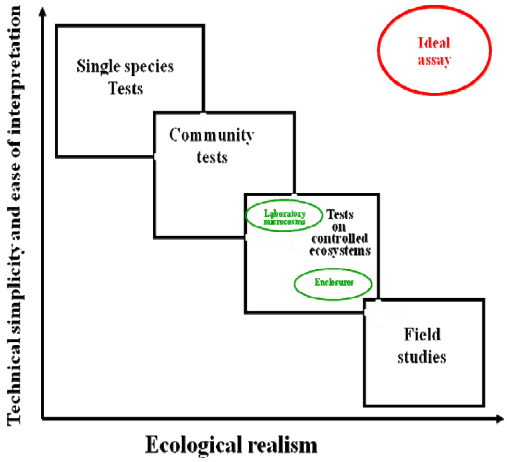

Figure 2. Relationship between ecological realism and simplicity in ecotoxicological testing. The ideal assay, which would be simple to reproduce and ecologically realistic, does not exist (modified after Blank *et al*., 1978).

1. Single species toxicity data. In spite of their low ecological realism, these are the data more frequently available in ecological risk assessment (ERA) procedures. Several European Directives on chemical control (REACH, pesticide directive etc.) require, as a base set, single species toxicity data on a number of selected organisms assumed as representative of the natural ecosystems (e. g., for freshwater, algae, Daphnia and fish). Therefore, for many chemicals, the base set, generally produced in GLP with officially accepted methodologies (ISO, OECD, etc.), represents the only information available. Besides official data produced in the framework of regulations, single species data may be taken from the open literature. The reliability of these data is usually evaluated as a function of the compliance with standard methodologies. In the official ERA procedures, the uncertainty and the lack of ecological realism of these kinds of data is covered by the use of application factors.

2. Species sensitivity distribution (SSD). The SSD procedure is based on single species toxicity data but represents an attempt for quantifying the variability of the responses among the different species of a biological community. Moreover, it allows a probabilistic assessment of uncertainty, instead of a deterministic approach applicable to the base set of toxicity data. The approach does not consider the interactions among the species of a

community or the indirect ecological effects. The minimum requirements for the acceptability of a SSD approach are described in the standard ERA procedures (EC, 2003). Besides these minimum requirements, the major rules for evaluating the validity and reliability of a SSD are:

- the number of species tested
- the distribution of the species in different taxonomic groups
- the comparability of methods and endpoints

3. Higher tier testing. This includes microcosm and mesocosm experiments and represents the attempt to reproduce, in controlled conditions, the structure of natural communities and ecosystems. The level of complexity of experimental structures is extremely variable, from small aquaria in laboratory (microcosms) to large outside structures reproducing ponds and rivers or terrestrial ecosystems. Even if, in any case, they represent a simplification of natural ecosystems, the ecological realism is strongly improved by the possibility of studying the interactions among species, the indirect ecological effects and the combined effects of variable environmental conditions. The major drawback of the approach is the difficulty of reproducing experimental conditions. There are no standard procedures or methods for micro and mesocosm testing. The validity and reliability of the approach should be evaluated case-by-case, relying on expert judgement. For a better regulatory use of higher tier testing, the development of standard, officially accepted, procedures is recommended.

4. Field studies. The direct study of natural ecosystems offers, obviously, a more authentic view of the ecology. However, the precise assessment of cause-effect relationships is very difficult, if not impossible, in most cases. The use of field studies data, if available, may be important in WoE. Nevertheless, their usefulness must be considered very carefully, on a case-by-case basis.

5. Biomarkers and other sub-individual approaches. Biomarkers and other approaches based on molecular biology (omics) are widely used in ecotoxicology as indicators of exposure to toxicants and effects of stress factors. However, the usefulness of these approaches in ecotoxicology and, in particular, their ability to provide unambiguous and ecologically relevant information on exposure to or effects of toxicants has been challenged (Forbes *et al.,* 2006). The debate in the scientific community on the ecological relevance of sub-individual parameters is still on-going, as well as on their usefulness in perspective and retrospective ecological risk assessment (Forbes & Calow, 2012). There are several reasons for this debate. First, at present, our knowledge of the relationships between effects measured at the sub-individual level and the consequences at the community level (the actual goal of environmental protection) is very poor and must be better investigated (Forbes & Calow, 2012). Moreover, knowledge of the natural variability of biochemical and physiological parameters (practically complete for man) is very poor for natural populations. It has been demonstrated that their variability as a function of environmental conditions, unaltered by anthropogenic activities, may be very high (Ippolito *et al.,* 2016; Scarduelli *et al.,* 2017). Therefore, these data should be used with care in a WoE. They may be useful as early warning exposure indicators in retrospective risk assessment and should not be used as indicators of ecologically relevant effects.

6.2 Exposure Assessment

Exposure assessment is the process of estimating or measuring the magnitude, frequency and duration of an individual's or population's exposure to a stressor or a mixture of stressors.

6.2.1 Human Exposure

The exposure assessment process requires identification of the potentially exposed population, establishing the pathways and routes of exposure and quantification of the

potential stressor effect to the body. It is the step in the human health risk assessment where questions about how chemicals come into contact with people and what amounts people could be exposed to are addressed. Human exposure to chemical stressors occurs through any or all of three potential exposure routes: inhalation, dermal uptake and ingestion. Exposure to a physical stressor can take place when humans find themselves in the respective field. The relevant exposure scenarios that may be of concern from a health perspective are identified at this stage.

The environmental concentration to which an organism may be exposed can be directly measured through experimental monitoring or through biomarkers. Exposure levels can be estimated by predictive models possibly taking into account environmental levels and personal behaviour, such as time-space activity. Monitoring data are valuable for confirming information for exposure assessment. Measured data are generally preferred over model estimates when available as studies based on model estimates can in some instances be less sensitive to specific bias. On the other hand, measured data may not be representative for the population exposed.

The scenarios refer to the specific conditions by which people could be exposed to stressors with consideration given to the sources, the nature and duration of the releases (i.e., intermittent vs. continuous), and other factors affecting the types and levels of exposure that could be experienced. This step is concerned with estimating the level of exposure to the stressors of concern that might be received by individuals via the various exposure pathways.

The pathways of exposure to chemicals are often distinguished as being primary or secondary in nature. The former pathways are dictated by the manner in which the chemicals of concern are emitted, discharged or released into the environment and represent direct avenues by which the chemicals can reach individuals (e.g., breathing in an air-borne chemical), whereas the latter pathways represent secondary routes by which the chemicals might reach people depending on the substance's environmental fate and behaviour (e.g., exposure via the food chain). The mode of exposure refers to the actual manner in which the substance can enter the body, with the principal modes being inhalation, ingestion and dermal contact.

The process of exposure assessment often relies on one or more forms of predictive modelling to arrive at the exposure estimates, with specific reliance on air dispersion modelling in the case of air-borne contaminants. Factors that can influence the amount of exposure received, such as the behaviour of the stressors of concern in the environment and the characteristics of individuals who may be exposed (e.g., body weight, breathing rate) are integrated into the assessment. Apart from estimating the exposures received from the selected emission sources under study, consideration is also often given to background exposures contributed by existing sources of the stressors of concern to arrive at estimates of cumulative exposures.

Distinction is made between exposures of a short-term (or 'acute') nature extending over a few minutes to several hours vs. long-term (or 'chronic') exposures lasting for several months or years, possibly up to a lifetime. Consideration is also given to people who might be especially vulnerable to exposure, including infants, young children, the elderly and individuals in poor health.

