TOX/2021/XX

COMMITTEE ON TOXICITY OF CHEMICALS IN FOOD, CONSUMER PRODUCTS AND THE ENVIRONMENT

COMMITTEE ON CARCINOGENICITY OF CHEMICALS IN FOOD, CONSUMER PRODUCTS AND THE ENVIRONMENT

Draft report on the synthesis and integration of epidemiological and toxicological evidence in risk assessments

Introduction

The Committees on Toxicity and Carcinogenicity (COT and COC) published a joint report on synthesising epidemiological evidence (SEES) in 2019. During their meetings the subgroup also discussed the approaches on the synthesis and integration of epidemiological and toxicological evidence and recognised that current approaches in risk assessment usually consider epidemiological evidence separately from toxicological evidence. Guidance on the integration of the two evidence streams is scarce.

Hence, a joint subgroup of the COT and the COC, that also included two members from Public Health England (PHE) for their specific expertise and relevant work with the independent expert advisory Committee on the Medical Effects of Air Pollutants (COMEAP) was formed in November 2019.

The aim of the Synthesis and Integration of Epidemiological and Toxicological Evidence (SETE) Subgroup was to report in a transparent fashion on the approaches taken by the Committees and to give (applicable) guidance on how to integrate the two evidence streams.

This report (Appendix 1) provides the considerations and deliberations of the SETE subgroup, including a practical and directly applicable guidance document to evidence integration in Annex 1.

Annex 2 of the report aims to provide practical examples applying the procedures for the integration of evidence (Section 4) and SETE guidance (Annex 1). However, due to time restraints the examples have not been finalised yet.

Questions to the Committee

i. Does the Committee agree that the section on assessing epidemiological evidence accurately reflects the approaches taken by the Committees?

ii. Does the Committee agree that the section on assessing toxicological evidence accurately reflects the approaches taken by the Committees?

iii. Does the Committee agree that the guidance provides practical and applicably advice on how to integrate epidemiological and toxicological evidence?

iv. Does the Committee have any other comments?

Secretariat
March 2021

Report of the Synthesis and Integration of Epidemiological and Toxicological Evidence Subgroup (SETE) of the Committee on Toxicity and the Committee on Carcinogenicity

This is a draft document.
It does not reflect the views of the Committee and should not be cited.

Report of the Synthesis and Integration of Epidemiological and Toxicological Evidence Subgroup (SETE) of the Committee on Toxicity and the Committee on Carcinogenicity

Table of Contents

Executive summary

1. The Synthesis and Integration of Epidemiological and Toxicological Evidence Subgroup (SETE) of the Committee on Toxicity of Chemicals in Food, Consumer Products and the Environment (COT) and Committee on Carcinogenicity (COC) was set up in 2019. Its aim was to review the approaches for synthesising and integrating epidemiological and toxicological evidence that are used by the COT and COC in chemical risk assessments and to provide a pragmatic guidance and transparent reflection of how the COT and COC review data.

2. The SETE subgroup identified scoping and problem formulation as the first (key) step in the process of evidence synthesis. This ensures that the right questions are asked, helps make the most efficient use of resources and identifies the most appropriate approaches to use in the assessment. An established system or guidance should be followed where feasible, for example published (systematic) reviews are commonly used by Committees.

3. The principles of evaluation of epidemiological studies and synthesis of evidence are well documented in the SEES report which should be read together with this document. Ideally, the design of epidemiological studies needs close collaboration between epidemiologists and toxicologists in order to take into account available information on exposure, toxicological and mechanistic information. This prior knowledge will improve the design of human studies to ensure they provide useful and relevant information. Collaboration and ongoing dialogue between epidemiologists, exposure experts and toxicologists is therefore strongly encouraged.

4. The advantage of observational epidemiological studies are larger sample size and duration, as well as a wider range of exposure. In addition, the route, dose and pattern of exposure are usually representative of the population concern. The quality of observational studies should be evaluated individually to identify and quantify possible biases, their direction and likely impact on estimated parameters. However, this should not necessarily lead to individual studies being excluded, since such a study may still be highly informative and it is recommended that all relevant studies should be included in evidence synthesis, using a weight of evidence approach.

5. The ability of a toxicological/experimental study to predict adverse human health effects, particularly in establishing a plausible causal relationship, is critical. In some respects, studies carried out under GLP are easy to review, owing to validation and standardisation. Non-standard studies however may add valuable insights into mode of action. As *in vitro* studies become more widely used, conclusions on chemical safety can sometimes be obtained by integrating data from multiple sources. When considering the conclusions of non-validated, non-standard studies it

is important to assess the quality of the evidence, especially if a test system is far removed from human. Physiologically based pharmacokinetic modelling may provide a means of bridging the exposure gap.

6. Relevance, reliability and adequacy of toxicological studies are determined by applying a number of criteria, including but not exclusively, identification of the chemical or mixture, test species or *in vitro* model, study design, presentation of test results, statistical analysis. The assessment should be iterative and flexible, as the nature of the problem becomes better defined. Useful, structured frameworks are available as a guide, and may be used appropriately alongside scientific/expert judgement.

7. For both, epidemiological and toxicological evidence a prescriptive checklist or scoring approach is not recommended. The decision-making process should be robust, transparent, evidence-based, defensible and documented and, importantly, it should be easy to use.

8. Information on mode of action (MOA) can be invaluable for evidence integration by enabling the qualitative and quantitative bridging between experimental data and observations in humans. MOA underpins weight of evidence considerations by providing the mechanistic link between empirical observation and biological plausibility.

9. The synthesis of epidemiological, toxicological and other evidence for risk assessment purposes is an integral part of the work conducted by scientific advisory groups. The majority of guidance documents and frameworks available on the use of epidemiological and toxicological information in chemical risk assessment assess these two evidence streams separately and subsequently bring them together qualitatively, applying expert judgement as required. Building on the limited frameworks available that provide practicable and applicable guidance combining epidemiological and toxicological evidence, the SETE subgroup aims to provide information on how different evidence streams should be integrated in a transparent manner, giving appropriate weight to both.

10. All lines of evidence should be considered, with no specific hierarchy *a priori*. However, initially assessing the strength of the lines of evidence separately will provide an indication of how reliable a line of evidence is. For example, it may be that the epidemiological evidence for a given compound is considered extremely robust, whereas the evidence from *in vivo* toxicological studies is considered very weak.

11. One way to clearly depict the influence of the different lines of evidence on the conclusion on causality is via visual representation. Conclusions can be drawn

based on where the causal interference appears on a graph. This can show whether a causal relationship in humans is likely or unlikely or if there is insufficient information to reach a conclusion. The impact of the different lines of evidence is influenced by several factors, including the impact of the strength or weakness in the data, the relative weighing of epidemiological and toxicological studies and the uncertainties associated with the data. The placement of the toxicological and/or epidemiological evidence can be easily adjusted when more information is added and/or becomes available.

12. The conclusion of the assessment should be stated, with an estimate of the overall uncertainty and, where appropriate, guidance on how data gaps could be filled

13. The SETE subgroup recognised that issues on which advice from the Committees is sought varies considerably and hence the guidance proposed should be sufficiently flexible to address this.

1. Introduction

14. Synthesis of epidemiological, toxicological and other evidence for risk assessment purposes is an integral part of the work conducted by scientific advisory groups.

15. Epidemiological studies can provide direct evidence of human health impacts of specific exposures. Thus, interspecies uncertainty factors used with toxicological studies are not necessary, but additional factors may be required to account for other sources of variability. For risk assessment, human studies are preferred if available. Experimental designs (for example, randomised controlled trials (RCTs), intervention studies, natural experiments, chamber studies, food challenge) can be particularly powerful in establishing causality and estimating the dose-response. However, epidemiological studies are often observational in design (cohort, case-control, cross-sectional, descriptive). It is often claimed that a cohort study, or case-control study nested within a cohort provides the most robust evidence, but this is not always the case – there is no rigid hierarchy of study designs, and the most appropriate type of evidence is highly context and question-specific. Some of the limitations of individual studies can be overcome by triangulation across contexts and study design. Observational studies are susceptible to potential biases and confounding effects which the design and analysis of the study attempts to mitigate. A common problem encountered is uncertainty about the exposure characterisation. For risk assessment, it is also important to consider whether results are generalizable from the study population to the population for whom the risk assessment is being carried out (e.g. general population, infants and toddlers). It has always been an aim of epidemiology to estimate causal effects, and a variety of methods have been used to do this. Recently, approaches based upon formal causal inference have been developed in which the aim of an observational study is to attempt to obtain the same effect estimate as would have been obtained with the RCT (Pearl, 2009; Pearl and Mackenzie, 2018). However, epidemiologists continue to also use a variety of other methods to assess causality, e.g. by triangulation across a variety of contexts and study designs, rather than relying on one ‘ideal’ study.

16. Toxicological studies provide mechanistic and experimental evidence of potential for causal associations and can form the basis of dose-response estimation if appropriate information is not available from human studies. Toxicological studies are, in general, planned experiments designed to answer specific scientific questions. Different treatments or interventions are imposed on experimental material and potential biases are controlled by aspects of design such as randomization. Effects can then be considered to have been caused by the experimental intervention. However, toxicological studies are not always good predictors of the impact of an exposure on the whole system in humans, including where the biologic response in humans may be affected by other concurrent

exposures (e.g. lifestyle factors, diet) or influenced by variability in toxicokinetics or the microbiome.

17. Current approaches usually consider epidemiological evidence separately from toxicological evidence, guidance on the integration of the two evidence streams is scarce. Hence, the Committee on Toxicity of Chemicals in Food, Consumer Products and the Environment (COT) decided to report in a transparent fashion on the approaches taken by the Committees and to give (applicable) guidance on how to integrate the two evidence streams.

18. Integration of information derived from epidemiological and toxicological studies requires an appreciation of the different scientific processes around the two disciplines to allow for an appropriate, evidence-based conclusion regarding causality. Therefore, a joint subgroup of the COT and the Committee on Carcinogenicity of Chemicals in Food, Consumer Products and the Environment (COC), that also included two members from Public Health England (PHE) for their specific expertise and relevant work with the independent expert advisory Committee on the Medical Effects of Air Pollutants (COMEAP) was formed in November 2019. Details of the subgroup membership and Members extensive experience of UK and international scientific advisory committees are given in the final section of this document.

19. The Synthesis and Integration of Epidemiological and Toxicological Evidence Subgroup (SETE) has met, predominantly virtually, eight times from November 2019 to March 2021, with additional shorter sub-meetings to tackle specific aspects of the documents in preparation.

20. This document provides the considerations and deliberations of the SETE subgroup, while the complementary guidance document (Annex 1), provides a practical and directly applicable approach to evidence integration.

21. It is hoped this initiative will prove of use to groups beyond COT, COC and COMEAP.

Aims and Objectives

The aim of this report was to review the approaches to synthesising epidemiological and toxicological evidence and to provide information on the approaches taken on integrating these two evidence streams by the COT and COC in chemical risk assessments and to make recommendations for guidance for COT, COC and other expert advisory committees.

The objectives were:

- To review the guidance on assessing epidemiological evidence.
- To review the guidance on assessing toxicological evidence.
- To review recent practises and frameworks on epidemiological and toxicological evidence, with a focus on integrating the two evidence streams.
- To develop pragmatic guidance to integrate epidemiological and toxicological evidence with a view to improving transparency in committee conduct, while accounting for the complexity and diversity of risk assessments conducted by COT and COC and the urgency of the work.

2. Problem formulation and literature retrieval

22. Problem formulation is the first stage in the assessment and underpins the whole process. The evidence needed for a risk assessment may vary depending on the nature of the issue. However, to fully evaluate the usefulness of studies/data, the issue to be addressed must be clearly understood. This ensures the efficient use of resources and the identification of the best method(s) in a given situation.

23. One of the key principles should be the effective dialogue and collaboration between epidemiologists and toxicologists but also exposure experts and experts in other relevant areas, ensuring a shared understanding of the question(s) to be addressed and the planned outputs of the risk assessment or other advice/evidence.

24. The questions to be addressed and their scope should be discussed and agreed by all Members and the problem formulation should be clear to ensure the evidence/data/studies assessed will be appropriate and cover the relevant issues. It is important that the right question(s) are asked, with the right context explained, so the evidence is sufficiently comprehensive and targeted appropriately to address the issue.

25. A number of considerations should be applied in the problem formulation process. Committees are asked to assess a wide range of questions, including but not exclusively, full risk assessments on a specific chemical, updates on a previously assessed compound, *ad hoc* answers to a specific issue, potential risks from a compound for specific age groups, information regarding a specific endpoint and establishment of health based guidance values. Therefore, it is important to consider why a review of evidence is required, as well as which population groups are at risk, be it all or whether there may be individuals/groups at higher risk.

26. For human risk assessments, especially for the integration of epidemiological and toxicological data, a key consideration is whether the chemical in question is absorbed in humans and hence might cause a systemic effect. If no absorption in humans occurs, effects from systemic exposure in animals would not be informative. Thus, considerations on exposure should be included at an early stage in the problem formulation process.

27. It is important that the scope of the assessment is achievable and considers the available resources. Hence, the initial problem formulation is important to determine resources needed to address the research question.

28. Systematic review is the formal optimal process to ensure all available evidence has been identified and rigorously assessed to provide the best estimate of the exposure-response relationship. It is frequently used for clinical and epidemiological studies but can also be applied to toxicology (Hoffman et al., 2017).

By selecting the appropriate databases and defining the search strategy, this first step (Stage 1) aims to provide a defined literature base on which to base the risk assessment. The overall route suggested for systematic review is outlined in Figure 1. It should be emphasised that this is an iterative process with flexibility and the need for expert knowledge built into the system.

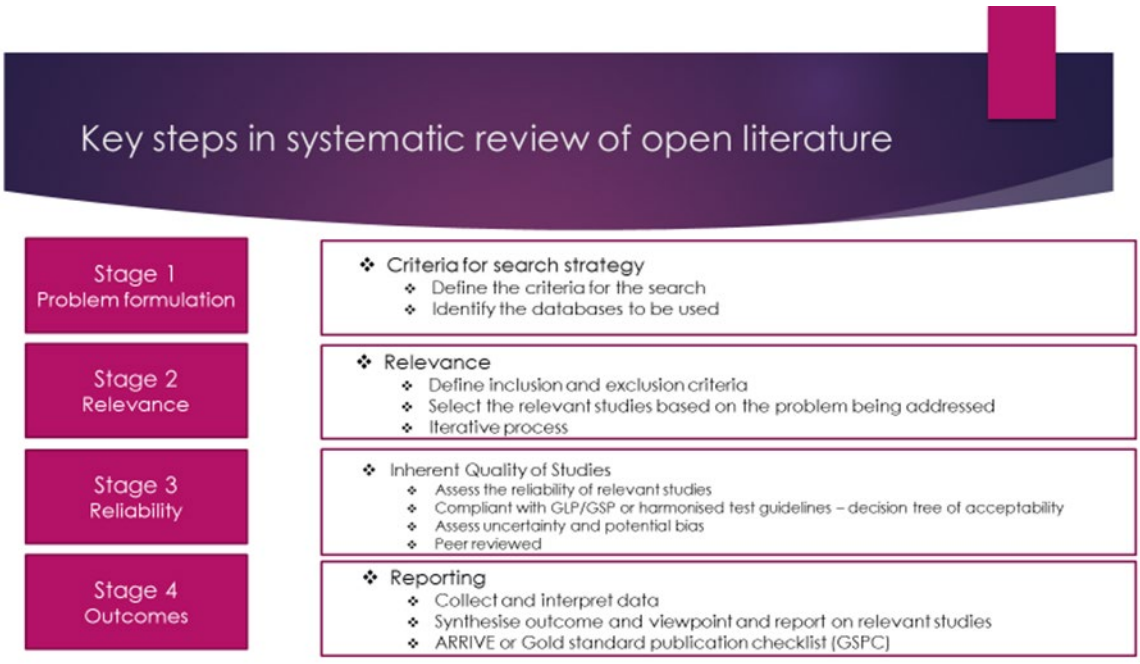

Figure 1: Key Stages in Systematic Review of the Literature

29. A new extensive systematic review would not be necessary in many situations encountered and published systematic reviews are commonly used by Committees. In addition to peer-reviewed literature, the Committees regularly utilise so called “grey literature” of information and data that have not been peer-reviewed, for example papers prior to publication, government reports or internal information provided by industry/companies. However, the greater the importance and consequence of an issue, the more likely a systematic review will be required. To ensure all relevant papers are identified a systematic review may also be more appropriate if the risk requires quantification. The time frame of the assessment plays a further important role in the consideration on the right method of literature retrieval. A thorough systematic review is complex process and can be time-consuming. The need for quick advice will limit the time available for the literature search and/or will require the use of existing reviews. Long-term, important issues may allow or require a new or updated systematic review. If a systematic review is not required or possible, it is important to consider what approach would be most appropriate and if there are any recent reviews available in the literature or by an authoritative body, for example International Agency for Research on Cancer (IARC), World Health Organisation (WHO), Food and Agriculture Organization (FAO), European Food Safety Authority (EFSA), on which to draw.

30. Some of the principles used for systematic review can help inform reporting of more limited reviews or other forms of literature.

31. The two most widely accepted over-arching guidance systems for both conducting and evaluating systematic reviews and meta-analyses of epidemiological studies come from Meta-analysis of Observational Studies in Epidemiology (MOOSE) and Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) (Steenland et al., 2020). A previous caveat to the use of these by COT/COC as discussed in the Synthesising Epidemiological Evidence Subgroup (SEES) report (COT/COC, 2015), is that they are generic i.e. not specific to the environmental and personal exposures that might be considered in COT and COC.

