OUTPUT CATEGORY

EFSA Journal

ADOPTED:

doi:10.2903/j.efsa.20YY.NNNN

Guidance on The Use of the Weight of Evidence Approach in Scientific Assessments

EFSA Scientific Committee,
Maged Younes (Chair), Jan Alexander, Emilio Benfenati, Qasim Mohammad Chaudhry, Peter Craig, Geoff Frampton, Matthias Greiner, Tony Hardy, Andrew Hart, Christer Hogstrand, Claude Lambre, Robert Luttik, David Makowski, Alfonso Siani , Christian Gortazar Schmidt, Helene Wahlstroem, Jaime Aguilera, Jean-Lou Dorne, Laura Martino, Nikolaos Georgiadis

ABSTRACT

(Max. 300 words, no paragraph breaks; no tables, footnotes, graphs or figures. Note that the abstract should end with the copyright)

Keywords: (risk assessment, weight of evidence, biological relevance, uncertainty, lines of evidence,)

Requestor: add requesting party

Question number: EFSA-Q-YYYY-NNNNN

Correspondence: xxx@efsa.europa.eu

Panel [or Scientific Committee] members: [add names in the format Name Surname, Name Surname and Name Surname].

Acknowledgements: The [Panel or Scientific Committee or EFSA] wishes to thank the following for the support provided to this scientific output: [staff members or others who made a contribution but are not eligible as authors]. The Panel [Panel/Scientific Committee/EFSA] wishes to acknowledge all European competent institutions, Member State bodies and other organisations that provided data for this scientific output.

Suggested citation: [EFSA ACRONYM Panel (EFSA Panel name)] [or EFSA (European Food Safety Authority)] [or EFSA Scientific Committee], [add individual author names in the same order as it is on the first page, followed by a comma, in the format: Surname Initial(s), Surname Initial(s) and Surname Initial(s)], 20YY. [Full title, including output category]. EFSA Journal 20YY;volume(issue):NNNN, 91 pp. doi:10.2903/j.efsa.20YY.NNNN

ISSN: 1831-4732

Summary

EFSA requested the Scientific Committee to develop a guidance document on the use of the weight of evidence approach in scientific assessments for use in all areas under EFSA's remit.

The guidance document addresses the use of the weight of evidence in scientific assessments using both qualitative and quantitative approaches. Several case studies covering the various areas under EFSA's remit are annexed to the guidance document to illustrate the applicability of the proposed approach.

In developing the guidance, the WG of the SC took into account other EFSA activities and related European and international activities to ensure consistency and harmonisation of methodologies in order to provide an international dimension to the guidance and avoid duplication of the work.

This guidance document is intended to guide EFSA panels and staff on the use of the weight of evidence approach in scientific assessments. It provides a flexible framework that is applicable to all areas within EFSA's remit, within which assessors can apply those methods which most appropriately fit the purpose of their individual assessment.

Weight of evidence assessment is a process in which evidence is integrated to determine the relative support for possible answers to a question. This document considers the weight of evidence assessment as comprising three basic steps: 1. assembling the evidence, 2. weighing the evidence, 3. integrating the evidence.

The present document defines reliability, relevance and consistency, in terms of their contributions to the weight of evidence assessment: Reliability is the extent to which the information comprising a line of evidence is correct. Relevance is the contribution a line of evidence would make to answer a specified question, if the information comprising the line of evidence was correct. Consistency is the extent to which the contributions of different lines of evidence to answering the specified question are compatible.

While no specific methods are prescribed, a list of criteria for comparing weight of evidence methods is provided to assist in evaluating the relative strengths and weaknesses of the different methods. The criteria do not necessarily have equal importance: their relative importance may be considered on a case-by-case basis when planning each weight of evidence assessment.

All EFSA scientific assessments must include consideration of uncertainties, reporting clearly and unambiguously what sources of uncertainty have been identified and what their impact on the assessment outcome is.

Reporting should be consistent with EFSA's general principles regarding transparency and reporting. In a weight of evidence assessment this should include justifying the choice of methods used, documenting all steps of the procedure in sufficient detail for them to be repeated, and making clear where and how expert judgement has been used. Reporting should also include referencing and, if appropriate, listing or summarising all evidence considered, identifying any evidence that was excluded; detailed reporting of the conclusions; and sufficient information on intermediate results for readers to understand how the conclusions were reached.

EFSA Panels and Units are encouraged to review their existing approaches to weight of evidence assessment in the light of the guidance document, and to consider in particular:

- Whether all pertinent aspects of reliability, relevance and consistency are addressed,
- How to ensure the transparency of weight of evidence assessments,
- Carry out some case studies to assess whether additional methods described in the guidance would add value to their scientific assessments.

Table of contents

1. Introduction

1.1. Background and Terms of Reference as provided by EFSA

Background

EFSA's Science Strategy 2012-2016 has identified four strategic objectives: 1. further develop excellence of EFSA's scientific advice, 2. optimise the use of risk assessment capacity in the EU, 3. develop and harmonise methodologies and approaches to assess risks associated with the food chain, 4. strengthen the scientific evidence for risk assessment and risk monitoring. In this context, the harmonisation and development of new methodologies for risk assessment and scientific assessments is of critical importance to deliver EFSA's science strategy. For this purpose, a number of projects have recently started at EFSA to address individual and cross-cutting methodological issues within the whole scientific assessment landscape. The Assessment Methodological unit (AMU) of EFSA has started a project (PROMETHEUS) with the objective of supporting the coordination and consistency of all EFSA projects that aim at developing or refining methodological approaches. Such an umbrella project will provide a definition of the guiding principles for evidence-based assessments and a collection of available approaches and will identify areas where methods or tools are needed to fulfil such guiding principles.

In July 2013, the Scientific Committee (SC) of EFSA published an opinion on "priority topics for the development of risk assessment guidance by EFSA's SC" which used a number of criteria to make recommendations for the preparation of new or the revision of existing guidance documents as follows:

- Across panel relevance
- Critical importance including urgency of topic to be addressed for several Panels
- Topic not being addressed by an individual Panel
- Sufficient information available to develop meaningful guidance
- International dimension

From this prioritisation exercise, the SC opinion identified three priority topics for 2014: uncertainty analysis, biological relevance, and the use of the weight of evidence (weight of evidence) in scientific assessments (EFSA Scientific Committee, 2013a).

The latter is the subject of this project. The weight of evidence has been defined by the WHO as "a process in which all of the evidence considered relevant for a risk assessment is evaluated and weighted" (WHO, 2009). The SC of EFSA used the WHO definition and pointed out that evidence can be derived from several sources such as white literature (peer reviewed scientific publications), grey literature (reports on websites of governmental, nongovernmental, intra-governmental agencies etc.) and black literature (confidential reports). In order to increase transparency in the risk and other scientific assessment processes, it is important to provide a methodology to select, weigh and integrate the evidence in a systematic, consistent and transparent way to reach the final conclusions and to identify related uncertainties (SCENIHR, 2012; EFSA

Scientific Committee, 2013a). In addition, the SC of EFSA noted that part of the overall weighing of the evidence deals with the evaluation of equivalent or similar questions performed by other international bodies and the adequacy of such evaluations should be judged by EFSA before taking them into account. This is particularly helpful in cases for which the information available is so extensive that it is beyond the capability of a single evaluation to judge each individual study, report, publication by itself. In addition, systematic reviews (SRs) may be very useful. However, the adequacy of the process, the pertinence to the risk assessment, the nature of the question and the inclusion and exclusion criteria should be transparently evaluated by EFSA before taking SRs into account (EFSA Scientific Committee, 2013a). Considering the example of chemical risk assessment, the Weight of evidence approach requires expert judgment of distinct lines of evidence (in vivo, in vitro, in silico, population studies, modelled and measured exposure data etc.) which may come from studies conducted according to official guidelines (e.g. OECD) or from non-standardised methodologies. In this context, data from all sources and categories of literature should be considered for the risk assessment processes, as appropriate to determine their quality and relevance. These considerations should then be reflected in the relative weight given to the evidence in the scientific assessment and transparently taken into account in the overall evaluation of uncertainty (EFSA Scientific Committee, 2013a). It is therefore proposed that the SC of EFSA develop guidance on the use of the weight of evidence approach in scientific assessments.

Terms of reference

EFSA requests the Scientific Committee to develop a guidance document on the use of the Weight of evidence approach in scientific assessments for use in all areas under EFSA's remit.

The guidance document should address the use of the weight of evidence in scientific assessments using both qualitative and quantitative approaches. Several case studies covering the various areas under EFSA's remit should be annexed in the guidance document to illustrate the proposed approaches.

In developing the guidance, the WG of the SC should take into account other EFSA activities and related European and international activities to ensure consistency and harmonisation of methodologies, to provide an international dimension to the guidance and avoid duplication of the work.

In line with EFSA's policy on openness and transparency, EFSA will publish a draft version of the guidance document for public consultation to invite comments from the scientific community and stakeholders. Subsequently, the guidance document and the results of the public consultation should be presented at an international event after publication.

1.2. Interpretation of the terms of reference

In the context of risk assessment, various formal definitions and synonyms have been offered by IPCS, US-EPA, WHO FAO, US National Research Council's Committee, SCHER, SCENIHR and SCCS, on Improving Risk Analysis Approaches for the phrase 'weight of evidence' or 'evidence synthesis'.

When addressing the mandate, the Scientific Committee acknowledged that the issue of weight of evidence in risk assessment encompasses aspects related to the reliability of the various pieces of evidence used in the assessment.

In order for the guidance document to address the use of the weight of evidence in scientific assessments using both qualitative and quantitative approaches, a list of the available approaches used globally has been provided together with several case studies from various areas under EFSA's remit to illustrate the proposed approaches.

In developing the guidance, the WG of the SC has taken into account other EFSA activities and related European and international activities to ensure consistency and harmonisation of methodologies, to provide an international dimension to the guidance and avoid duplication of the work.

In particular:

-Relevant guidance published by the SC on related subjects (transparency in risk assessment, uncertainty in exposure assessment, statistical significance and biological relevance (EFSA, 2006, 2009a, 2011) and the latest draft guidance documents on uncertainty and biological relevance that are being developed concomitantly.

The guidance on uncertainty analysis in risk assessment. It deals specifically with reporting and analysing uncertainties using qualitative and quantitative methods for all work within EFSA's remit. The overlaps with the weight of evidence approach should be carefully taken into account by both WGs to ensure that the weight of evidence and the uncertainty guidance documents are consistent, use harmonised methodologies, and do not duplicate the work.

- The guidance on biological relevance for all areas of work within EFSA's remit. It deals specifically with criteria to evaluate biological relevance in scientific assessments and the overlaps with the weight of evidence approach should be carefully taken into account by both WGs to ensure consistency, the use of harmonised methodologies, and to avoid duplication.

- The current work within EFSA on the approach to evidence-based risk assessment to ensure consistency, the use of harmonised methodologies, and to avoid duplication of the work.

The guidance was also expected to take into account other European and international activities:

-The work of the EC's non-food scientific committees and other agencies on weight of evidence approach, and where appropriate, seek their participation for the development of a harmonised guidance.

-The latest international developments in weight of evidence approaches should also be taken into account. Examples include best practices in weight of evidence methodologies, the use of systematic review in risk assessment, the WHO application of the weight of evidence approach in relation to the mode of action framework and other related international developments (EFSA, 2010; Rhomberg et al., 2013; Meek et al., 2014, Becker et al., 2015).

1.3. Aim and scope of the document

Weighing the evidence is an inherent part of every scientific assessment performed by EFSA. Experts review all available data, and come to conclusions based on an assessment of their overall confidence in the results of all reviewed studies. The approaches and methods used in conducting such an "informal", inherent weighing of the evidence are mostly not spelled out, however.

The aim of this guidance document is to provide a general framework for considering and documenting the approaches used to weigh the evidence in answering the main question of each scientific assessment or questions that need to be answered in order to provide, in conjunction, an overall answer. The document further indicates, in general terms, types of qualitative and quantitative approaches to weigh and integrate evidence, and lists individual methodologies with pointers as to where details of these can be found. Finally, the document provides suggestions for conducting and reporting of weight of evidence and evidence integration.

The document does not attempt to prescribe approaches or methods to be used, nor does it provide a comprehensive description of all methods that can be used.

1.4. Relation to other relevant EFSA guidance documents

The guidance on the use of the weight of evidence builds on the conceptual approach for scientific assessments as described in PROMETHEUS (ref.), which describes the overall process for dealing with data and evidence. The process has four steps as shown in figure 1:

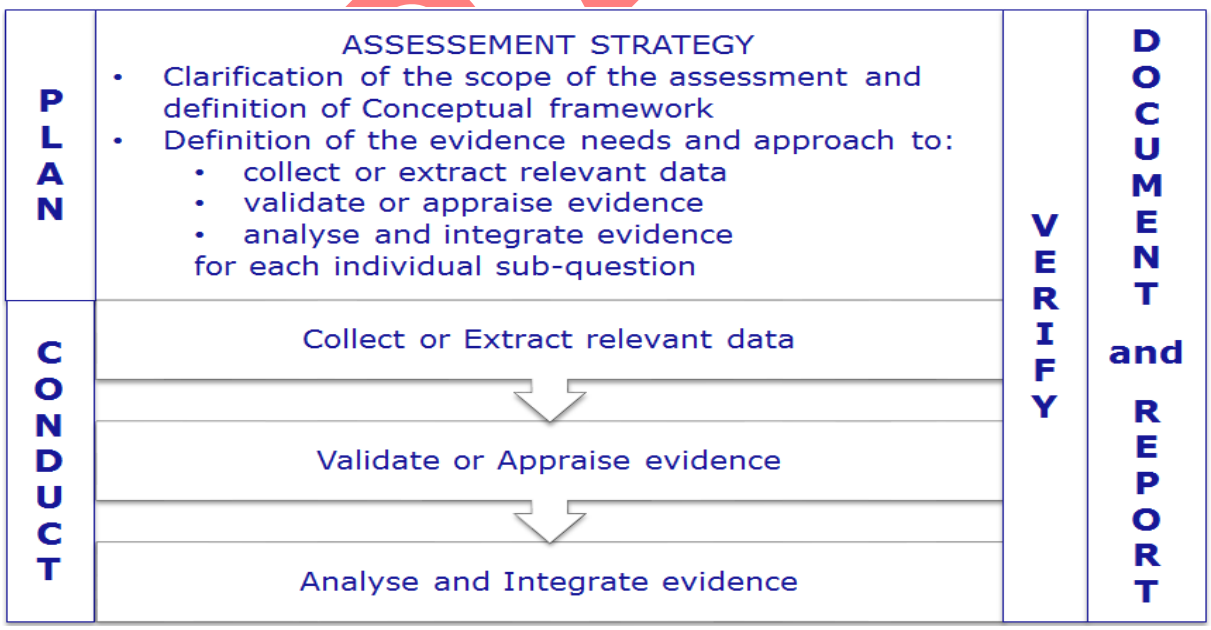


Figure 1: The process for dealing with data and evidence when conducting an assessment (EFSA, 2015b)

Transparent reporting of all assumptions and methods used, including expert judgement, is necessary to ensure that the assessment process leading to the conclusions is fully comprehensible.

306 'Open EFSA' aspires both to improve the overall quality of the available
307 information and data used for its scientific outputs and to comply with normative
308 and societal expectations of openness and transparency (EFSA 2009a, EFSA
309 2014). In line with this, EFSA is publishing three separate but closely related
310 guidance documents to guide its expert Panels for use in their scientific
311 assessments (EFSA, 2015). These documents address three key elements of the
312 scientific assessment: the analyses of Uncertainty, weight of evidence and
313 Biological Relevance.

314 The first document provides guidance on how to identify, characterise, document
315 and explain all types of uncertainty arising within an individual assessment for all
316 areas of EFSA's remit. The Guidance does not prescribe which specific methods
317 should be used from the toolbox but rather provides a harmonised and flexible
318 framework within which different described qualitative and quantitative methods
319 may be selected according to the needs of each assessment.

320 This current document on weight of evidence provides a general framework for
321 considering and documenting the approach used to evaluate and weigh the
322 assembled evidence when answering the main question of each scientific
323 assessment or questions that need to be answered in order to provide, in
324 conjunction, an overall answer. This includes assessing the relevance, reliability
325 and consistency of the evidence. The document further indicates the types of
326 qualitative and quantitative methods that can be used to weigh and integrate
327 evidence and points to where details of the listed individual methods can be
328 found. The weight of evidence approach carries elements of uncertainty analysis:
329 that part of uncertainty which is addressed by weight of evidence analysis does
330 not need to be reanalysed in the overall uncertainty analysis, but may be added
331 to.

332 The third document provides a general framework to addresses the question of
333 biological relevance at various stages of the assessment: the collection,
334 identification and appraisal of relevant data for the specific assessment question
335 to be answered. It identifies generic issues related to biological relevance in the
336 appraisal of pieces of evidence, in particular, and specific criteria to consider
337 when deciding on whether or not an observed effect is biologically relevant, i.e.
338 adverse (or shows a positive health effect).A decision tree is developed to aid
339 the collection, identification and appraisal of relevant data for the specific
340 assessment question to be answered. The reliability of the various pieces of
341 evidence used and how they should be integrated with other pieces of evidence
342 is considered by the weight of evidence guidance document.

343 EFSA will continue to strengthen links between the three distinct but related
344 topics to ensure the transparency and consistency of its various scientific
345 outputs while keeping them fit for purpose.

346 ## 1.5. Audience and degree of obligation

347 This Guidance is aimed at all those contributing to EFSA assessments and
348 provides a harmonised, but flexible framework that is applicable to all areas of
349 EFSA's work and all types of scientific assessment, including risk assessment. In
350 line with improving transparency, the Scientific Committee considers the

application of this guidance to be unconditional for EFSA. Each assessment must clearly and unambiguously document:

- what evidence was considered and how it was assembled into lines of evidence;
- how the evidence was weighed and integrated including consideration of reliability, relevance and consistency;
- the conclusion on the weight of evidence question in terms of the range and probability of possible answers. This can be expressed qualitatively or quantitatively, but should be quantified if possible when it directly addresses the Terms of Reference for the assessment.

The document provides guidance on the general principles of the weight of evidence approach but assessors have the flexibility to choose the appropriate methods for achieving the above, and the degree of refinement in applying them. The Scientific Committee considers that these should be fit for the purpose of the scientific assessment.

2. General framework and principles for weight of evidence assessment

This section provides a general framework and principles for weight of evidence assessment, including definitions of key concepts. Many scientific assessments involve weighing of evidence, though this may be implicit rather than explicit and is only sometimes described as 'weight of evidence assessment'. The aim of this guidance is to make weight of evidence assessment more explicit and transparent, and to provide a general framework of principles and approaches which is applicable to all areas of EFSA's work. Account is taken of approaches already used by EFSA, by other EU and international organisations, and in the scientific literature.

2.1 Weight of evidence and lines of evidence

WHO (2009) has defined weight of evidence assessment as 'a process in which all of the evidence considered relevant for a risk assessment is evaluated and weighted'. A recent review by ANSES (2016) defines weight of evidence assessment as 'the structured synthesis of lines of evidence, possibly of varying quality, to determine the extent of support for hypotheses'. Definitions and descriptions from a selection of other relevant publications are presented in Annex II, and reflect similar concepts to those of WHO and ANSES. The core of most definitions is that weight of evidence assessment is a process for integrating evidence to arrive at conclusions.

In practice, weighing[1] of evidence may occur when estimating quantities, as well as when assessing hypotheses, and both are relevant to EFSA's work. Therefore, this document uses a broader definition, as follows:

Weight of evidence assessment is a process in which evidence is integrated to determine the relative support for possible answers to a question.

It is often useful to organise evidence into groups or categories, which are often referred to as lines of evidence. ANSES (2016) defines a line of evidence as 'a set of relevant information of similar type grouped to assess a hypothesis'.

Various terms have been used to refer to distinct elements of information within a line of evidence, including 'studies' and 'pieces of evidence'. Piece of evidence is a more general term, as it could refer to a study (or to one of multiple outcomes of a study), or to other types of information including expert knowledge, experience, a model or even a single observation.

In some cases, a line of evidence may comprise only a single piece of evidence. Furthermore, what determines how far evidence should be sub-divided is how useful that is for the purpose of the scientific assessment. For example, in some assessments it might be sufficient to treat all human studies as a single line of evidence, whereas in other assessments it might be helpful to treat different types of human studies as separate lines of evidence. Therefore this document proposes the following definition, which remains compatible with the ANSES (2016) definition and the use of 'line of evidence' in the literature (see Annex II):

[1] The term 'weighing' is used in preference to 'weighting' in this document, as not all approaches to weighing evidence involve explicit assignment of 'weights'.

Lines of evidence are subsets of evidence which it is useful to distinguish when conducting a scientific assessment.

The definition for line of evidence is broadly-worded to accommodate different ways in which lines of evidence may contribute to answering a question. Different lines of evidence for the same question may be standalone, in the sense that each line of evidence offers an answer to the question without needing to be combined with other lines of evidence. It is important to distinguish these from complementary lines of evidence, which can only answer the question when they are combined. Multiple experiments measuring the same parameter are examples of standalone lines of evidence, whereas data on hazard and exposure are complementary lines of evidence for risk assessment because both are necessary and must be combined to assess risk. The distinction between complementary and standalone lines of evidence is important because it has practical implications in weight of evidence assessment (see Section 4). Note that a single question may be addressed by a combination of standalone and complementary lines of evidence.

2.2 When to use weight of evidence approaches

In general, the purpose of weight of evidence assessment is to answer a question, as implied in the preceding section. EFSA assessments address questions posed by their Terms of Reference. In some cases, a question in the Terms of Reference may be addressed directly, but in other cases it is beneficial to divide the primary question into two or more subsidiary questions (EFSA ,2015b).

Weighing of evidence is involved, either explicitly or implicitly, wherever more than one piece of evidence is used to answer a question. Weight of evidence assessment is not needed for questions where no integration of evidence is required.

Thus a single scientific assessment may comprise one or many questions, and none, some or all of those questions may require weight of evidence assessment.