To estimate human exposure to chemicals or mixtures, several exposure scenarios have been developed addressing the sources, pathways, frequency and routes of exposure. The OECD has published on its website Emission Scenario Documents (ESDs) for different industrial processes describing the sources, production processes, pathways and patterns for industrial, professional and private uses. ESDs aim to quantify the emissions of a chemical into water, air, soil and/or solid waste and to also address the frequency and duration of a task, as well as regional and climatic differences in their scenarios for workers and consumers.

For the risk assessment under REACH, guidance documents for the exposure assessment for workers and consumers have been developed (Guidance on Information Requirements and Chemical Safety Assessment Part D: Framework for exposure assessment). For the assessment of human exposure to biocidal products, ECHA published Technical Notes for Guidance (ECHA, 2007) as well as the Guidance for Human Health Risk Assessment, Volume III, Part B, Guidance on Regulation (EU) no 528/2012 concerning the marketisation and use of biocidal products (ECHA, 2013).

For the assessment of pesticides, EFSA published the Guidance on pesticides exposure assessment of operators, workers, residents and bystanders (EFSA, 2015).

In order to harmonise the approaches, the OECD Task Force on Environmental Exposure Assessment (TFEEA) has worked on the comparison of default values and assumptions used in different exposure models.

Different computational tools are available to calculate the human exposure for the different scenarios, e.g. a computerised database (BEAT) of exposure data (largely for occupational settings), the consumer exposure model ConsExpo for exposure of consumers to chemicals and the EUROPOEM, UK-POEM or BBA models for pesticides.

The SCHEER makes use of the different exposure scenarios suitable for the mandate for which it is tasked.

To accommodate the possible sources of uncertainty in the exposure assessment, conservatism is invariably incorporated to avoid overlooking or understating any potential health risks. The conservatism is commonly introduced through a combination of worst-case exposure scenarios and conservative assumptions with respect to exposure modelling parameters. If the conservative assessment indicates a risk, more realistic scenarios and parameter values can be introduced in the assessment.

6.2.2. Environmental Exposure

The environmental level of a stressor, e.g. concentration of a chemical, to which an organism may be exposed can be directly measured or estimated through predictive models.

In particular, for prospective risk assessment, predictive models represent the only possibility for estimating predicted environmental concentrations (PECs) in the different environmental compartment (air, groundwater, surface water and soil).

The criteria for evaluating the reliability of PECs depend on the type of model and on the quality of input data.

1. Type of models. A large number of multimedia models, generally based on the partitioning and fugacity concepts, have been developed and are suitable for application in different environmental conditions. The model selected should be adequate for the proper scale level (global, regional, local, site-specific). It should be developed for the specific ecosystem under study (terrestrial, aquatic lotic, aquatic lentic, etc.). It should be adequately validated in conditions comparable to those in which it will be applied. Several approaches to estimate the environmental concentration are currently applied in the European Union for regulatory purposes. There are for example, the EUSES system for substances under REACH and under the Biocidal Products Regulation (BPR) and the FOCUS-approach for pesticides. Other types of suitable models may be applied in specific situations.

2. Input data. Three types of input data are needed for the application of multimedia models: the description of the environmental scenario, the characteristics of the substance and the emission patterns. The environmental scenario should be adequately described in detail. If some information is lacking, default (worst case) data may be used, but with the awareness that this increases the uncertainty of the results. The physical-chemical properties of the substance (water solubility, vapour pressure, Kow, etc.) are generally available with enough precision. More difficult are the persistence

data in the different compartments that are not intrinsic properties of the substance but depend on environmental conditions. Finally, the quantity emitted, as well as the time and sites of emission are needed. This information is often the most difficult to obtain with enough precision and default data must be used.

For retrospective risk assessment both predicted and experimental data may be used. However, for ecological risk assessment, one must be aware that experimental data reported in the literature may be representative of a very specific situation in space and time, and not of a more generalised realistic exposure condition. Therefore, the European Technical Guidance Document on risk assessment (TGD) recommends a stepwise procedure based on a combination of experimental data and calculated PECs:

- Reliable representative data should be selected by evaluation of the sampling and analytical methods employed and the geographic and time scales of the measurement campaigns;
- The data should be assigned to local or regional scenarios by taking into account the sources of exposure and the environmental fate of the substance;
- The measured data should be compared to the corresponding calculated PEC; for risk characterisation a representative PEC should be decided upon based on measured data and a calculated PEC."

A detailed list of criteria to be fulfilled for the use of monitoring data in ERA and to reduce uncertainties is proposed by OECD (2000) and reported in the TGD for risk assessment (EC, 2003). Guidance documents are also provided by ECHA for chemicals under REACH (e.g. Guidance on information requirements and Chemical Safety Assessment, Chapter R.16: Environmental exposure assessment (https://echa.europa.eu/documents/10162/13632/information_requirements_r16_en.pdf or Guidance on Information Requirements and Chemical Safety Assessment Part E: Risk Characterisation" which also refers to uncertainty analysis. https://echa.europa.eu/documents/10162/13632/information_requirements_part_e_en.pdf/1da6cadd-895a-46f0-884b-00307c0438fd).

6.3 Characterisation of the dose-response function

A dose–response function describes the change of an effect on humans, experimental animals, other organisms, tissues or cell and subcellular systems caused by differing levels of exposure to a stressor. There are linear and non-linear dose–response relationships. The study of a dose-response relationship is crucial in order to determine any safe, hazardous or beneficial levels and dosages for a variety of stressors. The conclusions drawn are the basis for public policy. At this point, the role of modifying factors in defining a dose-response function should be underlined. A modifying factor is a variable that affects the shape and/or strength of the dose-response function between the stressor and the outcome. Taking into account a modifying factor is of great importance in correctly characterising the dose–response relationship.

A meta-analysis is considered the best approach to derive a dose-response relationship in human studies. Through the meta-analysis, a weighted estimate based on all studies deemed relevant and comparable is calculated together with the uncertainty around this point estimate due to the different studies' characteristics. A key benefit of meta-analyses is the aggregation of existing information, leading to increased statistical power and more robust point estimates than from any individual study. Thus, the characterisation of a dose-response function is considered more accurate when it comes from a meta-analysis of relevant studies.

In the extrapolation from animal experiments to humans, attention needs to be paid to obvious differences in, e.g., body mass, life expectancy, physiology, kinetics and metabolism between species and appropriate scaling factors have to be applied. Validity

of the animal model used should be considered – good animal models do not exist at present for all human diseases. Nevertheless, at a molecular level, many basic processes, such as DNA damage and repair, are similar in animals and humans, and animal studies have remained a cornerstone in evaluating the toxicity of chemical and physical agents.

6.4 Statistical analysis

The statistical analysis section of each research publication gathered should be carefully evaluated before being included in the Opinion. Evaluation should include appropriateness and completeness of statistical methods followed, accounting for potential confounding, mediation or moderation of the results, as well as for a statistical power analysis plan. It is possible that an observed effect or a lack of such an effect could be due to chance. This is a particular problem, especially in human studies with small, inadequate sample sizes or low exposure levels, but it may occur in environmental field studies, too. The presence or absence of statistical significance alone should not guide inclusion or exclusion of a study from the Opinion. In addition, presence of a statistically significant association does not alone constitute sufficient evidence for causality. Bias can produce a spurious association and can also mask existing associations. The effect size of the association, its related statistical uncertainty (e.g., confidence intervals of the effect estimates) and the internal consistency of the results should also be evaluated. Other important characteristics that are taken into consideration are the types of controls that have been used, any randomisation procedures and blinding to assure comparability of information and the degree to which replication studies have been performed. The number of statistical tests performed and the reliance on methods correcting for multiple comparisons will also need to be considered.