32. However, the literature search for all studies relevant to the endpoint in question, independently of the format, should be documented and any changes to the initial search criteria should be recorded. All studies that provide relevant data should be included at this point, bearing in mind that the process begins with a specific question.

33. The collection of available data/studies/evidence may lead to a change or refinement of the problem formulation or lead to additional questions being asked. Changes to the initial problem formulation should be recorded.

3. Quality assessment

34. To provide a comprehensive overview, the following section on quality assessment of epidemiological and toxicological evidence includes discussions on the practise of applying check lists to evidence synthesis. The SETE subgroup recommends against the use of a checklist approach for quality ranking of studies. Instead, the document(s) developed by SETE aim to provide guidance for experts and Committees to assess all information and apply good judgment transparently in a weight of evidence approach.

3.1 Assessing epidemiological evidence

3.1.1 General remarks

35. The principles of evaluation of epidemiological studies and synthesis of evidence are well documented in the SEES report which should be read together with this document. Ideally, the design of epidemiological studies needs close collaboration between epidemiologists and toxicologists to account for available information on exposure, toxicological and mechanistic information. This prior knowledge will improve the design of human studies to ensure they provide useful and relevant information. Collaboration and ongoing dialogue between epidemiologists, exposure experts and toxicologists is therefore strongly encouraged.

36. The advantage of observational epidemiological studies is larger sample size and duration, as well as a wider range of exposure. A further advantage is that route, dose and pattern of exposure are usually representative of the population of concern. The main limitations are the difficulty in reliable exposure assessment and the risk of confounding. However, there is limited scope for RCTs in assessing risks from many chemicals, as intentional exposures may not be ethical. Where they are possible, the range of exposure levels that can be ethically justified is very limited and such studies are often based on a small number of participants and short-term exposure.

3.1.2 Quality assessment

3.1.2.1 Overall approach

37. Synthesis of evidence from observational studies involves considering a wide variety of information. Observational studies may be evaluated individually to identify and quantify possible biases, their direction and likely impact on estimated parameters. However, this should not necessarily lead to a study being excluded, because such a study may still be highly informative. No single appraisal system (e.g. Grading of Recommendations Assessment, Development and Evaluation (GRADE), Promoting Methods for Evidence Use in Scientific assessments (PROMETHEUS) or MOOSE; Steenland et al., 2020) can provide a completely reliable assessment of quality because checklists and flowcharts are not flexible

enough, especially when a generic list of criteria is applied mechanically (Savitz et al., 2019). Scoring methods are difficult to replicate, are not transparent to the final user of the risk assessment and do not reflect the usefulness of an individual study. Moreover, even a study that 'scores low' may provide valuable evidence in the context of assessing a particular form of bias. The synthesis of evidence thus requires a broader approach than simply evaluating the quality of each individual study and weighting studies according to this assessment. Instead it should use the classical considerations for judging causality (Steenland et al., 2020), as outlined by Bradford Hill (1965) and others. Evidence synthesis should thereby consider the entire body of evidence available and not just individual studies in isolation.

38. One of Bradford Hill's considerations is 'specificity', i.e. a causal relationship may be more likely if the effects of a particular exposure are specific to a particular outcome (e.g. a particular dye increasing the risk of bladder cancer). However, if an exposure appears to be associated with many different outcomes, then this may indicate study bias rather than a true causal effect. There are, however, important exceptions to this – smoking and radiation cause many different diseases, and dioxin increases the risk of cancer in general. Thus, none of Bradford Hill considerations (apart from temporality – the cause should precede the effect) is essential – this is why Bradford Hill refers to causal 'considerations' rather than 'criteria'.

3.1.2.2 Assessing risk of bias

39. Risk of bias (RoB) assessments provide formal mechanisms to evaluate individual study quality regarding potential bias. Many RoB assessment tools use a hypothetical RCT as gold standard, but this is often not feasible, especially for occupational and environmental studies, but also for many other exposures. In these instances, RoB assessment tools are not appropriate and can be misleading. For example, an RCT is the definitive means of assessing vaccine efficacy, and identifying common side effects, but rare side effects are often not identified in an RCT, and their identification requires post-marketing surveillance involving large numbers. Observational studies are therefore not *a priori* weaker than RCTs for these situations (Eden et al., 2009; Steenland et al., 2020), and therefore RCTs should not be considered to be the gold standard against which to compare observational studies, as they are often not suitable to assess the relevant exposure and time-frame, and to detect rare events.

40. Thus, evidence synthesis should not start with a preferred hierarchy of study methods. Instead it should focus on the entire body of evidence – not just individual studies in isolation. As the evidence synthesis requires a wide range of expertise (epidemiology, toxicology, exposure – understanding of methods, research questions, biological/physiological relevance), it is best carried out in collaboration. Judgements should be made by considering the type, direction and magnitude of

potential biases identified across **all** studies, which is not possible when relying solely on scores from risk-of-bias instruments:

- When most studies suffer from the same type of bias, assessing the overall body of evidence by looking at individual tiers or score from each study is fully justified.
- Where studies have different types of bias, the type and direction of biases must be assessed in parallel.

41. For example: high risk of bias based on a scoring system might be due to very different types of bias with different directions in the individual studies – an assessment based on scores alone would simply downgrade the evidence, but in that case it would be highly unlikely that the consistently observed associations are due to these potential biases (since they would work in different directions, and must all be small if the studies are giving similar results). Such a scenario is plausible, and the approach of taking both type and direction of bias into account, compared to just looking at the risk of bias scoring, can lead to different conclusions (see Section 3.1.2.3 on triangulation).

42. There are, therefore, no universal rules for summarising the risk of bias in the body of evidence, and often the quality of studies often cannot be assessed easily using a generic scoring system based solely on their study design. It is, therefore, important to identify, describe and rank the biases present, in particular their direction, as this will allow use of a triangulation approach. Bias assessments should focus on identifying the most likely influential sources of bias, classifying each study on how effectively it has addressed each of these potential biases and determining whether results differ across studies in relation to susceptibility to each hypothesized source of bias (Savitz et al., 2019). This can lead to an overall informed judgement, taking all the evidence into account. As with any other scientific field, this involves expert judgement and needs to be made by appropriately qualified experts and committees. SETE recognised that, while this is the recommended approach, there needs to be some flexibility in how extensively it can be applied to a given problem, depending on any constraints consequent to problem formulation, e.g. rapid advice required, but this should be reflected in the advice given.

3.1.2.3 Triangulation

43. The use of evidence from different types of studies that may have different strengths and biases, has been termed “triangulation” (Lawlor et al., 2016). Triangulation is routinely used, e.g. by the IARC Monographs Programme which integrates epidemiological, animal and mechanistic evidence to infer causality for various potential carcinogens, including environmental carcinogens (an example is shown below where mechanistic evidence relating to the AhR receptor was used to strengthen the conclusions about the epidemiological evidence regarding dioxin and

cancer). If evidence from different epidemiological approaches all point to the same conclusion, this strengthens confidence that that is the correct causal conclusion, particularly when the key sources of bias of some of the approaches would predict that the findings would point in opposite directions.

44. In this approach, the combination of individual studies can provide strong evidence, even if individually they may have different uncertainties and biases. It is therefore important not to follow a mechanical risk of bias assessment, such as a scoring system, but to evaluate the totality of evidence.

3.1.3 Main issues in epidemiology

45. There are many different types of bias in epidemiology, which can be grouped into three major categories: confounding, selection and information bias. These arise from differences in baseline disease risk between the exposed and non-exposed sub-populations of the source population (confounding), biases arising from the selection of the study population from the source population (selection bias) and biases resulting from misclassification with respect to exposure or disease (information bias).

46. These biases are described briefly here. The likelihood of a bias occurring is highly specific to the study and the question of interest. For example, confounding by lifestyle (smoking, diet etc) can be of concern when investigating the health effects of exposure to pesticides in the community. However, there is evidence that most ‘blue collar’ occupation groups are similar - on average - with regards to lifestyle factors, including smoking (Checkoway et al., 2004).

3.1.3.1 Confounding

47. Confounding occurs when the two groups of interest in the source population, the exposed and non-exposed, are not comparable due to inherent differences in background risk due to differences in the distribution of risk factors. For example, when investigating the risk of heart disease in people who exercise or do not exercise, it is likely that those who exercised more smoked less. Thus, those who exercised more might have a lower risk of heart disease because they did not smoke and not because of exercising, i.e. smoking would be a confounding factor when assessing the possibly causal association between exercise and heart disease. Similar problems can occur in randomised trials when randomisation fails, leaving the treatment groups with different characteristics than the control group, which can not only affect baseline risk, but also differential loss and non-compliance. However, there is more concern in observational epidemiological studies because of the absence of randomisation. The concept of confounding generally refers to the source population, but it can also be introduced or removed by other processes, such as for example the selection of study participants (Pearce and Greenland, 2004), e.g. if

people who are both exposed, and have developed the disease, are more likely to participate than others.

48. In the absence of other biases, three conditions are necessary (but not sufficient) for a factor to be a confounder:

1. A confounder is predictive of disease in the absence of the exposure under investigation, (it does not have to be a causal association). Surrogate markers for causal factors can also be regarded as confounders.
2. A confounder is associated with exposure in the source population at baseline (or start of follow-up). In case control studies, a confounder thus tends to be associated with exposure among controls. An association can thus occur among cases simply because the study factor and a potential confounder are both risk factors for the disease, so this in itself does not indicate confounding.
3. A variable that is on the causal pathway between exposure and disease is not a confounder. It is therefore important to identify which factors are likely part of a causal pathway; it is not possible to do this statistically – it relies on knowledge of biological and social causation.

3.1.3.2 Selection bias

49. Confounding generally involves biases that are inherent in the source population, and therefore occur even if the entire source population were to take part in the study. In contrast, selection bias involves biases arising from the selection procedures in which the study participants are chosen from the source population. It is therefore not an issue in cohort studies with complete recruitment and follow-up, but it can occur when either participation or follow-up are incomplete. For example, UK Biobank (Sudlow et al., 2015) had an initial response rate of only 5.5%, and it is possible that participants were healthier, and had a different exposure profile, than the source population from which they were selected.

3.1.3.3 Effect modifiers (interaction)

50. Effect modification (interaction) occurs when the magnitude of the effect of the primary exposure on an outcome (i.e. the association) differs depending on the level of a third variable. In this situation, computing an overall estimate of association can be misleading. It is still possible to calculate an overall estimate of effect (which is the average population effect), but it may be valuable to also calculate the effect estimate separately for each level of the third variable. An example would be exposure to aflatoxin B1, hepatitis B virus (HBV) and hepatocellular carcinoma.

3.1.3.4 Information bias

51. Information bias refers to the people included in the study. It involves the misclassification of study participants with respect to disease or exposure status, as well as confounders, and is one common type of information bias.

- **Non-Differential Misclassification**
  When the probability of misclassification of exposure or disease is the same in cases and non-cases, it is described as non-differential misclassification. Non-differential misclassification of exposure often biases relative risk towards the null value of 1.0 and tends to produce false negative findings. It is therefore of particular concern in studies which find little or no associations.
- **Differential Misclassification**
  When the probability of misclassification of exposure or disease is different between diseases and non-diseased, or exposed or non-exposed persons. This can bias the observed effect estimates either towards or away from the null value of 1.0. For example, if cancer cases are more likely to recall past exposures than healthy controls, it would bias the estimated effect away from the null.

3.1.3.5 Exposure measurement and assessment in human studies

52. Human exposure to a substance is a dynamic process from the sources of the substance, through intake via different pathways (inhalation, ingestion, dermal), uptake and transport to one or more critical organs. However, an understanding of the toxicokinetics of a substance is often obtained through animal and *in vitro* studies. The human relevance should be evaluated for determining what would ideally be the most appropriate measures of exposure in human studies for assessing potential human risk. A lack of adequate exposure data has been reported to be the major limiting factor in preventing the identification of causal associations from human studies. Ideally the aim would be to measure exposure as close to the biological response as possible. However, in practice, surrogates of exposure are most often used. Critical issues include the:

- assessment method used: direct measurement e.g. personal monitoring or biomarkers; indirect methods e.g. exposure modelling, monitoring combined with time-activity data etc
- exposure patterns over time: duration; frequency; continuous or intermittent; critical time windows.
- relevance of the exposure metric to the exposure patterns: ever/never; duration of exposure; cumulative exposure; shorter-term intermittent exposure e.g. maximum/average intensity
- inclusion of all key sources of exposure via all possible routes and all media to give an aggregate exposure
- many exposures are part of mixtures and may therefore be highly correlated, this makes it difficult to evaluate the effects of individual components and ascertain potential interaction effects

53. Crucial to any evaluation of the exposure assessment is the characterisation of the uncertainties in the process and the potential impact on the outcomes. Exposure assessment uncertainty and inaccuracy can arise from: measurement error (e.g. instrument faults/misuse, poor execution of data collection protocol, data entry and analysis error); uncertainties due to subject characteristics (e.g. recall bias, day-to-day variability in biological characteristics); inappropriate exposure metric etc. Evaluation of uncertainties of these issues can range from qualitative discussion about the sources of uncertainty to quantitative approaches using sensitivity analyses. Measurement error can affect study sample size considerations and thus the power of a study.

54. It should be noted that data related to confounding and/or effect modifying factors may also be subject to measurement and assessment error; this can be investigated through exploration of the multivariate distribution of true and misclassified exposure and covariates but in practice the data are rarely available for this to be carried out.

3.2 Assessing toxicological evidence

55. This section is not intended to be a comprehensive summary of the literature but provides an overview, using selected references, on how to assess the relevance and reliability of both *in vivo* and *in vitro* toxicology studies so that they can be evaluated in a consistent and transparent manner. This requires a framework and a set of criteria to enable the systematic assessment of data and study quality. It is important to stress that the framework should not rely solely upon checklists but should be used as an 'aide' in developing a considered assessment. The process should be easy to use and sufficiently comprehensive so that, together with expert judgement, it will provide a robust evidence-based approach to risk assessment. This weight of evidence approach to the assessment of toxicological information is in some ways analogous to the concept of triangulation in epidemiological study evaluation, which has been discussed in Section 3.1.2.3.

56. In some respects, studies carried out under good laboratory practice (GLP) are easier to review as many of the key components are encoded in standardised study protocols and detailed reports. Scientific publications present a greater challenge. Although published studies are mostly peer reviewed, access to more detailed information on study design, conduct and reporting may not be available. It is helpful if the objectives and hypothesis are stated clearly so that the relevance of the study can be assessed. If pertinent, design and conduct using the principles of the Organisation for Economic Co-operation and Development (OECD) guidelines would also increase confidence in their utility. Transparency in the reporting of results with the potential to access raw data and any code used to carry out data analysis should also be encouraged.

57. Klimisch et al. (1997) identified reliability, relevance and adequacy as critical features of quality in studies. While over the years these criteria have been extended, revised and rewritten, the basis of the determination of a high-quality study remains unchanged. Recently, more structured frameworks have been developed with the criteria being assessed through a series of well-designed questions and sub-questions. For example, for specific chemicals (pesticides, Kaltenhäuser et al., 2017) or for specific situations (air quality, Goodman et al., 2020). Each criterion is assessed individually and for toxicology studies the questions have now also been separated into *in vivo* and *in vitro* sections to further refine the detail of the questions asked. Practical solutions have also been offered with tools such as the ToxRTool developed to assess the reliability of toxicological data (Schneider et al., 2009).

58. In order to fully evaluate the usefulness of a study, the issue to be addressed must be clearly understood and interpreted. As described in Section 2, this formulation of the problem is the first stage in the assessment and underpins the whole process. Once the search strategy is in place the relevance criteria for

inclusion and exclusion are established (Stage 2). This enables focussed retrieval of the studies appropriate to the problem defined in Stage 1. It is essential here to avoid initial bias and be comprehensive in the retrieval of the relevant literature.

59. In Stage 3, the objective is to assess the selected studies for relevance, reliability and quality of reporting to enable the most appropriate and highest quality studies to be used to address the specific questions raised in formulating the problem. Inherent in this approach and crucial for transparency is the need to specify why a study retrieved is excluded from the final assessment. This adds rigour to the assessment and aids in avoiding bias. The uncertainties within the assessment must also be described and assessed at this stage. The objectives here are to collect, extract, appraise, analyse, synthesise and integrate the evidence available.

60. In the final Stage 4, interpretation and analysis of the data extracted are assessed critically and, together with expert judgement, the final outcome is recorded.