Clarifying the questions posed by the Terms of Reference, and deciding whether to sub-divide them, is part of the first stage of scientific assessment, often referred to as problem formulation. This may show that the question is relatively simple and can be addressed directly, by a straightforward assessment. In many assessments, however, questions may need to be sub-divided to yield more directly answerable questions. In this manner, a hierarchy or tree of questions may be established. Assessment then starts at the bottom of the hierarchy. The evidence is divided into lines of evidence, as far as is helpful, assessed, weighed and integrated to answer each question at the bottom of the hierarchy. Integration continues upwards through the question hierarchy following similar principles, until full integration is reached to answer the main question defined by the problem formulation.

In some cases, the Terms of Reference for an assessment pose open questions, for example to review the state of science on a particular topic. These assessments also require weight of evidence assessment approaches, because their conclusions generally derive from weighing and integrating evidence.

2.3 Weight of evidence conclusions

As implied in the definition above, the purpose of weighing evidence is to assess the relative support for possible answers to a question. In some cases, it may be concluded that the evidence supports only one answer, with complete certainty. More usually, multiple answers remain possible, with differing levels of support. In such cases the conclusion should state the range of answers that remain possible, and not be reduced to a single answer unless a threshold level of support for conclusions has been agreed with decision-makers, because this involves risk management considerations.

When weight of evidence assessment directly addresses the conclusion of a scientific assessment, its output will be part of the response to the Terms of Reference for the assessment. In general, decision-makers need to know the range of possible answers to their questions, and how probable they are, because this may have important implications for decision-making (EFSA, 2016). Furthermore, it is important to express this quantitatively when possible, to avoid the ambiguity of qualitative expression (EFSA 2016, EFSA, 2012b).

When weight of evidence addresses an intermediate question in a larger assessment, the possible answers and their relative support needs to be taken into account in subsequent steps of the assessment. In these cases, relative support may be expressed either qualitatively or quantitatively, depending on what is convenient for use in the subsequent steps. Qualitative and quantitative approaches are discussed further in Section 3.

2.4 Steps in weight of evidence assessment

This document considers the weight of evidence assessment as comprising three basic steps:

1. Assembling the evidence,
2. Weighing the evidence,
3. Integrating the evidence.

This corresponds to the three basic steps distinguished by Suter and Cormier (2011, their Figure 3). The first step involves searching for and selecting evidence that is relevant for answering the question in hand, and deciding whether and how to group it into lines of evidence. The second step involves detailed evaluation and weighing of the assembled evidence. In the third step, the evidence is integrated to arrive at conclusions, which involves weighing the relative support for possible answers to the question.

Practical guidance for the three basic steps is provided in section 4. Relevant considerations to be taken into account in the weighing and integrating steps are discussed in section 2.5, while qualitative and quantitative methods for assessing those considerations are discussed in section 3.

The three steps of weight of evidence assessment, described above, may occur at one or more points in the course of a scientific assessment, wherever integration of evidence is required, as illustrated in Figure 2. The question to be addressed by each weight of evidence assessment is defined by problem formulation, which is a preceding step in the scientific assessment as a whole.

The output of weight of evidence assessment feeds either directly or indirectly into the overall conclusion of the scientific assessment. Although weight of evidence itself addresses some of the uncertainty affecting the scientific assessment (see below), a separate step of uncertainty analysis is still needed to take account of any other uncertainties affecting the overall assessment. Some assessments will also include a step of sensitivity analysis or influence analysis, to identify which evidence and uncertainties have most influence on the conclusion.

Any part of the overall assessment may be refined iteratively, when necessary, by returning from later steps to earlier steps, depending on which steps it is most useful to refine.

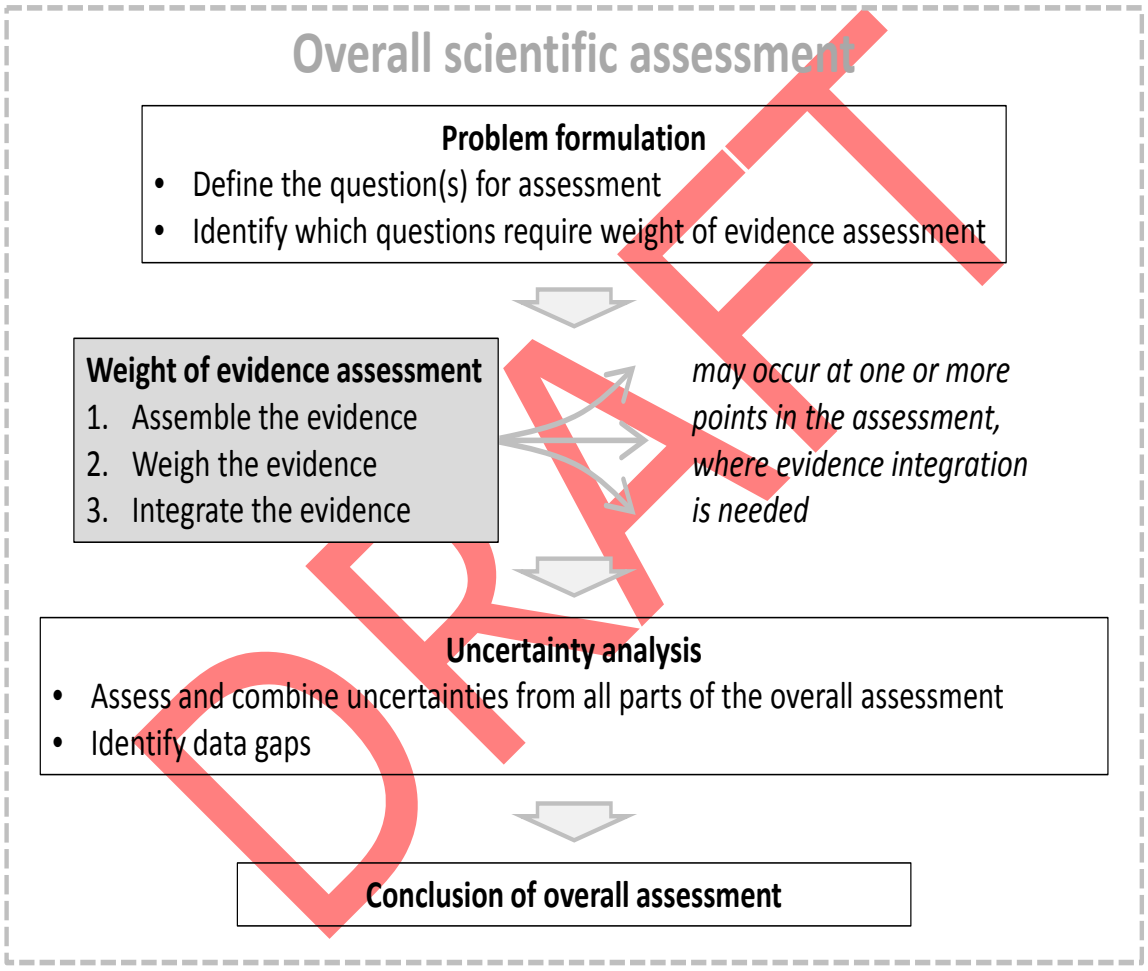


Figure 2: Diagrammatic illustration of weight of evidence assessment as a 3-step process which may occur at one or more points in the course of a scientific assessment.

2.5 Key considerations for weighing evidence

Reliability, relevance and consistency are mentioned in many publications on weight of evidence assessment (see Annex II). These can be seen as three basic considerations in weight of evidence: how applicable the evidence is to the question of interest, the quality of the evidence and how consistent it is with other evidence for the same question. How these three concepts relate to one another, to the three basic steps of weight of evidence assessment and to the

weight of evidence conclusion is illustrated in Figure 3. Note that relevance and reliability may be considered in both the first and second steps. First, relevance is considered when identifying evidence, and both relevance and reliability may be considered when selecting which of the identified evidence to include in the assessment (sometimes referred to as screening or filtering). However, the selected evidence will vary in both relevance and reliability and this will be considered in the second step, when weighing the evidence.

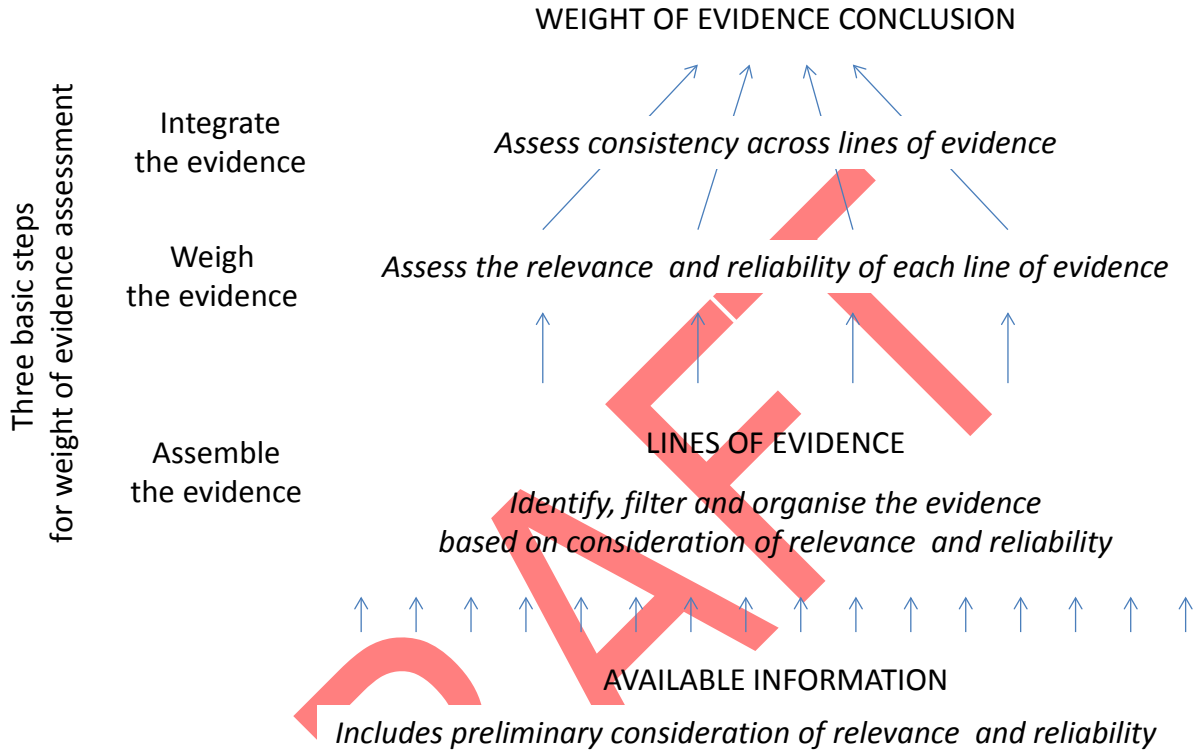


Figure 3: Relationship of relevance (including biological relevance), reliability and consistency to the three basic steps of weight of evidence assessment and to the conclusion for a weight of evidence question.

The present document defines reliability, relevance and consistency, in terms of their contributions to the weight of evidence assessment, as illustrated in Figure 3:

Reliability is the extent to which the information comprising a line of evidence is correct.

Relevance is the contribution a line of evidence would make to answer a specified question, if the information comprising the line of evidence was correct. This includes biological relevance (EFSA 2017) as well as relevance based on other considerations, e.g. temporal, spatial, chemical etc.

Consistency is the extent to which the contributions of different lines of evidence to answering the specified question are compatible.

These definitions are compatible with those found in other publications relating to weight of evidence assessment (e.g. ECHA 2010a, SCENIHR 2012, Vermeire et al. 2013). Different types of 'contribution' are discussed further below.

Other publications mention additional considerations relevant to weight of evidence assessment including, for example, quality, applicability, coherence, risk of bias, specificity, biological concordance, biological gradient and many others. Some of these are synonyms for reliability, relevance or consistency, some refer to combinations of reliability and relevance, and others refer to specific types of reliability, relevance or consistency that are important for particular areas of assessment (see Annex II). The reasons for giving particular emphasis to reliability, relevance and consistency are that they are generic considerations, applicable to every type of assessment, with defined relationships to the three basic steps of weight of evidence assessment, and a defined relationship to the weight of evidence conclusion, as illustrated in Figure 2. This makes them useful to assessors as a conceptual framework for identifying specific considerations relevant to particular assessments, and for assessing how they combine to determine the weight of evidence conclusion. How this can be applied in practice is discussed further in section 4.

2.6 Relation of weight of evidence and uncertainty

Weight of evidence and uncertainty are closely related. For example, SCENIHR (2012) state that 'strength of evidence is inversely related to the degree of uncertainty', while Suter and Cormier (2009) state that 'the weight of the body of evidence, based on the combined weights of individual pieces of evidence, may be used to express confidence or uncertainty in the results.'

Weight of evidence assessment is defined above as a process which determines the relative support for possible answers to a question. EFSA (EFSA, 2016) defines uncertainty as 'a general term referring to all types of limitations in available knowledge that affect the range and probability of possible answers to an assessment question'. Answers may refer to alternative hypotheses or estimates, and probability is one way of expressing relative support for possible answers. Thus the weight of evidence conclusion for a question and the uncertainty of the answer can be expressed in identical form: the range of possible answers and their relative degree of support or probability. This is explicit in meta-analysis, an evidence integration method producing conclusions in the form of estimates with confidence intervals, which express uncertainty.

However, expression of uncertainty for the conclusion of the scientific assessment as a whole should additionally include any uncertainties associated with the weight of evidence process itself. This may include uncertainties regarding, for example, the selection of evidence, assessment of reliability, relevance and consistency, and choice of weight of evidence methods. These should be taken into account by uncertainty analysis following the weight of evidence assessment (as indicated in Figure 1 above), and may modify the range and probability of answers to some degree.

Consistent with this, each of the three basic considerations in weight of evidence may be expressed in terms of uncertainty:

the reliability of a line of evidence can be expressed as the uncertainty of that evidence itself (i.e. how different the evidence might be if the information comprising it was correct);

the relevance of a line of evidence can be expressed as the uncertainty that would be associated with extrapolating from correct information, of the type provided by that evidence, to its contribution to answering the weight of evidence question;

limitations in consistency between lines of evidence contribute directly to uncertainty about the answer to the assessment question, and may imply uncertainty in the assessment of reliability and relevance.

Thus weight of evidence assessment carries elements of uncertainty analysis: that part of uncertainty which is addressed in the weight of evidence assessment does not need to be re-analysed in the uncertainty analysis, but may be added to.

Probability is not the only way to express relative support or uncertainty. It can also be expressed qualitatively, and this is essential for any uncertainties or aspects of weight of evidence that cannot be quantified (see Section 5.10 of EFSA, 2016). As explained above and by EFSA (EFSA, 2016), expressing the probability of possible answers is important for the conclusions of a scientific assessment, but need not apply to earlier steps in the weight of evidence process.

2.7 Relation of weight of evidence and variability

It is important also to consider how weight of evidence relates to variability. Variability is defined by EFSA (EFSA, 2016) as 'heterogeneity of values over time, space or different members of a population, including stochastic variability and controllable variability'. Note that 'values' could refer to values on a quantitative scale, or to alternative qualitative descriptors, and that 'population' is not restricted to biological populations but may also refer to other entities (e.g. variability in temperature at different points in time and space).

Variability is often important in a scientific assessment, e.g. variability in chemical occurrence in food and consumption are important in chemical exposure assessment. This needs to be dealt with when defining the questions for assessment, such that they refer to specific descriptors or summaries of the variable quantity, such as the average or 95th percentile exposure. If weight of evidence assessment was used as part of the exposure assessment, for example to integrate occurrence data from different countries, the reliability, relevance and consistency of the different lines of evidence regarding the variability of occurrence would be assessed. Thus variability can be the *subject* of a weight of evidence assessment, rather than a contributor to it.

Variability in data is a combination of variability of the quantity being measured (e.g. between individuals) and variability of the measurement process. Variability in data due to measurement error should be taken into account when assessing reliability.Variability between results reported by different studies might reflect differences in the reliability of those studies or differences in their relevance for the assessment. Such differences may lead to apparent inconsistencies in data that need to be considered when integrating evidence (Section 4.6).

3. Overview of qualitative and quantitative methods for weight of evidence assessment

3.1 Examples of weight of evidence approaches

3.1.1. Classification of weight of evidence approaches

Several reviews of weight of evidence approaches were published in the literature, especially by Chapman et al. (2002), Weed (2005), Linkov et al. (2009) and Lorenz et al. (2013). A systematic review was also recently published by ANSES (ANSES, 2016) on weight of evidence for hazard identification.

Instead of conducting another review of weight of evidence approaches, we only present here a few examples of approaches belonging to different categories. The implementation of these approaches will have to be adapted to the needs of each assessment. Our choice of examples is based on the classification proposed by Linkov et al. (2009). This classification covers weight of evidence approaches showing contrasted levels of complexity, and distinguishes the following categories:

- Listing evidence: Presentation of individual lines of evidence without attempt at integration
- Best professional judgment: Qualitative integration of multiple lines of evidence
- Causal criteria: A criteria-based methodology for determining cause and effect relationships
- Logic: Standardized evaluation of individual lines of evidence based on qualitative logic models
- Scoring: Quantitative integration of multiple lines of evidence using simple weighting or ranking
- Indexing: Integration of lines of evidence into a single measure based on empirical models
- Quantification: Integrated assessment using formal decision analysis and statistical methods.

In this classification system, approaches are classified by the degree to which they are quantitative (Linkov et al., 2009). The least quantitative approaches are categorized as “Listing evidence”, while the most quantitative ones fall within the category “Quantification” and are based on statistical approaches or on formal decision-analytical tools. Other categories correspond to intermediate situations.

In this guidance, we do not formally distinguish the categories “Logic”, “Scoring”, and “Indexing”. As these categories share many common characteristics, they are further considered as a single category named “Rating”. Thus we only consider five categories. Examples of approaches are listed below for each category. Some of them cover all three basic steps of the general weight of evidence process, while others are more specific and focus on one or two steps. Approaches restricted to problem formulation were considered as outside the scope of this guidance, and are not included.

We do not aim to cover here all existing approaches, but rather to give a brief overview of different types of approaches and to provide key references. Examples are briefly presented in the following sections. Categories described in sections 3.1.2 to 3.1.4 (i.e. listing evidence, best professional judgement and causal criteria) are collectively referred to in this guidance as qualitative approaches. Several criteria are then presented in section 3.2 to help risk assessors to choose among weight of evidence approaches. These approaches have been used in different EFSA scientific assessments (see for example Annex III). Note that assessments may use a combination of more than one of these categories.

3.1.2. Category "Listing evidence"

The approaches of this category do not attempt to integrate evidence. Pieces of evidence are simply listed in text or in a table, but no formal methodology for weighing the evidence is used. The origin of the evidence depends on the approach used for assembling evidence, for example:

- Several methods of evidence synthesis which do not involve quantitative integration could be classified under this category. These include extensive literature searches (EFSA, 2010; Higgins and Green 2011), systematic maps (Collaboration for Environmental Evidence 2016; James et al., 2016), and non-quantitative systematic reviews (i.e. those lacking a quantitative data synthesis step) (EFSA, 2010; Higgins and Green 2011).
- Evidence can be directly provided by applicants in reports and/or datasets rather than being selected by risk assessors.

3.1.3. Category "Best professional judgment"

This approach attempts to integrate evidence and lines of evidence based on expert judgments. No formal method is used for evidence integration. Instead, evidence and lines of evidence are used to form a conclusion by professional opinion via a discussion of the findings. Several examples are given in the review of Linkov et al. (2009), notably Staples et al. (2004).

3.1.4. Category "Causal criteria"

Approaches of this category provide a structure based on explicit criteria to evaluate relationships between cause and effect from one or several lines of evidence. Among these criteria, the Bradford Hill criteria (Hill 1965) are probably the most widely used, especially in epidemiology. They are often seen as the minimal conditions needed to establish a causal relationship between two items, and are frequently used as a checklist in risk assessments. Several variants were proposed in the literature. For example, Becker et al. (2013) proposed a template for weight of evidence of mode of action based on modified Bradford Hill considerations.

3.1.5. Category "Rating"

This category includes a variety of structure frameworks that can be considered under the categories "Logic", "Scoring", and "Indexing" defined by Linkov et al. (2009). Examples include GRADE[2] (Guyatt et al., 2011), WHO-IARC (IARC,

[2] GRADE stands for Grading of Recommendations, Assessment, Development and Evaluation

2006), WCRF/AICR (WCRF/AICR 2007), OSHA (2016), and guidance used to produce NTP Monographs (OHAT 2015, ORoC 2015). Guidance to assess and integrate evidence is based on several factors, often derived from the Bradford-Hill criteria. However, "Rating" builds on the approaches of the category "Causal criteria" in that more guidance is provided for appraising and integrating evidence.

These approaches usually relate to the second step of the weight of evidence process, including the appraisal of individual studies and rating confidence in the individual lines of evidence (e.g., "high confidence," "sufficient evidence"). Some of them also provide tools based on a matrix for integrating lines of evidence to reach hazard identification conclusions (WHO-IARC, OHAT, OSHA (2016)). None of these approaches use formal probabilistic techniques, but it is possible to combine application of the structured framework guidance with a more quantitative presentation of conclusions.

Some of these approaches (e.g., GRADE) are designed to be flexible for use in a variety of disciplines and able to be applied under different time and resource constraints in situations corresponding to different levels of urgency (Thayer and Schünemann, 2016).

In EFSA, the scheme presented in the "Guidance on a harmonised framework for pest risk assessment and the identification and evaluation of pest risk management options" (EFSA, 2010a) belongs to this category.

3.1.6. Category "Quantification"

This category covers a large diversity of approaches that can be used to integrate evidence into lines of evidence and/or to integrate different lines of evidence in order to reach a general conclusion.

This category includes standard statistical models used in meta-analysis, such as fixed-effect and random-effect models. These models are usually applied to estimate mean effect sizes, which can be interpreted as summary estimates of a quantity based on statistical integration of evidence from multiple primary studies. The results of meta-regression analysis may be interpreted in terms of variability or reliability depending on whether the regressor variable, for example, expresses a population characteristic or study quality issue, respectively. They are able to describe uncertainties through confidence intervals and probability distributions. Other types of statistical methods (e.g. Bayesian methods) are also useful for synthesizing multiple sources of evidence.

In addition to statistical methods, other approaches have been proposed, especially machine learning, in silico tools, and multi-criteria analysis. Linkov et al. (2015) consider that multi-criteria decision analysis can be used as a proxy for the Bayesian approach to weight of evidence when model formulation is restricted by data limitations.