6.5 Citing papers examined

As a consequence of in-depth evaluation, publications and any other sources of data used will be cited in the reference list in the Opinion or in a separate document in one of three categories:

- Publications that are relevant and of sufficient/suitable quality and were important for the development of the Opinion
- Publications that are relevant and of sufficient/suitable quality but were not judged to be necessary for the development of the Opinion
- Publications noted but not considered adequate (relevant or of sufficient quality) for the purposes of developing the Opinion; this group might be listed in an annex

6.6 How to present human studies

Several guidelines have been proposed on how to report the results of a study on human data (e.g. a meta-analysis, a clinical trial, an epidemiological study or a case-report). In this section, some basic instructions in presenting a meta-analysis, a clinical trial, an experimental study or an epidemiological study are given, based on widely adopted recommendations.

The main principles in presenting a meta-analysis, according to the QUORUM statement (Improving the quality of reports of meta-analyses of randomised controlled trials: the QUOROM statement, Lancet 1999; http://dx.doi.org/10.1016/S0140-6736(99)04149-5), include the following:

- The information sources (e.g., PubMed, Scopus, etc) and any other restrictions (e.g., years considered, publication status, language of publication)
- The criteria used for studies' selection or exclusion
- Study's design, participants' characteristics and measurements used

- Data abstraction
- Outcome(s) definitions
- Sources of heterogeneity and how it was handled
- Data synthesis and measures of effect (e.g., odds ratios or relative risks, etc.)
- Handling of missing data; any a-priori sensitivity and subgroup analyses; and
- Assessment of publication bias

Similarly, to QUORUM, PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) is an evidence-based set of items for reporting systematic reviews and meta-analyses, mainly of randomised trials, but it can also be used for reporting systematic reviews of other types of research, particularly evaluations of interventions (http://www.prisma-statement.org/). One of PRISMA's advantages is that it gives special emphasis to the reporting of harms through a checklist that contains extension items that must be used in any systematic review addressing harms, irrespective of whether harms are analysed alone or in association with benefits.

For clinical trials the following study's main characteristics should be reported, according to the CONSORT statement (http://www.consort-statement.org/):

- Description of trial design (parallel, cross-over, factorial), including allocation ratio
- Eligibility criteria for participants
- The interventions for each group with sufficient details to allow replication, including how and when they were actually administered
- Completely defined pre-specified primary and secondary outcome measures, including how and when they were assessed
- How sample size was determined
- Explanation of any interim analyses and stopping guidelines
- Type and method used to generate the random allocation sequence
- Methods for basic and additional analyses, such as subgroup and adjusted analyses

Main issues that should be evaluated in epidemiological studies, according to the STROBE statement (https://www.strobe-statement.org/index.php?id=strobe-home), are:

- Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection
- Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers or mediators. Give diagnostic criteria, if applicable
- Describe any efforts to address potential sources of bias
- Explain how the study size was calculated
- Describe all statistical methods, including those used to control for confounding
- Report numbers of outcome events or summary measures over time
- Report other analyses done, e.g., analyses of subgroups and interactions, and sensitivity analyses.

The SCHEER makes use of those criteria appropriate for evaluating the information

7. INTEGRATION OF DIFFERENT LINES OF EVIDENCE

In this third step of the WoE, the lines of evidence collected, evaluated and weighted during the previous steps have to be integrated to arrive at conclusions in order to evaluate how they contribute to the comprehensive assessment of exposure and effects, respectively. 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

Consistency with regard to a specific endpoint can be graded as:

- HIGH – most studies show similar findings
- MEDIUM – the studies result in mixed findings, some similar to each other and others supporting different outcomes
- LOW – little agreement between studies. This may be due to heterogeneity of results because of particular features of the studies considered or to effect modification, e.g. because of the presence of susceptible subgroups in the study. A thorough analysis of the causes of inconsistencies is recommended

The overall quality is determined by the combined merit of the relevance, reliability and validity. Based on expert judgement, the quality of lines of evidence can be considered to be high, medium, low.

The final result of the integration is the assessment of exposure and effects on the basis of the complete information available and critically evaluated. It can result in:

- Strong weight of evidence: Coherent evidence from a primary line of evidence (human, animal, environment) and one or more other lines of evidence (in particular mode/mechanistic studies) in the absence of conflicting evidence from one of the other lines of evidence (no important data gaps)
- Moderate weight of evidence: good evidence from a primary line of evidence but evidence from several other lines is missing (important data gaps)
- Weak overall weight of evidence: weak evidence from the primary lines of evidence (severe data gaps)
- Uncertain weight of evidence: due to conflicting information from different lines of evidence that cannot be explained in scientific terms
- Weighing of evidence not possible: No suitable evidence available

Table 2: The conclusion on the WoE based on consistency and quality

		Quality		
		high	medium	low
Consistency	high	strong	strong	moderate
	medium	strong	moderate	weak/uncertain/ not possible
	low	moderate	weak/uncertain/not possible	weak/uncertain/not possible

To draw conclusions, it is not recommended to simply add together weighting from individual lines of evidence. Integration implies that the combination of conclusions from different lines of evidence and their consistency (e.g. evidence for a MoA, explaining the occurrence of a certain effect in experimental animals and not in humans or vice-versa) is the final goal, but this demands an element of expert judgement. The severity of the effect/outcome and the likelihood of its occurrence in individuals or in the population at large is another factor to take into account at the integration level. One of the crucial points is the identification of the critical effect both in animal and human studies.

In weighing the lines of evidence, the type of question, described carefully in the problem formulation, determines what evidence can be considered as strong.

Currently, the results from *in silico* and specific *in vitro* tests generally tend to make a lesser contribution to the overall weighting. However this may change with time, if more experience is gained on the utility of such information for risk assessment purposes, also providing relevant supporting information for MoA and kinetics.

The key issues in the evaluation of human evidence are to assess whether the results demonstrate a true causal effect, to identify the affected population and to determine to what extent the adverse effects of the exposure might be avoidable.

This involves:

- Estimating the incidence and severity of adverse effects likely to occur in a population due to exposure to a substance
- Addressing several potential toxic effects and human (sub)populations, and considering each (sub)population's exposure by relevant exposure routes
- Focusing on the most critical effect(s) (with consideration of population, route, and time scale)
- Providing quantitative (or if not possible, qualitative) assessment of risk, and
- Characterising the sources and magnitude of uncertainties

For ecological risk assessment, WoE based procedures should overcome the traditional approach of risk characterisation based on the comparison of two numerical values (PEC and PNEC) obtained with univocal and deterministic procedures. As mentioned above, ideally, risk characterisation should be based on probabilistic values for exposure and assessment with statistically determined uncertainty.

If a complete data set would be available for exposure (emissions, modelling and monitoring data) and effects (laboratory data, SSD, higher tier data), the WoE should allow providing the information necessary for a more detailed characterisation of risk, capable to better describe the type of effect likely to occur (losses of biodiversity, reduction of ecosystem services, etc.), as well as the probability of its occurrence

evaluated as a function of the variability of exposed system (vulnerability, ecological value, etc.) and of the uncertainty of the results.

With this procedure, the output of risk assessment should be provided in a form that could be more useful for risk managers and decision makers, in order to better develop a socioeconomic analysis and to evaluate the risks and benefits of management.