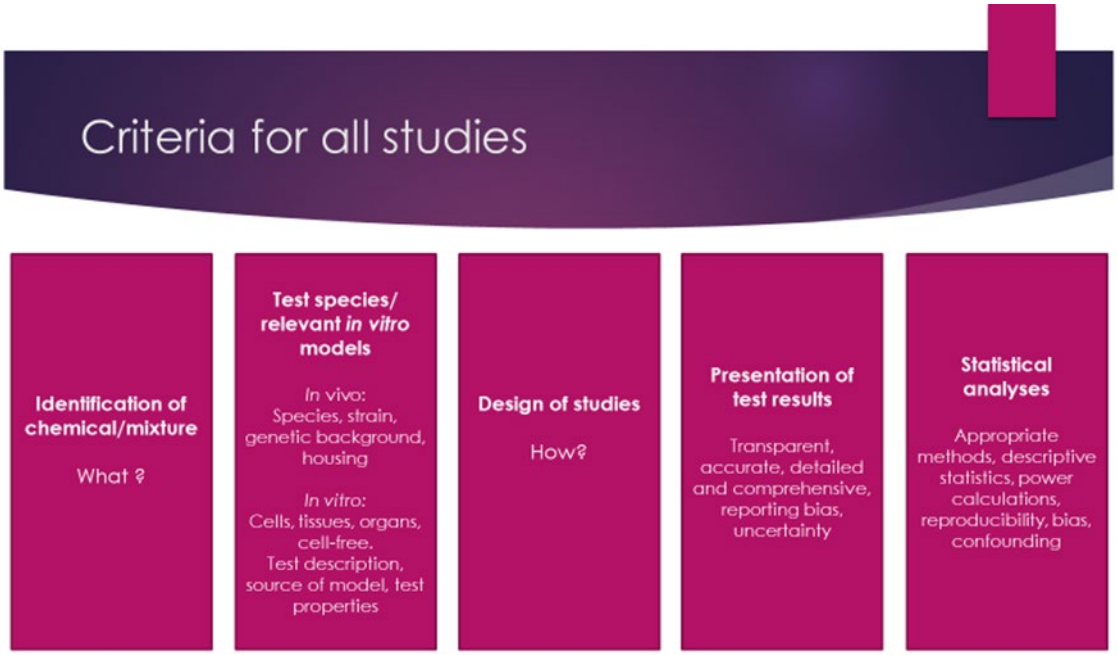

Figure 2: Criteria for Review of Studies

3.2.1 The assessment

61. Many researchers have developed ways of addressing each stage of the assessment and tables of specific questions are the most popular. Each stage of the review process is linked to a set of detailed questions that are relevant to the step being assessed. There are however several general headings that apply to all studies (Figure 2). These headings are the main pillars for the assessment of each study identified as relevant to the problem being assessed. Within each pillar there

are some general questions and then a range of sub-questions that aim to help with consistency in the assessment process.

62. By way of example, questions relating to each of the pillars in Figure 2 can be found in Tables 2-7 of Kaltenhäuser et al. (2017). Two examples are shown below in Figures 3 and 4. The questions are generic; more detailed and specific sub-questions can be added that are tailored and relevant to the assessment being undertaken. These specific questions aim to extract key details of the studies and should relate directly back to both the terms of reference and the problem formulation.

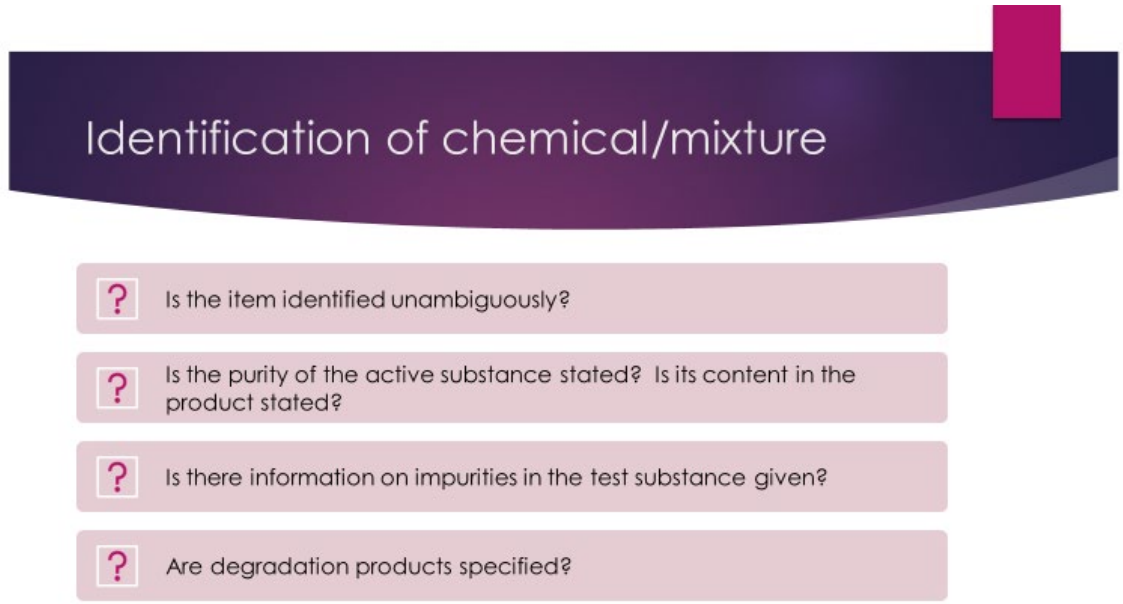

Figure 3: Chemical identification

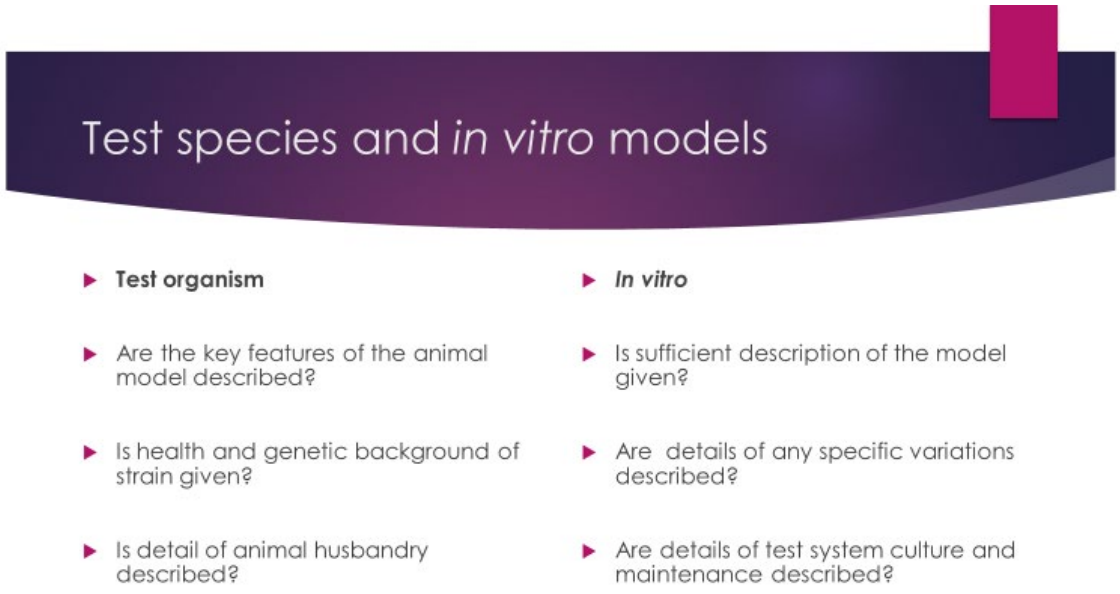

Figure 4: Test species and *in vitro* models

63. By using this tabular questioning approach greater consistency will be achieved between independent assessments. Allowing assessors to build their own sub-questions within each generic pillar gives flexibility to ensure the sub-questions are appropriate to the ongoing assessment.

64. Many guidelines and checklists exist for assessing the quality of scientific studies. For example, Nature journals have a check list[1] for their Life Sciences articles and state that "A condition of publication in a Nature Research journal is that authors are required to make materials, data, code, and associated protocols promptly available to readers without undue qualifications[2]". One specific aspect not completely confined to toxicology is the limitations imposed on studies both in costs and practicality but also for ethical animal usage (e.g. the revised Animal Research: Reporting of *In Vivo* Experiments (ARRIVE) guidelines[3] which have been endorsed by hundreds of scientific journals and societies (Percie du Sert, 2019)). Although well designed and carefully conducted studies are the expectation, assessments may also have to be based upon studies which have not been replicated but which nevertheless provide potentially relevant data.

65. The systematic assessment of *in vitro* toxicological studies is a relatively new process and presents a number of challenges.

[1] https://www.nature.com/documents/nr-reporting-summary-flat.pdf
[2] https://www.nature.com/nature-research/editorial-policies/reporting-standards
[3] https://www.biorxiv.org/content/10.1101/703181v1)

66. *In vitro* techniques are used extensively in academic research and vary widely, from the use of transformed cell monolayers to microphysiological systems (MPS). In recent years more *in vitro* techniques have gained regulatory approval. This has stemmed from public concern about the use of animals, the ban of the use of *in vivo* methods to characterise cosmetics, and the requirement of the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH) regulations to assess the safety of large numbers of existing industrial chemicals.

67. Although *in vitro* methods suffer from some limitations, e.g. limited or no xenobiotic metabolism, they do offer a number of important advantages. These include ease of manipulation, the ability to ask specific questions such as those relating to mode of action (MOA) or adverse outcome pathway (AOP), the speed with which an answer can be obtained, uniformity, the ability to control conditions and the surrounding environment, and certainty of exposure to a test substance, provided it enters the cell or cell matrix. These advantages mean that *in vitro* systems tend to have been used extensively to assess and characterise acute hazards such as cytotoxicity or genotoxicity, early in safety assessment.

68. *In vitro* systems can also be invaluable in providing information on toxicokinetics, e.g. metabolism and on mechanisms of toxicity. The development of AOPs (see below) enables interpretation of effects in many *in vitro* assays in terms of key events for an adverse outcome. Such information can add appreciably to weight of evidence. Interpretation relies on expert judgement, and formal guidelines are less helpful, as the results are not used as standalone regulatory endpoints.

69. Rapid, predictive screening is needed to be able to assess the safety of large numbers of existing industrial chemicals. Advances in the biological sciences have led to an ongoing paradigm shift in toxicity testing based on expanded application of high-through put *in vitro* screening and *in silico* methods to assess potential adverse health effects of environmental agents. This progresses the vision for toxicity testing elaborated by the US National Research Council (NRC) since the 2007 NRC report on Toxicity Testing in the 21st Century (Krewski et al., 2020). Among the principles for assessing the validity of these methods, such as (quantitative) structure−activity relationships (Q)SARs, is the need for a defined domain of applicability, i.e., identification of the range of compounds for which the method can be confidently applied for purposes of toxicity prediction. In the case of (Q)SARs, rules can be developed, based on organic reaction mechanistic principles, with particular emphasis on reactive toxicity for classifying reactive toxicants into their appropriate mechanistic applicability domains (Aptula and Roberts, 2006).

70. Models built using various sources of data can be used to predict adverse effects observed for drugs in humans. These models can be improved by adding a small set of targets to the current suite of *in vitro* human cell-based assay data. This

results in models that are reported to greatly outperform some built with the existing animal toxicity data (Huang et al., 2018).

3.2.2 Quality assessment of in vitro methods

71. *In vitro* assays are usually validated using the results of an *in vivo* animal study rather than against a human response and this may reduce the relevance of any effects seen. An experimental study can only support biological plausibility if the biological endpoint and system is relevant to humans.

72. If data generated with alternative approaches are ultimately used for decision-making on public health and the protection of the environment, the methods should have been developed and applied in a way that scientific integrity and quality is assured and demonstrated to be fit for purpose. Among the guidance documents outlining best practice for the development and implementation of *in vitro* methods for regulatory use in human safety assessment, two are briefly described.

73. The Guidance Document on Good *In Vitro* Method Practices (GIVIMP) was developed as a tool to avoid a reproducibility crisis in *in vitro* toxicological science (OECD, 2018). The aim is to reduce the uncertainties in cell and tissue-based *in vitro* method-derived predictions by applying all necessary good scientific, technical and quality practices from *in vitro* method development to *in vitro* method implementation.

74. Validation studies aim to characterise, assess and document transparently the underlying methods using an appropriate choice of methodology (Griesinger et al., 2019). This serves as a filter to ensure that only test methods able to produce data that help to address legislative requirements (e.g. EU's REACH legislation) are accepted as official testing tools. This creates a credible and transparent evidence base on test methods and provides the equivalent of a quality stamp. The reliability and relevance of the test method for a given purpose are also assessed. Relevance encapsulates the scientific basis of the test method, its capacity to predict adverse effects in the "target system" (i.e. human health or the environment) and its applicability for the intended purpose. At the core of these activities is the EU Reference Laboratory for alternatives to animal testing (EURL ECVAM) validation process.

75. While methods that have undergone the EURL ECVAM validation process are robust, transferable and widely trusted, it is a time- and resource intense procedure. It is therefore difficult to keep pace with the advances in biomedical science and technology driving the development of new approach methodologies. This, combined with the AOP initiative, has led to the recognition that alternative strategies to the current formal validation process will be needed if new approach methodologies are to be accepted for regulatory decision making. Hence, organisations such as the US Interagency Coordinating Committee on the Validation of Alternative Methods

(ICCVAM) have proposed that the focus of new method development should be on establishing fitness-for-purpose. Here, the emphasis is on 1) methodological reliability/performance; 2) the relevance of the method to biology/toxicology, for example by linkage to a key event; 3) interpretability for adverse effects *in vivo*. In addition to using *in vitro* methods standalone, conclusions on chemical safety can sometimes be obtained by integrating data from multiple sources of information using a methodology known as Integrated Approaches to Testing and Assessment (IATA). It follows from the description of the aims of validation, that the principles of a validated test should be followed, if appropriate.

76. According to OECD (2016), an IATA is an "approach based on multiple information sources used for the hazard identification, hazard characterisation and/or safety assessment of chemicals. An IATA integrates and weighs all relevant existing evidence and guides the targeted generation of new data, where required, to inform regulatory decision-making regarding potential hazard and/or risk. Within an IATA, data from various information sources (i.e. physicochemical properties, *in silico* models, grouping and read-across approaches, *in vitro* methods, *in vivo* tests and human data) are evaluated and integrated to draw conclusions on the hazard and/or risk of chemicals." An overview of existing guidance on IATAs and their component parts has been published recently (OECD, 2020). The quality of the evidence required to establish a plausible cause-effect in humans needs to be assessed.

77. The use of *in vitro* methods for risk assessment is greatly facilitated by a quantitative understanding of the key events leading to an AOP. The aim is to build up a cause-effect relationship. Ultimately the AOP knowledge derived from testing multiple chemicals may be extrapolated to predict the toxicity of all chemicals that trigger the same Molecular Initiating Event (MIE) or series of Key Events (KEs).

3.2.3 Assessing relevance of results from in vitro studies to predict risk in vivo

78. The aim of risk assessment based on experimental studies is to protect public health. This means that the further that the biological test system is removed from a human, the more careful the consideration of aspects that may influence its relevance need to be. Transformed mammalian-derived cell lines may lack functional p53, for example, and so are not arrested at a cell cycle checkpoint. Cell cycle arrest provides an opportunity for DNA repair or for progression to apoptosis. Such features increase the sensitivity of cells to the effects of a substance and may result in a response that would not occur in humans.

79. Assessing risk using *in vitro* systems is more problematic than assessing hazard. An IATA approach requires quantitative *in vitro* to *in vivo* extrapolation (QIVIVE) of cell-based toxicity assay results. The metabolites to which human organs are exposed may not be those which are generated in vitro and the profile and quantities are likely to be different. Quality aspects to consider, include the

presence or source of exogenous metabolism in the *in vitro* test systems, whether human cells are used and whether the cells are transformed or primary, thus retaining some residual metabolic capability. Physiologically based pharmacokinetic (PBPK) modelling provides a framework for such an extrapolation (Yoon et al., 2012). Substantial technological advances in *in vitro* methodologies have facilitated the study of *in vitro* metabolism and the further use of such data for *in vivo* prediction. However, extrapolation to *in vivo* with a degree of confidence, requires continuous innovation and progress in the field to address challenges such as e.g. *in vitro* evaluation of chemical–chemical interactions, accounting for individual variability and also analytical challenges for ensuring sensitive measurement technologies (Wilk-Zasadna et al., 2015). Bell et al. (2018) reviewed progress in the use of IVIVE for high throughput prioritization and regulatory decision-making and outlined their capabilities and limitations. Based, in part, on case studies of the uses of IVIVE in safety assessments they produced a set of conclusions and recommendations to support their use.

3.2.4 Extrapolation of results from animal studies to predict risk in humans

80. Toxicology studies in animals provide much of the information that regulatory authorities use to assess risk to humans. This includes information on various biological effects, including their reversibility and severity, the identity of target organ(s) and a pattern of toxicity information.

81. The set of characteristics including pathology, clinical chemistry and clinical signs comprising the toxicity pattern will be more apparent when multiple studies are reviewed. The concordance or otherwise between studies and between results in different sexes, species and strains, dose routes and whether exposure is acute or chronic can be assessed.

82. For many adverse effects, when animals are exposed to a range of doses, there is a threshold on the dose-effect curve below which the adverse effect is unlikely to occur, owing to protective mechanisms *in vivo*. A value for the dose levels where the effect is indistinguishable from “normal” background levels can be identified or estimated (for more information see Section 3.2.6).

83. From such values, uncertainty (safety) factors are imposed to estimate a dose likely to be without appreciable effect in humans. Typically, this is a 100-fold default factor to take into account inter-species extrapolation and the existence of sensitive human individuals. This is generally regarded to be protective and may be reduced if reliable scientific evidence is available (Dourson et al., 1996).

84. Irrespective of the quality of animal studies, the process of extrapolating the results to humans is an uncertain one (ECETOC, 2009).