When a quantitative model is used for weight of evidence, several authors recommend performing a sensitivity analysis to study the stability of the main conclusions to the model assumptions, e.g., to the model equations, or to the parameter values (Linkov et al., 2011; Borenstein et al., 2009).

Examples of quantitative approaches and several key references are listed below:

- Statistical methods for integrating data provided by several studies sharing similar characteristics (classic fixed-effect and random-effect models used in meta-analysis, and Bayesian hierarchical models). Many textbooks and methodological papers are available on these methods, for examples Borenstein et al. (2009) on classic techniques, and Sutton and Abrams (2001) and Higgins et al. (2009) on Bayesian methods in the context of meta-analysis. Increased use of these methods in EFSA assessments could be beneficial, and would be facilitated by provision of guidance and training.
- Statistical methods for integrating different types of studies in order to allow decisions based on all available evidence and to analyse uncertainty (Small, 2008; Turner et al. 2009; Gosling et al. 2013)
- Quantitative expert judgement including multi-criteria decision analysis for integrating different types of studies (Linkov et al., 2011; Linkov et al., 2015)
- Machine learning techniques (Li and Ngom, 2015)
- In silico tools including QSAR, PBTK-TD (ECHA, 2016).

3.2. Choosing weight of evidence methods

A challenge when planning a weight of evidence assessment is to determine which assessment method(s) to select, given the variety of different methods available. A single easy-to-use weight of evidence method that covers all the basic steps of the weight of evidence process and enables transparent quantification of uncertainty may not be available. A pragmatic approach is therefore recommended for identifying the most suitable method, or combination of methods, for the weight of evidence assessment. A list of criteria for comparing weight of evidence methods is suggested in Table 1, to assist in evaluating the relative strengths and weaknesses of the different methods. This list is not exhaustive but based on discussions during the development of the current guidance and has not been formally tested. These criteria may also be helpful for transparently recording and reporting the decision-making process used for weight of evidence method selection, in keeping with EFSA's requirement for transparency in the conduct and reporting of scientific assessments (EFSA, 2006)

As with any type of evidence synthesis, weight of evidence methods face a potential trade-off between what would be ideal in terms of resource requirement (i.e. rapid, cheap, methods) and scientific rigour (i.e. methods that transparently display uncertainty at all steps of the weight of evidence process). Careful consideration will be needed early on in the planning process (in problem formulation) to ensure that adequate resource (time, staff expertise) is available to achieve the desired level of scientific rigour. In EFSA weight of evidence assessments which have pre-specified and fixed resources, the criteria in Table 1 could be used to judge the optimal scientific rigour that could be achieved within

the available resources. Alternatively, if resource availability for a weight of evidence assessment is negotiable, or if a weight of evidence assessment is at a preliminary scoping phase, these criteria may be helpful for estimating the resource needs for the assessment.

Table 1: Criteria for assessing the relative strengths and weaknesses of weight of evidence methods

Criterion	Key considerations
Availability of guidance	Guidance on the weight of evidence method should be readily available in the public domain and, ideally, should be endorsed, e.g. through peer-review and/or wide acceptance. Guidance should document the rationale of the method, the full process, and how to interpret the results. Ideally, access to help and support facilities should be readily available (e.g. any relevant tools such as tutorials, software programs or modules). Availability of guidance is important both for the conduct and the critical appraisal of weight of evidence assessments. Weight of evidence methods that lack adequate guidance would rate poorly on this criterion.
Expertise needed	Weight of evidence methods are likely to vary in the level of technical skill required to conduct them. Some quantitative methods, for example, may require specific skills in statistics and/or programming. The expertise requirement should be considered in relation to availability of expertise and tools and, if necessary, whether available resources would support the outsourcing of expertise or provision of training. The level of expertise required has implications both for the conduct and the critical appraisal of weight of evidence assessments.
Ease of understanding for the non-specialist	Weight of evidence methods are likely to vary in how easy they are to understand by non-specialists, and this may be related to the expertise needed, as well as the availability of adequate guidance. In the interests of facilitating clear communication of the weight of evidence methods and results to a wide range of stakeholders (as recommended by EFSA), principles of the methods that can be readily understood by non-specialists are preferable.
Time needed	Weight of evidence methods that can be conducted quickly would be preferable where urgent weight of evidence assessments are required. However, rapid methods risk sacrificing scientific rigour. When estimating the time required for a particular weight of evidence method, consideration should be given to the availability of team expertise, since this could influence the time required for a weight of evidence assessment.
Transparency and reproducibility	Transparency and reproducibility are fundamental principles required by EFSA in its scientific assessments. Transparency should apply to all parts of the weight of evidence method, meaning that it should be possible to follow clearly how the input data for the assessment are analysed to produce the conclusions. Reproducibility is defined such that consistent results should be expected if the same method were to be repeated using the same input data (but note that results are unlikely to be identical, dependent on the degree to which expert opinion is involved).
Variability and uncertainty	Weight of evidence methods should, ideally, explicitly report and analyse both variability and uncertainty at all steps of the assessment, and propagate them appropriately through the assessment. Quantitative expression of variability and uncertainty is preferable to qualitative expression. Careful consideration may be needed to ensure that the weight of evidence method can include all relevant sources of variability and uncertainty.

The criteria in Table 1 should be considered together. This is because the strengths and weaknesses of weight of evidence methods are multi-dimensional, and individual criteria alone may not be able to capture important trade-offs, e.g. between resource availability and scientific rigour. Note that the criteria do

not necessarily have equal importance: their relative importance may be discussed on a case-by-case basis when planning each weight of evidence assessment. The criteria in Table 1 are not exhaustive. Other criteria which may be useful include the strength and scope of the theoretical basis for a method and the extent to which the output of the method is in a form which can be tested.

4. Practical guidance for conducting weight of evidence assessment

This section contains practical guidance for applying weight of evidence approaches within EFSA scientific assessments. The guidance is advisory, not mandatory. Assessors should choose the specific approaches that are best suited to the needs and context of their assessments.

Three types of assessment are distinguished, which require different approaches:

- assessments following standardised procedures, which include pre-specified approaches to weight of evidence assessment;
- case-specific assessments, where there is no pre-specified procedure and assessors need to choose and apply approaches on a case-by-case basis;
- emergency procedures, where the choice of approach is constrained by limitations on time and resources.

Standardised assessment procedures have been established in many areas of EFSA's work, especially for regulated products. Standardised procedures are generally defined in documents, e.g. EU regulations or EFSA guidance documents. They specify what questions should be addressed, what evidence is required, and what methods of assessment should be applied to it. They generally include standardised elements that are assumed to provide adequate cover for uncertainty (EFSA, 2016).

Where a standard assessment involves questions that require integration of evidence, the methods for doing this will be specified in the standard procedure. For example, in human health risk assessment of chemicals in food, the outcome may often be based on one of the available studies, which is considered to provide the highest level of protection for the consumer. While not generally thought of as weight of evidence, this is a procedure for integration which, after considering all the evidence, in effect gives all the weight to a single study (sometimes referred to as the critical study).

In assessments following a standardised procedure, the default approach should be to integrate evidence using the methods as specified by the procedure. If the methods are specified in detail they may be sufficient to conduct the assessment without further guidance; where the methods are not fully specified, the assessor may benefit from the guidance in Sections 4.1-4.6. If an assessment that would normally be addressed by a standard procedure includes weight of evidence issues that are not adequately addressed by the standard procedure, a case-specific approach will be needed for that part of the assessment, following the guidance in Sections 4.1-4.6.

In **case-specific assessments**, for which there is no standard procedure, evidence integration will need to be conducted case-by-case, following the guidance in Sections 4.1-4.6.

Emergency assessments are required in situations where there are exceptional limitations on time and resources. If an emergency assessment involves questions that require integration of evidence, assessors should first consider whether any standard procedure exists that can be applied within the

time and resources available. If not, then the assessor should conduct a case-specific assessment, choosing options from Sections 4.1-4.6 that are compatible with the time and resources available.

4.1 Define the questions for weight of evidence assessment

A single scientific assessment may comprise one or more questions, and none, some or all of those questions may require weight of evidence assessment. Interpreting the questions posed by the Terms of Reference, and deciding whether to sub-divide them, is part of the first stage of scientific assessment, often referred to as problem formulation. General guidance on problem formulation for EFSA's scientific assessments is provided in other documents (EFSA 2015, EFSA 2015b, EFSA 2006), and the need to ensure questions are well-defined is further discussed by EFSA (2016). Weight of evidence assessment does not involve any additional requirements or considerations for specifying the questions for assessment, so the reader is referred to the documents referred to above for details.

Problem formulation also includes planning the strategy and methods for assessment (EFSA, 2015). As part of this, the assessors should identify which of the questions in the assessment will require integration of evidence and therefore the use of weight of evidence approaches.

The output of problem formulation should therefore include a list of the questions that require weight of evidence assessment. Each question that requires weight of evidence assessment should then be addressed by applying the basic steps described in the three following sub-sections. When the assessment involves a hierarchy of questions, start with the questions at the lowest level of the hierarchy, as the conclusions of these will inform lines of evidence for higher questions. It is sometimes necessary to return to and revise the problem formulation later, if additional questions are identified in the course of the assessment.

4.2 Assemble the evidence

This is the first of the three basic steps of weight of evidence assessment for an individual question (see Section 2.4). Guidance for this and the following two steps is provided as a series of numbered (sub-)steps, which may be considered in sequence. As noted above, this guidance is advisory, not mandatory. For every step, assessors should choose approaches that are appropriate to the needs and context of the assessment in hand, including any limitations on time and resources.

1) **Identify potentially relevant evidence**. In some EFSA assessments, the evidence to be used is defined by regulations or guidance, and/or submitted by applicants. This applies especially in standard assessment procedures. In other assessments, the assessors define the strategy and criteria for identifying and accessing potentially relevant evidence. Procedures for this, with varying degrees of formality, are described in EFSA guidance on extensive literature searching and systematic review (e.g. EFSA 2015b, 2010), which is designed to increase coverage and reduce potential biases in evidence gathering.

2) **Select evidence to include in the weight of evidence assessment**. In principle, all evidence identified as potentially relevant in step 1 should be taken into account, but limitations on time and resources may require the assessment to focus primarily on the most relevant and/or most reliable evidence. This subset of evidence may identified by filtering or screening using appropriate criteria for relevance and/or reliability (EFSA 2015b, 2010a). Evidence that was considered potentially relevant but not included should be retained separately, for example as a list of references or folder of documents, so that the impact of excluding it can be considered as part of uncertainty analysis (below).

3) **Group the evidence into lines of evidence**, i.e. subsets of evidence which the assessors find useful to distinguish when conducting the assessment. There are no fixed rules for how to form lines of evidence, but it may be helpful to distinguish those which are standalone and those that are complementary (Section 2.1). If the lines of evidence are complementary, they may be grouped according to the contribution they make to answering the question (e.g. exposure, hazard, etc.). Standalone lines of evidence may comprise evidence on the same aspect of the assessment but generated by different methods (e.g. different study types), with different subjects (e.g. species, chemicals, etc.) and in different conditions. This will tend to group evidence that has similar relevance and/or reliability. The lines of evidence and the rationale for constructing them should be documented, identifying which are standalone and which are complementary.

4.3 Weigh the evidence

Five broad categories of methods for weight of evidence assessment are presented in Section 3, together with suggestions for choosing between them: listing of evidence; best professional judgement; causal criteria; rating; and quantitative methods. Assessors should first consider the possibility of using quantitative methods, because an appropriate and well-conducted quantitative analysis will generally be more rigorous than other methods. For example, when it is possible and appropriate to combine multiple studies by meta-analysis, this will be more rigorous than integrating them by expert judgement. However, there are two important caveats to this. First, quantitative methods may not be appropriate for various reasons, e.g. not applicable to the nature or heterogeneity of the evidence to be integrated, not practical within the time and resources available, etc.. Second, quantitative methods may not address all the considerations that are relevant for weighing the evidence. For example, common approaches to meta-analysis only capture those aspects of reliability and consistency that are represented in the variability of the data, although some forms of meta-analysis can also take account of relevance and additional aspects of reliability (e.g. Turner et al. 2009).

Therefore, the approach proposed below is to check first whether quantitative methods are practical and appropriate, and then complement them with qualitative methods (categories to ensure all relevant considerations are addressed. When quantitative methods are not practical or appropriate, only qualitative methods can be used. However, assessors should start by deciding what considerations are relevant for weighing the evidence, as this may have implications for the choice of methods.

1) **Decide what considerations are relevant for weighing the evidence**. The general considerations for weighing evidence are reliability, relevance and consistency, as explained and defined in Section 2.5. Assessors may choose to work with these three basic considerations, or use more specific criteria appropriate to their area of work, especially if these have already been established in guidance or the scientific literature. If using pre-established criteria, assessors should check that they cover all aspects of reliability, relevance and consistency that are relevant for the assessment in hand, and define any additional criteria that are needed.

    a. **Reliability** is the extent to which the information comprising a line of evidence is correct. It may be assessed by considering the uncertainty of the evidence, i.e. how different it might be if the information comprising it was correct. Anything that contributes to that uncertainty should be included when assessing reliability.

    b. **Relevance** is the contribution a line of evidence would make to answering a specified weight of evidence question, if the information comprising the line of evidence were correct. Anything that contributes to the need for extrapolation, and its uncertainty, should be included when assessing relevance.

        i. For a standalone line of evidence, consideration of relevance involves thinking about how well that evidence would answer the question, if the information comprising it were correct. How much extrapolation is involved, between the subjects and conditions the evidence relates to and those relevant for the question, and how uncertain is that?

        ii. For a complementary line of evidence, consideration of relevance involves identifying what that evidence contributes to the conceptual model or argument for answering the question, and considering what extrapolation is required to provide that contribution.

    c. **Consistency** should be considered when integrating evidence (below).

2) **Decide on the method(s) to be used for weighing and integrating the evidence**. Refer to the categories and criteria in Section 3. Some of the methods for weighing evidence also perform the integration step (e.g. meta-analysis), or limit the choice of methods for integration, so both steps should be considered when choosing between methods. The choice of methods may also be affected by whether the lines of evidence are standalone or complementary. For example, meta-analysis can be used to integrate standalone lines of evidence, whereas complementary lines of evidence require a quantitative model of the relationships between the lines of evidence and the answer to the question (e.g. the relationships between exposure, hazard and risk).

    a. **Consider whether quantitative methods are practical and appropriate for the needs and context of the assessment**. If it is decided to use a quantitative method, identify which aspects of reliability, relevance and consistency it will address, and which it will not.

b. **Choose one or more qualitative methods to address those aspects of reliability, relevance and consistency that are not treated quantitatively**. This could include methods from one or more of the non-quantitative categories presented in Section 3.

c. **Check that the chosen methods (quantitative and/or qualitative) address all pertinent aspects of reliability, relevance and consistency (identified in step 1 above)**.

d. **If more than one method is chosen for weighing evidence, consider whether their results can be combined directly when integrating the evidence**. Some methods are capable of incorporating the outputs of other methods: e.g. Doi (2014) has developed methods for incorporating quality scores into meta-analysis, while Turner et al. (2009) have proposed methods for incorporating quantitative expert judgements about the effects of study limitations (which may include reliability and relevance) into meta-analysis. Such methods may be used, if they are appropriate and practical for the needs and context of the assessment.

3) **Apply the chosen methods for weighing the evidence and summarise the results in a form that is helpful for integration**. If more than one method has been used (e.g. a quantitative method combined with a qualitative method), it is recommended to find a way of presenting the results together in a concise tabular or graphical summary. For example, estimates and confidence intervals from quantitative methods can be plotted on a graph alongside symbols or text showing the results of qualitative methods (e.g. EFSA 2015a, 2016a). This provides a useful overview of the evidence, which may be helpful for the assessors in the integration step and also for others, who read the finished assessment.

4.4 Integrate the evidence

In this step, the evidence is integrated to arrive at the conclusion, taking account of the reliability and relevance of the evidence, assessed in the preceding step, and also the consistency of the evidence.

1) **Consider the conceptual model for integrating the evidence.** Integration always involves a conceptual model, even if this is not made explicit. Integrating standalone lines of evidence requires a conceptual model of how evidence of differing weight is combined. Integrating complementary lines of evidence additionally requires a conceptual model of the contributions made by the different lines of evidence and how they combine to answer the question. Assessors may make find it helpful to make the conceptual model explicit, e.g. as a flow chart or list of logical steps. This should help assessors to take appropriate account of the relationships between lines of evidence and with the question being assessed, both when the integration of evidence is done by expert judgement and when it is done using a quantitative model. Making the conceptual model explicit also contributes importantly to the transparency of the assessment, unless it is simple and self-evident.

2) **Assess the consistency of the evidence**. Consistency is the extent to which the contributions of different lines of evidence are compatible (Section

2.5). Limitations in consistency arise in part from limitations in the relevance and reliability of different lines of evidence. If a question is well-defined, only a single correct answer should be possible, and any apparent inconsistencies in the evidence should be explicable in terms of differences in reliability and/or relevance. Assessors should not, however, simply conclude that inconsistent evidence is unreliable or irrelevant. Rather, assessors should consider whether, after taking differences in reliability and relevance into account, the lines of evidence still appear inconsistent. If so, this may imply the presence of additional limitations in relevance and reliability, beyond those already taken into account, or limitations in the conceptual model for integrating the lines of evidence. Alternatively, it may imply there is more than one possible answer to the question, in which case it may need to be more precisely defined or split into two or more separate questions. Any remaining inconsistency should be considered as part of the uncertainty affecting the weight of evidence conclusion.

3) **Apply the method(s) chosen for integrating the evidence** (the methods chosen in step 2 of Section 4.5, above). As already mentioned, one or more methods may be used, from one or more of the categories of methods described in Section 3.

4) **Develop the conclusion for the weight of evidence assessment**. If there is no single method that can take into account all the pertinent aspects of reliability, relevance and consistency, it will be necessary to integrate the results of the different methods by expert judgement. This may be done within the process for reaching a conclusion, as follows:

   a. **Summarise all the results up to this point in a concise tabular or graphical format**. This should comprise all the results of weighing the evidence, as in step 3) of Section 4.5, together with the results of any integration that has been done (e.g. integrated estimates and confidence intervals from meta-analysis, or integrated scores from scoring methods).

   b. **Define how the conclusion of the weight of evidence assessment (range of possible answers and their relative support) will be expressed**. The range of answers may be expressed on an appropriate quantitative scale, or as alternative qualitative statements or propositions. Relative support for the possible answers may be expressed quantitatively, e.g. as probabilities, or qualitatively.

      i. When weight of evidence assessment directly addresses the conclusion of a scientific assessment, the range of possible answers and how probable they are should be expressed quantitatively if possible. Any considerations that cannot be included in the quantitative expression imply that the conclusion will be subject to unquantified uncertainties, which should be described qualitatively (see EFSA, 2016). If no quantitative expression is possible, this implies that each probability could be anywhere between 0 and 100%, and the assessor should consider whether the evidence supports any conclusion at all (see also Sections 5.10 and 5.11 of EFSA 2016).

ii. When weight of evidence addresses an intermediate question in a larger assessment, the range of answers and their relative support may expressed either qualitatively or quantitatively, depending what is convenient for use in subsequent steps of the assessment. One method could be taken as the primary method of integration, for use in subsequent steps of the assessment, and additional considerations (not covered by the primary method) could be carried over to the uncertainty analysis at the end of the scientific assessment as a whole. For example, if meta-analysis was used to integrate occurrence data in a chemical risk assessment, the output of the meta-analysis could be used in subsequent steps of exposure and risk assessment. Any other considerations regarding the quality and relevance of the occurrence data and the assumptions of the meta-analysis would be addressed as part of combined uncertainty at the end of the risk assessment (EFSA 2016, Section 12).

c. **Where expert judgement is required, use an appropriate procedure for this**. This could be formal expert knowledge elicitation (EFSA, 2014a), or semi-formal expert knowledge elicitation or expert discussion (EFSA, 2016). Assessors should choose a procedure that is appropriate for the needs, timeframe and context of their assessment. For example, if the judgement is likely to be critical for decision-making, that would be a reason for more formal methodology, if time and resources allow.

4.5 Uncertainty and influence analysis

All EFSA scientific assessments must include consideration of uncertainties, reporting clearly and unambiguously what sources of uncertainty have been identified and what their impact on the assessment outcome is. 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, and that any uncertainties that cannot be included in this should be described qualitatively (EFSA, 2016). These recommendations apply to weight of evidence approaches, as well as other types of scientific assessment.

Weight of evidence assessment contributes to uncertainty analysis, as explained in Section 2.6. However, weight of evidence conclusions expressed as the range of possible answers and their relative support (e.g. an estimate and confidence interval from a meta-analysis) may not incorporate all the uncertainty affecting the weight of evidence assessment. First, the assessors may have omitted some considerations regarding the evidence from the integration process, leaving them to be addressed in the uncertainty analysis, as described in step 5)b.ii. of Section 4.6. Second, there will often be additional uncertainties associated with the identification of evidence (including the choice of search criteria), the impact of any potentially relevant evidence that was excluded from detailed assessment, and the choice and implementation of methods for assembling, weighing and integrating data. This includes 'uncertainties in the judgement used' in weight of evidence assessment (SCENIHR, 2012). Assessors should systematically document all identifiable uncertainties affecting the weight of evidence assessment, and take them into account in the assessment of combined uncertainty for the overall scientific assessment (EFSA 2016, Section 12).

Influence analysis or sensitivity analysis is an optional part of scientific assessment (EFSA, 2016). It can be valuable in identifying which sources of uncertainty contribute most to the uncertainty of assessment conclusions, and hence in targeting refinement of the assessment when this is required. When applied to assessments that include weight of evidence approaches, this could help decide whether and where to refine the weight of evidence assessment (see below). In meta-analysis, it is good practice to study the influence of individual primary studies on the effect size estimates (e.g. leverage plots, jack-knife procedure) and to study the impact of the modelling approach (e.g. random effect model vs. fixed effect model). In addition, meta-regression has been recommended to study the sensitivity of the effect size estimates to explanatory factors related to the study characteristics. In case of heterogeneity in the results of primary studies this may support identification of major sources of uncertainties and variability. Sensitivity analysis was also applied with other weight of evidence approaches, especially with quantitative multi-criteria analysis. For example, the last step of the quantitative multi-criteria analysis framework described by Linkov et al. (2011) explicitly deals with sensitivity analysis. It is useful to explore analogies between these formal approaches for influence analysis and similar approaches applicable in less formal methods for evidence integration. For example, it can be recommended to study the effect of leaving out individual lines of evidence or, if applicable, individual sources (e.g. primary studies) on the weight of evidence conclusions. Likewise, alternative methods for evidence integrating could be used and their influence on the weight of evidence outcome be demonstrated.