8. DESCRIPTION OF UNCERTAINTY IN WoE

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. In its uncertainty analysis, the SCHEER will rely mainly on methods described by EFSA (EFSA, 2017 and EFSA toolbox). It does not mean that all uncertainties will be quantified using the most sophisticated scientific methods available (e.g. a fully probabilistic analysis); this would be inefficient in cases where simpler methods of quantification would provide sufficient information on uncertainty for decision making. In line with EFSA (2017), it is recommended that the combined impact of as many as possible of the identified uncertainties be expressed quantitatively, in terms of the range and probability of possible answers to the assessment question, and that any uncertainties that cannot be included in this should be described qualitatively.

8.1 Identification of significant sources of uncertainty

Sources of uncertainty can be classified as random (those that cannot be predicted) or as systematic (those that are related to personal, procedural, and/or instrumental uncertainty). The random uncertainties decrease the precision of the analysis whilst the systematic uncertainty decreases the accuracy of the analysis.

An important objective of any uncertainty analysis is the description of the critical sources of uncertainty and characterisation of their impact on the Opinion for the formulated problem. Further characterisation of the uncertainties is useful for identifying the priorities for generating more or better data. In the WoE, the uncertainty attributed to individual lines of evidence must be considered, and in the final integration, uncertainty must also be part of the output and communication of the Opinion. The uncertainties in the WoE largely relate to the determination of the relevancy, reliability, validity and consistency (Section 4.2.3). In addition to considering the uncertainties involved in the evidence and its synthesis, it is important to identify significant uncertainties in the judgement used (Meek *et al.*, 2014). These assessments of uncertainty in the WoE should also take into account any other uncertainties affecting the overall assessment. Uncertainties in the risk assessment (human health and environment) are discussed in section 8.2. Qualitative and quantitative expressions of uncertainty are considered in section 8.3. The influence of the uncertainty assessment on risk management is briefly described in section 8.5.

8.2 Uncertainties in risk assessment

8.2.1 Uncertainties in human health risk assessment

Exposure assessment relies on direct measures (experimental personal monitoring or biomarkers of exposure) and/or predictive modelling (e.g. using environmental levels, personal behavior patterns, dispersion modelling and emission scenarios). Many exposure assessments rely on limited data and, in the absence of adequate data, deterministic approaches are used to assess exposures, often using standard default values in the modelling, which are often conservative to ensure that exposure is not underestimated. Thus, (i) worst-case exposure scenarios are used; (ii) conservative assumptions and parameters in exposure modelling are used; and (iii) when deriving numerical values, uncertainty factors are introduced (see, for e.g. IPCS, 2008 for a comprehensive review of uncertainties in exposure assessment).

Uncertainty or scaling factors are also used to take into account the lack of knowledge generated in experimental systems (e.g. need to extrapolate from animals to humans, from *in vitro* to *in vivo*, from high to low doses). When using human data, variability will introduce further uncertainties (e.g. susceptible groups; different exposure routes).

8.2.2 Uncertainties in ecological risk assessment

Experimental monitoring and/or predictive models can be used to assess environmental exposures. In experimental monitoring, uncertainties should be accounted for arising from both the sampling and the analytical phase. The analytical phase may provide a quantitative estimate of precision. Modern technology can produce analytical data with relatively low margins of uncertainty.

The planning of a monitoring campaign should ensure that samples are as representative as possible of the matrix sampled and of its time and space variability. Therefore monitoring should be based on emission patterns and on the environmental characteristics that may affect exposure (rain events, wind, water flow, etc.). In absence of this information, samples may not be representative of the actual variability and the results may be misleading (see, for example, Bonzini *et al.,* 2006).

For modelling approaches, (section 7.2.2), sources of uncertainty may be of different origins and will depend on the selection of (i) the most appropriate model; (ii) the environmental scenario (worst case, best case, most probably realistic case); and (iii) input data (e.g. the properties of the chemical and emission patterns).

A more detailed evaluation of the needs for reducing uncertainties is proposed by Di Guardo and Hermens (2013) and in a SCHER, SCENIHR, SCCS document (EC, 2013).

The major source of uncertainty in the use of (eco)toxicity data is the extrapolation from the hierarchical levels at which tests are performed (in many cases individual) to those ecologically relevant (communities and ecosystems). Most regulations on chemical risk require single species tests on selected indicator organisms assumed as representative of the ecosystem. To cover the uncertainties intrinsic in this simplification, traditional approaches are based on the deterministic use of an application factor (e.g. 1000, 100, 10) decreasing in function of the increase of available information.

The use of species sensitivity distribution (SSD) may reduce and statistically quantify the uncertainty due to the different sensitivity levels of the various species.

Higher tier testing procedures (microcosms, mesocosms, and enclosures) allow a higher ecological realism accounting for ecological interactions and indirect effects. However, they suffer for high variability and poor replicability of the results.

Predictive approaches for effect assessment (QSARs) are not as widely accepted in ERA as for exposure with only relatively simple QSAR approaches being accepted as predictive tools in European regulations (e.g. REACH), which are mainly based on hydrophobicity properties and limited to narcotic and polar narcotic modes of action (EC, 2003; Vighi *et al.,*2008).

8.3 Expression of the uncertainties

A simple scheme is required that is readily understood by both risk assessors and risk managers. Generally, uncertainty may be expressed in several ways, namely using:

- Standardised terms or phrases. Various terms are used by the EU Scientific Committees. However as noted in the SSC Opinion on harmonisation of risk assessment (2000, 2003), there is no consistency in how different terms are used.
- Tabular forms
- Quantitative expression. This is only appropriate if the risk assessment is expressed in probabilistic terms.

These three ways of expressing uncertainty may be regarded as the tiered approach already presented. If there is limited data, the use of standardised terms may be the only one suitable.

8.3.1 Expressing the uncertainty for individual lines of evidence

The tabular presentation from table 3 (EFSA, 2017) is an example of a tabular presentation of uncertainties in the WoE of categorical questions and of expressing the uncertainty of the conclusion. The symbols used are identified in the text just below the table.

Table 3: Expression of uncertainty for individual lines of evidence

Aspect	Nature of the uncertainty	Influence on conclusion	Importance of the uncertainty to the risk assessment
Line of evidence 1 Key aspects: * *		↑, ↑↑, ↑↑↑	
Line of evidence 2 **Key aspects:** * *		↑, ↑↑, ↑↑↑	
Line of evidence x **Key aspects:** * *		↑, ↑↑, ↑↑↑	
Conclusion after integration of lines of evidence		↑, ↑↑, ↑↑↑	

Key to symbols: ↑ , ↑ ↑ , ↑ ↑ ↑ represent minor, intermediate and strong upward influence on probability respectively (see EFSA, 2017).

8.3.2 Qualitative expression of uncertainty

EFSA also suggested using a tabular approach to list and describe sources of uncertainty in parameters and evaluate their nature, magnitude and individual and combined impacts on the assessment outcome quantitatively (EFSA, 2017; Edler *et al.,* 2013). A simplified version was described in SCENIHR (2012).

The table proposed (Table 4) should indicate:

- The direction of any uncertainties, i.e. are they equally distributed or are they most likely to be over- or underestimations of the risk. This requires considering the degree of conservatism used in modelling, etc.
- The magnitude of any uncertainties, i.e. are they likely to be small or large
- The importance of each uncertainty in the overall level of confidence in the conclusions of the risk assessment

It may be helpful to use the following symbols to simplify the expression of the analysis:

Direction of uncertainties

The direction of uncertainties (i.e. whether there is a trend towards an over- or underestimation of the risks) should be expressed by the use of + and – values as follows:

+ The risk could be higher due to the uncertainty

- The risk could be lower due to the uncertainty

+/- There is an equal chance of the uncertainty producing a risk estimate that is either too high or too low

? The direction of the impact of the uncertainty cannot be reasonably estimated

Table 4: Sources and ranges of uncertainty and influence on conclusion

Parameter	Nature of uncertainty	Direction of uncertainty for individual parameter	Influence on conclusion
1	-	+/-/?	
2	-	+/-/?	
X	-	+/-/?	
Assessment output			+/-/?