3.2.5 Measurement of exposure

85. Measurement or estimation of exposure is critical to assessing the relevance of studies to risk assessment and to bridging the gap between epidemiology and toxicology studies. This includes likely extent (or intensity) and duration, frequency, possible accumulation, likelihood of reaching the target organs, route of exposure and the exogenous metabolism for *in vitro* studies. As part of *in vitro* and *in vivo* assays carried out according to GLP, the concentrations of test substance are generally measured generally in the dosing formulations. Toxicokinetic assessments of aspects of absorption, distribution, metabolism and excretion are normally integrated into non-clinical *in vivo* toxicity studies, in order to assess the systemic exposure to the test substance. This is needed to be able to extrapolate from animals to humans. Toxicity to the target organ shows exposure but toxicokinetic modelling or satellite studies are substitutes.

86. To extrapolate from *in vitro* to *in vivo* the internal dose is the key parameter. The concentration of free chemical available to reach target cells will depend on properties such as lipophilicity and affinity e.g. protein binding capacity. *In vitro*, factors such as evaporation, precipitation and adherence to surfaces will also influence what is available to the cell as an internal concentration (Yoon et al, 2012).

3.2.6 Measurement of concentration-effect

87. Classifying chemicals solely on hazard identification, for example for their ability to cause malformations or dermal sensitisation, will lead to placing chemicals with different potencies into the same category (Boobis et al., 2016). However, it is often useful to distinguish such chemicals from one another on the basis of their potency, as this can be necessary to enable appropriate advice to be provided, e.g. level of concern from accidental exposure.

88. For *in vivo* toxicology studies it is becoming widely accepted that the BMD approach is a scientifically more advanced method compared to the no observed adverse effect level (NOAEL) approach for deriving a Reference Point (EFSA, 2016). However, there remain discussions regarding its application. The EFSA Scientific Committee is committed to reconsider test guidelines given the expected wide application of the BMD approach (EFSA, 2016).

89. The further development of the threshold of toxicological concern (TTC) (EFSA, 2019) for endpoints such as mutagenicity has been recommended (e.g. Cohen et al., 2019) to limit unnecessary testing where exposures are expected to be low. This could be extended to areas such as foods and food ingredients (Blaauboer et al., 2016).

90. Such an approach will depend on being able to accurately measure and confidently predict the exposure, and in particular its upper bound, of humans to a chemical from all sources and the quality of this becomes critical.

3.2.7 Concluding remarks

91. This section provides considerations on how to assess the relevance, reliability and adequacy of both *in vivo* and *in vitro* toxicology studies, so that they can be evaluated in a consistent and transparent manner.

92. The formulation of the problem underpins the process. The issue to be addressed must be clearly understood and interpreted. This informs the search strategy and the appropriate databases to be used. This should be comprehensive and unbiased.

93. Relevance criteria for inclusion or exclusion are generated, the objective being to focus on the most appropriate, but manageable number of studies for further assessment.

94. Relevance, reliability and adequacy are determined by using a tabular questioning approach. Criteria for all studies include the:

- identification of the chemical or mixture,
- test species or *in vitro* model,
- study design,
- presentation of test results
- statistical analysis.

95. More detailed criteria can be added, relevant to the problem to be addressed. The quality of reporting is important, particularly information on exposure and concentration-effect relationships, to bridge the gap between toxicology and epidemiology studies. The aim is to select the most appropriate and highest quality studies to use in the assessment.

96. The ability of a study to predict adverse human health effects, particularly in establishing a plausible causal relationship, is critical. In some respects, studies carried out under GLP are easy to review, owing to validation and standardisation. Non-standard studies may add valuable insights into mode of action. As *in vitro* studies become more widely used, conclusions on chemical safety can sometimes be obtained by integrating data from multiple sources (IATA). When considering the conclusions of non-validated, non-standard studies it is important to assess the quality of the evidence, especially if a test system is far removed from human. A QIVIVE is a means of assessing relevance to humans. Physiologically based pharmacokinetic modelling may provide a means of bridging the gap.

97. The assessment should be iterative and flexible, as the nature of the problem becomes better defined. Useful, structured frameworks are available as a guide, and may be used appropriately alongside scientific/expert judgement. A prescriptive checklist or scoring approach is not recommended. The decision-making process should be robust, transparent, evidence-based, defensible and documented and, importantly, it should be easy to use.

3.3 Mode of Action

98. The concept of a MOA for adverse health outcomes evolved over several decades but was formalised into a framework for chemical risk assessment by the International Programme on Chemical Safety (Boobis et al, 2006; 2008; Meek et al, 2018; Sonich-Mullin et al., 2001). In addition to providing guidance on how to establish a MOA, based on weight of evidence, the framework also provides a systematic approach to assess the qualitative and quantitative relevance of a MOA to humans. A MOA comprises a sequence of events responsible for a toxicological effect of a test substance. The events are those considered “key” (i.e. necessary and measurable) for the MOA.

99. While it is often difficult, and sometimes not possible, to assess the causal relationship between exposure to a chemical substance and adverse health outcomes in human populations (see above for details), the key events in a MOA provide a feasible approach to assess weight of evidence for causality in human epidemiology studies. Thus, a MOA and its associated key events provide a powerful bridge between experimental studies (in animals, *in vitro* or *in silico*) and observations in human populations.

100. Mode of action-human relevance assessment comprises a number of well-defined steps.

- Is there a substance related adverse effect (adverse outcome) in an experimental system? This requires considerations of study quality, consistency and weight of evidence as described above.
- Is there sufficient evidence in experimental studies to establish a MOA for this adverse effect? This requires assessment of weight of evidence using considerations modified from those proposed by Bradford Hill.
- If so, is it possible that the MOA would be operative in humans? This requires qualitative consideration of the biology underlying the key events. For example, does a key event depend on a biological process operating only in the experimental species, with no functional equivalent in humans?
- If it is considered possible that a MOA would be operative in humans, considering kinetic and dynamic differences, how probable is it that the MOA would be operative in humans? This requires a quantitative concordance analysis of the key events in the experimental animals and in humans (or human-derived systems, such as isolated cells).
- If it is not possible to dismiss human relevance of a MOA, how can qualitative and quantitative information on the key events be used to inform the risk assessment, for example in the choice of uncertainty factors.

101. Identification of a MOA for an adverse effect in experimental animals that is considered relevant to humans would add appreciable weight to the assessment of causality underlying an association observed in human epidemiology studies. A conclusion that a MOA is not relevant to humans would argue against causality for the specific outcome in exposed subjects. Even if a MOA is considered relevant, quantitative analysis of key events in experimental animals and humans may strengthen or weaken the likelihood of a causal relationship underlying an association observed in epidemiology studies.

102. Here, dose/concentration-response assessment plays a key role. Observation of a key event in studies in experimental animals or *in vitro* will contribute little weight to establishing a causal relationship if it occurs only at much higher doses/concentrations than those observed in humans, with appropriate allowance for the possible range of human exposures. This emphasises the importance of *in vitro-to in vivo* extrapolation (and PBPK) and of comparison of the toxicodynamic response in animals and humans, using appropriate biological targets (i.e. comparison of key events). For example, if the MOA for dioxin-like compounds involves activation of the Ah receptor as a key event, and potency for the human receptor is less than for the rat receptor, observations of effects in the rat at systemic exposures only at many times those occurring in humans, would provide little support for a causal relationship between human exposure and those outcomes. It would not necessarily weaken the case for a causal relationship, but it would not strengthen it. In contrast, if effects were observed at relevant concentrations in the rat, allowing for toxicodynamic differences, this would clearly strengthen the case.

103. Similar considerations apply to effects observed *in vitro*. How does the concentration at which these are observed compare with the predicted target tissue concentration *in vivo*? Is the effect observed *in vitro* specific, or is it secondary to general cytotoxicity? This is a relatively common occurrence at high concentrations.

104. Over the last decade or so, there has been considerable interest in the use of AOPs as a means of organising biological and toxicological information, and to guide the development of novel methods for assessing chemical toxicity. In many ways, AOPs are conceptually analogous to modes of action. Both comprise a series of intermediate key events that are necessary, but usually in themselves not sufficient to produce an adverse outcome. Both depend on weight of evidence, based on the Bradford Hill considerations, for assessing the causal role of a key event. There is a greater focus with AOPs on forward prediction from assays for key events, usually *in vitro*. Hence, as with MOAs, AOPs can provide an important link between non-animal methods and assessing possible adverse health effects in humans. The OECD has a major programme on AOPs and their website should be consulted for details. The use of AOPs in risk assessment is still at an early stage and hence their current

application is largely case-by-case. However, the COT is developing separate guidance on this (as of February 2021).

4. Integration of evidence

105. In assessing risks to human health from exposure to chemical substances, relevant evidence comes from both animal and human research. Toxicological data can be used to provide mechanistic information, such as biological plausibility, to support epidemiological findings; combining both toxicological and human data helps in establishing causality (EFSA, 2017). Current approaches usually consider epidemiological evidence separately from toxicological evidence, and then combine information at the end, but a common dose response relationship is often difficult to establish. There are several methods available for quantitative synthesis of epidemiological studies, which were reviewed in the SEES report. However, there are few methods for toxicological studies or for combining epidemiological and toxicological studies. Some work on how to integrate epidemiological and toxicological evidence has been conducted on international level and brief summaries have been provided in the following paragraphs.

106. EFSA and the evidence-Based Toxicology Collaboration (EBTC) organised a colloquium in 2018 to develop an understanding of best practice, challenges and needs for evidence integration in chemical risk assessment, focusing on hazard identification and combining multiple studies and end-points for dose-response modelling.

107. The US Environmental Protection Agency (EPA) uses an Integrated Risk Information System (IRIS)[4] in their risk assessment approach, namely in the first two steps of the risk assessment process, hazard identification and dose-response assessment. The diagram on the EPAs website shows the integration of evidence for each health outcome as part of the draft development stage, however no further details or guidance are given on the practical application of the evidence integration.

108. The OECD applies the IATA, relying on the integrated analysis of existing information and the integration of new information, taking into account the acceptable uncertainties. IATA are pragmatic approaches for chemical hazard characterisation, that can include a combination of methodological approaches, such as (Q)SAR, read across, *in chemico*, *in vitro*, *ex vivo*, *in vivo* or omic technologies. There is no one overall guidance, however numerous guidance documents on specific endpoints such as skin sensitisation and non-genotoxic carcinogenicity are available.

109. In 2012, the National Toxicology Program (NTP) Office of Health Assessment and Translation (OHAT)[5] started developing an approach for the implementation of systematic review methodologies to carry out evaluations about potential human

[4] Review of the IRIS approach, Chapter 6 focuses on the evidence integration and hazard identification. The IRIS Handbook is currently being updated (as of February 2021).
[5] Review of the OHAT framework

health hazards. The updated handbook (2019) provides procedures to integrate multiple evidence streams, and of specific interest here, a section on evidence integration to develop hazard identification conclusions (Step 7). Ideally, human data providing a high level of evidence are considered together with the conclusions drawn from animal data with a high level of evidence or mechanistic data, if they provide support for biological plausibility. The OHAT hazard identification labels are similar to the labels used in the Globally Harmonised System of Classification and Labelling of Chemicals (GHS).

110. The International Programme of Chemical Safety (IPCS) produced a unified Mode of Action Framework for cancer and non-cancer risk assessments to provide a generic approach to analyse data and to contribute to harmonization. The frameworks start with the concept that it is sometimes possible to establish a causal path for a series of key events, whereby the key events are involved in the MOA. Once the MOA is established, qualitative and quantitative comparison of each key event between animal and human data enables a conclusion regarding the relevance of the MOA to human risk.

111. A framework for the systematic review and integrated assessment (SYRINA) of endocrine disrupting chemicals was published in 2016, which included authors involved in the Navigation Guide and from the US EPA, IARC and university departments in a number of countries. The framework builds on existing methodologies and evaluates the evidence from individual studies, followed by the evaluation of each evidence stream and finally the integration of evidence across all streams. The framework aims to provide the evidence base needed to draw conclusions, make recommendations, evaluate the uncertainties and support decision making.

112. The Preamble to the IARC Monographs describes the objectives and scope of the programme as well as the general principles and procedures for a transparent synthesis of different evidence streams and integration of those streams for the assessment /identification of carcinogenic risks/hazards.

113. Furthermore, a number of papers have been published over the years focusing on integration of epidemiological and toxicological evidence.

114. Adami et al. (2011) propose a five step “Epid-Tox” process, bringing together the data and analysis from epidemiological and toxicological studies with the aim to provide a view on an adverse causal relationship between an agent and a disease. The process includes the quality assessment of each individual study, the assignment of scalable conclusions regarding the biological plausibility and evidence and the placement of the findings on a causal relationship grid. The framework also aims to identify and show the influence additional data can have on the potential outcome.

115. Negri et al. (2017) applied the integrated approach by Adami et al. (2011) to the assessment of perfluorooctanoic acid (PFOA) and perfluorooctane sulfonate (PFOS) exposure to fetal growth. The authors assigned scalable conclusions to the epidemiological and toxicological evidence separately, taking into consideration the plausibility and association of an effect with the chemical, the risk of bias of the epidemiological studies and performing meta-analysis. The results of the epidemiological and toxicological data showed a reduced body weight in both, humans and rodents, however, the effective extrapolated serum concentrations in animals were 10^2-10^3 times higher than those in humans. The authors therefore concluded, based on the integrated data, that the toxicological data does not support the epidemiological association, thus reducing the biological plausibility of a causal relationship.

116. Lavelle et al. (2012) proposed a framework for evaluating and integrating human and animal data in chemical risk assessment. The process includes a step wise determination and assessment of the quality of the available human and animal data. The evaluation of human data includes various quality elements and the nature and specificity of the critical effect, the evaluation of the animal data includes data quality assessment and relevance to humans. The integration of the human and animal data involves the comparison of the various quality ratings and the determination of which data can be used to create the risk assessment based on a set of principles, such as 1) best quality data should be applied, independent of human or animal origin, 2) human studies of high quality should take precedence, 3) several considerations if human and animal data are of equal quality and are concordant or not. The framework draws on previously proposed guidelines and provides a number of case studies.

117. A publication by Boyes et al. (2005) looked at the integration of human (experimental and epidemiological) and animal data to evaluate the potential risk to human health from chronic exposure, focusing on neurotoxicity. The authors suggested that the comparability and the consistency of outcomes across studies could be improved by considering functional domains rather than individual test measures. Currently, only the abstract is available, no details regarding the practical application could be given.

118. All of the frameworks and publications described above have certain aspects or steps in common, such as 1) problem formulation 2) (systematic) literature review, including exclusion and inclusion criteria for data/literature extraction 3) quality assessment of studies among the different endpoints, species, toxicological and epidemiological data etc and 4) quality assessment across studies. The last two steps mainly apply specific criteria and take into account factors such as dose-response, biological plausibility, coherence and consistency, and finally the integration of all data (animal and human) to conclude on the effect or lack thereof.

119. However, there are a small number of frameworks, namely the work by Adami et al. (2011), Lavelle et al. (2012) and the Food Standards Agency (FSA) funded project on evaluation and expression of uncertainty in risk assessment (Hart et al., 2010), that provide practical and applicable guidance on combining epidemiological and toxicological studies to reach a conclusion on causality.

120. The following sections provide an overview of the considerations and deliberations by the SETE subgroup on the integration of epidemiological and toxicological evidence, building upon the frameworks/publications discussed above. For practical application and guidance on how to integrate epidemiological and toxicological evidence please see the accompanying guidance document (Annex 1).

4.1 Exposure

121. In assessing exposure, the emphasis is on assessing the totality of the available information, which includes different sources and routes of exposure, the assumptions and extrapolations made and the uncertainties that remain in the resulting estimates.

There are three key considerations:

- How the concentrations in studies in experimental animals relate to human exposures.
- How effect levels in studies *in vitro* can be extrapolated to doses *in vivo.*
- How higher levels and different patterns of exposure in studies of occupational exposure compare with those of the general population.

122. In addressing these considerations, it is necessary to consider the context and the data available:

Context

- What information is available on exposure for a particular substance?
- How variable are the exposure data?
- How do they inform the questions being asked of the Committee?

Data available

- What is the aim of the evaluation, e.g. is it for a specific sub-population such as infants?
- What exposure information is available for humans?
- How was exposure measured or estimated?
- What information is available on exposure in experimental studies *in vivo* and *in vitro* (e.g. concentrations, applied doses or amounts, internal doses)?

- What information is available to enable comparison between humans and experimental animals (e.g. cell levels, background levels, kinetics, dynamics)?

123. Often there will be data from both humans and experimental animals, as well as data from *in vitro* studies and in evaluating each study, consideration should have been given to the relevance of the exposure conditions in each study. When integrating the lines of evidence, it is necessary to consider the overall picture. Studies with unrealistic or unlikely exposure conditions for the general population may still provide insights into findings observed (or lack of) in epidemiological studies under more relevant conditions (e.g. if effects are significant/likely/possible only when protective mechanisms are depleted or overwhelmed).

124. During evidence integration, the rationales and reasons for the choice of exposure information used for a given substance are provided and identifies the consequences and uncertainties of these choices for the overall assessment.

125. For example: an association is observed in several epidemiological studies between an effect at environmentally relevant, albeit relatively low, exposures (say 10 µg/kg bw per day). However, in a study in experimental animals much higher exposures (say 500 mg/kg bw per day) are required to produce a similar effect. The question then is whether the animal data are supportive of causation in humans or not. Some means of comparing these different observations is needed. The animal data suggest that the association could be mechanistically credible at some dose. The next question is to consider the dose response and sample sizes in both the animal and epidemiological studies. For this, consideration needs to be given as to whether there is some species-specific difference that is relevant (e.g. a difference in metabolism or receptors). The power of the respective studies needs to be compared, how many individuals were in the respective studies and what is the minimum effect that could be observed. Could there be a sensitive sub-population, e.g. due to a genetic polymorphism? A high dose would be necessary to replicate effects in a large human population using a small number of animals. However, this will depend on the nature of the effect and the response metric, i.e. presence/absence of effect, magnitude of a continuous effect, a combination of these. The integration includes consideration of mechanistic understanding and other study aspects to make a judgement on whether the effect observed is credible at exposures of concern in the human population. The weight given to these may be influenced by the protection goals in the assessment.