4.6 Iterative refinement of the assessment

Iterative refinement is an option in any type of scientific assessment. It is generally aimed either at reducing uncertainty or improving the characterisation of uncertainty, in those areas of the assessment that contribute most to the uncertainty of the assessment conclusions as identified by influence analysis or sensitivity analysis (EFSA 2016). In general, assessment should start at a level of refinement the assessors consider appropriate to the needs and context of the assessment, and then be refined as far as is necessary to inform decision-making or until the agreed time and resources are expended.

When refinement is needed in parts of an assessment where weight of evidence approaches are used, this can be achieved by returning to earlier steps of the process (illustrated in Figure 1), depending on what contributes best to refining the assessment. In some cases, it may be sufficient to refine one or more of the basic steps of the weight of evidence assessment, whereas in other cases it may be beneficial to return to problem formulation and reformulate the questions to be addressed. For example, if a question involving complementary lines of evidence contributed significantly to overall uncertainty, consideration could be given to further sub-dividing the question, addressing each complementary part as a separate sub-question, and then combining their conclusions (with the option of using a quantitative model).

5. Reporting weight of evidence assessment

If the weight of evidence assessment has been conducted following a standardised procedure previously established for use in this area of EFSA's work, the weight of evidence assessment may be reported in the manner that is normal for that standardised procedure, provided this is transparent. The standardised procedure should be referenced and its applicability to the case in hand should be explained if it is not self-evident.

All other weight of evidence assessments should be reported as described below, although the level of detail may be reduced due to time constraints in emergency assessments.

Reporting should be consistent with EFSA's general principles regarding transparency (EFSA 2006, 2009a) and reporting (EFSA 2014a, 2015b). In a weight of evidence assessment this should include justifying the choice of methods used, documenting all steps of the procedure in sufficient detail for them to be repeated, and making clear where and how expert judgement has been used. Where the assessment used methods that are already described in other documents, it is sufficient to refer to those. Reporting should also include referencing and, if appropriate, listing or summarising all evidence considered, identifying any evidence that was excluded; detailed reporting of the conclusions; and sufficient information on intermediate results for readers to understand how the conclusions were reached.

Weight of evidence assessment is part of the wider process of scientific assessment, as illustrated earlier in Figure 1. Guidance on reporting other parts of the wider procedure, including evidence review, problem formulation and uncertainty analysis, is provided elsewhere (e.g. EFSA 2014b, EFSA 2015b, 2016). This section focusses on the reporting for the three basic steps of weight of evidence assessment: assembling the evidence, weighing the evidence and integrating the evidence. These steps should be reported separately for each question or sub-question that is addressed.

To aid transparency and accessibility for readers it may be useful to summarise weight of evidence assessment in a tabular form. A suggested format is shown in Table 1. If a tabular format is not used, then all the information listed in the templates must be included in the assessment report, in a location and format that can easily be located by the reader (e.g. identifiable from section headings in the table of contents). If the information is presented in tabular form it should be concise (ideally not more than 1 page per table) and refer the reader to the text of the opinion for details.

Table 2: Optional tabular format for summarising weight of evidence assessment.

Question		*Insert text of question here*
Assemble evidence	Select evidence	*Briefly summarise the methods used to search, select and extract the evidence (see Note 1).*
	Lines of evidence	*List the line(s) of evidence into which the evidence were assembled for assessment (see Note 2).*
Weigh the evidence	Methods	*Briefly summarise the method(s) used to weigh the lines of evidence (see Note 3).*
	Results	*Give a reference to the section of the assessment where the results of weighing the lines of evidence are presented (see Note 4).*
Integrate the evidence	Methods	*Briefly summarise the methods used to integrate the lines of evidence (see Note 5).*
	Results	*State the conclusions of integrating the evidence for this question (see Note 6).*

Italic descriptions are for guidance only and should be deleted once the table is completed..

Notes cited in the table are presented below.

Notes to Table 2:

Note 1. The summary of the methods used to search, select and extract the evidence should include, for example, 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.

Note 2. When listing the lines of evidence, give enough information for the reader to understand what they contain and how they differ. Present them as numbered bullets for ease of reference. State whether the lines of evidence are complementary, or standalone, or a mixture of both (see section 2.1). Identify lines of evidence that were generated by (are conclusions from) preceding weight of evidence questions (if any).

Note 3. When summarising the method(s) used to weigh the lines of evidence, give enough information to make clear the type of method involved (see types of method in section 3). If weighing and integration was done within lines of evidence, without being assessed in preceding sub-questions, briefly summarise also the methods used for that. Refer the reader to the sections of the assessment where details of each method are provided.

Note 4. The detailed results of weighing the evidence must be presented together, 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). If they can be summarised briefly, include them in Table 1.

Note 5. Briefly summarise the methods used to integrate the lines of evidence, giving enough information to make clear the type of method involved (see types of method in section 3 of Guidance).

Note 6. State the conclusion of integrating the evidence for this question in a form that expresses the range of possible answers and their relative support.

- If the weight of evidence assessment directly addresses the conclusion of a scientific assessment, results of weighing and integration that have been conducted by different methods (e.g. a combination of qualitative and quantitative methods), should integrated into a single conclusion on the

relative support for different answers to the question, and expressed quantitatively to the extent that is possible. Any considerations that remain unquantified should be described qualitatively.

- If the weight of evidence assessment addresses an intermediate question in a larger scientific assessment, results of weighing and integration that were conducted by different methods may be expressed either qualitatively or quantitatively. They may either integrated into a single conclusion here, or carried forward separately to later stages of the scientific assessment.

6. Way forward and recommendations

This guidance document is intended to guide EFSA panels and staff on the use of the weight of evidence approach in scientific assessments. It provides a flexible framework that must be used in all areas within EFSA's remit, within which assessors should apply those methods which most appropriately fit the purpose of their individual assessment.

EFSA Panels and Units must ensure that their existing approaches to weight of evidence assessment are in line with the unconditional requirements of the current guidance document. In particular they should consider:

- Whether all pertinent aspects of reliability, relevance and consistency are addressed,
- How to ensure the transparency of weight of evidence assessments.

It is further recommended that:

- EFSA identify areas of its work where weight of evidence assessment is especially needed and initiate further work to apply suitable approaches from this guidance in those areas. This might include, for example, the integration of different types of evidence in chemical risk assessment, including in vivo, in vitro, in silico, omics, PBPK modelling as well as the Mode of Action and Adverse Outcome Pathway concepts.
- EFSA identify specific weight of evidence approaches that may provide added value in EFSA's work (especially quantitative methods, e.g. meta-analysis) and consider whether further guidance or training on them would facilitate uptake.
- EFSA should explore approaches to apply weight of evidence in rapid scientific assessments, where time and resources are limited.

In implementing all the aforementioned recommendations, it is suggested that EFSA continues to collaborate at the European and international level to harmonise developments in this area.

7. Glossary (to be completed in the final document)

Weight of evidence assessment

Weight of evidence assessment is a process in which evidence is integrated to determine the relative support for possible answers to a question.

Lines of evidence

Lines of evidence are subsets of evidence which it is useful to distinguish when conducting a scientific assessment.

Relevance

Relevance, in terms of its contribution to the weight of evidence assessment, is the contribution a line of evidence would make to answer a specified question, if the information comprising the line of evidence was correct.

Reliability

Reliability, in terms of its contribution to the weight of evidence assessment, is the extent to which the information comprising a line of evidence is correct.

Consistency

Consistency, in terms of its contribution to the weight of evidence assessment, is the extent to which the contributions of different lines of evidence to answering the specified question are compatible.

Standalone line of evidence

Standalone line of evidence is the line of evidence which offers an answer to the question without needing to be combined with other lines of evidence.

Complementary line of evidence

Complementary line of evidence is the line of evidence which can only answer the question when it is combined with other line(s) of evidence.

Standard assessment

Standard assessment is the procedure where the evidence to be used is defined by regulations or guidance, and/or submitted by applicants. Standard assessments follow standardised procedures, which include pre-specified approaches to weight of evidence assessment.

Case-specific assessment

case-specific assessment is the procedure where the assessors need to choose and apply approaches on a case-by-case basis due to the lack of any pre-specified procedure;

Emergency assessment

Emergency assessment is the procedure where the choice of approach is constrained by limitations on time and resources.

8. References

Anses (2016). Evaluation du poids des preuves à l'Anses: revue critique de la littérature et recommandations à l'étape d'identification des dangers. Rapport d'expertise collective. Saisine 2015-SA-0089.

Ashby D, Smith AF (2000). Evidence-based medicine as Bayesian decision-making. Statistics in medicine, 19(23), 3291-3305.

Borenstein M., Hedges L.V., Higgins J.P.T., Rothstein H.R. 2009. Introduction to meta-analysis. Wiley.

Peter M. Chapman , Blair G. McDonald & Gary S. Lawrence. 2002. Weight-of-Evidence Issues and Frameworks for Sediment Quality (And Other) Assessments, Human and Ecological Risk Assessment: An International Journal, 8:7, 1489-1515, DOI: 10.1080/20028091057457.

Codex, 2015. Codex Alimentarius Commission. Procedural Manual. 23rd edition.

Collaboration for Environmental Evidence. Systematic mapping. 2016. http://environmentalevidence.org/wp-content/uploads/2014/05/EE_InstructionsforAuthors_SYSTMAPS.pdf.

Suhail A. R. Doi (2014). Evidence Synthesis for Medical Decision Making and the Appropriate Use of Quality Scores. doi:10.3121/cmr.2013.1188.

Dubois, Didier, and Henri Prade. Possibility theory. Springer, 1988.

EC (European Commission), 2000. Communication from the Commission on the precautionary principle. COM(2000) 1 final. European Commission, Brussels.

EC (European Commission), 2006. Report of the First Meeting of the Chairs of Scientific Committees of Community bodies involved in Risk Assessment (7 and 8 December 2005). C7/MM D(2006) 370004. Brussels, 28 February 2006.

ECHA, 2016. Practical guide. How to use and report (Q)SARs. Version. 3.1. https://echa.europa.eu/documents/10162/13655/pg_report_qsars_en.pdf.

EFSA, 2006. EFSA (European Food Safety Authority), 2006. Transparency in risk assessment carried out by EFSA: Guidance document on procedural aspects. The EFSA Journal (2006), 353, 1-16.

EFSA, 2009. EFSA (European Food Safety Authority), 2009. Guidance of the Scientific Committee on Transparency in the Scientific Aspects of Risk Assessments carried out by EFSA. Part 2: General Principles. The EFSA Journal (2009), 1051, 1-22.

EFSA, 2010. EFSA (European Food Safety Authority), 2010. Application of systematic review methodology to food and feed safety assessments to support decision making. EFSA Journal 2010; 8(6):1637.

EFSA, 2010a. EFSA (European Food Safety Authority), 2010. Guidance on a harmonised framework for pest risk assessment and the identification and evaluation of pest risk management options. EFSA Journal 2010; 8(2):1495.

EFSA, 2011. EFSA (European Food Safety Authority), 2011a. EFSA Scientific Committee; Statistical Significance and Biological Relevance. EFSA Journal 2011;9(9):2372. [17 pp.] doi:10.2903/j.efsa.2011.2372.

EFSA, 2012. EFSA (European Food Safety Authority), 2012. Guidance on selected default values to be used by the EFSA Scientific Committee, Scientific Panels and Units in the absence of actual measured data. The EFSA Journal, 2579, 1-32.

EFSA, 2013. EFSA SC (EFSA Scientific Committee), 2013. Scientific Opinion on Priority topics for the development of risk assessment guidance by EFSA's Scientific Committee. EFSA Journal 2013;11(8):3345, 20 pp. doi:10.2903/j.efsa.2013.3345.

EFSA, 2014. EFSA European Food Safety Authority), 2014. Discussion Paper - Transformation to an "Open EFSA". www.efsa.europa.eu.

EFSA, 2014a. EFSA (European Food Safety Authority), 2014. Guidance on Expert Knowledge Elicitation in Food and Feed Safety Risk Assessment. EFSA Journal 2014; 12(6):3734.

EFSA, 2014b. EFSA (European Food Safety Authority), 2014. Guidance on Statistical Reporting. EFSA Journal 2014;12(12):3908, 18 pp.

EFSA, 2014c. EFSA (European Food Safety Authority), 2014. Systematic review guidance.

EFSA, 2015. EFSA (European Food Safety Authority), 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.

EFSA, 2015a. EFSA (European Food Safety Authority), 2015. Scientific Opinion on the risks to public health related to the presence of bisphenol A (BPA) in foodstuffs – Part II: Toxicological assessment and risk characterisation. EFSA Journal 2015;13(1):3978.

EFSA, 2015b. EFSA (European Food Safety Authority), 2015. Scientific report on Principles and process for dealing with data and evidence in scientific assessments. EFSA Journal 2015;13(5):4121, 35 pp. doi:10.2903/j.efsa.2015.4121.

EFSA, 2016. EFSA (European Food Safety Authority) Scientific Committee, 2016. Guidance on Uncertainty in EFSA Scientific Assessment - Draft version for internal testing. Available at https://www.efsa.europa.eu/en/topics/topic/unvertainty.

EFSA, 2016a. EFSA BIOHAZ Panel (EFSA Panel on Biological Hazards), 2016. Scientific opinion on the evaluation of the safety and efficacy of Listex™ P100 for reduction of pathogens on different ready-to-eat (RTE) food products. EFSA Journal 2016;14(8):4565, 94 pp. doi:10.2903/j.efsa.2016.4565.

Glenn W. Suter II, Susan M. Cormier. 2011. Why and how to combine evidence in environmental assessments: Weighing evidence and building cases. Science of the Total Environment. 409 (2011) 1406–1417.

Good IJ (1950). Probability and the Weighing of Evidence. Charles Griffin.

Good IJ (1952). Rational Decisions. Journal of the Royal Statistical Society, Series B (14), 107--114.

Good IJ (1983). Good Thinking: The Foundations of Probability and Its Application. Minneapolis, MN: University of Minnesota Press.

Good IJ (1988). Statistical evidence. In Kotz S and Johnson NL (Eds.), Encyclopedia of Statistics (Vol. 8, pp. 651–655). New York: John Wiley and Sons.

Gosling, J P , Andy Hart, H Owen, M David, J Li, et C MacKay. 2013. "A Bayes linear approach to 25 weight-of-evidence risk assessment for skin allergy." Bayesian Analysis 8 (1):169-186.

Guyatt G, Oxman AD, Akl EA, et al. 2011. GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. J Clin Epidemiol. 64(4):383-394.

Higgins and Green 2011. Cochrane 5 Handbook for Systematic Reviews of Interventions.

Higgins JP, Thompson SG, Spiegelhalter DJ. 2015. A re-evaluation of random-effects meta-analysis.

Hill AB. 1965. The environment and the disease: association or causation? Proc R Soc Med. 58:295-300.

IARC. 2006. IARC Monographs on the Evaluation of Carcinogenic Risks to Humans - Preamble. Lyon, France: World Health Organization International Agency for Research on Cancer.

IPCS (International Programme on Chemical Safety), 2004. IPCS Risk Assessment Terminology. International Programme on Chemical Safety. WHO, Geneva.

James KL, Randall NP, Haddaway NR. A methodology for systematic mapping in environmental sciences. Environ Evid. 2016;5(1):1.

Jones DR (1995). Meta-analysis: Weighing the evidence. Statistics in medicine, 14(2), 137-149.

Krimsky S (2005). The weight of scientific evidence in policy and law. American Journal of Public Health, 95(S1), S129–S136.

Li Y., and A. Ngom, "Data integration in machine learning," Bioinformatics and Biomedicine (BIBM), 2015 IEEE International Conference on, Washington, DC, 2015, pp. 1665-1671. doi: 10.1109/BIBM.2015.7359925 URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7359925&isnumber=7359638.

1494 Linkov I, Loney D, Cormier S, Satterstrom FK, Bridges T. 2009. Weight-of-evidence
1495 evaluation in environmental assessment: review of qualitative and quantitative
1496 approaches. Sci Total Environ. 407(19):5199-205.

1497 Linkov, I., Welle, P., Loney, D. et al. (2011). Use of multicriteria decision analysis to
1498 support weight of evidence evaluation. Risk Analysis 31, 1211-1225.

1499 Linkov I, Massey O, Keisler J, Rusyn I, Hartung T. 2015. From "weight of evidence" to
1500 quantitative data integration using multicriteria decision analysis and Bayesian methods.
1501 ALTEX. 32:3-8.

1502 Oxford Dictionaries, 2015. Accessed 17 February 2015.
1503 http://www.oxforddictionaries.com/ definition/english/.

1504 Lorenz R. Rhomberg, Julie E. Goodman, Lisa A. Bailey, Robyn L. Prueitt, Nancy B. Beck,
1505 Christopher Bevan, Michael Honeycutt, Norbert E. Kaminski, Greg Paoli, Lynn H.
1506 Pottenger, Roberta W. Scherer, Kimberly C. Wise, and Richard A. Becker. 2013. A survey
1507 of frameworks for best practices in weight-of-evidence analyses. Critical Reviews In
1508 Toxicology Vol. 43 , Iss. 9,2013.

1509 M E (Bette) Meek, Christine M Palermo, Ammie N Bachman, Colin M North, and R Jeffrey
1510 Lewis. 2014. Mode of action human relevance (species concordance) framework:
1511 Evolution of the Bradford Hill considerations and comparative analysis of weight of
1512 evidence. J Appl Toxicol. 2014 Jun; 34(6): 595–606. Published online 2014 Feb 10. doi:
1513 10.1002/jat.2984.

1514 NTP. 2015. Handbook for Preparing Report on Carcinogens Monographs July 20, 2015.
1515 https://ntp.niehs.nih.gov/ntp/roc/handbook/roc_handbook_508.pdf.

1516 OHAT. 2015. Handbook for Conducting a Literature-Based Health Assessment Using
1517 OHAT Approach for Systematic Review and Evidence Integration. Research Triangle Park,
1518 NC: 30 OHAT.

1519 OSHA. 2016. Guidance on Data Evaluation for Weight of Evidence Determination:
1520 Application to the 2012 Hazard Communication Standard.

1521 Richard A. Becker, Gerald T. Ankley, Stephen W. Edwards, Sean W. Kennedy, Igor
1522 Linkov, Bette Meek, Magdalini Sachana, Helmut Segner, Bart Van Der Burg, Daniel L.
1523 Villeneuve, Haruna Watanabe, Tara S. Barton-Maclaren. 2015. Increasing Scientific
1524 Confidence in Adverse Outcome Pathways: Application of Tailored Bradford-Hill
1525 Considerations for Evaluating Weight of Evidence. Regulatory Toxicology and
1526 Pharmacology. Volume 72, Issue 3, August 2015, Pages 514–537.

1527 SCHER, SCENIHR, SCCS, 2013. Making risk assessment more relevant for risk
1528 management. Scientific Committee on Health and Environmental Risks, Scientific
1529 Committee on Emerging and Newly Identified Health Risks, and Scientific Committee on
1530 Consumer Safety. March 2013.

1531 SCENIHR (Scientific Committee on Emerging and Newly Identified Health Risks),
1532 Memorandum on the use of the scientific literature for human health risk assessment
1533 purposes – weighing of evidence and expression of uncertainty, 19 March, 2012.

1534 Small M.J. 2005. Methods for assessing uncertainty in fundamental assumptions and
1535 associated models for cancer risk assessment. Risk analysis 28, 1289-1307.

1536 Small MJ. Methods for assessing uncertainty in fundamental assumptions and associated
1537 models for cancer risk assessment. Risk Anal. 2008 Oct;28(5):1289-308. doi:
1538 10.1111/j.1539-6924.2008.01134.x. Epub 2008 Sep 20.

1539 Smith EP, Lipkovich I, Ye, K (2002). Weight-of-evidence (weight of evidence):
1540 quantitative estimation of probability of impairment for individual and multiple lines of
1541 evidence. Human and Ecological Risk Assessment, 8(7), 1585-1596.

1542 Spiegelhalter D and Knill-Jones R (1984) Statistical and knowledge-based approaches to
1543 clinical decision support systems, with an application in gastroenterology. Journal of the
1544 Royal Statistical Society, (Series A), (147), 35--77.

1545 Staples C., E. Mihaich, J. Carbone, K. Woodburn, G. Klecka. 2004. A weight of evidence
1546 analysis of the chronic ecotoxicity of nonylphenol ethoxylates, nonylphenol ether
1547 carboxylates, and nonylphenol. Hum Ecol Risk Assess, 10, 999–1017.

1548 Sutton A.J., Abrams K.R. (2001). Bayesian methods in meta-analysis and evidence
1549 synthesis. Stat Methods Med Res. 2001 Aug;10(4):277-303.

1550 Thayer K.A., Schünemann H.J. (2016). Using GRADE to respond to health questions with
1551 different levels of urgency. Environment International 92_93, 585-589.

1552 Turner R.M., David J Spiegelhalter, Gordon C S Smith, and Simon G Thompson. 2009.
1553 Bias modelling in evidence synthesis. J R Stat Soc Ser A Stat Soc. 2009 Jan; 172(1):
1554 21–47.

1555 US Environmental Protection Agency, 2000. Risk characterization handbook. Science
1556 Policy Council, U.S. EPA, Washington DC.

1557 Vermeire T, Aldenberg T, Buist H, Escher S, Mangelsdorf I, Pauné E, Rorije E, Kroese D.
1558 2013.OSIRIS, a quest for proof of principle for integrated testing strategies of chemicals
1559 for four human health endpoints. Regul Toxicol Pharmacol. 2013 Nov;67(2):136-45. doi:
1560 10.1016/j.yrtph.2013.01.007. Epub 2013 Feb 4.

1561 Vose D, 2008. Risk analysis: a quantitative guide. 3rd edition. Wiley.

1562 Walley, Peter (1991). Statistical Reasoning with Imprecise Probabilities. London:
1563 Chapman and Hall. ISBN 0-412-28660-2.

1564 WCRF / AICR 2007 Second expert report - Food, Nutrition, Physical Activity and the
1565 Prevention of Cancer: a Global Perspective Systematic Literature Review - Specification
1566 Manual.