8.3.3 Quantitative expression of uncertainty

Quantitative approaches of uncertainty express the range of possible outcome and/or the range of the probabilities of the different outcomes.

A complete quantitative expression of uncertainty would specify all the outcomes that are considered possible, including their probabilities. This approach requires comprehensive information and is time-consuming, requiring the appropriate resources. Partial quantitative expression provides only partial information on the probabilities and in some cases partial information on the possibilities (specifying a selection of possible outcomes). Partial quantitative expression requires less information and fewer judgements but may be sufficient in some cases, e.g. for decision-making.

According to EFSA (2017) Quantitative uncertainty can be expressed as

1. Individual values:

Uncertainty partially quantified by specifying some possible values, without specifying what other values are possible or setting upper or lower limits.

2. Bound:

Uncertainty partially quantified by specifying either an upper limit or a lower limit on a quantitative scale, but not both.

3. Range:

Uncertainty partially quantified by specifying both a lower and upper limit on a quantitative scale, without expressing the probabilities of different values within the limits.

4. Bound/Range with probability:

Uncertainty partially quantified by specifying a bound or range with an accompanying probability which may itself be expressed as a bound (bounded probability).

5. Distribution:

Uncertainty fully quantified by specifying the probability of all possible values on a quantitative scale.

High emphasis is given on probabilistic approaches in risk assessment as they provide the most information. 'Probabilistic' is used to express uncertainty and the mathematics of probability for combined uncertainties. Methods for probabilistic evaluation of uncertainty include:

- Methods for obtaining probabilities by statistical analysis of data (confidence intervals, the bootstrap, and Bayesian inference),
- Methods for making probability calculations to combine uncertainties expressed probabilistically (probability bounds analysis, Monte Carlo, and approximate calculations)

According to EFSA, uncertainties should not necessarily be quantified using the most sophisticated scientific methods available (e.g. a fully probabilistic analysis). Moreover, it is important to provide sufficient information on uncertainty for decision making.

8.4 Overall influence of the uncertainties

The expression of the significance of the uncertainties associated with a particular risk assessment taking into account both weighting of evidence and judgemental factors should follow the terminology shown in Table 5. This terminology corresponds to that recommended by EFSA.

Other terms to express certainty and uncertainty should not be used without a supporting text.

Table 5: Scale proposed by EFSA's Guidance on the weight of evidence (2017) for harmonised use in EFSA to express the probability of uncertain outcomes

Probability term	Subjective probability range
Extremely likely	99-100%
Very likely	90-99%
Likely	66-90%
As likely as not	33-66%
Unlikely	10-33%
Very unlikely	1-10%
Extremely unlikely	0-1%

This table is based on calibrated language for describing quantified uncertainty from the Intergovernmental Panel on Climate Change (IPCC). IPCC used ranges of subjective probabilities indicating the chance that a result is true in its synthesis report on Climate Change in 2001 (IPCC, 2001). Acknowledging people's difficulties in understanding numeric probability estimates, IPCC added a qualitative description of these probability ranges. Likelihood may be based on statistical or modelling analyses, elicitation of expert views, or other (semi-)quantitative analyses.

8.5 Explanation of actions on uncertainties to risk managers

The purpose of the risk assessment may be 2-fold: (1) to inform about risk management options for, and prioritisation and decisions by a regulatory agency and (2) to provide regulatory officials with information for communicating with stakeholders and the public

about current risks and the expected risks and benefits after taking a particular action (based on WHO, 2014: EFSA).

Risk assessors need to assess uncertainties in the underlying data as well as in their own interpretations of these data (Levin, *et al.,* 2004). Characterisation of the uncertainties in a risk assessment is important for transparency and should also be a valuable aid to risk managers in determining how to respond to risk management advice. In addition, it is a useful way of indicating priorities for further work to improve the robustness of risk assessments. However, if not clearly and suitably described, the expression of uncertainty may result in inappropriate concerns and/or actions. The degree to which characterisation of uncertainty (and variability) is needed will depend on the risk assessment and risk management contexts as determined in the questions asked, i.e. problem formulation.

An important issue in weighing of evidence is the influence of values on expert judgement due to differences in ideological views. This needs to be recognised and made explicit as far as possible. It may introduce a further degree of subjectivity that is difficult to quantify but which can impact the uncertainties in the risk assessment and consequently the advice to risk managers (Van der Sluijs, 2003 and 2005).

In the explanation of the uncertainties in risk assessment to risk managers, the following questions may be addressed (based on Wardekker *et al.,* 2013):

- What are the type and degree of uncertainties that would be relevant for each target audience, taking into account:
- The questions, problems, tasks and policy challenges it faces;
- The policy phase of the issue being studied;
- Any situations that render uncertainty particularly relevant;
- Possible future developments that should be anticipated in the communication.
- What are the possible implications of uncertainty for the study and for policy?
- What are the main messages of the Opinion and what are the main assumptions?
- What were these assumptions based on? How robust are the Opinion's conclusions in the light of these assumptions and uncertainties?
- Which actions are recommended to reduce the uncertainty in a risk assessment or make allowances for it, such as:

  - The expected imminent need for relevant data,
  - Specific research recommended to substantially reduce the uncertainty,
  - No options for significant reduction of uncertainty in the foreseeable future,
  - Options for precautionary risk management measures recommended to avoid further exposure.

9. CONCLUSIONS/REPORTING

Conducting scientific assessments based on a weight of evidence approach demand a structured and clearly documented process if the outcome of the scientific assessment is to be communicated unambiguously to decision makers, the wider scientific community and stakeholders. This process will help to clearly focus on key issues and allow reproducibility of the assessments between expert groups (Hardy *et al.,* 2015).

The weight of evidence assessment should report the method used for the selection of the individual lines of evidence to ensure that the iterative process leading to the conclusions is fully comprehensible and reproducible. To better achieve this goal, it is advisable that the weight of evidence assessment is reported in a standardised way concerning the choice of methods and all assumptions used, including expert judgement (EFSA Guidance on WoE, 2017). Biological relevance and associated uncertainty should also be addressed and reported as part of the weight of evidence assessment (EFSA Guidance on Biological Relevance of data in scientific assessments, 2017). Where the assessment methods used are already described in other documents, it is sufficient to cross-reference.

To draw conclusions, it is not recommended to simply add together weighting from individual lines of evidence. Integration implies that the combination of conclusions from different lines of evidence and their consistency (e.g. evidence for a MoA, explaining the occurrence of a certain effect in experimental animals and not in humans or vice-versa) is the final goal, but this demands an element of expert judgement.

According to this Memorandum, reporting should be performed at each step of the iterative process of building the WoE:

Step 1 - Problem formulation should be purpose oriented and conducted with the correct understanding of the relevant questions. The major issues are to identify the risk assessment context and the target users.

Step 2 - Identification, collection and selection of the possible sources of evidence: the documentation of the search strategy should be presented for transparency. In this sense, the summary of the methods used to search, select and extract the lines of evidence should include whether an extensive literature search or systematic review was conducted, and whether any of the evidence was obtained by expert elicitation, and if so, by which method. Information of what is taken forward for the assessment including impact of what is ignored should be described for derivation of confidence levels/remaining uncertainty. Any lines of evidence that are required (e.g. by legislation or guidance documents) but that are missing should also be identified, i.e. data gaps.