126. The pharmacokinetics (PK) of a xenobiotic in the body is a complex process, governed by a variety of factors, including the physicochemical properties of the xenobiotic substance, organ and tissue blood flow rates, the permeability of various cell membranes, tissue composition and the affinity of tissues for the xenobiotic. In drug development, the ability to characterise and predict PK has been recognised as of utmost importance for several decades (Nestorov, 2003). More recently, the

importance of PK in the risk assessment for environmental pollutants is also being recognised (Aylward, 2018; Bartels et al., 2012; McNally et al., 2012). In the pharmaceutical arena most PK models are based on concentration-time profiles of drug in blood and sometimes other easily accessible body fluids (e.g. urine, faeces, breast milk). In contrast, the risk assessment of environmental chemicals, which is a human data poor area, is based on the biological monitoring of media that can be collected by non-invasive techniques e.g. exhaled breath, urine and to a lesser extent breast milk and faeces (Bevan et al., 2012; Cocker and Jones, 2017). However, an important consideration in the interpretation of biomarker concentration-time profiles is that the site of xenobiotic action is usually at the tissue level. Therefore, the biomarker concentration-time profile is only a 'surrogate marker' of the concentration-time profile at the site of action. Further, the relationship between the biomarker and tissue-concentration-time profiles may not be simple and straightforward. A modelling technique that can describe the PK of a xenobiotic in blood and various body tissues and fluids simultaneously is called PBPK modelling'.

127. PBPK models are quantitative, mathematical descriptions of the interplay between the key determinants of absorption, distribution, metabolism and excretion (ADME) of chemicals in biological systems. The biological basis of PBPK model provides a suitable platform for the integration of epidemiological and toxicological data.

128. A key strength of this approach to chemical safety assessment is the ability to describe the level of biological detail considered appropriate in order to provide a model that is fit for purpose. PBPK models typically include organ and tissue masses, regional blood flow rates, chemical specific parameters such as partition coefficients, the description of non-linear biological mechanisms and processes such as, enzyme and cell membrane transporter activity and receptor binding. All of these parameters and processes interact to provide a powerful means of estimating tissue dose, and consequently, the correlation with health effects (Clewell and Andersen, 1985; Krishnan and Andersen, 1994; Rostami-Hodjegan and Tucker, 2007). Parameters for the models can be obtained either experimentally or, increasingly, by computational prediction, for example based on the physicochemical properties of the xenobiotic.

129. Tissue dosimetry has several advantages over other measures of exposure. Tissue dose is not necessarily linearly related to external exposure, is a composite measure of multiple routes of exposure, and is determined by differences in individual behaviour (e.g., personal hygiene), work rate (characterized by different respiration rates), anatomy, physiology, metabolism and hence susceptibility (Boogaard et al., 2011).

130. In addition to the inclusion of *in vivo* mechanistic, pharmacokinetic, and toxicological information PBPK models are particularly suitable tools for integrating

data generated using *in vitro* and *in silico* methods. The availability of mechanistic, universal models for the calculation of steady state tissue:plasma partition coefficient (Peyret et al., 2010; Poulin and Haddad, 2012; Rodgers et al., 2005; Rodgers and Rowland, 2006; Schmitt, 2008) and the measurement *in vitro* of metabolism or clearance in isolated human cells has led to the development and application of "bottom up" PBPK models (Tsamandouras et al., 2015). These models are based on a broader understanding of the human body and its mechanisms and have been applied with considerable success in pharmaceutical development (Gobeau et al., 2016; Jamei et al., 2009; Rostami-Hodjegan and Tucker, 2007) and increasingly for environmental chemical safety assessment (McNally, et al., 2012; McNally et al., 2019; Moreau et al., 2017; Pendse et al., 2019). However, uncertainty and sensitivity analysis must be conducted during the model development phase, as well as on the final model output. This is an important way of evaluating the sufficiency and relevance of the biological and toxicological mechanisms described in the model. The sensitivity of model output to all input parameters and to *in vitro* and *in silico* derived parameters, in particular, must be quantified in order to provide confidence for use in chemical risk assessment (McNally et al., 2011).

131. An OECD Harmonised Template (OHT) for PBPK models providing clear guidance on the critical elements of model evaluation for regulatory application is in preparation (Tan et al., 2020). The guidance will assist public health agencies receiving PBPK model submissions. The guidance will include essential components such as, a description of the modelling purpose and strategy, summary of data used for model development, calibration and evaluation, model equations, parameters, simulations and uncertainty and sensitivity analysis and software used. The source of all data and parameters will be provided along with electronic files and supporting documents.

4.2 Integration of the different lines of evidence

132. Rather than following a set of rules, establishing cause and effect across studies is a subjective process in which all of the evidence needs to be considered. Furthermore, it is important to establish the confidence in the different lines of evidence. Rarely is the process unequivocal, where all evidence either supports or discounts a causal relationship. More often information from epidemiological and toxicological data is ambiguous and hence evaluating all evidence to reach a conclusion on an effect or lack thereof requires a systematic and transparent approach.

133. To establish the strength and weaknesses of the data it is important to look not only at the strength of the effect but also at the consistency, specificity and coherence of the effect within and across studies. Establishing the strength and weaknesses of the lines of evidence in turn allows for an informed decision on how a specific data set will influence the overall conclusion.

134. Both toxicology and epidemiology produce quantitative data. Such data has some degree of variability and uncertainty associated with it. The IPCS (2004) have defined both terms. Uncertainty is imperfect knowledge concerning the present or future state of an organism, system or (sub)population under consideration, variability is observable diversity in biological sensitivity or response, and in exposure parameters (IPCS, 2004). More information may be obtained, if feasible, to reduce uncertainty while variability is an intrinsic characteristic of biological organisms but can be better characterised by careful study design and conduct, which helps reduce uncertainty about the extent of variability.

135. Quantitative estimates from these studies should be accompanied by measures that can be used to assess the uncertainty associated with them. One approach is the presentation of results with the central (or 'best') estimate and some measure of uncertainly such as the 95% confidence (or for Bayesian, credible) interval. Although the correct interpretation and precise definition of confidence intervals is somewhat arcane it is often (simplistically) interpreted as a range of values within which the true value should lie i.e. with the upper and lower bounds as the most optimistic or the worst-case results. Sometimes this range is misinterpreted as each value within the interval having an equal chance of occurring. However, this range is, in fact, a representation (or simplification) of an underlying distribution of possible outcomes such that the probability of a result occurring close to the central estimate is much higher than in the tails where the upper and lower bounds are located and, potentially, but relatively rarely outside the bounds.

136. This can be shown in various diagrammatic forms by, for instance, regression lines with the distribution represented by shaded areas around the line, with the depth of the shading representing the density of the distribution. A cross-sectional slice at a point in the line would show the distribution, such as a normal or other distribution, at that point. van der Bles et al. (2019) have published a paper on “Communicating uncertainty about facts, numbers and science” which covers the communication of uncertainty. Their Figure 5 illustrates expressing uncertainty related, for instance, for mean and confidence interval type data derived from a Cochrane summary.

137. Another source for expressing uncertainty is the Probability Yardstick from the UK Government Professional Development Framework (PHIA, 2019). "The Professional Head of Intelligence Assessment Probability Yardstick splits the probability scale into seven ranges. Terms are assigned to each probability range. The choice of terms and ranges was informed by academic research and they align with an average reader’s understanding of terms in the context of what they are reading." Note that the divisions used for the Yardstick do not match up with those from van der Bles et al. (2019).

138. Building on the work above, such as that of Adami et al., the SETE working group established a number of key points to be considered when integrating epidemiological and toxicological lines of evidence. Main aspects in these considerations are whether or not the data indicate robust evidence of an effect in animals and whether the same effect has been reported in human/epidemiological data. If the same effect has been reported in both animal and human studies, considerations should be given as to how the effect levels compare. If possible, active site concentrations should be compared, together with the relative sensitivities of the molecular target. Does the effect concentration in the experimental studies reflect a realistic exposure scenario in the general population? *In vitro* data can provide further support for key events, if occurring at plausible concentrations, and are important to include in the integration considerations, together with any other mechanistic data. Information on AOPs or MOAs can further strengthen the association between animal and human data and support a biologically plausible mechanism. As an example, experimental and mechanistic evidence of the effects of dioxin at the AhR receptor make it plausible that it could increase the risk of cancer in general – a hypothesis supported by the epidemiological evidence.

139. If a predominantly positive answer can be given to the main considerations provided above and covered in more detail in the guidance document (Annex 1) then the weight of evidence strongly supports causality. However, it is important to establish the strength and robustness of the evidence for each line of evidence and reflect on how the uncertainties may influence the weight of evidence. Taken together these should provide information on how the various lines of evidence influence the overall conclusion, increasing or decreasing the likelihood of a conclusion of causality. For example, *in vitro* data demonstrating that a key event occurs at the same tissue concentrations as estimated in the exposed population would add weight to a conclusion of causality, whereas the absence of effects in occupationally exposed populations at or above levels at which effects are observed in experimental animals would reduce the weight of a conclusion of causality. Considerations should be given to whether or not a line of evidence is considered sufficient by itself or provides a significant contribution to the overall weight of evidence.

140. One way to clearly depict the influence of the different lines of evidence on the conclusion on causality is via visual representation. A graphical approach is recommended, similar to that of Adami et al. (2011) (Figure 5). In this, the relative impacts of epidemiological and toxicological evidence are plotted against each other.

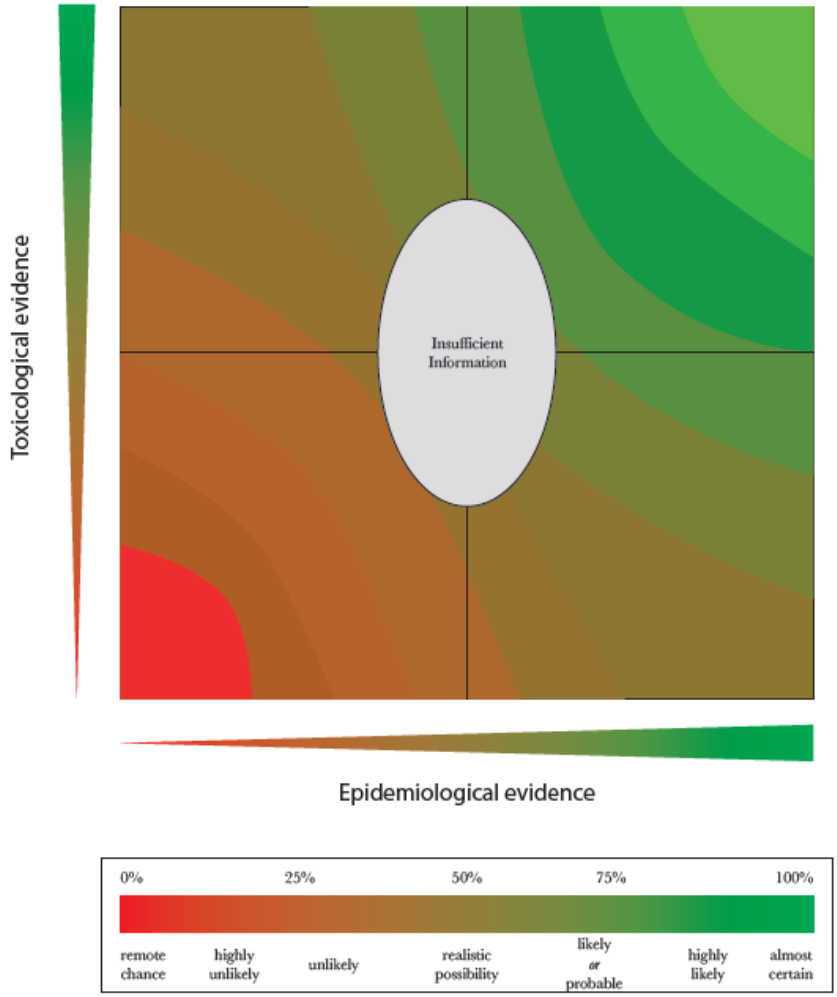

Figure 5: Example for the visual representation of the likelihood of a causal relationship, considering both epidemiological and toxicological data.

141. It is important to begin with the initial estimate of causal interference at the centre of the graph. Depending on whether a line of evidence supports or discounts (or has no clear influence) a conclusion of causality, placement on the graph is then moved accordingly, either in a positive or negative direction. The movement itself is influenced by several factors, including the impact of the strength or weakness of the evidence, any relative weighing given to epidemiological and toxicological studies and the uncertainties associated with the data. As more information is included in the process and/or becomes available, the placement of the toxicological and/or epidemiological evidence can be easily adjusted.

142. Annex 2 of the report provides practical examples of applying the aforementioned considerations and the SETE guidance document to the different lines of evidence on caffeine, cadmium and tropane alkaloids (TAs) and on reaching conclusions on the likelihood of a causal relationship.

PLEASE NOTE

Annex 2 has not been finalised.

Background information on caffeine, cadmium and tropane alkaloids are included, however conclusions on causality by the SETE subgroup have not been finalised and hence the visual representation has not been included in this first draft version of the report.

However, an example of the envisioned graphical representation is provided in Figure 6. The colours thereby represent the likelihood of a causal relationship, vectors have been included to indicate the influence of the different evidence streams and uncertainties on the final conclusion. The overall conclusion and the influence of the lines of evidence on the conclusion are currently based on the assessment of caffeine by Hart et al. (2002).

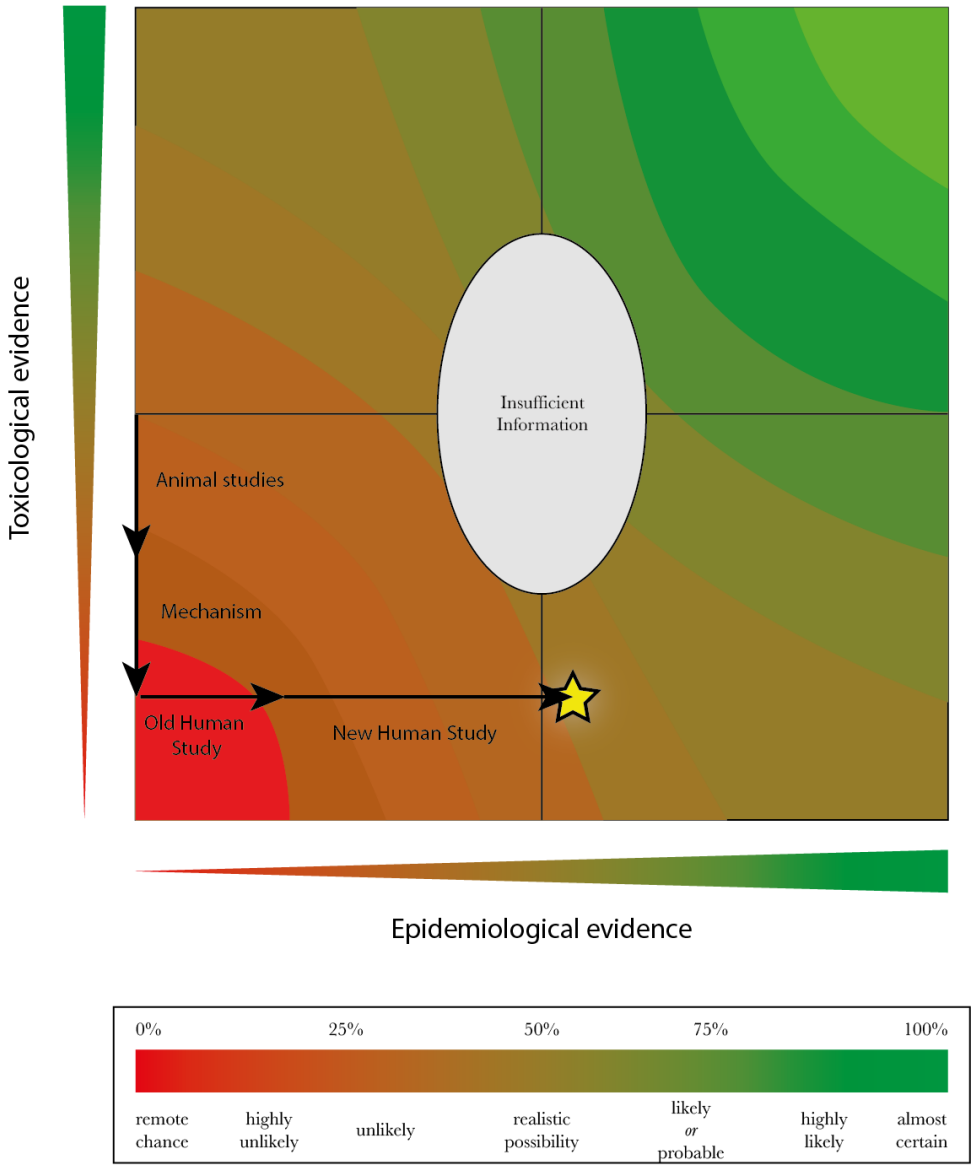

Figure 6: Visual representation of the causality of caffeine intake and an increased risk of foetal growth restriction.