1567 Weed D. 2005. Weight of evidence: a review of concept and methods. Risk Anal,
1568 25:1545–e1557.

1569 WHO, 2009. WORLD HEALTH ORGANIZATION. Food Safety. Project to update the
1570 principles and methods for the assessment of chemicals in food. Principles and methods
1571 for the risk assessment of chemicals in food. EHC 240. ISBN 978 92 4 157240 8.

ANNEX I. Illustration of the proposed Approach to assess the Weight of evidence: Problem formulation, hierarchy of questions and mapping lines of evidence for chemical risk assessment

Annex I provides examples of the application and reporting of the weight of evidence approaches used in assessing chemical risks to human health and the environment in the context of EFSA's mandate. In the introductory part, general principles of conducting chemical risk assessments are presented, and a decision tree provided to demonstrate data availability and how it drives the applied approaches. This is followed by examples demonstrating different situations with regard to data availability in human health risk assessment and an example of ecological risk assessment. In all cases, specific questions are addressed.

A.1. Problem formulation

Problem formulation refers to framing of the scientific question(s) that are generally posed by a decision maker or a stakeholder. It is important to ensure that the question(s) fully encompass the issue(s) that need to be addressed, and are clear and agreed prior to the start of the assessment (Hardy et al., 2015). The WHO (2009) has regarded problem formulation as an iterative process involving risk assessors and risk managers that determines the need for and the extent of a risk assessment. According to the US-EPA (2007), problem formulation is "a systematic planning step to identify the major factors to be considered in a particular assessment in relation to preliminary hypotheses with regards to hazard assessment (i.e. likelihood and severity of adverse effects which might occur or have occurred) and exposure assessment (i.e. likelihood and significance of exposure)".

In a food safety context, a typical problem formulation may include a description of the intended application (e.g. a food additive) and the commodities involved; issues expected to be affected (e.g. human health), and potential consequences; consumer perception of the hazards or risks; and distribution of possible risks among different segments of the population. The desired outcomes of problem formulation are in the form of clear questions that need answering, identification of the resources and the timeframe that would be needed for the assessment. In relation to EFSA's work, problem formulation is often defined in the terms of reference (ToR) provided by risk managers from the European Commission or a Member State. Data requirements for pre-market authorisation of a substance are defined in a number of regulation and guidance documents.

With regard to applying weight of evidence in such assessments, it is important to note that problem formulation is a pre-requisite and precedes the assessment, including the weighing and the integration of the evidence. In the case of food chemicals, the problem may be too complex to be addressed in a single question, and may need to be divided into more lower-tier questions that can be answered directly, and, by combining the answers to these, the main

question can be addressed. In doing this, a hierarchy of questions is defined. Relevant data/information can be collected, assessed, weighed and integrated into separate lines of evidence that would answer each question at the bottom of the hierarchy. Integration continues upward the question hierarchy until an overall integration can be reached to respond to the main question.

A.2. Chemical risk assessment (human health)

The basic steps of a chemical risk assessment involve a structured way to address the hierarchy of the question(s) for each step of the process – i.e. hazard assessment, exposure assessment and risk characterisation. This gives the assessor a starting point to map the evidence needed for the assessment in a particular context (e.g., regulated chemicals/ products or contaminants for human health or environmental risk assessment).

Human risk assessment of chemicals applied by EFSA to the food and feed safety area may include the following generalised elements for hazard identification and characterisation. For each of those elements, different lines of evidence may be available and can be integrated depending on the purpose of the assessment (e.g. a standard procedure, rapid risk assessment, chemical-specific assessment), and availability of data, time and resources.

Hazard identification

- Genotoxicity: In vitro studies such as bacterial Ames test and mammalian micronucleus assay; in vivo studies; in silico models.
- Toxicokinetics: In vivo studies on absorption, distribution, metabolism, excretion (ADME); in vitro studies; in silico models.
- Toxicity/ Toxicodynamics: Epidemiological and clinical studies; case reports; in vivo studies (acute, sub-chronic, chronic toxicity, carcinogenicity, reproductive/developmental toxicity); in vitro studies; in silico models.

Hazard characterisation (dose response relationship to derive Health Based Guidance Values (HBGV) or Reference Values (RVs) for Margin of Exposure (MOE).

- Toxicokinetics: In vivo and/ or in vitro studies on absorption, distribution, metabolism, excretion (ADME), in silico models (toxicokinetic and/ or physiologically based models).
- Acute toxicity: In vivo studies, in vitro studies, case reports.
- Sub-Chronic/Chronic/Carcinogenicity: Epidemiological and clinical studies; in vivo studies including pathological investigations, clinical chemistry; in vitro studies; in vivo and/ or in vitro OMICs studies (transcriptomics, proteomics, metabolomics, exposomics); in silico models; read - across extrapolations; default values (e.g. Threshold of Toxicological Concern (TTC)).

- Other studies (as necessary): e.g. studies on reproductive/developmental toxicity, neurotoxicity, immunotoxicity.

Exposure assessment

- Occurence: concentration of the chemical in food (total diet study, food monitoring, food composition tables, etc) or other media using results of analytical methods (High Performance Liquid Chromatography (HPLC), Gas Chromatography (GC), Mass Spectrometry (MS), etc.), default values (e.g. Maximum residue levels (MRLs), maximum limits, etc..), *in silico* models.

- Food Consumption: (consumption surveys, food consumption and composition databases e.g. EFSA food consumption and composition databases, budget method, volume of production, default value), *in silico* models.

Risk characterisation

- Compounds with threshold effects: comparison of health-based guidance value (e-g. Acceptable Daily Intake (ADI), Tolerable Daily Intake (TDI)) with exposure estimates.

- Compounds with non-threshold effects (e.g genotoxic and carcinogenic compounds) MOE approach.

The cases described below relate to human risk assessment of regulated products for pre-market authorisation, as well as re-evaluation of regulated compounds and contaminants.

A.2.1. Human health risk assessment of regulated products for pre-market authorisation

For regulated substances (e.g. food and feed additives, flavourings, food contact materials, pesticides), minimum data requirements are provided in the respective EC regulations and relevant guidance documents. The information is required to be submitted in a dossier by any applicant who is seeking authorisation of a substance prior to placing it on the market (see, for example, EFSA 2012d). In these cases, problem formulation is often set by the EC regulations and guideline documents, defining the hierarchy of the questions for the purpose of the assessment. Where there are data gaps in relation to hazard identification/ characterisation, data/information derived from other methods, such as *in silico* modelling and read across, may be useful in filling the gaps. Similarly, in the absence of comprehensive data on the occurrence in different food groups, exposure assessment may be based on point estimates.

A.2.2. Re-evaluation of human health risk assessment of existing regulated products

For re-evaluation of regulated substances, available data may include an original dossier, and in some circumstances in addition, historical assessments, and/or published studies in scientific literature that can be used to address each question at different steps of the risk assessment process. The availability of more detailed information would provide more options for applying quantitative probabilistic methodologies.

A.2.3. Human health risk assessment of contaminants

Contaminants may include regulated chemicals that may have been removed from the market, e.g., non-authorised pesticides, brominated flame retardants, or anthropogenic contaminants e.g., metals, persistent organic pollutants. Some contaminants may also result from food and feed processing, e.g., acrylamide, PAHs, or from natural sources such as toxins of plant, fungal, or marine organism origin. Although, in the case of chemicals, the hierarchy of the question(s) still follows the structured steps of risk assessment, the nature of the data available may vary enormously. This depends on whether an assessment had been performed previously, or if the contaminant is new and emerging with limited safety data. Consequently, methods to combine evidence may range from basic description of the evidence in data-poor situations to full probabilistic assessment in data-based situations.

A.2.4. A tiered approach to map the level of knowledge and evidence available for human health risk assessment of chemicals

Under the scenarios described above, a tiered approach has been illustrated in Figure 4 with the aim to map the level of knowledge and the type of evidence/data necessary to address the questions for hazard characterisation of chemicals in human health risk assessment. The tiered approach has been adapted from the WHO, the US-EPA and EFSA (Meek et al., US-EPA, 2007, EFSA, 2013). The reader should consider each step of the risk assessment independently since the level of knowledge can be any combination of data-poor and data-based situation for the exposure and/or hazard identification/ characterisation and consequently risk characterisation.

A.2.4.1. Weighing evidence for human risk assessment of chemicals

Once, in the problem formulation, the questions for data - poor and data - rich situations have been defined, and available data sources and data gaps have been identified, a hierarchy of questions can be developed in relation to assembling, weighing, and integrating the available evidence.

A.2.4.2. Hazard Identification and characterisation

The first step involves searching for and selecting the evidence that is relevant for answering the question(s) at hand and deciding if and how to group it/them

into lines of evidence as needed. The second step involves detailed evaluation and weighing of the assembled evidence. In the third step, the evidence is integrated to arrive at conclusions. Each weight of evidence assessment is associated with specific uncertainties that contribute to the overall uncertainty assessment (see Section 2.6 below).

Figure 4 provides a decision tree which summarises each step of the WoE analysis for hazard identification and hazard characterisation of chemicals in food in data rich and data poor situations. These examples of WoE assessment for hazard characterisation of a chemical in these two data situations are given since these represent both ends of the spectrum EFSA panels may encounter, when dealing with human health risk assessment of chemicals. In the context of authorisation of regulated compounds, data needs are determined by the relevant guidance and may not cover all endpoints listed e.g. derivation of a reference point (point of departure) and a health-based guidance value. In some cases, data gaps for specific endpoints may be encountered. In such cases, empirical data can be combined with estimates generated from *in silico* tools.

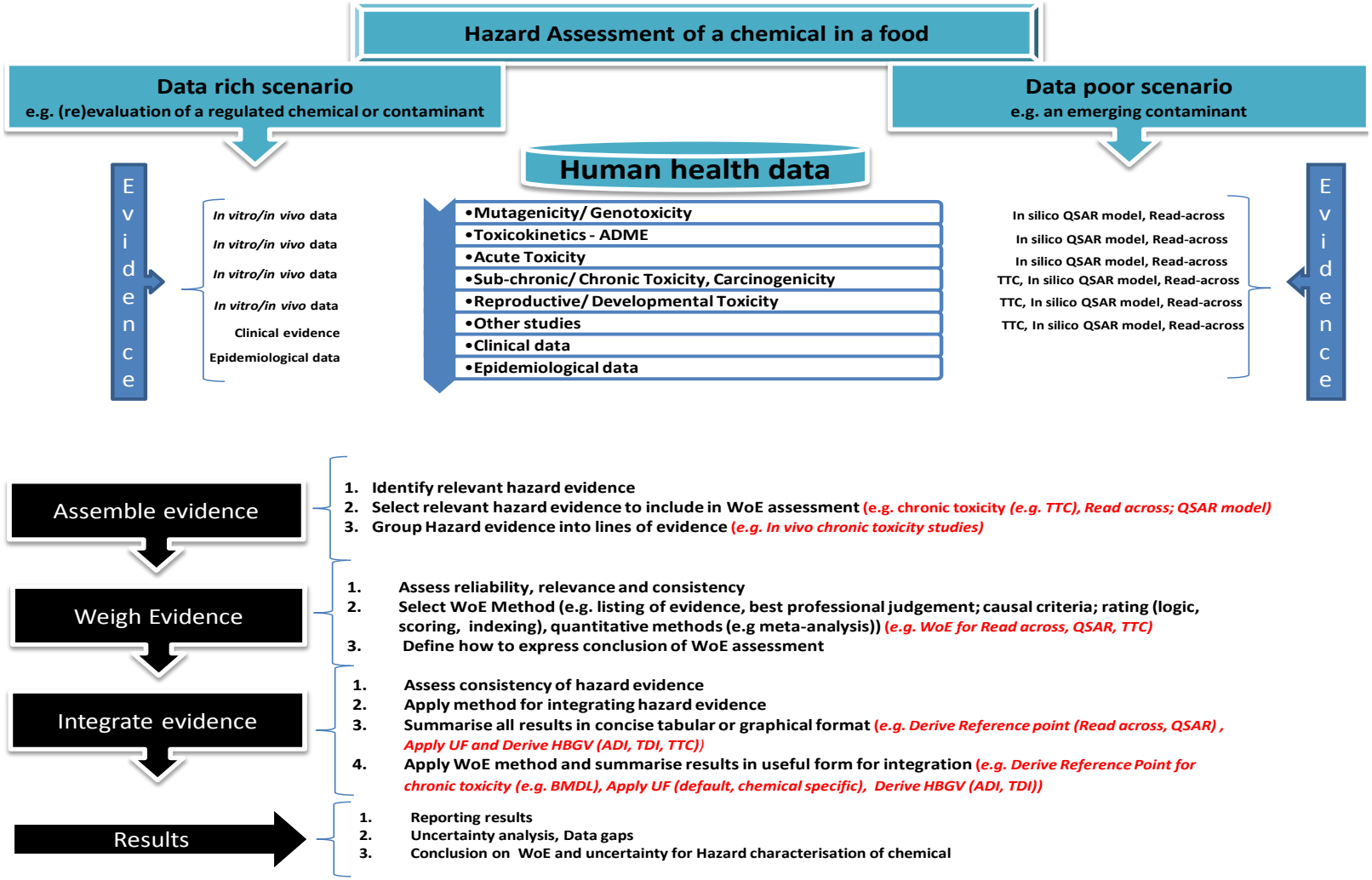


Figure 4: Generic decision tree for hazard identification and hazard characterisation of chemicals in food in data rich and data poor situations

A.2.4.3. Hazard characterization of an emerging contaminant for human health

The example presented below describes the use of weight of evidence in emerging contaminants. They can be anthropogenic contaminants, e.g., brominated contaminants or result from food and feed processing, e.g., newly identified Maillard reaction products or even from natural sources such as toxins of plant, fungal, or marine organism origin.

Table 3: Optional tabular format for summarising weight of evidence assessment for an emerging contaminant.

Question:Hazard characterisation of an emerging contaminant		
Assemble evidence	Select evidence	No toxicity data available: use read across from already-tested similar compounds, in silico tools (QSAR) to predict toxicity
	Lines of Evidence	Identify lines of evidence for potential effect (s) from the presence of a structural alert or QSAR models, read across from similar compounds,.
Weigh the evidence	Methods	Evaluate the reliability, relevance and consistency of the QSAR models. This can include weighing model results on a statistical basis (e.g. likelihood of a compound with a structural alert to express (a) toxic property(ies) .
	Results	Toxicity value for each line of evidence, with associated assessment of reliability (e.g. through the applicability domain of the models used).
Integrate the evidence	Methods	If the estimates from the different models converge, the level of uncertainty regarding the toxic property (ies) can be evaluated (e.g. through the applicability domain of the models used). If the estimates do not converge, further modelling for the toxic property (ies) could undertaken to evaluate whether the results can be improved.
	Results	Integrated the toxicity value and uncertainty factor to derive a health based guidance value for the emerging contaminant: Summary Table.

Preferably use multiple tools for the assessment, for instance, multiple QSAR models. Different in silico tools will have different levels of transparency, and may be more or less relevant for the target compound. Preference should therefore be given to the model(s) that also provide an assessment of their applicability domain(s). A general strategy for weighing the evidence from different tools should consider the following criteria:

1) Does the chemical possess a (structural/functional) feature that indicates potential for toxicity (e.g. the presence of a structural alert)?
2) Are there other factors that may negate this feature (e.g. an exception rule to the structural alert, or other influencing factors such as the lack of systemic absorption)?

3) Are there similar compounds with the same feature as the target compound (again considering any counteracting factors)?

   a. Do the similar compounds have the expected toxicity, as suggested by the presence of the alert?

   b. If there are no similar compounds with the relevant toxicity feature, are there other structurally- or functionally-similar compounds that have been tested?

4) Do the results of the QSAR models support the focussed studies listed above?

Note that a compound may still be toxic even if it is not associated with any known alert, or not toxic even if it contains a structural alert for toxicity.

In straight forward cases, there is generally an agreement between the presence of a toxicity alert, read-across from similar compounds, and predicted toxicity by QSAR model(s). In more complex situations, however, some alerts may conflict with the evidence from different methods, e.g. similar compounds used in a read-across may have conflicting values, or the output from different QSAR models may show disagreement. The use of more sophisticated in silico tools (such as the OECD toolbox, VEGA, etc.) may help the Expert in identifying any toxicity alerts in the target compound, and/or the presence of the alerts in other similar compounds, as well as providing a measure of the uncertainty associated with each alert and result of read-across/QSAR modelling.

A.2.4.4. Example of the use of non-testing methods within a weight of evidence framework

The example presented below describes the use of weight of evidence when information is derived from non-testing methods (NTM), such as in silico models and read across. Three in silico platforms - VEGA (www.vega-qsar.eu), T.E.S.T. (www.epa.gov/nrmrl/std/qsar/qsar.html#TEST), and Toxtree (http://toxtree.sourceforge.net/download.html) - have been used to estimate toxicity of the target compound. A read-across software - ToxRead (www.toxread.eu) - has also been used to provide additional support for the results where necessary.

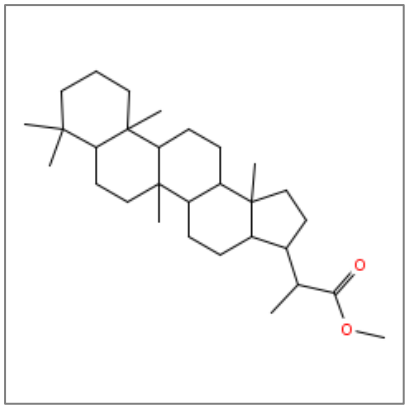


Figure 5: shows chemical structure of the compound used in in silico assessment of toxicity

Table 4: Summary of the results obtained by different non-testing methods

Software	Model / Method	Results	Applicability Domain Index
Toxtree	In vitro mutagenicity (Ames test) alerts by ISS	Non-mutagenic	Not available
T.E.S.T.	Consensus method Hierarchical method FDA method Nearest neighbor method	Mutagenicity Negative Mutagenicity Negative Mutagenicity Negative Mutagenicity Negative	Internally checked
VEGA	Consensus model CAESAR[3] SARpy[3] ISS[3] KNN[3]	Non-mutagenic Mutagenic Non-mutagenic Non-mutagenic Non-mutagenic	 0.719 0.853 0.786 0.819
ToxrRead	Read across	Non-mutagenic	Not available

As shown in Table 4 all models predicted the substance as non-mutagenic, except the CAESAR model within the VEGA platform. Indeed, the target compound does not have a structural alert for mutagenicity. On the contrary, seven alerts associated with the lack of affect have been reported by SARpy within the VEGA platform. Most of the similar compounds are also not-mutagenic, and the majority of the QSAR models used also indicated non-mutagenicity. Thus, the three main elements in the evaluation using non-testing

[3] These are the names of models as found in VEGA (www.vega-qsar.eu)

methods (structural alerts, read-across from similar compounds, and QSAR models) agree on the lack of mutagenic effect. These results are based on experimental values for structurally-similar compounds and on the results of the QSAR models, using both rule- and statistically-based models. Since there is no alert for mutagenicity, reasoning about the mechanism is not applicable. The read across tool (ToxRead) confirms that the substance should not be regarded mutagenic.

However, two indications for mutagenicity need further explaining. First, the CAESAR model indicates mutagenicity, and there is a structurally similar compound (CAS 1250-95-9) that is mutagenic as reported by both VEGA and T.E.S.T. The applicability domain of the CAESAR model for this compound is nevertheless weak, and the software indicates a particular warning highlighting this. This coupled with the fact that similar compounds in the training sets are also non-mutagenic, suggest that the model prediction for mutagenicity of the target compound is not correct. The VEGA software clearly indicates that the very likely reason of the mutagenic effect of the structurally-similar compound is the presence of an epoxide moiety. However, this moiety is not present in the target compound and therefore this conflicting prediction can be disregarded.

The read-across software (ToxRead) provides further support to this conclusion. Most of the closely related substances (Figure 6) show a lack of mutagenicity, whereas the one similar compound which is mutagenic (indicated in red circle) is in fact the same as discussed above.

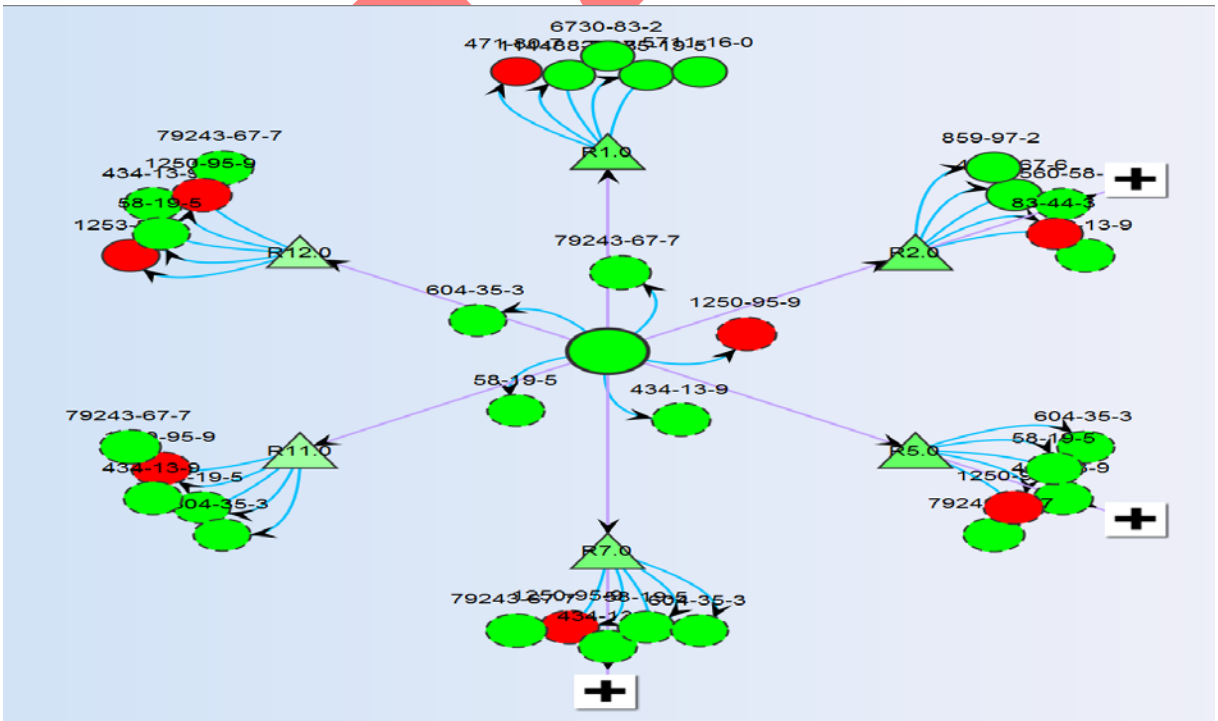


Figure 6: ToxRead chart (www.toxread.eu). The numbers refer to CAS identifiers. Straight arrows link the target chemical to alerts, while curved arrows link to chemicals.