Step 3 – Assessment and weighing of individual lines of evidence: a partial weighing of evidence occurs firstly at the step of assessment of the individual lines of evidence. Relevance, validity and reliability, as described in 5.2.3, are used in weighing the evidence for the purpose of the assessment.

The weighing of evidence occurs at the step of the assessment, combining the overall quality with consistency and plausibility of the contributing lines of evidence. Methodologies may vary in terms of complexity and range from general schemas/frameworks consisting of a set of questions to more elaborate methodologies. Examples of frameworks that contain elements of weighing evidence include Integrating Testing Strategy (ITS), WHO/IPCS Mode of Action (MoA) and Advanced Outcome Pathways (AOPs). The detailed results of weighing the evidence must be presented in an appropriate part of the assessment report, in a format that helps the reader to compare the results for the different lines of evidence (e.g. a tabular listing). In case of conflicting study results, the weight allocated to each study will be case-dependent (depending on the test method, quality of the data and the endpoint under consideration).

Step 4 - Integration of different lines of evidence: the methods used to integrate the lines of evidence should be briefly summarised, giving enough information to make clear

the type of method involved. If weighing and integration were done as per an iterative process, the reader should be referred to where that is described. The conclusion of integrating the evidence for a specific question should be stated in a form that expresses the range of possible answers and support for those answers.

Step 5 – Description of uncertainties: when developing a Scientific Opinion, it is important to communicate what is known and what is not known and to state potential uncertainties. Uncertainty levels can be usually reported qualitative as: high, medium, low. The use of narrative forms backed up with diagrams (where appropriate) can be useful. It is advisable, if possible, to quantify the uncertainties impact on the Opinion for the formulated problem. However, the uncertainty analysis should be formulated for the public audience, so that it can be well communicated and perceived.

As a concluding remark, this memorandum is a living document and is intended to make explicit the approach used by the SCHEER for determining the weight of evidence and the uncertainties involved in the development of its Opinions. It involves a staged approach. The approach draws on a number of schemes that have been developed by various national and international bodies. However, it introduces a number of additional elements that are considered to benefit both transparency and consistency.

Particular attention has been paid to ensuring that the format can be applied to a wide range of lines of evidence and types of publication.

9.1 Reporting of the weight of evidence

In the process of integrating lines of evidence, information could be evaluated by attributing different weights to the available data.

A well-defined and consistent framework for reporting can make it easier to reach conclusions in risk assessment and indicate confidence in the findings. If the risk assessment covers both human and environmental risks, separate tables should be constructed.

The report should document all steps of the procedure in sufficient detail for them to be repeated (starting from collection of the relevant information, to weight and integration of the different line of evidence) making clear how expert judgement has been used in a transparent way. The weighing of the overall evidence may be presented in a table, to help people external to the Scientific Committee understand how conclusions have been reached. In the Annex, an example is given for the reporting in tabulated format. However, only those lines of evidence, relevant for the specific opinion need to be filled in. Moreover, any deviation from reporting schemes which enhances transparency and comprehensibility of the WoE process is possible.

Strong weight of evidence: coherent evidence from a primary line of evidence (human, animal, environment) and one or more other lines of evidence (in particular mode/ mechanistic studies) in the absence of conflicting evidence from one of the other lines of evidence (no important data gaps).

Moderate weight of evidence: good evidence from a primary line of evidence but evidence from several other lines is missing (important data gaps).

Weak weight of evidence: weak evidence from the primary lines of evidence (severe data gaps).

Uncertain weight of evidence: uncertain evidence due to conflicting information from different lines of evidence that cannot be explained in scientific terms.

Weighing of evidence not possible: no suitable evidence available.

In each case, free text is required to explain the assignment. It is important to identify studies that appear to have been well conducted but generate findings that are very different (outliers) from those of other studies in the same line of evidence. Inconsistencies between apparently very similar, good quality studies also need to be addressed in the final risk assessment along with comments on possible unknowns.

10. ABBREVIATIONS AND GLOSSARY OF TERMS

AOPs	Adverse outcome pathways (AOPs)
BEAT	a computerised database of exposure data (largely for occupational settings)
ECETOC	European Centre for Ecotoxicology and Toxicology of Chemicals
ERA	Ecological Risk Assessment
FAO	Food and Agriculture Organization of the United Nations
GHS	Globally Harmonised System of Classification and Labelling of Chemicals
GLP	Good Laboratory Practices
GRADE	Grading of Recommendations Assessment, Development and Evaluation
IARC	International Agency For Research On Cancer
ILSI	International Life Science Institute Europe
JECFA	The Joint FAO/WHO Expert Committee on Food Additives
MoA	Mode of action
NIEHS	National Institute of Environmental Health Sciences
OECD	The Organisation for Economic Co-operation and Development
PECs	predicted environmental concentrations
RCT	Randomised clinical trials
SSD	Species Sensitivity Distribution
TFEEA	OECD Task Force on Environmental Exposure Assessment
TGD	European Technical Guidance Document on risk assessment
WHO	World Health Organisation
WMO	World Meteorological

	Organization
WWF	World Wildlife Fund

11 REFERENCES

Administration, OSHA Occupational Safety and Health. (2016). "Guidance on Data Evaluation for Weight of Evidence Determination: Application to the 2012 Hazard Communication Standard."

Agerstrand, M., and A. Beronius. (2015). "Weight of evidence evaluation and systematic review in EU chemical risk assessment: Foundation is laid but guidance is needed." Environ Int. doi: 10.1016/j.envint.2015.10.008.

Balshem H., Helfand M., Schünemann H.J. (2011). GRADE guidelines: 3. Rating the quality of evidence". J Clin Epidemiol. 2011, 64 (4): 401–06).

Becker R. A., Ankley G. T.,. Edwards S. W, Kennedy S. W., Linkov I., Meek B., Sachana M., Segner H., Van Der Burg B., Villeneuve D. L., Watanabe H., and Barton-Maclaren T. S.. (2015). "Increasing Scientific Confidence in Adverse Outcome Pathways: Application of Tailored Bradford-Hill Considerations for Evaluating Weight of Evidence." Regul Toxicol Pharmacol 72 (3):514-37. doi: 10.1016/j.yrtph.

Blank H., Dave G., Gustafsson K. (1978). An annotated literature survey of methods for determination of effects and fate of pollutants in aquatic environments. Swedish Board for Technical Development. Solna, Sweden.

Bonzini, S., Verro, R., Otto, S., Lazzaro, L., Finizio, A., Vighi, M. (2006). Experimental validation of a GIS-based procedure for predicting pesticide exposure in surface water. Environ. Sci. Technol., 40, 7561-7569.

Boobis A.R., Cohen S.M., Dellarco V., McGregor D., Meek M.E., Vickers C., Willcocks D. and Farland W. (2006). IPCS framework for analysing the relevance of a cancer mode of action for humans Crit Rev Toxicol 36 781-792.

Boobis A.R., Doe J.E., Heinrich-Hirsch B., Meek M.E., Munn S., Ruchirawat M., Schlatter J., Seed J. and Vickers C. (2008). IPCS framework for analysing the relevance of a noncancer mode of action for humans Crit Rev Toxicol 38, 87-96.

Borgert, C. J., E. M. Mihaich, L. S. Ortego, K. S. Bentley, C. M. Holmes, S. L. Levine, and R. A. Becker. (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. doi: 10.1016/j.yrtph.2011.07.007.

Borgert, C. J., E. V. Sargent, G. Casella, D. R. Dietrich, L. S. McCarty, and R. J. Golden. (2012). The human relevant potency threshold: reducing uncertainty by human calibration of cumulative risk assessments. Regul Toxicol Pharmacol 62 (2):313-28. doi: 10.1016/j.yrtph.2011.10.012.