5. Conclusions and recommendations

TO BE FINALISED

Annex 1 Guidance on evidence synthesis

Following the work of the Synthesising Epidemiological Evidence Subgroup (SEES) of the COT and COC, the subgroup identified the following over-arching guidance on the synthesis of epidemiological and toxicological evidence. It was recognised that issues on which advice from the committees is sought vary considerably and, hence, the guidance proposed should be sufficiently flexible to address this. For example, in some situations (e.g. risk from exposure to a relatively new product) studies in experimental animals may provide the most valuable, and perhaps even the only, information, whereas in other situations (e.g. long-term and significant exposure to an environmental contaminant), epidemiological studies may provide the most relevant information. For both epidemiological and toxicological information, a weight of evidence approach is proposed, the details differing, depending on the type of information available.

Problem formulation and literature retrieval

The first step in the process of evidence synthesis is scoping and problem formulation. This ensures that the right questions are asked, helps make the most efficient use of resources and identifies the best approaches to use in the assessment. Problem formulation is developed by the risk manager (e.g. FSA) in discussion with the committee. The following points should be considered.

- Has the issue been addressed previously by the committee?
- Why is a review of the evidence needed now?
- How urgent is the review?
- Which sub-populations are of potential concern?
- Is there any systemic exposure (determines the need for an assessment and if so, are data on systemic or local exposure of most concern)?
- Is a systematic review required?
    - If advice is needed urgently, a formal systematic review will not be possible, so what form will the review take, e.g. a focused literature search or use of a review by another authoritative body or from the published literature? If the issue is of major, long-term significance, a new or updated systematic review may be appropriate.
    - Is qualitative (hazard) or quantitative (risk) advice needed? If the latter, a systematic review is most likely to be necessary, to ensure all risk estimates are identified and included.
- Has the issue been addressed recently by another authoritative body (e.g. JECFA, EFSA, IARC)?
    - If yes, does this serve the needs of the committee, e.g. is it systematic and of satisfactory quality?
    - Does the review only need updating, or is a new review necessary?
    - Is the starting date for literature retrieval adequate, or could useful older literature be missing?
    - Was the characterisation of risk appropriate to the needs of the committee (e.g. were both acute and chronic risks addressed; were risks in the sub-population of concern assessed)?

    - Is there in an existing meta-analysis, and if so, does it need to be updated?
- As information is retrieved and evaluated, this may necessitate some change or refinement of the problem formulation or lead to additional questions being asked. Any such changes should be agreed with the risk manager (e.g. FSA) and clearly recorded.

Overarching principles

- An established system or guidance should be followed where appropriate (e.g. for a systematic review; quality assessment of toxicological studies).
- The evidence synthesis should include an expression of uncertainty to the extent possible.
- Potential conflicts of interest should be identified and considered, including for published papers and reviews.

Information retrieval

- What information is being sought (e.g. potential adverse health effects of substance X in the general population?
- What are the constraints on the search for information, if any, e.g. within a specific time frame; for a specific geographic region?
- How extensive will be the search for information (e.g. systematic review, focused review)?
- What are the potential sources of information (e.g. bibliographical databases, proprietary information from food producers)?
- What search strategy will be used for open literature, i.e. search terms?
- Will the grey literature be searched, and if so, how will this be done?
- How will other potential sources of information be searched, if necessary?

Epidemiological information

The Report of the Synthesising Epidemiological Evidence Subgroup (SEES) of the Committee on Toxicity and Committee on Carcinogenicity (provide link) provides detailed information and guidance for the committees on the evaluation of epidemiological information. The current guidance summarises and updates the recommendations of the SEES report.

Focused literature search

As a minimum, this should include the details described under Information Retrieval, above: i.e. purpose of search; information sources searched (e.g. PubMed); period covered (e.g. < Jan 2010; > June 2019); search terms and their combinations.

The results of the search should be summarised, as follows:

- Numbers of papers identified, and numbers included in the review
- Reasons for exclusion of papers (e.g. not covering health effects of the substance of concern)
- Extraction of key information from relevant literature in narrative, graphical and/or tabular format. It can be particularly useful to determine what

information is needed for the committee assessment (e.g. effects of substance X on developmental outcomes) and to tabulate relevant information from each paper on this (e.g. exposure metrics, outcomes, affected population). Guidance such as the Meta-analysis of Observational Studies in Epidemiology (MOOSE) and Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA), as adapted in Appendix A1 of the SEES report, should be consulted for the types of information that would be value.

Evaluating an existing systematic review

Details are provided in the SEES report.

- As a minimum, an adapted checklist, such as from MOOSE or PRISMA (see SEES report) should be consulted to assess the completeness of the information available in the literature used in an existing systematic reviews and meta-analyses.
- The evidence synthesis method and any scoring system used by the authors should be described, and the potential implications on the conclusions noted, e.g. GRADE gives lower weight to evidence from observational studies.

Conducting a new systematic review

The recommended approach for reviewing the open literature comprises four stages.

- Scoping: Criteria for the search strategy
    - Define the criteria for the search (see above)
    - Identify the information sources to be used
- Relevance
    - Define inclusion and exclusion criteria
    - Select studies relevant to the assessment
- Reliability: Quality of the studies
    - Assess the reliability of the studies
    - Compliance with appropriate guidelines (e.g. Good Epidemiological Practice (GEP))
    - Assessment of uncertainty and potential bias
    - Peer review
- Outcomes: Reporting
    - Collect and interpret the evidence
    - Evidence synthesis
    - ARRIVE or GOLD publication checklist

An adapted checklist such as from the MOOSE guidelines (SEES Report Appendix A1) should be used. SEES recommends a number of additions to the published MOOSE checklist:

- Include a flow chart for the identification of papers at the different stages of the systematic review
- Assess the adequacy of study data presentation
- Describe how data were extracted
- Use forest plots to illustrate findings from the studies reviewed

    - Include patterns of association and confidence intervals where possible

Assessment of epidemiological evidence

All available studies, e.g. observational studies, meta-analyses, should be evaluated individually to identify potential sources of confounding, possible biases, their direction and likely impact on estimated parameters; nature of exposure; outcomes; and conclusions. This should not necessarily lead to individual studies being excluded, since such a study may still be highly informative and it is recommended that all relevant studies should be included in evidence synthesis, using a weight of evidence approach (see below). A description of possible sources of confounding and bias in epidemiological studies is provided in the SETE report. These include confounding, selection bias, effect modifiers and information bias.

- Assess risk of bias. The type, direction and magnitude of potential biases identified across all studies should be considered
    - When most studies suffer from the same type of bias, assessing the overall body of evidence by looking at individual tiers or score from each study is fully justified.
    - Where studies have different types of bias, the type and direction of biases must be assessed in parallel.
    - Identify the most likely influential sources of bias, classifying each study on how effectively it has addressed each of these potential biases, and determine whether results differ across studies in relation to susceptibility to each potential source of bias
- Exposure assessment
    - Assessment method used: direct measurement (e.g. personal monitoring, biomarkers) or indirect methods (e.g. exposure modelling, food consumption pattern)
    - Exposure patterns over time: duration; frequency; continuous or intermittent; critical time windows
    - Relevance of the exposure metric to the exposure patterns: ever/never; duration of exposure; cumulative exposure; shorter-term intermittent exposure (e.g. maximum/average intensity)
    - Have all key sources of exposure, via all possible routes, via all relevant media, been included, to give an estimate of aggregate exposure?
    - Many exposures are part of mixtures and may therefore be highly correlated, making it difficult to evaluate the effects of individual substances
    - If possible, determine whether uncertainty in exposure assessment in a study is likely to under- or overestimate the exposure.
- Outcome assessment
    - Nature of adverse health effects, e.g. testicular cancer, decreased birth weight
    - Affected population, e.g. all exposed individuals, young children
    - Any possible differences in sensitivity, which cannot be accounted for by exposure, e.g. atopic individuals

    - Strength of the effect in terms of severity and number of individual affected, as a fraction of the exposed population
    - Uncertainty associated with effect estimates
- Conclusions of the study or review
    - Risk metric, e.g. relative risk, odds ratio, incremental risk
    - Statistical significance of findings
    - Confidence intervals in risk estimates
    - Likelihood findings were by chance, e.g. confidence intervals, number and types of exposure-effect comparisons
    - Power of the study, e.g. minimum detectable effect given the size of the population studied

Triangulation. Even if individual studies have different uncertainties and biases, the totality of the evidence should be evaluated, to determine whether the combination of individual studies can overcome the different biases and provide suitable evidence.

Assessment of toxicological evidence

Following information retrieval as described above, toxicological data should be evaluated using a weight of evidence approach, analogous to the triangulation approach described above for the assessment of epidemiological data. This should include assessment of uncertainty, both qualitative e.g. the toxicological significance of an effect observed, and quantitative, e.g. dose without observable effects. It is recommended that a framework for the systematic assessment of data and study quality should be used for this purpose. This should be sufficiently comprehensive that, together with expert judgement, it provides a robust evidence-based approach to risk assessment, whilst being easy to use.

In vivo studies

- Assess the quality of each study, using the criteria proposed by Klimisch et al (1997) for reliability, relevance and adequacy. Published modifications to the scheme proposed by Klimisch et al (1997) may be more appropriate for a given assessment (e.g. Schneider et al, 2009; Kaltenhäuser et al, 2017; Goodman et al, 2020).
    - Does the study comply with Good Laboratory Practice (GLP) or the principles of GLP?
    - Was the study conducted according to an accepted guideline (e.g. OECD, EPA)?
- For each study, consider the following
    - Was the test material clearly identified and defined?
    - Was the experimental system (e.g. test species, strain, husbandry) appropriate?
    - Was the study suitably designed (e.g. route of administration, dose selection)?
    - Was exposure suitably assessed?
    - Was the dose expressed appropriately, i.e. what was the dose metric?
    - Were the results reported adequately (e.g. sufficient detail)?

- Were the statistical analyses of the results appropriate (e.g. expression of uncertainty (e.g. CIs) where necessary, power calculations, assumptions on data distribution)

There are a number of published schemes that provide details on how to check the quality of scientific studies, e.g. Nature journals' checklist for Life Sciences articles, ARRIVE guidelines.

In vitro studies

In vitro studies should be evaluated using similar principles to those above for *in vivo* studies. However, relatively few *in vitro* methods have been fully validated for use in regulatory toxicity testing (e.g. no OECD guideline). Hence, reliability needs to be assessed using other approaches.

- Was the test material clearly identified and defined?
- Has the method used been formally validated (e.g. EURL ECVAM)?
- Is there a guideline for the method from an authoritative body (e.g. OECD)?
- Was the study conducted according to the OECD (2018) Good In Vitro Method Practices (GIVIMP)?
- Is sufficient information provided to assess the relevance of the method?
    - Is the endpoint used being measured reliably (e.g. specificity, variability, metabolic capacity)?
    - Was exposure assessed suitably?
    - Is the endpoint measured biologically/toxicologically relevant (e.g. cell line, culture conditions, duration of exposure)?
    - Is it possible to extrapolate the findings *in vitro* to a mode of action or adverse outcome pathway *in vivo* (e.g. known relationship to adversity *in vivo*)*?*
        - Qualitatively (e.g. key event in an AOP, known relationship to adversity *in vivo*)?
        - Quantitatively (e.g. suitable PBPK extrapolation available)?
- Were the statistical analyses of the results appropriate (e.g. expression of uncertainty (e.g. CIs) where necessary, power calculations, assumptions on data distribution)?

Assessment of mode of action

Information on mode of action can be invaluable for evidence integration by enabling the qualitative and quantitative bridging between experimental data and observations in humans. Mode of action (MOA) underpins weight of evidence considerations by providing the mechanistic link between empirical observation and biological plausibility. The WHO IPCS has developed a well-established framework for assessing mode of action and its implications of human health risk assessment (Boobis et al, 2006, 2008; Meek et al, 2014).

The key elements in assessing a mode of action are as follows:

- Is there a substance related adverse effect (adverse outcome) in an experimental system? This requires considerations of study quality, consistency and weight of evidence as described above.

- Is there sufficient evidence in experimental studies to establish a MOA for this adverse effect? This requires assessment of weight of evidence using considerations modified from those proposed by Bradford Hill (1965).
- If so, is it possible that the MOA may occur in humans? This requires qualitative consideration of the biology underlying the key events. For example, does a key event depend on a biological process operating only in the experimental species, with no functional equivalent in humans?
- If it is considered possible that a MOA would be operative in humans, considering kinetic and dynamic differences, how probable is it that the MOA would be operative in humans? This requires a quantitative concordance analysis of the key events in the experimental animals and in humans (or human-derived systems, such as isolated cells).
- If it is not possible to dismiss human relevance of a MOA, how can qualitative and quantitative information on the key events be used to inform the risk assessment?

Adverse outcome pathways (AOPs) are in many ways conceptually analogous to modes of action. However, there is a greater focus on forward prediction from assays for key events, usually *in vitro*. Hence, AOPs provide an important link between non-animal methods and assessing possible adverse health effects in humans. The OECD has a major programme on AOPs and their website should be consulted for details. The use of AOPs in risk assessment is still at an early stage and hence their current application is largely case-by-case. However, the Committees are developing separate guidance on this.

Evidence integration

All lines of evidence should be considered, with no specific hierarchy *a priori*. However, assessment of the strength of evidence from a particular approach, as described above, will provide an indication of how reliable a line of evidence is. For example, it may be that the epidemiological evidence for a given compound is considered extremely robust, whereas the evidence from *in vivo* toxicological studies is considered very weak. This should be reflected in how the respective lines of evidence are weighted. This is different from consideration of the nature of the evidence.

The guidance provided here has been developed from published approaches, such as the “Epid-Tox” process developed by Adami et al (2011). For each question some upper and lower estimate of uncertainty should be made.

- Epidemiological evidence
    - How strong is the evidence that exposure to the substance of concern causes an adverse health effect in humans?
    - Are the exposures at which effects are reasonably anticipated to occur in humans realistically achievable in the population(s) of concern?

    - Is the same adverse health effect observed in toxicological studies, recognising that some effects are not produced in toxicological studies?
    - Are there any modifying factors in sub-populations that increase or decrease susceptibility, consistent with the mode of action (see below) (e.g. genetic polymorphisms in molecular targets for the AOP, differences in life-stage sensitivity)?
- Experimental evidence
    - How strong is the evidence that the substance of concern causes an adverse outcome on administration to experimental animals?
    - Is the adverse outcome observed relevant to humans (e.g. known species or strain specific sensitivity to a class of compounds)?
    - Is the same adverse outcome observed in exposed human populations?
- Mechanistic data/MOA
    - Is there sufficient information to establish a MOA?
    - Is there evidence that the key events (precursor events) observed experimentally occur is exposed humans?
    - Is there evidence from other information (e.g. pathophysiology) that should a key event occur in humans it will lead to the adverse outcome?
- Exposure
    - Is the exposure in experimental models (laboratory species, *in vitro*) at which adverse effects are observed achieved in the subjects of an epidemiological study? If not, it may be difficult to draw conclusions on causation, as no effects would be expected at this exposure level.
    - Is the predicted (e.g. using PBPK modelling) or measured internal exposure at which adverse effects are observed in humans consistent with that at which adverse outcomes are observed in experimental animals?
    - Is the predicted (e.g. using PBPK modelling) target site concentration at which adverse effects are observed in humans consistent with the predicted concentration at which adverse outcomes are observed in experimental animals?
    - Is the predicted (e.g. using PBPK modelling) target site concentration at which adverse effects are observed in humans consistent with the predicted concentration at which adverse effects are observed *in vitro*?
    - If the relative sensitivity of the molecular target in humans and experimental models (e.g. laboratory species, cell line *in vitro*) is known, is the dose/concentration-effect relationship in humans consistent with the experimental observations?

Combining the evidence

- Integration of the lines of evidence
    - A graphical approach similar to that of Adami et al (2011) is recommended. However, the two axes should be “Epidemiological evidence” (x-axis) and “Experimental evidence” (y-axis).

    - Start with a clear hypothesis relating exposure to the substance of concern to adverse health effect(s) in humans (e.g. caffeine during pregnancy causes low birthweight). This forms the initial estimate of causal inference and should be placed centrally in the grid.
- Assess the impact of each line of evidence on confidence in the initial estimate, using expert judgment to assign a number of units of movement along the axes.
- Where possible, include an estimate of uncertainty to provide a range (likely, upper and lower bound of impact)
- Epidemiological Evidence
    - Consider how the answer to each question would affect confidence in the initial estimate and move the estimate accordingly leftwards or rightwards along the x-axis, as appropriate.
- Experimental Evidence
    - Include all other lines of evidence under this heading
    - Consider how the answer to each question would affect confidence in the initial estimate and move the estimate accordingly upwards or downwards along the y-axis, as appropriate.
- Conclusion on the evidence
    - Based on where the estimate of causal inference appears on the graph, after taking account of all lines of evidence, one of several conclusions is possible:
        - A causal relationship in humans is likely
        - A causal relationship in humans is unlikely
        - A causal relationship in humans is possible, but lacks strong experimental support
        - A causal relationship in humans is possible, but lacks strong epidemiological support
        - There is insufficient information to reach a conclusion on the possibility of a causal relationship

Reporting

- The problem being assessed should be clearly stated, together with why it is being reviewed by the committee
- Each step of the procedure should be clearly described
- Information sources should be documented, including the databases searched, details of the search terms used, criteria for selection of papers and the papers identified
- All lines of evidence should be described, together with their identified uncertainties.
- A clear conclusion on how each line of evidence affects the estimate of causal inference should be provided, together with the associated uncertainty
- Tabulation of this information may be of value
- A graphical presentation of evidence integration should be provided

- The conclusion of the assessment should be stated, with an estimate of the overall uncertainty and, where appropriate, guidance on how data gaps could be filled

Annex 2 Examples of epidemiological and toxicological data integration

PLEASE NOTE THIS SECTION HAS NOT BEEN FINALISED.