The overall conclusion that can be drawn from this exercise is that the substance in question is most likely not mutagenic (90% of probability).

This example illustrates how different tools can be used to derive toxicity estimates for a given compound in the absence of experimental data. Some of these tools may be rule-based (e.g. Toxtree), statistically based (e.g. T.E.S.T), and tools which offer both these approaches and thus go beyond simple prediction, and provide elements for explicit reasoning. For example, some programmes (e.g. VEGA) provide necessary details that facilitate the reasoning, and thus the acceptance, of the results. The read-across tool, ToxRead, provides a weight of evidence program that can integrate results from QSAR and read across. Other tools do not provide a built-in means to assess reliability of the prediction they provide. In all cases, such tools must not be used as a 'black box', and the final assessment must be carried out by an expert.

1855 **Table 5: Optional tabular format for summarising weight of evidence assessment of an**
1856 **emerging contaminant**

Question		*Hazard characterisation of an emerging contaminant*
Assemble evidence	Select evidence	Ten QSAR models from three in silico platforms and a program for read across were used to estimate genotoxicity potential of the target compound.
	Lines of evidence	Except two, all estimates indicated the compound to be non-mutagenic. The exception were the QSAR model CAESAR within VEGA platform that predicted the compound as mutagenic, and the read-across programme ToxRead that showed one out of five similar compounds to be mutagenic.
Weigh the evidence	Methods	VEGA provides a quantitative measurement of reliability and values higher than 0.8 ADI[4] are considered more reliable. T.E.S.T. applies a filter to eliminate not reliable predictions. The results obtained from these platforms in this case are therefore reliable. ToxTree does not allow assessing the reliability. ToxRead indicates the alerts associated with the effect and similar compounds. In case of chemicals with the toxicity value conflicting with the alert, the user should check if there are alerts present only in the similar compound and not in the target, explaining the conflicting toxicity value.
	Results	T.E.S.T. results consistently indicated non-mutagenicity. The VEGA models called SARpy and KNN showed higher indices for reliability, also predicted non-mutagenicity. The CAESAR and ISS models within the VEGA models showed relatively lower reliability. ToxRead results show that most of the compounds similar to the target compound were not mutagenic. The only structural alert for mutagenicity found in one similar compound is not present in the target compound, and therefore is not relevant.
Integrate the evidence	Methods	The in silico estimates have been integrated while considering the reliability and relevance of the individual values, together with the consistency of all the predicted values, to make an informed expert judgement about the probability that the target compound is not mutagenic.
	Results	The large majority of the in silico values are in concordance for non-mutagenicity of the target compound. One conflicting estimate is less reliable whereas the other is not relevant to the target compound. Considering all the evidence from this in silico assessment, it was concluded by informed expert judgement that the target compound is most likely (about 90% probability) to be non-mutagenic.

1857

1858 ### A.2.4.5. Example of the use of human epidemiological data within a
1859 ### weight of evidence framework: The cadmium example

1860 The CONTAM panel performed a human risk assessment for cadmium in food in
1861 2009 (EFSA 2009, Amzal et al., 2009) and its assessment is an example of the
1862 use of a WOE approach using human epidemiological data to derive a Reference
1863 Point based on Benchmark Dose Limit. Cadmium (Cd) is a heavy metal found as
1864 an environmental contaminant, both through natural occurrence and from
1865 industrial and agricultural sources. Foodstuffs are the main source of cadmium
1866 exposure for the non-smoking general population. Cadmium absorption after
1867 dietary exposure in humans is relatively low (3–5 %), but cadmium is efficiently

[4] Applicability Domain Index

retained in the kidney and liver in the human body, with a very long biological half-life ranging from 10 to 30 years.

The CONTAM panel considered human studies relating to urinary cadmium and urinary biomarkers of toxicity for kidney toxicity (N-acetyl-f3-glucosaminidase, Beta–microglobulinuria (B2-M), alpha 1-microglobulinuria, Urinary retinol-binding protein, Proteinuria) and bone effects (bone mineral density, alkaline phosphatase activity, serum calcium, parathyroid hormone) using a systematic review of the literature. Based on the literature availability at the time, expert judgement and previous international risk assessments (JECFA, ATSDR), the CONTAM panel concluded that B2-M was the most reliable, relevant and consistent urine biomarker for Cd-induced renal tubular toxicity with 35 studies reporting continuous variables as preferable for Benchmark Dose modelling.

A Bayesian meta-analysis and hierarchical modelling was performed to build an overall dose-effect relationship accounting for inter-study heterogeneity and for inter-individual variability of dose and effect. Subsequently, a BDML was evaluated, using a hybrid approach for various cut-offs. As a lower and more protective cut-off level, the panel proposed a biological cut-off for B2M of 300 µg/g creatinine from expert judgment and clinical evidence that exceeding such a threshold is associated with an accelerated age-related decline renal function together with increased mortality (Amzal et al., 2009). The CONTAM Panel selected an overall group-based BMDL5 of 4 µg cadmium / g creatinine. The use of 300 g B2M / g creatinine as critical effect of cadmium exposure to base the risk assessment leads to a possible overestimation of the risk, but is protective of the most sensitive groups of the population. A summary of the WoE assessment for deriving a BDML for Cadmium is presented in Table 5.

Table 6: Summary of the weight of evidence assessment for the derivation of a Human Reference point (Benchmark Dose Limit) for Cadmium in food

Question		Deriving a Reference Point (Benchmark Dose Limit) For Cadmium In humans (Hazard characterisation)
Assemble evidence	Select evidence	Systematic review of Human studies relating urinary cadmium and excretion of biomarkers of toxicity Select relevant papers: biomarkers of kidney and bone effects with continuous outcome to include in WoE assessment
	Lines of evidence	Urinary cadmium and Renal effects LOE 1: N-acetyl-f3-glucosaminidase (NAG) LOE 2: Beta– microglobulinuria (B2-M) LOE 3: Alpha 1-microglobulinuria LOE 4 :Urinary retinol-binding protein (RBP)

		LOE 5: Proteinuria Urinary cadmium and Bone effects LOE 1: Bone mineral density (bone MD) LOE 2:Alkaline phosphatase activity (bALP) LOE 3: Serum calcium LOE 4: Parathyroid hormone (PTH)
Weigh the evidence	Methods	1. Assess reliability, relevance and consistency by expert judgement. 2. Select WoE Method: Quantitative method as meta-analysis using Bayesian hierarchical mixed effect model to build dose-response relationship between urinary cadmium and urinary B2-M. 3. Conclusion for WoE Assessment: Dose-effect relationship accounting for inter-study heterogeneity and for inter-individual variability of urinary cadmium and excretion of B2-M selected for the modelling
	Results	Overall 35 studies B2-M studies (LOE1) were selected as most relevant, consistent and reliable including previous meta-analysis and assessments from ATSDR, WHO. Urinary cadmium and B-2M selected for meta-analysis based on biological relevance of the biomarker for cadmium toxicity, reliability and consistency of the human database.
Integrate the evidence	Methods	Meta-analysis of the dose-effect relationship between urinary cadmium and B2-M accounting for inter-study heterogeneity and for inter-individual variability of dose and effect using Bayesian hierarchical mixed effect model. Expert judgement was used to select biologically relevant cut off values for urinary B2-M reflecting renal tubular damage (300 µg B2-M/g creatinine). Exceeding such a cut off value has been associated with an accelerated age-related decline of renal function together with increased mortality. Derive the Reference Point based on Benchmark Dose Limit (BMDL5) modelling for urinary B2-M using Hybrid approach and urinary cadmium
	Results	The meta-analysis and dose response modelling based on B2-M as a marker of tubular effect, identified an overall group based BMDL5 for a 5 percent increase of the prevalence of elevated B2-M of 4 µg Cadmium/g creatinine. This is selected as Reference Point.

1895

A.2.4.6. Example of the use of weight of evidence framework in derivation of an acceptable daily intake (ADI) for a regulated chemical

EFSA may derive an acceptable daily intake (ADI) for regulated products that are likely to be consumed by humans or animals. The ADI is an estimate of the amount of a chemical that can be consumed on a daily basis over a lifetime without health risk. Within the context of EFSA risk assessments, ADIs are derived for food and feed additives, and pesticide residues. An applicant wishing to market a regulated product is required to demonstrate its safety by providing data from relevant toxicity studies, from which an ADI can be derived. For example, the EFSA Panel on Food Additives and Nutrient Sources added to Food (ANS) uses a tiered approach with increasing data requirement in higher tiers reflecting greater potential risk (EFSA, 2012d). The data set needs to address toxicokinetics, genotoxicity, sub-chronic toxicity, chronic toxicity, carcinogenicity, reproductive toxicity and developmental toxicity. These studies may be performed in several species such as rat. mouse, dog, rabbit and each type of study per species constitutes a line of evidence (Table 6).

For each Line of Evidence (LOE), effect data from the different species may be compared qualitatively to identify the No-Observed-Adverse-Effect-Level (NOAEL) for the most sensitive species and endpoint taking into account biological relevance, reliability and consistency of the data. Alternatively, benchmark dose modelling could be performed on the results from each study using model averaging according to the guidance of EFSA scientific Committee (EFSA 2017). Quantitative comparison of the results of the BDML for each study and LOE can then be carried out to determine a BMDL for the most sensitive species and endpoint taking into account biological relevance, reliability and consistency of the data.

An ADI is finally derived by dividing the NOAEL or BDML by an appropriate uncertainty factor to account for differences in TK and TD between experimental animals and human (10-fold), and variability among humans (10-fold). An additional uncertainty factor may in some cases be applied to account for severity of the effect or deficiency in the data.

Table 7: Summary of the weight of evidence assessment for the derivation of an acceptable daily intake (ADI) for a regulated non-genotoxic chemical

Question		Derivation of an acceptable daily intake (ADI) for a regulated non-genotoxic chemical
Assemble evidence	Select evidence	In the context of a regulated compound, toxicity studies that may be used to derive an ADI are illustrated below (this list is not exhaustive). These studies may be performed in several species namely rat. mouse, dog, rabbit and each type of study per species constitute a line of evidence

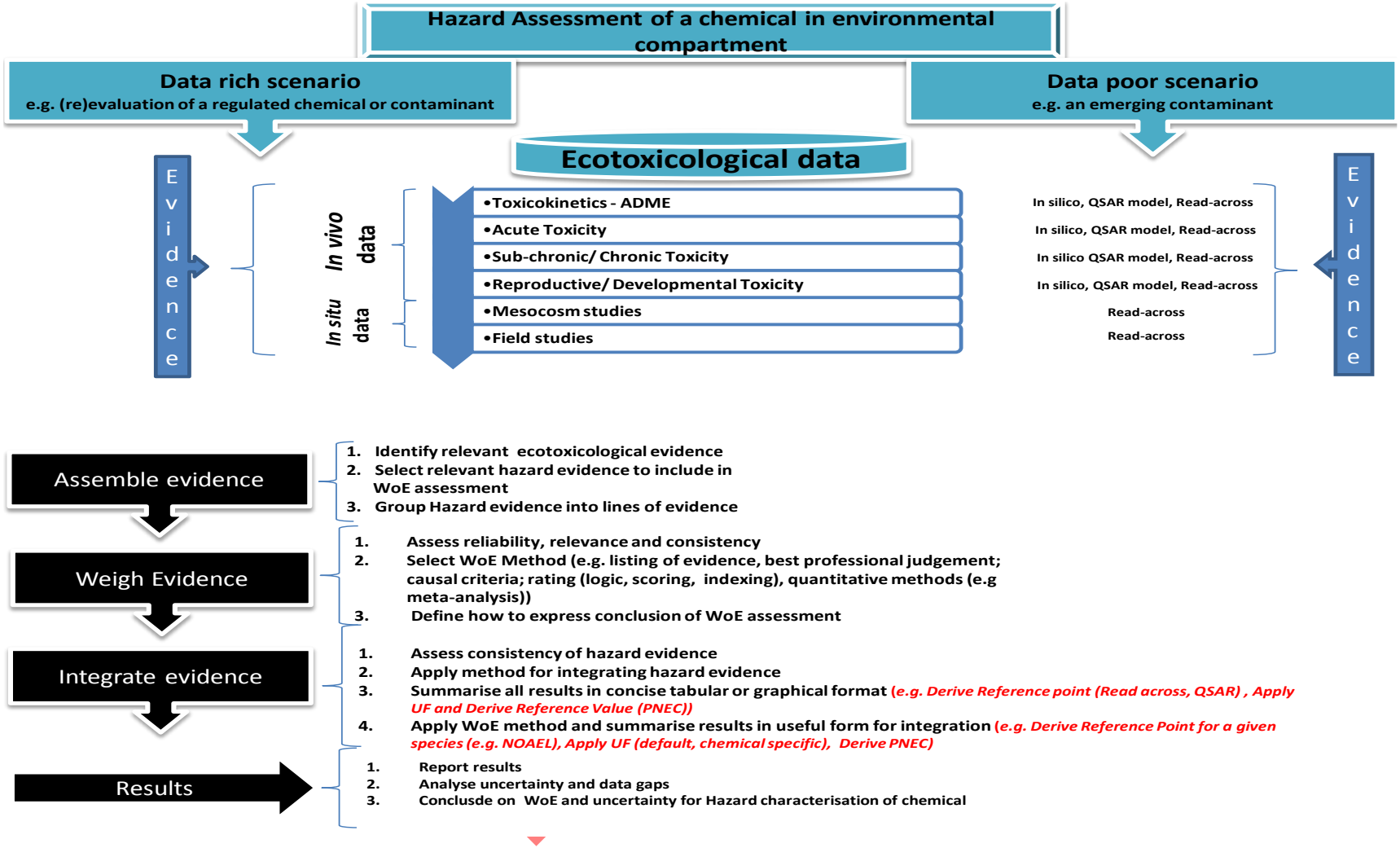
	Lines of evidence	Examples include LOE 1: Sub-chronic toxicity study 28 days, LOE 2: Sub-chronic toxicity study 90 days, LOE 3: Reproductive toxicity study LOE 4: One generation developmental toxicity study) LOE 5: Chronic toxicity studies (e.g 1 year toxicity study) LOE 6: Two year carcinogenicity study
Weigh the evidence	Methods	A number of options may be available depending on the specific assessment. Generic examples are illustrated below: 1. Qualitative comparison of each LOE per species to derive a reference point such as a No-Observed-Adverse-Effect-Level (NOAEL) for the most sensitive species and endpoint taking into account biological relevance, reliability and consistency of the data. 2. Benchmark dose modelling using model averaging according to the guidance of EFSA scientific Committee (EFSA SC, 2017). Quantitative comparison of the results of the BDML for each study and LOE to determine a BMDL for the most sensitive species and endpoint taking into account biological relevance, reliability and consistency of the data. 3. Expert judegment to assess whether an extra UF is needed for the severity of the effect or uncertainty in the DB for the NOAEL or the BMDL.
	Results	Possible outcomes include: • Reference Point (NOAEL or BDML) value for the regulated compound for the most sensitive species and endpoint NOAEL or BMDL selected as Reference Point • The quality of the studies in terms of relevance, reliability and consistency do not allow the derivation of a Reference Point. • No effects were observed at the highest dose tested and there is no need to derive a numerical reference point. • Decision as to whether an extra UF should be

		applied and how large it should be.
Integrate the evidence	Methods	Quantitative methods: combination of reference point and default uncertainty factors
	Results	A number of options may be available depending on the specific assessment. Four options are illustrated below: 1. Derivation of the ADI applying the default uncertainty factor of 100 to the NOAEL, taking into account species differences (10-fold) and human variability (10-fold). 2. Derivation of the ADI applying the default uncertainty factor of 100 to the BDML taking into account species differences (10-fold) and human variability (10-fold). 3. Derivation of the ADI applying the default uncertainty factor of 100 and an extra uncertainty factor for the severity of the effect (e.g. carcinogenicity) or uncertainty in the database to the NOAEL or BDML. 4. Derivation of the ADI is not possible since regulatory requirements are not met and additional toxicity studies are required. 5. Derivation of a numerical ADI is not needed since regulatory requirements are met and no adverse effects were observed at the highest dose tested.

A.3. Chemical risk assessment (Environment)

While environmental risk assessment of chemicals follows basically the same conceptual framework like that for human health, it is not identical. For hazard identification, all available studies are assessed, and the most sensitive are selected. Available data are integrated to derive a guidance value such as a regulatory acceptable concentration. In the risk characterization step, actual environmental exposure levels are compared with this guidance value.

Figure 7 provides a decision tree for environmental hazard characterization of a chemical in data-based and data poor situations.



1940

1941 **Figure 7: Decision tree for environmental risk assessment of a chemical in data rich and data poor situations**

A.3.1. Ecotoxicological hazard characterization of a regulated substance.

Ecotoxicity studies from the dossier: in vivo studies from the dossier (lethality, developmental, reproductive, acute, sub-chronic and chronic toxicity). Select most sensitive biologically relevant endpoints from a study or all studies when available for a specific group of animals (see note 1). Where more studies are available for the ecotoxicological characterization than the basic dossier requirements, select most sensitive biologically relevant endpoints by comparing all studies. Use default assessment factor (see note 2) when only dealing with dossier studies, or assess whether there is a need to divert from the default assessment factor on the basis of the available information for assessing the uncertainty and/or variability. Integrate ecotoxicity values and assessment values to derive environmental based guidance value (or regulatory acceptable concentration) for a regulated compound.

The Table below shows an example assessment according to the Guidance on tiered risk assessment for plant protection products for aquatic organisms in edge-of-field surface waters (EFSA, 2013)

Table 8: Summary of the weight of evidence assessment for plant protection products for aquatic organisms in edge-of-field surface waters

Question		Assessing a regulatory acceptable concentration (RAC) for fish for compound Idefix
Assemble evidence	Select evidence	Take the ecotoxicological information from the dossier
	Lines of evidence	Acute toxicity values: LOE 1: 5 day LC50 study with Oncorhynchus mykiss of 12 mg/l LOE 2: 5 day LC50 study with Lepomis macrochirus of 47 mg/l Chronic toxicity studies: LOE 1: 35 day NOEC study (ELS) with Oncorhynchus mykiss of 0.8 mg/l LOE 2: 35 day NOEC study (ELS) with Lepomis macrochirus of 0.5 mg/l Standard/default assessment factors: LOE 1: for acute situation assessment factor is 100 LOE 2: for chronic situation assessment factor is 10
Weigh the evidence	Methods	Take lowest value according to the method described in the guidance document and apply appropriate assessment factor. If this results in concern being raised, consider by expert judgement whether to use an alternative value, e.g.

		the geometric mean (EFSA 2005)
	Results	For the acute situation the lowest toxicity value is 12 mg/l For the chronic situation the lowest toxicity value will 0.5 mg/l
Integrate the evidence	Methods	Quantitative combination of a point estimate with a default value.
	Results	For the acute situation the RAC is 12/100 = 0.12 mg/l For the chronic situation the RAC is 0.5/10 = 0.05 mg/l

Note 1.

Here available studies are part of a dossier. In case of a new compound/product, the dossier will contain only the ecotoxicological studies required by legislation. This could be one or (in a few cases) two studies. Standard ecotoxicological studies are performed on behalf of, or by, the applicant, which implies that there are also standard endpoints which are in most cases considered as biologically relevant. There are also standard species tested, e.g. Daphnia magna as the standard species for crustaceans or the rainbow trout as the standard species for fish. The weight of evidence in those cases has been applied before, choice of representative species or biological relevant endpoints (Commission Regulation (EU) No 283/2013, Commission Regulation (EU) No 284/2013). The uncertainties around the hazard c.q. risk assessment are assumed to be covered by the assessment factor that is used for decision making on whether or not to allow the compound on the market or for going to higher tier risk assessment steps.

Note 2.

It is assumed that the assessment factors are only to be used for the uncertainty inherent to the ecotoxicity values and that the uncertainty around the exposure value is included in the assessment of the exposure value, for instance by using the 90th percentile of the exposure distribution.

References

EFSA 2009. Scientific Opinion of the Panel on Contaminants in the Food Chain on a request from the European Commission on cadmium in food. The EFSA Journal (2009) 980, 1–139.

Amzal B, Julin B, Vahter M, Wolk A, Johanson G, Akesson A. Environ Health Perspect. 2009 Population toxicokinetic modeling of cadmium for health risk assessment. Aug;117(8):1293-301. doi: 10.1289/ehp.0800317. Epub 2009 May 6.

EFSA Panel on Food Additives and Nutrient Sources added to Food (ANS); 2012d. Guidance for submission for food additive evaluations. EFSA Journal 2012;10(7):2760. [60 pp.] doi:10.2903/j.efsa.2012.2760.

Meek, M.E.(Bette), Boobis, A.R., Crofton, K.M., Heinemeyer, G., Raaij, M.V., Vickers, C., Risk assessment of combined exposure to multiple chemicals: A WHO/IPCS framework. 2011. Regulatory Toxicology and Pharmacology (2011), doi: 10.1016/j.yrtph.2011.03.010.

USEPA. 2007. Concepts, Methods and Data Sources for Cumulative Health Risk Assessment of Multiple Chemicals, Exposures and Effects: A Resource Document. EPA/600/R-06/013F. Washington, D.C.: SEPA. http://cfpub.epa.gov/ncea/risk/recordisplay.cfm?deid=190187.

EFSA (European Food Safety Authority), 2013.International Framework Dealing with Human Risk Assessment of Combined Exposure to Multiple Chemicals. EFSA Journal 2013;11(7):3313, 69 pp. doi:10.2903/j.efsa.2013.3313.