Bui-Klimke T., Wu F. (2014). Evaluating weight of evidence in the mystery of Balkan endemic nephropathy. Risk Anal 34 (9):1688-705. doi: 10.1111/risa.12239.

Collier Z. A., Gust K. A., Gonzalez-Morales B., Gong P., Wilbanks M. S., Linkov I., Perkins E. J. (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.

DFG. (2010). List of MAK and BAT values. Wiley.VCH, Weinheim, Germany.

Di Guardo A., Hermens J. (2013). Challenges for exposure prediction in ecological risk assessment. Integr. Environ. Assess. Manag., 9, e4–e14.

Dunjo J., Fthenakis V., Vilchez J.A. and Arnaldos J. (2010). Hazard and operability (HAZOP) analysis. A literature review J Hazard Mats 173, 19-32.

European Commission. (2003). Technical Guidance Document (TGD) on Risk Assessment of Chemical Substances (2nd edition). European Commission, European Chemical Bureau, Joint Research Centre, EUR 20418 EN/2

EC (European Commission). (2013). SCHER, SCENIHR, SCCS Opinion on: Addressing the New Challenges for Risk Assessment. European Commission: Brussels, Belgium, p. 154.

ECHA. (2010). Practical guide 2: How to report weight of evidence. ISBN-13: 978-92-9217-028-8.

ECHA. (2017). Weight of Evidence/Uncertainty, Available at: https://echa.europa.eu/support/guidance-on-reach-and-clp-implementation/formats).

ECHA. (2008). Guidance on information requirements and chemical safety assessment. Part D: Exposure scenario Building. ECHA, Helsinki, Finland. www.echa.europa.eu

EFSA. (2017). Guidance on the use of the weight of evidence approach in scientific assessments, EFSA Journal 2017;15(8):4971 [69 pp.], DOI: 10.2903/j.efsa.2017.4971

EFSA. (2006). Guidance of the Scientific Committee on a request from EFSA related to uncertainties in dietary exposure assessment. The EFSA Journal 438:1-54.

EFSA. (2015). Principles and process for dealing with data and evidence, EFSA Journal 2015;13(5):4121 [36 pp.], DOI: 10.2903/j.efsa.2015.4121

EFSA. (2010). Systematic review methodology and food and feed safety risk assessment, EFSA Journal 2010; 8(6):1637, DOI: 10.2903/j.efsa.2010.1637

Forbes V.E., Calow P. (2012). Promises and problems for the new paradigm for risk assessment and an alternative approach involving predictive systems models Environ Toxicol Chem 31:2663-2671.

Forbes V.E., Palmquist A, Bach L. (2006). The use and misuse of biomarkers in ecotoxicology. Environ Toxicol Chem 25:272-280.

Gamble C and Hollis S (2005) Uncertainty method improvement on best-worst case analysis in a binary meta-analysis J Clin Epidemiol 58, 579-588

Gee D. (2008). Establishing evidence for early action: the prevention of reproductive and developmental harm. Basic Clin Pharmacol Toxicol 102:71-72.

Hardy A., Dorne JLCM, Aiassa E., Alexander J., Bottex B., Chaudhry Q., Germini A., Nørrung B., Schlatter J., Verloo D., Robinson T. (2015). Editorial: Increasing robustness, transparency and openness of scientific assessments. EFSA Journal 2015;13(3):e13031. 3 pp. doi:10.2903/j.efsa.2015.e13031

Hart A, Roelofs W, Hardy AR, Macleod A. (2007). Comparative review of risk terminology. A comparative review of terminology and expressions used by the three scientific committees established by Commission Decision 2004/210/EC and by their predecessors established by Commission Decision 97/579/EC (repealed by Commission Decision 2004/210/EC). Report No. S12.454739. The Central Science Laboratory, DEFRA, Sand Hutton, York, UK.

Hommen U., Baveco J.M., Galic J.N., van den Brink P.J. (2010). Potential Application of Ecological Models in the European Environmental Risk Assessment of Chemicals I: Review of Protection Goals in EU Directives and Regulations. Integrated Environmental Assessment and Management, 6, 325–337

Hristozov, D. R., Gottardo S., Cinelli M., Isigonis P., Zabeo A., Critto A., Van Tongeren M., Tran L., and Marcomini A.. (2014). Application of a quantitative weight of evidence approach for ranking and prioritising occupational exposure scenarios for titanium dioxide and carbon nanomaterials. Nanotoxicology 8(2):117-31. doi: 10.3109/17435390.2012.760013.

IARC. (2006). The Preamble to the IARC Monographs. IARC Monographs on the Evaluation of Carcinogenic Risks to Humans, Lyon, France.

IPCC. (2001). Thirty-third session of the IPCC, Abu Dhabi, 10-13 May 2011 IPCC-XXXIII/Doc. 12, Add.2 (28.IV.2011)

IPCS. (2009). Risk assessment of combined exposures to multiple chemicals: a WHO/IPCS framework. WHO/IPCS, Geneva, Switzerland. Harmonization Project DRAFT Document for Public and Peer Review

IPCS. (2014). Guidance document on evaluating and expressing uncertainty in hazard characterization. ISBN 978 92 4 150761 5,

IPCS. (2008). Uncertainty and data quality in exposure assessment. Part 2: Hallmarks of data quality in chemical exposure assessment. WHO/IPCS, Geneva, Switzerland. IPCS Harmonization Project Document No. 6. www.who.int/ipcs

Ippolito A, Giacchini R, Parenti P, Vighi M. (2016). Natural variability of enzymatic biomarkers in freshwater invertebrates. Environ Sci Pollut Res DOI 10.1007/s11356-016-7833-4

Kalberlah F, Schneider K and Schuhmacher-Wolt U. (2003). Uncertainty in toxicological risk assessment for non-carcinogenic health effects Regulat Toxicol Pharmacol 37, 92-104

Klimisch HJ, Andreae M and Tilmann U. (1997). A systematic approach for evaluating the quality of experimental toxicological and ecotoxicological data Regul Tox and Pharmacl 25 1-5

Levin R, Jansson SO, Rudén C. (2004). Indicators of uncertainty in chemical risk assessments. Regul Toxicol Pharm 39:33-43.

Linkov I., Loney D., Cormier S., Satterstrom F.K. and Bridges T. (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., Welle P., Loney D., Tkachuk A., Canis L., Kim J. B., and Bridges T. (2011). Use of multicriteria decision analysis to support weight of evidence evaluation. Risk Anal 31 (8):1211-25. doi: 10.1111/j.1539-6924.2011.01585.x.

Linkov I., Massey O., Keisler J., Rusyn I., and Hartung T. (2015). From "weight of evidence" to quantitative data integration using multicriteria decision analysis and Bayesian methods. Altex 32 (1):3-8. doi: http://dx.doi.org/10.14573/altex.1412231.

Lutter R., Abbott L., Becker R., Borgert C., Bradley A., Charnley G., Dudley S., Felsot A., Golden N., Gray G., Juberg D., Mitchell M., Rachman N., Rhomberg L., Solomon K., Sundlof S., and Willett K.. (2015). Improving weight of evidence approaches to chemical evaluations. Risk Anal 35 (2):186-92. doi: 10.1111/risa.12277.

Meek M.E., Boobis A., Cote I., Dellarco V., Fotakis G., Munn S., Vickers C. (2014). New developments in the evolution and application of the WHO/IPCS framework on mode of action/species concordance analysis. J. Appl. Toxicol. 2014; 34: 1–18. doi: 10.1002/jat.2949. Epub 2013 Oct 25.