The following sections provide practical examples applying the above procedures for the integration of evidence (Section 4) and SETE guidance (Annex 1).

Example 1: Caffeine

The lines of evidence for the data integration were drawn from the COT Statement on the reproductive effects of caffeine (2008), the FSA funded project on evaluation and expression of uncertainty in risk assessment (Hart and Gosling, 2010) and EFSAs Scientific Opinion on the safety of caffeine (2015). For background information and a full review of caffeine, please see the respective statements.

The potential reproductive effects of caffeine have been studied in a wide range of animal species. Significant reductions in birth weight have been reported in rats repeatedly exposed to caffeine, yet it was not possible to determine whether the reduced birth weight was a direct effect on the foetus or if it was secondary to a maternal effect (decreased maternal bodyweight). A study in monkeys reported high rates of still births and miscarriages after exposure to caffeine. However, the serum metabolite of caffeine and metabolic enzyme for caffeine differs between monkeys and humans, hence for a given dose of caffeine the systemic exposure for monkeys is likely to be higher and the study may be less relevant for the assessment of caffeine in humans.

In human studies caffeine consumption was reported to be associated with foetal growth restriction (FGR), reduced birth weight, miscarriages or an increased risk of still birth at caffeine consumption ≥ 300 mg per day. However, the data gave no indication of a threshold level of exposure below which there was no risk, hence a conclusion on the relation of risk to level of exposure (dose/response) is not possible. Most of the available human studies assessed caffeine intakes at various stages of pregnancy through dietary questionnaires and calculated intakes by multiplying the number of servings by an estimated caffeine content, potentially influencing the accuracy of the data.

Table X Summary of the strengths and weaknesses of the data on caffeine and the influence of the lines of evidence on the overall conclusion.

Lines of evidence and their main strengths (S) and weaknesses (W)	**Influence on Conclusion**
Animal data	Due to the limitations in the study design and the uncertainty about the relevance of the

S – Available data indicate an effect (on FGR) occurred in conjunction with maternal toxicity W – The relevance of the effects in experimental animals to humans is uncertain	findings for humans the data is not very informative for assessing the risk of caffeine intake
Human data (COT/Hart et al.) S – reliable and good study design (prospective study rather than retrospective analysis), allows for better subject control S –high rate of completion, details of exposure and outcome, selection of a robust endpoint (FGR), assessment of metabolic phenotype for caffeine metabolism W – Residual cofounding always possible, caffeine intake may have been a surrogate for other lifestyle factors W – Although exposure assessment thorough, potential errors as reliant on subject recall (particularly during the first trimester)	Weaknesses of this study also occur in other epidemiological studies, influence on conclusion not as strong as the strengths of the study design and robust endpoint
Human data (EFSA) S – Reliable study design, human intervention and prospective cohort studies with adequate control for cofounding variables, reducing the risk of reverse causality and recall bias. W – prospective cohort studies cannot provide evidence for causal association between caffeine and adverse birth related outcomes. Given consistency of dose-response observed and plausibility of mode of action, EFSA assumes relationship is causal	
Mechanistic data W – limited date was available; the data that were available were unable to identify a plausible biological mechanism for an effect of caffeine on FGR.	While the data does not provide information on causality, its value/influence is limited due to the lack of data/investigation in itself The absence of evidence is not evidence of absence

From the available data and the assessment of the strengths and weaknesses of the various lines of evidence it is not possible to be confident that there is a causal relationship between caffeine intake and increased FGR. While it can be assumed that the key events for adverse effects of caffeine in animals and humans would be the same, (experimental) data is currently lacking. However, the epidemiological and to a degree the toxicological evidence does allow the conclusion that there is an association between caffeine intake during pregnancy and increased risk of FGR. It

is not possible to define a level of caffeine intake below which there is no risk, however it seems likely that the risk increases with intakes in the order of 200 mg.

INSERT GRAPH

Figure X Visualisation of the causality of caffeine intake and an increased risk of FGR.

Example 2: Cadmium

The lines of evidence for the data integration were drawn from the EFSA Opinion on cadmium in food (2009) and the EFSA Statement on tolerable weekly intake for cadmium (2011). For further background information and a full review of cadmium, please see the published statements. For full reviews of the toxicity of cadmium in experimental animals please see WHO-IPCS, 1992; ATSDR, 1999; JECFA, 2006; EC, 2007.

Cadmium primarily effects the kidney, especially the proximal tubular cells, where it accumulates and may cause renal dysfunction. Cadmium can also cause bone demineralisation (direct through bone damage or indirect through renal dysfunction). After prolonged and/or high exposure tubular damage may progress to decreased glomerular filtration rate and eventually renal failure. Exposure data in the general population have also been associated with an increased risk in cancer (lung, endometrium, bladder, breast) and IARC has classified cadmium as a human carcinogen (group 1) based on occupational studies, the European Commission has classified it as a possible carcinogen (Carcinogen Category 2). The latter concluded that there is currently no evidence that cadmium acts as a carcinogen following oral exposure, however the Weight of Evidence collected on genotoxicity testing/long-term animal experiments/epidemiological studies suggests cadmium oxide as a suspected inhalation carcinogen.

Cadmium effects both gene transcription and gene translation. Cadmium also plays a role in controls various transduction pathways by playing the role of an alternative signalling molecule. Furthermore, it regulates the internal cell concentration of calcium and may interfere with calcium homeostasis by its ability to modulate extracellular calcium sensing receptors (CaSR). Calcium may therefore profoundly affect the function of cells expressing CaSR such as kidney cells. Cadmium itself is not a redox-active metal but induces the production of reactive oxygen species (ROS) by indirect processes. By modulation of gene expression and signal transduction cadmium can therefore affect cell proliferation, differentiation, apoptosis and other cellular activities. These changes as well as its capacity to inhibit DNA repair enzymes may contribute to its genotoxic and carcinogenic effect.

The target organs (kidney, lung) and the toxicokinetics after oral exposure are similar among species, however the estimated absorption of cadmium in rodents is lower compared to humans, especially after prolonged exposure. In addition, species specific differences in metallothionein, cadmium kinetics and toxicity have also been well established.

Table X Summary of the strengths and weaknesses of the data on cadmium and the influence of the lines of evidence on the overall conclusion.

Lines of evidence and their main strengths (S) and weaknesses (W)	**Influence on Conclusion**

Animal data S – Number of studies on toxicity of cadmium in experimental animals, the target organs (kidney, lung) and the toxicokinetics after oral exposure are similar among species (including humans) W - Estimated absorption of cadmium in rodents is lower compared to humans, especially after prolonged exposure	While there are species specific differences in metallothionein, cadmium kinetics and toxicity, these differences are well established and the animal data (target organs/endpoints) are in support of human findings
Human data S – Consistent evidence that cadmium targets kidney after chronic exposure W – Cross sectional studies results effected by some degree of imprecision, while this can cause underestimation of true cadmium toxicity it may be less significant but bias in the same direction W – No firm conclusion on reversibility of renal damage, some data indicate possibility, others note glomerular dysfunction to progress even after contaminated soil replacement W – Imprecisions in cancer studies likely to have biased findings towards no effect; limited number of exposed workers, sparse historical data	Strong evidence that elevated levels of several biomarkers of renal dysfunction associated with cadmium burden, less agreement about the significance of these changes
Mechanistic data S – Link between the MoA and human data	

From the available data and the assessment of the strengths and weaknesses of the various lines of evidence

INSERT GRAPH

Figure X Visualisation of the causality of

Example 3: Tropane alkaloids (TAs)

The lines of evidence for the data integration were drawn from the Joint FAO/WHO Expert meeting on tropane alkaloids (2021). For further background information and a full review of tropane alkaloids, please see the published statement.

Effects caused by TAs, more specifically hyoscyamine and scopolamine are due to their competitive antagonistic binding/inhibition to muscarinic acetylcholine receptors (M_1-M_5) in the central nervous system (CNS) and autonomic nervous system (ANS). However, they differ in the ability to affect the CNS, (-)-scopolamine having a more prominent effect on the CNS.

However, hyoscyamine and scopolamine can also act as competitive antagonists at 5-hydroxytryptamine type-3 (5-HT3) receptors, which are excitatory, ligand-gated ion channels located throughout CNS and PNS. The application of atropine or scopolamine with 5-hydroxytryptamine (i.e., serotonin: a 5-HT3 receptor agonist) resulted in concentration-dependent inhibition of the serotonin-evoked response. Additionally, at high concentrations, atropine and scopolamine can also inhibit nicotinic acetylcholine receptors, binding affinities for atropine and scopolamine were much lower for the nAChRs compared to the mAChRs.

In contrast to atropine, scopolamine has been investigated in some detail in experimental animals and due to structural similarities and a common mode of action hyoscyamine is expected to exhibit similar pharmacological/toxicological effects. The LOAEL in mice and rats were 10.4 mg/kg bw per day and 0.69 mg.kg bw per day for short term and chronic oral toxicity, respectively, based on pupillary dilation; and in case of short term toxicity also decreased bw. Scopolamine was not carcinogenic in mice and/or rats and based on the available *in vitro* (negative mutagenicity and SCE, weakly positive clastogenicity at highest concentration) and *in vivo* (negative clastogenicity) data, the weight of evidence suggests that it is unlikely to exhibit genotoxicity *in vivo*. There is no clear evidence of developmental toxicity for scopolamine in mice or rats in the absence of maternal toxicity.

In humans, toxic effects of (-)-hyoscyamine and (-)-scopolamine include inhibition of saliva, bronchial and sweat gland secretion, dilation of pupils and paralysis of accommodation, change in heart rate, inhibition of urination, reduction in GI tone and inhibition of GI secretion. In extreme cases, toxic effects can include hallucination, delirium and coma. Death due to CNS depression, circulatory collapse and hypotension are rare but may also occur. Overtly toxic reactions to atropine, including death, have been reported following doses of approximately 100 mg or less in adults and 10 mg in children. Reported oral lethal doses of atropine sulphate have also been suggested as 10-20 mg/kg bw for adults and from 1-10 mg/kg bw for children.

Clinical applications of hyoscyamine/atropine and scopolamine include uses as mydriatic agent, to reduce secretion, as an anti-spasmodic for GI conditions, to

reduce excess salivation and to treat bradycardia and motion sickness. Recommended (maximum) therapeutic doses of atropine are 0.5 mg (children) and 1.5-3.0 mg (adults), recommended doses for scopolamine are 0.25-0.8 mg for adults and children. A LOAEL of 2 µg/kg bw for a single dose of scopolamine was identified in humans, based on a reduction in heart rate, similar effects were observed at 7 µg/kg bw for atropine (sulphate). Data from the use of scopolamine and atropine during pregnancy do not indicate any adverse developmental effects or significant fetotoxicity at therapeutic doses.

Table X Summary of the strengths and weaknesses of the data on TAs and the influence of the lines of evidence on the overall conclusion.

Lines of evidence and their main strengths (S) and weaknesses (W)	**Influence on Conclusion**
Animal data S – W –Pharmacological and toxicological studies with atropine largely uninformative, due to route of exposure, nature of effects and magnitude of dosing	
Human data S –	
Human poisoning data S – Confirmation of effects seen elsewhere W – Generally lack quantitative dose-response data W – Only provide conformation of the presents of plants parts in the food with self-reported intake estimates	While data not ideal due to lack of dose-response and predominantly self-reported intakes, in support of overall human data
Clinical data	
Mechanistic data S – Clear mode of action in experimental animal and human data; effects of TAs considered to be due to their competitive antagonistic binding/inhibition to muscarinic acetylcholine receptors (M1-M5) in the peripheral and central nervous system.	Key event can be linked to experimental animal and human data Common mechanism of action between hyoscyamine and scopolamine

From the available data and the assessment of the strengths and weaknesses of the various lines of evidence

INSERT GRAPH

Figure X Visualisation of the causality of caffeine intake and an increased risk of FGR.

References

Adami H-O, Sir Berry CL, Breckenridge CB, Smith LL, Swenberg JA, Trichopoulos D, Weiss NS, Pastoor TP (2011). Toxicology and epidemiology: Improving the science with a framework for combining toxicological and epidemiological evidence to establish causal interference. Toxicological Science, 122(2): 223-34.

Aptula AO and Roberts DW (2006). Mechanistic applicability domains for nonanimal-based prediction of toxicological end points: General principles and application to reactive toxicity. Chemical Research in Toxicology, 18(8):1097-1105.

Aylward, L L (2018). Integration of biomonitoring data into risk assessment. Current Opinion in Toxicology, 9: 14-20.

Bartels M, Rick D, Lowe E, Loizou G. Price P, Spendiff M, Arnold S, Cocker J, Ball N (2012). Development of PK- and PBPK-based modelling tools for derivation of biomonitoring guidance values. Computational Methods and Programs in Biomedicine, 108(2): 773-88.

Bell SM, Chang X, Wambaugh JF, Allen DG, Bartels M, Brouwer KLR, Casey WM, Choksi N, Ferguson SS, Fraczkiewics G, Jarabek AM, Ke A, Lumen A, Lynn SC, Paini A, Price PA, Ring C, Simon TW, Sipes NS, Sprankle CS, Strickland J, Troutman J, Wetmore BA, Kleinstreuer NC (2018). In vitro to in vivo extrapolation for high throughput prioritization and decision making. Toxicology in Vitro, 47: 213-27.

Bevan R, Angerer J, Cocker J, Jones K, Koch H M, Sepai O, Schoeters G, Smolders R, Levy L (2012). Framework for the development and application of environmental biological monitoring guidance values. Regulatory Toxicology and Pharmacology, 63(3): 453-460.

Blaauboer BJ, Boobis AR, Bradford B, Cockburn A, Constable A, Daneshian M, Edwards G, Garthoff JA, Jeffery B, Krul C, Schuermans J (2016). Considering new methodologies in strategies for safety assessment of foods and food ingredients. Food and Chemical Toxicology, 91: 19-35.

Van der Bles AM, van der Linden S, Freeman ALJ, Mitchell J, Galvao AB, Zaval L, Spiegelhalter DH (2019), Communicating uncertainty about facts, numbers and science. The Royal Society.
https://royalsocietypublishing.org/doi/10.1098/rsos.181870

Boobis AR, Cohen SM, Dellarco V, McGregor D, Meek ME, Vickers C, Willcocks D, Farland W (2006). IPCS framework for analyzing the relevance of a cancer mode of action for humans. Critical Reviews in Toxicology, 36(10):781-92.

Boobis AR, Doe JE, Heinrich-Hirsch B, Meek ME, Munn S, Ruchirawat M, Schlatter J, Seed J, Vickers C (2008). IPCS framework for analyzing the relevance of a noncancer mode of action for humans. Critical Reviews in Toxicology, 38(2):87-96.

Boobies AR, Cohen SM, Dellarco VL, Doe JE, Fenner-Crisp PA, Moretto A, Pastoor TP, Schoeny RS, Seed JG, Wolf DC (2016). Classification schemes for carcinogenicity based on hazard-identification have become outmoded and serve neither science nor society. Regulatory Toxicologty and Pharmacology, 82: 158-66.

Boogaard P J, Hays S M, Aylward L L (2011). Human biomonitoring as a pragmatic tool to support health risk management of chemicals - Examples under the EU REACH programme. Regulatory Toxicology and Pharmacology, 59(1): 125-32.

Boyes WK, Moser VC, Geller AM, Benignus VA, Bushnell PJ, Kamel F (2007). Integrating epidemiology and toxicology in neurotoxicity risk assessment. Human and Experiemntal Toxicology, 26(4): 283-93.

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

Chechoway H, Pearce N, Kriebel D (2004). Research methods in occupational epidemiology. 2nd ed. New York: Oxford University Press.

Clewell H J, Andersen M E (1985). Risk Assessment extrapolations and physiological modelling. Toxicology and Industrial Health, 1(4): 111-31.

Cocker J, Jones K (2017). Biological Monitoring Without Limits. Annals of Work Exposures and Health, 61(4): 401-5,

Cohen SM (2019). Chemical carcinogenicity revisited 3: Risk assessment of carcinogenic potential based on the current state of knowledge of carcinogenesis in humans. Regulatory Toxicology and Pharmacology, 103: 100-5.

COT (2008). Statement on the reproductive effects of caffeine. https://webarchive.nationalarchives.gov.uk/20200803134928/https:/cot.food.gov.uk/cotstatements/cotstatementsyrs/cotstatements2008/cot200804

COT, COC (2015). Report of the synthesising epidemiological evidence (SEES) of the Committee on Toxicity and Committee on Carcinogenicity. https://cot.food.gov.uk/COTjointreports

Dourson ML, Felter SP, Robinson D (1996). Evolution of science-based uncertainty factors in noncancer risk assessment. Regulatory Toxicology and Pharmacology, 24(2 Pt 1):108-20.

ECETOC (2009). Framework for the integration of human and animal data in chemical risk assessment. Technical Report 104 https://www.ecetoc.org/uploads/Publications/documents/TR%20104.pdf

Eden J, Wheatley B, McNeil B, Sox H (2008). Knowing what works in health care: A roadmap for the nation. Washington, DC: The National Academies Press.