EFSA 2017. EFSA Scientific Committee, Hardy A, Benford D, Halldorsson T, Jeger MJ, Knutsen KH, More S, Mortensen A, Naegeli H, Noteborn H, Ockleford C, Ricci A, Rychen G, Silano V, Solecki R, Turck D, Aerts M, Bodin L, Davis A, Edler L, Gundert-Remy U, Sand S, Slob W, Bottex B, Abrahantes JC, Marques DC, Kass G and Schlatter JR, 2017. Update: EFSA EFSA 2017.Guidance on the use of the benchmark dose approach in risk assessment. EFSA Journal 2017;15(1):4658, 41 pp. doi:10.2903/j.efsa.2017.4658.

EFSA 2005. EFSA Panel on Plant Protection Products and their Residues, 2005. Opinion of the Scientific Panel on Plant protection products and their residues (PPR) on a request from EFSA related to the evaluation of pirimicarb, EFSA Journal 2005;3(8):240, 21 pp. doi:10.2903/j.efsa.2005.240.

Commission Regulation (EU) No 283/2013 of 1 March 2013 setting out the data requirements for active substances, in accordance with the Regulation (EC) No 1107/2009 of the European Parliament and of the Council concerning the placing of plant protection products on the market. OJ L 93, 3.4.2013, p. 1–84.

Commission Regulation (EU) No 284/2013 of 1 March 2013 setting out the data requirements for plant protection products, in accordance with the Regulation (EC) No 1107/2009 of the European Parliament and of the Council concerning the placing of plant protection products on the market. OJ L 93, 3.4.2013, p. 85–152.

ANNEX II. Examples of weight of evidence definitions, criteria and outputs from the scientific literature

This Annex contains examples referred to in Section 2 of the Guidance, on three topics:

Examples of definitions and descriptions of weight of evidence (Table 8).

Examples of definitions and descriptions of 'Line of evidence' from the published literature (Table 9).

Examples of criteria for weighing evidence from the published literature, mapped onto the three general concepts of reliability, relevance and consistency (Table 10).

All three tables contain examples from a selection of publications on weight of evidence assessment. They do not comprise a systematic or comprehensive review.

Some additional explanation is needed for Table 10. Many publications on weight of evidence approaches specify criteria to be considered when weighing evidence. Some of these criteria are very general, applying to virtually any context. Others are expressed in ways that are more specific to particular problem areas, e.g. the essentiality of key events (Collier et al. 2016). As may be expected, in some cases different publications use different words to express what appear to be similar criteria. Table 10 aims to show how the various criteria used in these papers relate to the three basic concepts introduced in Section 2.5 of this Guidance: reliability, relevance and consistency.

Criteria referring to the quality of studies or data, risk of bias, imprecision and sensitivity are all factors affecting the reliability of a piece of evidence in itself. They are all aspects of the way a study (for example) was conducted or reported, which affect the reliability of the resulting evidence as a correct representation of what actually occurred in the study.

Relevance concerns the relation between evidence and a purpose for which it is being used, i.e. the question it is being used to address. Confounding is a factor affecting relevance, because it concerns the possibility that an observed effect may have been caused by agents or factors other than the one of interest for the question in hand. The same applies to other criteria relating to attribution of an effect, e.g. specificity, temporality, essentiality, experimentation and randomised trials. Criteria such as spatial and temporal representativeness concern the relevance of the conditions in which evidence was generated to the conditions for the question being assessed: for example, if the assessment question refers to the EU as a whole, old data from one EU Member state is less relevant than new data from an EU-wide study (other things being equal).

Some criteria refer to dose-response and other forms of association or correlation (including spatial and temporal correlation). These criteria can contain elements of both reliability and relevance. For example, when an observed effect is large in magnitude and/or shows a consistent trend, then it is more likely to be real (reliability of the finding) and more likely to be caused by the agent of interest (relevance to the question).

Some other types of criteria can also influence both reliability and relevance. For example, the use of standard methods, the clarity and completeness of reporting and the extent of evaluation and peer review all influence judgements about both the reliability and relevance of evidence.

Many publications include criteria relating to the consistency of evidence. Consistency intrinsically includes the notions of quantity and diversity, because consistency has more weight when seen in a larger body of evidence, and/or when the evidence is of diverse types. Many publications include explicit criteria for consistency, quantity or diversity of evidence, or obviously related criteria such as coherence, reproducibility and replicability. Some of the other criteria in Table 10 refer particular types of consistency: for example, plausibility, concordance and analogy all refer to whether there is consistency between the evidence being considered and established knowledge or theory, and hence with the evidence on which the knowledge or theory is based. Similarly, 'experimental verification' refers to one piece of evidence being supported by another.

Criteria listed in the right hand column of Table 10 are of a different nature. 'Adequacy' is mentioned by several publications but refers to standards against which reliability, relevance and consistency should be judged, rather than being a separate criterion. SCENIHR (2012) propose that assessors should 'identify uncertainties in the judgement used' in the weight of evidence process: this refers to the need to consider uncertainties affecting the assessment of reliability, relevance and consistency, rather than being a separate criterion.

Table 9: Examples of definitions and descriptions of weight of evidence.

Publication	Definitions or descriptions given for weight of evidence
Agerstrand and Beronius (2015).	'In general terms, weight of evidence and systematic review are processes of summarising, synthesising and interpreting a body of evidence to draw conclusions…these processes differ from the traditional method for risk assessment by promoting the use and integration of information from all the available evidence instead of focusing on a single study'.
ANSES (2016)	Defines weight of evidence as 'the structured synthesis of lines of evidence, possibly of varying quality, to determine the extent of support for hypotheses'.

Beronius et al. (2014)	States that 'The meaning of weight of evidence intended here is the collective summary and evaluation of all existing evidence after a certain 'weight' has been attributed to individual studies, e.g. by evaluating reliability and relevance'.
Collier et al. (2016)	Describes weight of evidence as 'a term used in multiple disciplines to generally mean a family of approaches to assess multiple lines of evidence in support of (or against) a particular hypothesis, although (it) tends to be used inconsistently and vaguely across disciplines'.
ECHA (2015a) [Guidance on the Application of the CLP Criteria]	'A weight of evidence determination means that all available information bearing on the determination of hazard is considered together.'
ECHA (2015b) [Guidance on the Biocidal Products Regulation]	'A weight of evidence assessment involves the consideration of all data that is available and may be relevant to reproductive toxicity.'
ECHA(2010),	"One definition for weight of evidence is: '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'".
EFSA (2013) [PPR Aquatic Ecol RA guidance doc]	States that the 'process of combining available lines of evidence to form an integrated conclusion or risk characterisation is frequently referred to as weight-of-evidence assessment. This term reflects the principle that the contribution of each line of evidence should be considered in proportion to its weight.'
EPA (2003)	Describes weight of evidence as an 'approach (which) considers all relevant information in an integrative assessment that takes into account the kinds of evidence available, the quality and quantity of the evidence, the strengths and limitations associated with each type of evidence and explains how the various types of evidence fit together'.
Good (1979, 1985)	Defined weight of evidence as the logarithm of the ratio of the likelihood of a given hypothesis to the likelihood of an alternative hypothesis. This expression corresponds to the Bayes factor.
Hope and Clarkson (2014)	Refers to Good for quantitative definition. Describes weight of evidence as 'basically the process of considering the strengths and weaknesses of various pieces of information in order to inform a decision being made among competing alternatives'.

Hull and Swanson (2006)	Describe weight of evidence as 'approaches (that) integrate various types of data (e.g., from chemistry, bioassay, and field studies) to make an overall conclusion of risk'
Linkov et al. (2009)	Defines weight of evidence as 'a framework for synthesising individual lines of evidence, using methods that are either qualitative (examining distinguishing attributes) or quantitative (measuring aspects in terms of magnitude) to develop conclusions regarding questions concerned with the degree of impairment or risk'.
NRC (2009)	States that 'The phrase weight of evidence is used by EPA and other scientific bodies to describe the strength of the scientific inferences that can be drawn from a given body of evidence.'
Rhomberg et al. (2013)	Defines 'weight of evidence Framework' as 'approaches that have been developed for taking the process from scoping an assessment and initial identification of relevant studies through the drawing of appropriate conclusions'.
Schleier et al. (2015)	Describe weight of evidence as 'approaches in which multiple lines of evidence can be considered when estimating risk'
Suter and Cormier (2011)	'In sum, weighing evidence is a synthetic process that combines the information content of multiple weighted pieces of evidence. The information may be dichotomous (supports or not), quantitative values (e.g., an exposure or risk estimate), qualitative properties (e.g., large, medium or small), or a model. The weights that are applied to the information may express various properties that affect its credibility or importance and the weights themselves may be qualitative or quantitative. The combining of evidence may be a simple quantitative operation (e.g., weighted averages of concentration estimates) but more often involves difficult qualitative judgments.'
Vermeire et al. (2013)	Implicit definition: 'The different and possibly contradictory information is weighted and the respective uncertainties taken into account in a weight of evidence approach'.
Weed (2005)	Identifies three characteristic uses of the term weight of evidence: metaphorical, methodological and theoretical. Does not propose a definition but recommends that authors using weight of evidence should define the term and describe their methods.
WHO (2009)	Defines weight of evidence as 'a process in which all of the evidence considered relevant for a risk assessment is evaluated and weighted'.
SCENIHR (2012), Meek et al. (2014)	Use the term weight of evidence but do not include an explicit definition or summary description.

Rooney et al. (2014) (OHAT), Morgan et al. (2016) (GRADE)	These publications do not use the term weight of evidence but rather use other terms including 'evidence synthesis' and 'evidence integration'.

2102

2103

2104 **Table 10: Examples of definitions and descriptions of 'line of evidence' from the**
2105 **published literature.**

Agerstrand and Beronius (2015).	Line of evidence used only when quoting EFSA (2013) Guidance on assessment of pesticide risks to aquatic organisms.
ANSES (2016)	Defines line of evidence as 'A set of relevant information of similar type grouped to assess a hypothesis'.
Bradford Hill (1965)	Not used.
Collier et al. (2016)	Uses line of evidence frequently. Does not provide an explicit definition, but Table 3 appears to imply that the evidence relating to the molecular initiating event, the adverse outcome, and to each key event is considered as one line of evidence in each case.
ECHA(2010),	Line of evidence not mentioned as such. Refers to weighing pieces of evidence; also refers to weight of evidence providing 'the opportunity to make use of less reliable information/studies when they are pooled together with other information'. However 'pooling' here may refer to the body of evidence as a whole rather than to creating subsets of evidence.
EFSA (2013) PPR Panel Aquatic RA Guidance	States that 'the contribution of the multiple assessment approaches (multiple lines of evidence) in reducing overall uncertainty can … be evaluated by weight of evidence in the final risk characterisation', implying that in this context an line of evidence is an 'assessment approach'.

EPA (1998)	Two page section on lines of evidence. 'The development of lines of evidence provides both a process and a framework for reaching a conclusion regarding confidence in the risk estimate.' 'Confidence in the conclusions of a risk assessment may be increased by using several lines of evidence to interpret and compare risk estimates. These lines of evidence may be derived from different sources or by different techniques relevant to adverse effects on the assessment endpoints, such as quotient estimates, modeling results, or field observational studies.' 'The phrase lines of evidence is used to de-emphasize the balancing of opposing factors based on assignment of quantitative values to reach a conclusion about a "weight" in favor of a more inclusive approach, which evaluates all available information, even evidence that may be qualitative in nature. It is important that risk assessors provide a thorough representation of all lines of evidence developed in the risk assessment rather than simply reduce their interpretation and description of the ecological effects that may result from exposure to stressors to a system of numeric calculations and results.'
EPA (2003),	Refers to use of line of evidence by USEPA (1998) guideline for ecological RA, but does not use the term further.
Hope and Clarkson (2014)	Line of evidence used extensively but no explicit definition.
Hull and Swanson (2006)	Does not define line of evidence but refers to toxicity tests and population or community survey measures as examples of lines of evidence.
Linkov et al. (2009)	Uses line of evidence several times. Does not provide own definition but when reviewing the USEPA (1998) guidance for ecological RA says that 'a weight of evidence evaluation treats each assessment and measurement endpoint as an individual line of evidence'.
Meek et al. (2014)	Line of evidence not used.
Morgan et al. (2016) (GRADE)	Line of evidence not used.
NRC (2009)	Not used.
Rhomberg et al. (2013)	Line of evidence used several times but no definition found

Rhomberg et al. (2015)	Line of evidence used in one place without definition. Also uses 'lines of argument (or hypotheses)'.
Rooney et al. (2014) (OHAT)	Line of evidence not used. Text refers to studies and body of evidence.
SCENIHR (2012)	Uses 'lines of evidence' throughout. Does not provide an explicit definition but page 9 gives lists of lines of evidence, e.g. for hazard assessment the list comprises epidemiology studies, human volunteer studies, other human data, animal studies, in vitro studies, in silico studies, mathematical modelling, and mechanistic/mode of action studies; while for exposure assessment the list comprises exposure measurements, mathematical modelling and toxicokinetics.
Schleier et al 2015	Do not define line of evidence but refer to 'integrating multiple lines of evidence from different study types'
Suter and Cormier (2011)	Uses 'Categories of evidence' in similar manner to line of evidence. Does not provide an explicit definition for Categories. Includes 'lines of evidence' in list of keywords and uses it several times but does not explain how it relates to 'Categories'. Also uses 'body of evidence' which is defined in their section 3.3.1 as all of the weighted categories of evidence for a hypothesis, but is also used in the first part of their section 3 to 10 studies of the same type, i.e. 'body' also can refer to multiple pieces of evidence in a single category.
Vermeire et al. (2013)	Line of evidence not used.
Weed (2005)	Line of evidence used once, when quoting a US EPA document.
WHO (2009)	Line of evidence not in glossary or cumulative index.

2106

Table 11: Examples of criteria for weighing evidence from the published literature, mapped onto the three general concepts of reliability, relevance and consistency.

Publication	Reliability	Relevance	Combination of Reliability & Relevance	Consistency	Other
Bradford Hill (1965)		Temporality Experimentation Specificity	Strength of association Biological gradient	Consistency of association Biological plausibility Coherence Analogy	
Collier et al. (2016)	Uncertainty & variability (treatment of)	Applicability & utility Essentiality of key events	Soundness Evaluation and peer review (extent of) Clarity & completeness (of reporting)	Consistency Biological concordance Concordance of empirical observations among key events Analogy (to other chemicals)	
ECHA(2010)	Reliability	Relevance		Quantity (in particular if contradictory info is present)	Adequacy for classification and RA

Publication	Reliability	Relevance	Combination of Reliability & Relevance	Consistency	Other
EPA (1998)	Adequacy and quality of data Degree and type of uncertainty associated with the evidence	Relationship of the evidence to the risk assessment questions			
EPA (2003)	Uncertainty & variability (treatment of)	Applicability & utility	Soundness Clarity & completeness (of reporting) Evaluation and peer review (extent of)		
Hope and Clarkson (2014)	Study quality	Site specificity Spatial representativeness Temporal representativeness Specificity to stressor	Use of standard methods Endpoint/attribute association Exposure/response function Sensitivity to stressor Quantification of response		

Publication	Reliability	Relevance	Combination of Reliability & Relevance	Consistency	Other
Hull and Swanson (2006)		Specificity of cause	Magnitude Biological gradient/strength Uncertainty Spatial correlation Temporal correlation	Plausibility: mechanism Plausibility: stressor Consistency of association Experimental verification	
Meek et al. (2014)		Essentiality of key events		Consistency Biological concordance Concordance of empirical observations among key events Analogy (to other chemicals)	

Publication	Reliability	Relevance	Combination of Reliability & Relevance	Consistency	Other
Morgan et al. (2016) (GRADE)	Risk of bias Imprecision Publication bias	Indirectness Confounding Study design (randomised or observational)	Effect size Dose response	Inconsistency	
Rhomberg et al. (2013)	Study design Bias/chance Reliability Statistical methods Strength of association Internal consistency	Confounders Temporality Relevance	Strengths & weaknesses Dose response Predictivity	Replicability (if observed) Biological plausibility	Adequacy (Also recommends Klimisch, ToxRTool, AMSTAR, and GRADE criteria)
Rooney et al. (2014) (OHAT)	Risk of bias (15 sub-questions) Imprecision Publication bias Rare outcomes	Indirectness Residual confounding	Effect magnitude Dose response	Consistency	'Other'

Publication	Reliability	Relevance	Combination of Reliability & Relevance	Consistency	Other
SCENIHR (2012)	Quality/validity/reliability	Relevance/potential importance The characterization of the stressor The relevance of the set of data for a particular endpoint	Utility (combining quality and relevance) Soundness and appropriateness of the methodology used The extent to which the full details of methodology are provided	The reproducibility of findings between experiments Consistency	Uncertainties in the judgement used
Suter and Cormier (2011)	Performance Statistical analysis Potential for bias	Relevance Inherent weights of study types (e.g. randomised vs. observational, field vs. lab)	Study design Reporting Strength	Number of pieces Coherence Diversity	Case-specific criteria
Vermeire et al. (2013)	Sensitivity Reliability	Relevance Specificity	Validity/predictivity		Adequacy

References

Good, I. J. 1979. "Studies in the History of Probability and Statistics. XXXVII A. M. Turing's Statistical Work in World War II." Biometrika 66 (2):393-6.

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

Gosling, J P , Andy Hart, H Owen, M David, J Li, et C MacKay. 2013. "A Bayes linear approach to weight-of-evidence risk assessment for skin allergy." Bayesian Analysis 8 (1):169-186.

Guyatt, G. H., A. D. Oxman, R. Kunz, D. Atkins, J. Brozek, G. Vist, P. Alderson, P. Glasziou, Y. Falck-Ytter, et H. J. Schunemann. 2011. "GRADE guidelines: 2. Framing the question and deciding on important outcomes." J Clin Epidemiol 64 (4):395-400. doi: 10.1016/j.jclinepi.2010.09.012.

Guyatt, G. H., A. D. Oxman, Regina Kunz, Jan Brozek, Pablo Alonso-Coello, David Rind, P. J. Devereaux, V. M. Montori, Bo Freyschuss, Gunn Vist, Roman Jaeschke, J. W. Williams, Jr., M. H. Murad, David Sinclair, Yngve Falck-Ytter, Joerg Meerpohl, Craig Whittington, Kristian Thorlund, Jeff Andrews, et H. J. Schünemann. 2011. "GRADE guidelines 6. Rating the quality of evidence - imprecision." Journal of Clinical Epidemiology 64 (12):1283-1293. doi: 10.1016/j.jclinepi.2011.01.012.

Guyatt, G. H., A. D. Oxman, Regina Kunz, James Woodcock, Jan Brozek, Mark Helfand, Pablo Alonso-Coello, Paul Glasziou, Roman Jaeschke, E. A. Akl, Susan Norris, Gunn Vist, Philipp Dahm, V. K. Shukla, Julian Higgins, Yngve Falck-Ytter, et H. J. Schünemann. 2011. "GRADE guidelines: 7. Rating the quality of evidence - inconsistency." Journal of Clinical Epidemiology 64 (12):1294-1302. doi: 10.1016/j.jclinepi.2011.03.017.

Guyatt, G. H., A. D. Oxman, Victor Montori, Gunn Vist, Regina Kunz, Jan Brozek, Pablo Alonso-Coello, Ben Djulbegovic, David Atkins, Yngve Falck-Ytter, J. W. Williams, Jr., Joerg Meerpohl, S. L. Norris, E. A. Akl, et H. J. Schünemann. 2011. "GRADE guidelines: 5. Rating the quality of evidence - publication bias." Journal of Clinical Epidemiology 64 (12):1277-1282. doi: 10.1016/j.jclinepi.2011.01.011.

Guyatt, G. H., A. D. Oxman, Shahnaz Sultan, Paul Glasziou, E. A. Akl, Pablo Alonso-Coello, David Atkins, Regina Kunz, Jan Brozek, Victor Montori, Roman Jaeschke, David Rind, Philipp Dahm, Joerg Meerpohl, Gunn Vist, Elise Berliner, Susan Norris, Yngve Falck-Ytter, M. H. Murad, et H. J. Schünemann. 2011. "GRADE guidelines: 9. Rating up the quality of evidence." Journal of Clinical Epidemiology 64 (12):1311-1316. doi: 10.1016/j.jclinepi.2011.06.004.

Guyatt, G. H., A. D. Oxman, Gunn Vist, Regina Kunz, Jan Brozek, Pablo Alonso-Coello, Victor Montori, E. A. Akl, Ben Djulbegovic, Yngve Falck-Ytter, S. L. Norris, John Williams, Jr, David Atkins, Joerg Meerpohl, et H. J. Schünemann. 2011. "GRADE guidelines: 4. Rating the quality of evidence - study limitations (risk of bias)." Journal of Clinical Epidemiology 64 (4):407-415. doi: 10.1016/j.jclinepi.2010.07.017.

Guyatt, Gordon H., Andrew D. Oxman, Regina Kunz, James Woodcock, Jan Brozek, Mark Helfand, Pablo Alonso-Coello, Yngve Falck-Ytter, Roman Jaeschke, Gunn Vist, Elie A. Akl, Piet N. Post, Susan Norris, Joerg Meerpohl, Vijay K. Shukla, Mona Nasser, et Holger J.

2154 Schünemann. 2011. "GRADE guidelines: 8. Rating the quality of evidence - indirectness."
2155 Journal of Clinical Epidemiology 64 (12):1303-1310. doi: 10.1016/j.jclinepi.2011.04.014.

2156 Guyatt, Gordon, A. D. Oxman, Shahnaz Sultan, Jan Brozek, Paul Glasziou, Pablo Alonso-
2157 Coello, David Atkins, Regina Kunz, Victor Montori, Roman Jaeschke, David Rind, Philipp
2158 Dahm, E. A. Akl, Joerg Meerpohl, Gunn Vist, Elise Berliner, Susan Norris, Yngve Falck-
2159 Ytter, et H. J. Schünemann. "GRADE guidelines: 11. Making an overall rating of
2160 confidence in effect estimates for a single outcome and for all outcomes." Journal of
2161 Clinical Epidemiolog 66 (2):151-157.

2162 Hill, Austin Bradford. 1965. "The Environment and Disease: Association or Causation?"
2163 Proceedings of the Royal Society of Medicine 58 (5):295-300.

2164 Hope, B. K., et J. R. Clarkson. 2014. "A Strategy for Using Weight-of-Evidence Methods
2165 in Ecological Risk Assessments." Human and Ecological Risk Assessment 20 (2):290-315.
2166 doi: 10.1080/10807039.2013.781849.