Moermond C.T.A., Kase R., Korkaric M., Ågerstrand M. (2016). CRED: Criteria for reporting and evaluating ecotoxicity data. Environmental Toxicology and Chemistry. Doi: 10.1002/etc.3259.

Morgan R. L., Thayer K. A., Bero L., Bruce N., Falck-Ytter Y., Ghersi D., Guyatt G., Hooijmans C., Langendam M., Mandrioli D., Mustafa R. A., Rehfuess E. A., Rooney A. A., Shea B., Silbergeld E. K., Sutton P., Wolfe M. S., Woodruff T. J., Verbeek J. H., Holloway A. C., Santesso N., Schunemann H. J.. (2016). GRADE: Assessing the quality of evidence in environmental and occupational health.

Nendza, M., Aldenberg., Benfenati, E., Begnini, R., Cronin, M.T.D., Escher, S., Fernandez, A., Gabbert, S., Giralt, F., Hewitt, M., Hrovat, M., Jeram, S., Kroese, D., Madden, J.C., Mangelsdorf, I., Rallo, R., Roncaglioni, A., Rorije, E., Segner, H., Simon-Hettich, B., Vermeire, T. (2010). Data quality assessment for in silico methods: a survey of approaches and needs. Chapter 4 in: Cronin, M.T.D., Madden, J.C. (eds.) In silico

toxicology, principles and applications. RSC Publishing, Cambridge, UK. ISBN 978-1-84973-004-4

NTP/OHAT. Handbook for Conducting a Literature-Based Health Assessment Using OHAT Approach for Systematic Review and Evidence Integration. (2015). P 1-94.

OECD (2000). Report of the OECD Workshop on Improving the Use of Monitoring Data in the Exposure Assessment of Industrial Chemicals. OECD Environmental Health and Safety Publications. Series on Testing and Assessment No. 18, Paris.

OECD (2014). New guidance document on an integrated approach on testing and assessment (iata) for skin corrosion and irritation. Series on testing and assessment. No. 203, 2014

OECD (2016). Draft guidance document on an integrated approach on testing and 8 assessment (iata) for serious eye damage and eye irritation, 2016

OECD (2007). Guidance document on the validation of (Q)SARs. Organisation of Economic Cooperation and Development, Paris, France. Series of Testing and Assessment 69

Perrier F., Giorgis Allemand L., Slama R., (2016). Within-subject pooling of biological samples as a way to reduce exposure misclassification in biomarker-based studies of chemicals with high temporal variability. Epidemiology 27(3):378-388.

Rhomberg L. (2015). Hypothesis-Based Weight of Evidence: An Approach to Assessing Causation and its Application to Regulatory Toxicology. Risk Anal 35 (6):1114-24. doi: 10.1111/risa.12206

Rorije E., Aldenberg T., Buist H., Kroese D., Schuurmann G. (2013). The OSIRIS Weight of Evidence approach: ITS for skin sensitisation. Regul Toxicol Pharmacol 67 (2):146-56. doi: 10.1016/j.yrtph.2013.06.003.

Sassi G. and Ruggeri B. (2008). Uncertainty evaluation of human risk analysis (HRA) oc chemicals by multiple exposure routes Risk Analysis 28 1343-1356.

Scarduelli L., Giacchini R., Parenti P., Migliorati S., Di Brisco A., Vighi M. (2017). Natural variability of biochemical biomarkers in the macro-zoobenthos: dependence on life stage and environmental factors. Environ Toxicol Chem. (in press)

SCCS (Scientific Committee on Consumer Safety). (2016). Notes of Guidance for the Testing of Cosmetic Ingredients and their Safety Evaluation 9th revision, 29 September 2015, SCCS/1564/15, revision of 25 April 2016

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

Linder S. H., & Sexton K. (2011). Conceptual Models for Cumulative Risk Assessment. American Journal of Public Health, 101(Suppl 1), S74–S81. http://doi.org/10.2105/AJPH.2011.300318 Sperber WH (2005) HACCP and transparency Food Control, 16, 505-509

Thayer K. (2012). Systematic Review and New Tools for Information Management and Data Display. National Toxicology Program (NTP), Office of Health Assessment and Translation, 1p.

U.S. Preventive Services Task Force (August 1989). Guide to clinical preventive services: report of the U.S. Preventive Services Task Force. DIANE Publishing. pp. 24–. ISBN 978-1-56806-297-6

Van de Meent D. De Bruijn JHM. (2007). Environmental exposure assessment. In: Van Leeuwen CJ, Vermeire, TG., Risk assessment of chemicals: an introduction. Springer, ISBN 978-1-4020-6101-1

Van der Sluijs JP, Craye M, Funtowicz S, Kloprogge P, Ravetz J, Risbey J. (2005). Combining quantitative and qualitative measures of uncertainty in model-based environmental assessment: the NUSAP system. Risk Anal 25:481-492.

Van der Sluijs JP, Risbey JS, Kloprogge P, Ravetz JR, Funtowicz SO, Quintana SC, Pereira AG, De Marchi B, Petersen A, Janssen PHM, Hoppe R, Huijs SWF. (2003). RIVM/MNP guidance for uncertainty assessment and communication. Detailed Guidance. Utrecht University, Utrecht, The Netherlands, ISBN 90-393-3536-2.

Verhaar, H.J.M., van Leeuwen, C.J., Hermens, J.L.M. (1992). Classifying environmental pollutants. structure–activity relationships for prediction of aquatic toxicity. Chemosphere 25, 471–491.

Vighi M., Migliorati S., Monti G. S. (2008) QSAR equation for narcotic and polar narcotic toxicity on the luminescent bacterium Vibrio fisheri (Beijerinck). Ecotox. Environ. Saf. 72, 154-162

Wardekker J.A., Kloprogge P., Petersen A.C., Janssen P.H.M., van der Sluijs J.P. (2013) Guide for uncertinty cmmunication. PBL, Netherlands Environmental Assessment Agency.

Weed D.L. (2005). Weight of Evidence: a review of concept and methods. Risk Anal 25:1545-1557.

WHO/IPCS. (2014). Guidance document on evaluating and expressing uncertainty in hazard characterization. IPCS harmonization project document; no. 11. World Health Organisation, International Programme on Chemical Safety. ISBN 978 92 4 150761 5.

Williams, P. R., G. S. Dotson, and A. Maier. (2012). "Cumulative Risk Assessment (CRA): transforming the way we assess health risks." Environ Sci Technol 46 (20):10868-74. doi: 10.1021/es3025353.

Annex

Example for reporting the *contribution of the different lines of evidence to the weight-of-evidence*. Only lines relevant for the specific question need to be considered and filled in. Any deviation from reporting schemes which enhances transparency and comprehensibility of the WoE process is possible.

	Strong	**Moderate**	**Weak**	**Uncertain**	**Not possible**
A. Human Health					
Exposure assessment					
External Exposure measurement					
Internal exposure measurement (Biomonitoring)					
Exposure modelling					
Hazard assessment					
Epidemiologic studies Human volunteer studies Kinetics studies (ADME) Animal studies *In vitro* studies Mathematical models, structure activity and other *in silico* data Studies on MoA,					
B. Environment					
Exposure assessment					
Emission studies					
Environmental fate modelling					
Quantitative exposure modelling					
Experimental monitoring data					
Effect assessment					
Single species ecotoxicity studies					
QSARs					
SSD					
Higher tier (micro and mesocosm) studies					
Field studies					
Conclusion from the totality of evidence (based on expert judgement, short description)					