2167 Klimisch, H. J., M. Andreae, et U. Tillmann. 1997. "A systematic approach for evaluating
2168 the quality of experimental toxicological and ecotoxicological data." Regul Toxicol
2169 Pharmacol 25 (1):1¬5. doi: 10.1006/rtph.1996.1076.

2170 Linkov, I., D. Loney, S. Cormier, F. K. Satterstrom, et T. Bridges. 2009. "Weight-of-
2171 evidence evaluation in environmental assessment: review of qualitative and quantitative
2172 approaches." Sci Total Environ 407 (19):5199-205. doi:
2173 10.1016/j.scitotenv.2009.05.004.

2174 Meek, M. E., A. Boobis, I. Cote, V. Dellarco, G. Fotakis, S. Munn, J. Seed, et C. Vickers.
2175 2014. "New developments in the evolution and application of the WHO/IPCS framework
2176 on mode of action/species concordance analysis." Journal of Applied Toxicology 34 (1):1-
2177 18.

2178 Meek, M. E., C. M. Palermo, A. N. Bachman, C. M. North, et R. Jeffrey Lewis. 2014.
2179 "Mode of action human relevance (species concordance) framework: Evolution of the
2180 Bradford Hill considerations and comparative analysis of weight of evidence." J Appl
2181 Toxicol 34 (6):595-606. doi: 10.1002/jat.2984.

2182 OHAT. 2015. Handbook for Conducting a Literature-Based Health Assessment Using
2183 OHAT Approach for Systematic Review and Evidence Integration. Research Triangle Park,
2184 NC: OHAT.

2185 Rhomberg, L. R., J. E. Goodman, L. A. Bailey, R. L. Prueitt, N. B. Beck, C. Bevan, M.
2186 Honeycutt, N. E. Kaminski, G. Paoli, L. H. Pottenger, R. W. Scherer, K. C. Wise, et R. A.
2187 Becker. 2013. "A survey of frameworks for best practices in weight-of-evidence
2188 analyses." Crit Rev Toxicol 43 (9):753-84. doi: 10.3109/10408444.2013.832727.

2189 Rhomberg, Lorenz. 2015. "Hypothesis-Based Weight of Evidence: An Approach to
2190 Assessing Causation and its Application to Regulatory Toxicology." Risk Analysis 35
2191 (6):1114-1124. doi: 10.1111/risa.12206.

2192 Rooney, A A, A L Boyles, M S Wolfe, Bucher J R, et Thayer KA. 2014. "Systematic review
2193 and evidence integration for literature-based environmental health science assessments."
2194 Environmental Health Perspectives (122):711-718. doi:
2195 http://dx.doi.org/10.1289/ehp.1307972.

SCENIHR. 2012. Memorandum on the use of the scientific literature for human health risk assessment purposes – weighing of evidence and expression of uncertainty. Brussels: European Commission.

Schleier Iii, J. J., L. A. Marshall, R. S. Davis, et R. K. Peterson. 2015. "A quantitative approach for integrating multiple lines of evidence for the evaluation of environmental health risks." PeerJ 3:e730. doi: 10.7717/peerj.730.

US EPA. 1998. Guidelines for Ecological Risk Assessment. Washington D.C: U.S. Environmental Protection Agency.

US EPA. 2003. A summary of general assessment factors for evaluating the quality of scientific and technical information. Washington, DC: U.S. Environmental Protection Agency.

Annex III. Examples of the application of approaches for assessing weight of evidence in different areas of work of EFSA

C.1. NDA Panel example

Summary of the weight of evidence assessment for the Scientific Opinion on the substantiation of a health claim related to vitamin D and risk of falling pursuant to Article 14 of Regulation (EC) No 1924/2006

Background information

Information about how the evidence is weighted in scientific assessments for health claims other than those related to well-established functions of essential nutrients can be extracted from existing guidance documents to applicants (EFSA NDA Panel, 2016a and b). For nutrition claims, and for health claims related to well-established functions of essential nutrients, the scientific evidence is generally not weighed.

The main question to be addressed is always the same. Is a health claim related to a specific food/constituent[5] and to a specific health effect scientifically substantiated? This main question can be broken down into three sub-questions, namely:

1. Is the food/constituent sufficiently characterised?
2. Is the claimed effect a beneficial physiological effect (relevant to human health) for the target population and can it be measured *in vivo* in humans?)
3. Is there a cause and effect relationship between the consumption of the food/constituent and the claimed effect (for the target population, under the proposed conditions of use)?

A negative answer to any of the above-mentioned questions could stop the scientific evaluation by the NDA Panel, leading to a negative conclusion (i.e. the proposed health claim is not scientifically substantiated).

If a cause and effect relationship is considered to be established, an additional question should be addressed in order to establish the conditions of use for the claim:

4. What is the (lowest) effective dose and the pattern of consumption required to obtain the claimed effect?

Human studies are needed for the scientific substantiation of health claims. Whereas animal and *in vitro* studies can be used to assess the biological plausibility of the effect, they alone cannot substantiate a health claim, and thus guidance documents contain little information on how to appraise these studies individually or weigh them within the totality of the evidence.

[5] Food/constituent means a food category, a food or a food constituent (e.g. a nutrient or other substance, or a fixed combination of nutrients/other substances)

Each relationship between a food/constituent and a claimed effect is assessed by the NDA Panel separately on a case by case basis for specific claim applications. Pertinent human studies are an absolute requirement for the scientific substantiation of health claims, and pertinent human efficacy studies are at the top of the hierarchy that informs decisions on substantiation. However, there is no pre-established rule as to how many or which types of studies are needed for substantiation. The reproducibility of the effect of the food/constituent, as indicated by the consistency of the findings (within and across studies), and the biological plausibility of the effect are also considered.

A hierarchy of evidence for substantiation is given as follows (in decreasing order of importance):

a) Human intervention studies.

b) Human observational studies: prospective cohort studies, nested-case control or case-cohort studies, cross-sectional studies, ecological studies.

c) Summary studies (systematic reviews, meta-analysis).

There are a series of questions to be considered for each type of human study to decide on whether to include them or not among the totality of evidence which will be pertinent to the claim (i.e. on whether or not they will be considered in the weighing of the evidence). For human intervention studies, aspects related to their relevance, such as how the intervention relates to the food/constituent that is the subject of the health claim, how the study population relates to the target population for the claim, and how the outcome variables relate to the claimed effect, are considered first. Relevant (pertinent) human intervention studies are then assessed in relation to their reliability (risk of bias) by carefully considering aspects such as randomisation, appropriateness of the control group, the use of a placebo, blinding, whether the duration of the intervention is sufficient to observe the expected changes, and whether the statistical analysis of data is appropriate. For observational studies, appropriate control for confounders is an important aspect. No scoring system is in place, however, to rate the overall risk of bias of individual studies.

If a summary publication (including systematic reviews and meta-analysis) is provided for the scientific substantiation of a claim, the Panel reviews the primary data to ensure that all the individual studies included are relevant (pertinent) to the claim. Meta-analysis can provide information about the reproducibility and consistency of the effect across studies and study groups, about the dose–response relationship, and about the minimum effective dose of the food/constituent which is required to obtain the claimed effect (i.e. to establish conditions of use). The NDA Panel, however, has not relied so far on the results of meta-analyses to make a scientific judgement on whether a cause and effect relationship between the consumption of the food/constituent and the claimed effect has been established (EFSA NDA Panel, 2016a).

Health claim related to vitamin D and risk of falling

This is a health claim pursuant to Article 14 of Regulation (EC) No 1924/2006 and falls under the scope of disease risk reduction.

In order to assess the scientific substantiation of a health claim related to vitamin D and risk of falling, the Panel considered the following in relation to the first two questions (EFSA NDA Panel, 2011):

1. The food/constituent proposed by the applicant, vitamin D (D_2 and D_3), was sufficiently characterised.
2. The claimed effect was "reduces the risk of falling. Falling is a risk factor for fractures", and the proposed target population for the claim was men and women 60 years of age and older. Risk of falling is an established risk factor for fractures in the target population, and can be assessed in human studies as the number of falls per person per observation time (incidence), the total number of falls and/or the number of subjects falling at least once).

To complete the assessment of the scientific substantiation of the claim, two main questions remain to be assessed:

3. Is there a cause and effect relationship between the consumption of Vitamin D and the risk of falling in men and women 60 years of age and older?

If so,

4. What is the (lowest (4a)) dose of Vitamin D and the pattern of consumption required to reduce the risk of falling in men and women 60 years and older (4b)?[6]

Tables 11 and 12 below summarise how the NDA Panel weighed and integrated the evidence in order to answer questions 3 and 4, respectively. Different sections of the scientific opinion (EFSA NDA Panel, 2011) are cross-referenced.

[6] It should be noted that question 4 is conditional to question 3 (i.e. it is only addressed if the answer to question 3 is positive)

Table 12: Summary of how the NDA Panel weighed and integrated the evidence in order to answer question 3.

Question		Is there a cause and effect relationship between the consumption of vitamin D and the risk of falling in men and women 60 years of age and older?
Assemble evidence	Select evidence	The opinion is based on the evidence provided by the applicant. Details on the literature search and inclusion/exclusion criteria applied by the applicant are given in the first paragraph of section 3 of the opinion. The Panel selected the studies/meta-analysis relevant (pertinent) to the claim by considering whether they were designed to address the specific question (see second paragraph of section 3 of the opinion and first paragraph under "*randomized controlled trials*").
	Lines of evidence	**LOE 1.** Six RCTs investigating the effects of vitamin D supplementation on the risk of falling in the target population. **LOE 2.** Five observational studies investigating the association between vitamin D supplementation and/or vitamin D status (as a surrogate marker of total vitamin D intake) and risk of falling in the target population. **LOE 3.** Data (from RCTs and observational studies) and background expert knowledge on the mechanisms by which vitamin D could reduce the risk of falling (biological plausibility of the effect).
Weigh the evidence	Methods	**LOE 1.** Narrative based on expert discussion **LOE 2.** Narrative based on expert discussion **LOE 3.** Narrative based on expert discussion
	Results	**LOE 1.** Five RCTs showed an effect of vitamin D on the risk of falling in the target population at daily doses of 800-1000 I.U. (20-25 μg); one four-arm study using vitamin D doses of 200-800 I.U. (5-20 μg) did not show an effect, but it might have been underpowered for that outcome (see section 3 of the opinion, penultimate paragraph under "*randomized controlled trials*" **LOE 2**. Results from the observational studies provided were inconsistent; residual confounding could not be excluded (see section 3 of the opinion under "*observational studies*") **LOE 3.** Given the well-established role of vitamin D on muscle function, it is biologically plausible (but still to be established) that vitamin D supplementation could improve muscle strength, physical performance and body balance in the target population (see section 3 of the opinion, first paragraph under "*mechanisms of action*")
Integrate the evidence	Methods	**LOE 2** was dismissed, rather than integrated with **Line 1**. Integration of **LOEs 1** and **3** was done by expert discussion as explained in the last two paragraphs but one of section 3 in the opinion.
	Results	The Panel concludes that a cause and effect relationship has been established between the intake of vitamin D and a reduction in the risk of falling (section 3 of the opinion, last paragraph) in the target population for the claim, which is men and women 60 years of age and older (section 5 of the opinion).

Table 13: Summary of how the NDA Panel weighed and integrated the evidence in order to answer question 4.

Questions		What is the (lowest (4a)) dose of Vitamin D and the pattern of consumption required to reduce the risk of falling in men and women 60 years and older (4b)?
Assemble evidence	Select evidence	The applicant provided a meta-analysis of eight RCTs which aimed to investigate the efficacy of supplemental vitamin D with or without calcium in preventing falls among older individuals. Two of the RCTs, however, were not considered pertinent to the claim, and thus secondary analyses were conducted to assess the dose-response relationship and the risk of publication bias (see section 3 of the opinion under "*meta-analysis of randomized controlled trials*").
	Lines of evidence	**LOE 1.** Six RCTs pertinent to the claim **LOE 2.** Funnel plot for risk of publication bias analysis
Weigh the evidence	Methods	**LOE 1.** Meta-analysis of the six RCT pertinent to the claim **LOE 2.** Assessment of the risk of publication bias
	Results	**LOE 1.** No conclusions can be drawn from the meta-analysis with respect to the effect of vitamin D supplementation on the risk of falling at doses of 200-600 I.U./day (5-15 µg/day). No dose-response can be identified. The meta-analysis consistently shows, however, a significant effect of vitamin D supplementation on the risk of falling (RR=0.83; 95 % CI: 0.75-0.92) at doses of 800-1000 I.U./day (20-25 g/day). **LOE 2.** No significant publication bias was identified. The risk of bias was quantified.
Integrate the evidence	Methods	The Copas model was used to adjust the effect of vitamin D supplementation at doses of 800-1000 I.U./day (20-25 g/day) on the risk of falling for possible publication bias (RR=0.85; 95 % CI: 0.75-0.96).
	Results	The available data do not provide information about the lowest effective dose of vitamin D needed to obtain the claimed effect (**sub-question 4a**). In order to obtain the claimed effect, 800 I.U. (20 µg) of vitamin D from all sources should be consumed daily (**sub-question 4b**) (see section 3 of the opinion under "*meta-analysis of randomized controlled trials*", last paragraph; see also section 5 of the opinion on conditions and restrictions of use)

References

EFSA Panel on Dietetic Products, Nutrition and Allergies (NDA); Scientific Opinion on the substantiation of a health claim related to vitamin D and risk of falling pursuant to Article 14 of Regulation (EC) No 1924/2006. EFSA Journal 2011;9(9):2382. [18 pp.]. doi:10.2903/j.efsa.2011.2382.

EFSA NDA Panel (EFSA Panel on Dietetic Products, Nutrition and Allergies), 2016a. General scientific guidance for stakeholders on health claim applications. EFSA Journal 2016;14(1):4367, 38 pp. doi:10.2903/j.efsa.2016.4367

EFSA NDA Panel (EFSA Panel on Dietetic Products, Nutrition and Allergies), 2016b. Turck D, Bresson J-L,Burlingame B, Dean T, Fairweather-Tait

S, Heinonen M, Hirsch-Ernst KI, Mangelsdorf I, McArdle HJ, Naska A, Neuhäuser-Berthold M, Nowicka G, Pentieva K, Sanz Y, Sjödin A, Stern M, Tomé D, Van Loveren H, Vinceti M, Willatts P, Martin A, Strain JJ,Heng L, Valtueña Martínez S and Siani A, 2017. Scientific and technical guidance for the preparation and presentation of a health claim application (Revision 2). EFSA Journal 2017;15(1):4680, 31 pp. doi:10.2903/j.efsa.2017.4680

C.2. PPR Panel

Summary of the weight of evidence assessment for the substantiation of pesticidial active substances to be included in Cumulative Assessment Groups (CAGs)

Background

In 2013 the PPR Panel developed a methodology to identify pesticidal active substances to be included in Cumulative Assessment Groups (CAGs). It was assumed that compounds belonging to the same CAG can be treated in cumulative risk assessment as if they were simple dilution of one other and will follow the dilution principles of dose addition (EFSA, 2013).

The methodology was intended to address cumulative effects in relation to maximum residue limits (MRLs) setting. Four levels of grouping were proposed, each indicating a refinement step in the methodology.

CAG level 1: common toxicological target organ

CAG level 2: common specific phenomenological effect

CAG level 3: common mode of action (if available)

CAG level 4: common mechanism of action (if available).

Since information on mode and mechanism of action is frequently not available, refinement to CAG level 3 and 4 was inconclusive for most of the CAGs and the induction of the same phenomenological effect was deemed sufficient for accepting similar action and therefore justifies dose addition. The methodology for grouping substances in CAG level 2 involves the identification and characterisation of the specific effect. Identification of the specific effect was based on the following criteria: exclusion of local effect, exclusion of non-adverse effects, exclusion of effects not relevant to humans, evaluation of unambiguous nature of the effect, identification of non-specific effect. The characterisation of the specific effect is described by supporting indicators e.g. histological, biochemical or clinical indicators. The methodology developed by the PPR Panel was substantiated by expert knowledge (EFSA, 2013).

Following establishment of the criteria, a data collection including authorized and non-authorized substances was performed by collecting data from the dossiers submitted for the authorization procedure and the active substances matching with the established criteria were included in the CAGs. These CAGs were defined at Level 2, and often contain large numbers of substances.

Assuming that all the substances in a Level 2 CAG combine by dose addition may lead to a large overestimation of risk if some or many of them do not, in fact, share the same mechanism of action. Therefore it is relevant for the risk assessors and the risk managers to consider the level of confidence or certainty that a given substance belongs to a CAG at Levels 3 or 4 and consequently contributes to the cumulative risk assessment (CRA) by dose addition.

Example

The following example is proposing a weight of evidence approach using a pre-established CAG level 1 Nervous System and CAG level 2 Autonomic response (acute). The methodology also includes the selection of a Reference Compound which was selected on the consistency and robustness of the database (e.g. dose related effect, consistently observed across the studies, known pesticidal mode of action, known toxic mode of action). Four lines of evidence were identified for assessing whether each substance should qualify for combining with the reference compound by dose addition, and criteria were defined for weighing each line of evidence on scales expressed as 0/+ or 0/+/++ (see below for details). The lines of evidence were then integrated by expert judgement, expressing the conclusion for each substance in terms of the probability that it qualified for dose addition, expressed on a scale of 0-100%. This can then be used in cumulative risk assessment to take account of the confidence that each substance should be included or excluded in dose addition, using probability theory, whereas there would be no such theoretical basis if the conclusion was expressed qualitatively (e.g. as the number of + scores).

This approach requires that the question addressed by the probability for each substance should be well-defined. It was considered that dose addition could be assumed for a given substance (Y) if it causes a significant effect on the autonomic nervous system (CAG level 2) and has a key event (e.g. biochemical effect) that is (a) in common with substance **X** (the reference compound) and (b) has a causal relation to the adverse outcome (AO). The approach to the weight of evidence assessment for this question is summarised in Table 13, and results for a selection of substances are shown in Table 14. As a final step, the elicited probability was assigned to one of the categories on a probability scale suggested in the draft EFSA Guidance on Uncertainty (see Table 15).

Table 14: Summary of proposed approach.

Question	Does substance **Y** cause a significant effect on autonomic (CAG level 2) and have a key event (e.g. biochemical effect) that is (a) in common with

		substance **X** (the reference compound) and (b) has a causal relation to the AO and therefore justifies dose addition? (This question is assessed separately for each substance Y in the Level 2 CAG, except for the reference substance X).
Assemble evidence	Select evidence	Data were collected from the authorization dossiers and draft assessment reports (DAR) using criteria defined in EFSA 2013. Data collection is available on EFSA 2012 and 2015.
	Lines of evidence	LOE 1: Specificity of the effect/Dose Relationship LOE 2: Clinical observation LOE 3: Biochemical observation LOE 4: MoA Note that Lines 1 and 2 relate to CAG Level 2, whereas LOEs 3 and 4 relate to CAG Level 3 or 4.
Weigh the evidence	Methods	LOE 1: 0 No dose relationship, + Effect observed at the high dose only, ++ Effect showing a dose relationship LOE 2: 0 No, + Yes LOE 3: 0 Not observed, + biochemical read-out observed at the high dose only, ++ Biochemical read-out showing a dose relationship LOE 4: 0 Not established, + presumed, ++ known
	Results	See Table 2
Integrate the evidence	Methods	Expert knowledge elicitation (EKE) was the selected methodology to integrate the evidence and express the conclusion as a probability. The methodology followed the principles of the EFSA (2014) EKE guidance, modified for application to the case in hand (a binary question) and streamlined to be practical for application to multiple substances.
	Results	The conclusion was expressed as the probability that substance Y causes an autonomic effect and has a shared KE with the reference substance. This can be used in a cumulative risk model to take account of the degree of confidence that Y and X should be combined by dose addition.

Table 15: Example of preliminary results. CAG: Autonomic division, Acute

Active substance	Indicator of specific effect	NO(A)EL	LO(A)EL	Mode/ mechanism of action	Study	Lines of evidence					
		mg/kg bw	mg/kg bw			Is the effect specific and therefore dose related? 0=not specific +=only observed at high dose ++=dose related	Is the effect defined at clinical level? 0=No, +=Yes	Is the effect defined at biochemical level? 0=No, +=Yes, ++= Yes, dose related	Is the effect supported by a mechanism of action? 0=No +=presume ++=Yes	Expert Knowledge Elicitation	Probability scale
Oxamyl (reference compound)	Salivation, urination	0.1	0.75	**Known,** inhibition of AChE	Acute neurotoxicity rat (Malley, 1997)	++	+	++	++	100%	Extremely Likely
Acetamiprid	Urination	10	30	**Known**, Nicotinic acetylcholine receptor (nAChR)	Acute neurotoxicity rat (Hughes, 1997a)	++	+	0	++	50%	As likely as not

				agonist							
Beta-Cyfluthrin	Salivation	2	10 (only observed at high dose)	**Presumed**, Type II (α-cyano) pyrethroid	Acute neurotoxicity rat (Sheets, Gilmore & Hamilton, 1997)	+	+	0	+	33%	Unlikely
Dicofol	Lacrimation salivation	25	250	Unknown	Acute neurotoxicity rat (Krzywicki, K., Bonin, R. 1985)	++	+	0	0	40%	As likely as not
Tebuconazole	Salivation	250	500 (only observed at high dose)	Unknown	Acute neurotoxicity rat (Sheets, 1997)	+	+	0	0	30%	Unlikely

Table 16:Probabillity scale suggested by EFSA (2016).

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%

References

EFSA 2013 (Scientific Opinion on the identification of pesticides to be included in cumulative assessment groups on the basis of their toxicological profile, EFSA Journal 2013;11(7):3293).

EFSA 2014 (Guidance on Expert Knowledge Elicitation in Food and Feed Safety Risk Assessment. EFSA Journal 2014;12(6):3734).

EFSA 2012; Grant agreement CFT/EFSA/PRAS/2012/07 awarded to the consortium ANSES/ICPS/RIVM: AS screened for neurotoxicity, liver toxicity, repro&development.

EFSA 015; Grant agreement GP/EFSA/PRAS/2013/02 awarded to the consortium ANSES/ICPS/RIVM: active substances screened for effects on the nervous system, thyroid, liver, repro&development, adrenals, eye.

EFSA 2016. Guidance on uncertainty in EFSA scientific assessment. Revised draft for internal testing. https://www.efsa.europa.eu/en/topics/topic/uncertainty-assessment