EFSA (2009). Scientific Opinion of the Panel on Contaminants in the Food Chain on a request by the European Commission on cadmium in food. The EFSA Journal, 980: 1-139. https://efsa.onlinelibrary.wiley.com/doi/epdf/10.2903/j.efsa.2009.980

EFSA (2011). EFSA Panel on the Contaminates in the Food Chain (CONTAM): Scientific Opinion on tolerable weekly intake from cadmium. EFSA Journal, 9(2): 1975. https://efsa.onlinelibrary.wiley.com/doi/epdf/10.2903/j.efsa.2011.1975

EFSA (2015). EFSA Panel on Dietetic Products, Nutrition and Allergies (NDA): Scientific Opinion on the safety of caffeine. EFSA Journal, 13(5): 4102. https://efsa.onlinelibrary.wiley.com/doi/epdf/10.2903/j.efsa.2015.4102

EFSA (2016). Update on the use of the benchmark dose approach in risk assessment. EFSA Journal, 15(1): 4658. https://efsa.onlinelibrary.wiley.com/doi/epdf/10.2903/j.efsa.2017.4658

EFSA (2017). Scientific Opinion of the PPR Panel on the follow-up of the findings of the External Scientific Report 'Literature review of epidemiological studies linking exposure to pesticides and health effects'. EFSA Journal 15(10): 5007 https://www.efsa.europa.eu/en/efsajournal/pub/5007

EFSA (2019). Guidance on the use of the threshold of toxicological concern approach in food safety assessment. Technical Report, EFSA Journal, EFSA Supporting Publication, 16(6). https://efsa.onlinelibrary.wiley.com/doi/10.2903/j.efsa.2019.5708

FAO and WHO (2020). Joint FAO/WHO Expert meeting on tropane alkaloids. Food Safety and Quality Series No 11, Rome. http://www.fao.org/documents/card/en/c/cb1857en

Goodmann JE, Prueitt RL, Harbison RD, Johnson GT (2020). Systematically evaluating and integrating evidence in national ambient air quality standards reviews. Global Epidemiology, 2, 100019.

Griesinger C, Casey W, Coecke S, Desprez B, Zuang V (2019). Chapter 4: Validation of alternative in vitro methods to animal testing: Concepts, challenges, processes and tools. https://ec.europa.eu/jrc/en/publication/chapter-4-validation-alternative-vitro-methods-animal-testing-concepts-challenges-processes-and

Gobeau N, Stringer R, De Buck S, Tuntland T, Faller B (2016). Evaluation of the GastroPlus™ advanced compartmental and transit (acat) model in early discovery. Pharmaceutical Research, 33(9): 2126-39.

Guo X and Mei N (2018). Benchmark dose modelling of in vitro genotoxicity data: A reanalysis. Toxicological Research, 34(4), p. 303-10.

Hart A, Gosling JP, Boobis A, Coggon D, Craig P, Jones D (2010). Development of a framework for evaluation and expression of uncertainties in hazard and risk assessment.

Hoffmann TC, Oxman AD, Ioannidis JP, Moher D, Lasserson TJ, Tovey DI, Stein K, Sutcliffe K, Ravaud P, Altman DG, perera R, Glasziou P (2017). Enhancing the usability of systematic reviews by improving the consideration and description of interventions. The British Medical Journal. https://core.ac.uk/reader/111028561?utm_source=linkout

Huang R, Xia M, Sakamura S, Zhao J, Lynch C, Zhao T, Zhu H, Austin CP, Simeonov A (2018). Expanding biological space coverage enhances the prediction of drug adverse effects in human using in vitro activity profiles. Scientific Reports, 8 (3783).

Jamei M, Marciniak S, Feng K, Barnett A, Tucker G, Rostami-Hodjegan A (2009). The Simcyp((R)) Population-based ADME Simulator. Expert Opinion Drug Metabolism Toxicology, 5(2): 211-23.

Kaltenhäuser J, Kneuer C, Marx-Stoelting P, Niemann L, Schubert J, Stein B, Solecki R (2017). Relevance and reliability of experimental data in human health risk assessment of pesticides. Regulatory Toxicology and Pharmacology, 88: 227-37.

Klimisch HJ, Andreae M, Tillmann U (1997). A systematic approach for evaluating the quality of experimental toxicological and ecotoxicological data. Regulatory Toxicology and Pharmacology, 25, 1-5.

Krewski D, Andersen ME, Tyshenko MG, Krishnan K, Hartung T, Boekelheide K, Wambaugh JF, Jones D, Whelan M, Thomas R, Yauk C, Barton-Maclaren T, Cote I (2020). Toxicity testing in the 21st century: Progress in the past decade and future perspectives. Achieves of Toxicology, 94(1), p 1-58.

Krishnan K, Andersen M E (1994). Physiologically based pharmacokinetic modeling in toxicology. In Principals and Methods of Toxicology, A. W. Hayes, 3 ed., pp. 149-88. Raven Press Ltd., New York.

Lavelle KS, Schnatter AR, Travis KZ, Swaen GMH, Pallapies D, Money C, Priem P, Vrijhof H (2012). Framework for integrating human and animal data in chemical risk assessment. Regulatory Toxicology and Pharmacology. 62: 302-12.

Lawlor DA, Tilling K, Davey Smith G (2016). Triangulation in aetiological epidemiology. International Journal of Epidemiology, 45, 1866-86.

McNally K, Cotton R, Cocker J, Jones K, Bartels M, Rick D, Price P, Loizou G (2012). Reconstruction of Exposure to m-Xylene from Human Biomonitoring Data Using PBPK Modelling, Bayesian Inference, and Markov Chain Monte Carlo Simulation. Journal of Toxicology,18.

McNally K, Cotton R, Loizou G (2011). A workflow for global sensitivity analysis of PBPK models. Frontiers in Pharmacology: Predictive Toxicity 2, Article 31, 1-21.

McNally K, Sams C, Loizou G (2019). Development, Testing, Parameterization, and Calibration of a Human Physiologically Based Pharmacokinetic Model for the Plasticizer, Hexamoll® Diisononyl-Cyclohexane-1, 2-Dicarboxylate Using In Silico, In Vitro, and Human Biomonitoring Data. Frontiers in Pharmacology, 10(1394).

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

Moreau M, Leonard J, Phillips K A, Campbell J, Pendse S N, Nicolas C, Phillips M, Yoon M, Tan Y-M, Smith S, Pudukodu H, Isaacs K, Clewell H (2017). Using exposure prediction tools to link exposure and dosimetry for risk-based decisions: A case study with phthalates. Chemosphere, 184: 1194-1201,

Negri E, Metruccio F, Guercio V, Tosti L, Benfenati E, Bonzu R, La Veccia C, Moretto A (2017). Exposure to PFOA and PFOS and fetal growth: A critical merging of toxicological and epidemiological data. Critical Reviews in Toxicology, 47(6):489-515.

Nestorov I. (2003). Whole Body Pharmacokinetic Models. Clinical Pharmacokinetics, 42(10): 883-908,

OECD (2016) Integrated Approaches to Testing and Assessment (IATA). http://www.oecd.org/chemicalsafety/risk-assessment/iata-integrated-approaches-to-testing-and-assessment.htm

OECD (2018). Guidance document on good in vitro method practice (GIVIMP). https://www.oecd-ilibrary.org/environment/guidance-document-on-good-in-vitro-method-practices-givimp_9789264304796-en;jsessionid=uS8qy0tkRsTA9CQtT7ZodxOm.ip-10-240-5-141

OECD (2020). Overview of concepts and available guidance related to integrated approaches to testing and assessment (IATA). Series on Testing and Assessment, No. 329, Environment, Health and Safety, Environment Directorate. http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=ENV/JM/MONO(2020)25&docLanguage=en

Pearl J (2009) Causality, Models and Inference. 2nd Edition, Cambridge University Press, Cambridge.

Pearl J, Mackenzie D (2018) The Book of Why. Penguin Random House, London

Pendse S N, Efremenko A, Eric Hack C, Moreau M, Mallick P, Dzierlenga M, Nicolas CI, Yoon M, Clewell HJ, McMullen PD (2019). Population Life-course Exposure to

Health Effects Model (PLETHEM): An R package for PBPK modelling. Computational Toxicology, 13: 100115.

Percie du Sert N, Hurst V, Ahluwalia A, Alam S, Avey MT, Maker M, Browne WJ, Clark A, Cuthill IC, Dirnagl U, Emerson M, Gerner P, Holgate ST, Howells DW, Karp NA, Lidster K, MacCallum CM, Macleaod M, Petersen O, Rawle F, Reynolds P, Rooney K, Sena ES, Silberberg SD, Steckler T, Würbel H (2019). The ARRIVE guidelines 2019: updates guidelines for reporting animal research. PLOS BIOLOGY, 18(7). https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.3000410

Peyret T, Poulin P, Krishnan K (2010). A unified algorithm for predicting partition coefficients for PBPK modelling of drugs and environmental chemicals. Toxicology and Applied Pharmacology, 249(3): 197-207.

Poulin P, Haddad S (2012). Advancing prediction of tissue distribution and volume of distribution of highly lipophilic compounds from a simplified tissue-composition-based model as a mechanistic animal alternative method. Journal of Pharmaceutical Science, 101(6): 2250-61.

Rodgers T, Leahy D, Rowland M (2005). Physiologically based pharmacokinetic modeling 1: predicting the tissue distribution of moderate-to-strong bases. Journal of Pharmaceutical Science, 94(6): 1259-76.

Rodgers T, Rowland M (2006). Physiologically based pharmacokinetic modelling 2: predicting the tissue distribution of acids, very weak bases, neutrals and zwitterions. Journal of Pharmaceutical Science, 95(6): 1238-57.

Rostami-Hodjegan A, Tucker GT (2007). Simulation and prediction of in vivo drug metabolism in human populations from in vitro data. Nature Reviews Drug Discovery, 6(2): 140-8.

Savitz, DA, Wellenius GA, Trikalinos TA (2019). The problem with mechanistic risk of bias assessments in evidence synthesis of observational studies and a practical alternative: Assessing the impact of sources of potential bias. American Journal of Epidemiology, 188(9), p 1581-85.

Schneider K, Schwarz M, Burkholder I, Kopp-Schneider A, Edler L, Kinsner-Ovaskainen A, Hartung T, Hoffmann S (2009). "ToxRTool", a new tool to assess the reliability of toxicological data. Toxicology Letters, 189: 138-44.

Schmitt W (2008). General approach for the calculation of tissue to plasma partition coefficients. Toxicology In Vitro, 22(2): 457-67.

Sonich-Mullin C, Fielder R, Wiltse J, Baetcke K, Dempsey J, Fenner-Crisp P, Grant D, Hartley M, Knaap A, Kroese D, Mangelsdorf I, Meek E, Rice JM, Younes M (2001). IPCS conceptual framework for evaluating a mode of action for chemical carcinogenesis. Regulatory Toxicology and Pharmacology, 34(2):146-52.

Steenland K, Schubauer-Berigan MK, Vermeulen R, Lunn RM, Straif K, Zahm S, Stewart P, Arroyave WD, Mehta SS, Pearce N. (2020). Risk of bias assessments and evidence syntheses for observational epidemiologic studies of environmental and occupational exposures: strengths and limitations. Environmental Health Perspectives. 112(13).

Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, Downey P, Elliott P, Green J, Landray M, Liu B, Matthews P, Ong G, Pell J, Silman A, Young A, Sproson T, Peakman T, Collins R (2015). UK Biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLOS Medicine, 12(3).

Tan C, Chan M, Chukwudebe A, Domoradzki J, Fisher J, Hack E, Hinderliter P, Hirasawa K, Leonard J, Lumen A, Paini A, Quian H, Ruiz P, Wambaugh J, Zhang F, Emcry M (2020). PBPK model reporting template for chemical risk assessment applications. Regulatory Toxicology and Pharmacolology, 115, 104691.

Tsamandouras N, Rostami-Hodjegan A, Aarons L (2015). Combining the 'bottom up' and 'top down' approaches in pharmacokinetic modelling: fitting PBPK models to observed clinical data. British Journal of Clinical Pharmacology, 79(1): 48-55.

Wilk-Zasadana I, BernasconiC, Pelkonen O, Coecke S (2015). Biotransformation in vitro: An essentiel consideration in the quantitative in vitro-to-in vivo extrapolation (QIVIVE) of toxicity data. Toxicology, 332, p. 8-19.

Yoon M, Campbell JL, Andersen ME, Clewell HJ (2012). Quantitative in vitro and in vivo extrapolation of cell-based toxicity assay results. Critical Reviews in Toxicology, 42(8), p. 633-52.

Glossary

Adverse outcome pathways (AOPs)	An AOP is a sequence of key events linking a molecular initiating event (MIE) to an adverse outcome (AO) through different levels of biological organisation. Construction of an AOP can: • Organize information about biological interactions and toxicity mechanisms into models that describe how exposure to a substance might cause illness or injury. • Suggest cell- or biochemical-based tests for pathway elements that could be used to develop testing strategies for targeted toxicity. • Identify steps in a toxicity mechanism that need improved characterization.
ARRIVE Guidelines (Animal Research: Reporting of In Vivo Experiments)	The ARRIVE guidelines are a checklist of recommendations to improve the reporting of research involving animals – maximising the quality and reliability of published research, and enabling others to better scrutinise, evaluate and reproduce it.
Benchmark dose modelling (BMDL)	An alternative quantitative approach to dose-response assessment using more of the data than the NOAEL process. This approach utilises mathematical models to fit all available data points and uses the best fitting model to interpolate an estimate of the dose (benchmark dose) that corresponds to a particular level of response (a benchmark response). A measure of uncertainty is also calculated, and the lower confidence limit on the benchmark dose is called the BMDL. The BMDL accounts for the uncertainty in the estimate of the dose-response that is due to characteristics of the experimental design such as sample size and biological variability. The BMDL can be used as the point of departure (see POD) for derivation of a health-based guidance value or a margin of exposure. (COT Glossary)
Bias	This is a specific term in epidemiology relating to problems in the study design that may affect the observed measure of association in the statistical analysis. Bias cannot be removed by including larger numbers and it cannot be adjusted for in the statistical analysis. The two main types of bias in epidemiological studies are selection bias and information bias (i.e. measurement error). For example, a study relying on occupational health records to investigate a specific exposure, will not have information on those who developed disease after they left their job (selection bias).
Biological plausibility	The causal consideration that an observed, potentially causal association between and exposure and a health outcome may plausibly be attributed to causation on the basis of existing biological and medical knowledge.

Biological relevance	An effect considered by expert judgement as important and meaningful for human, animal, plant or environmental health. It therefore implies a change that may alter how decisions for a specific problem are taken.
Bradford Hill considerations	Sir Austin Bradford Hill established a set of 'principles' (not to be taken as 'criteria') that may be used to assist in the interpretation of associations reported from epidemiological studies: - Strength – The stronger the association the more likely it is causal. The COC has previously noted that the relative risks of <3 need careful assessment for effects of bias or confounding. - Consistency – The association has been consistently identified by studies using different approaches and is also seen in different populations with exposure to the chemical under consideration. - Specificity – Limitation of the association to specific exposure groups or to specific types of disease increases likelihood that the association is causal. - Temporality – The association must demonstrate that exposure leads to disease. The relationship of time since first exposure, duration of exposure and time since last exposure are all important in assessing causality. - Biological gradient – If an association reveals a biological gradient or dose response curve, then this evidence is of particular importance in assessing causality. - Plausibility – Is there appropriate data to suggest a mechanism by which exposure could lead to concern? However, even if an observed association may be new to science or medicine it should not be dismissed. - Coherence – Cause and effect interpretation of data should not seriously conflict with generally known facts. - Experiment – Can the association be demonstrated experimentally? Evidence from experimental animals may assist in some cases. Evidence that removal of the exposure leads to a decrease in risk may be relevant. - Analogy – Have other closely related chemicals been associated with the disease?
Case-control studies	Case control studies compare individuals with a specific disease or outcome of interest (cases) to individuals from the same population that don't have that disease or outcome (controls). Studies aim to find associations between the disease or outcome and prior exposure to a particular risk factor but are prone to various biases. (Cochrane glossary, 2018)
(Social) Causation	TO BE FINALISED

Cohort	A defined population that continues to exist through a period of time, e.g. a group of individuals who had a specific occupation.
Cohort studies	A cohort study is an observational study in which a defined group of individuals (the cohort) is followed over time. The outcomes of individuals in subgroups of the cohort are compared, to examine individuals who were exposed or not exposed (or exposed at different levels) to a particular intervention or other factor of interest. These can be prospective or retrospective in nature. (Cochrane glossary, 2018)
Combined exposure	Exposure to multiple chemicals by a single or multiple route at the same or different times.
Confounder	A confounder is a factor that is independently associated with both an intervention (or exposure) and the outcome of interest. Failure to account for this will distort the observed measure of association in the statistical analysis. For example, if people in the experimental group of a controlled trial (or the exposed group) are younger than those in the control group, it will be difficult to decide whether a lower risk of death in one group is due to the intervention (or exposure) or the difference in ages. (Cochrane glossary, 2018).
Cross-sectional studies	For example, a survey. Information on outcome and exposures is taken at the same point in time. These are relatively easy to conduct, but it is more difficult to ascribe causality than in a cohort study.
Cytotoxicity	The quality of being toxic to cells.
Descriptive studies	A descriptive study describes the characteristics or health status of a sample of individuals. In this type of study, the investigators do not actively intervene to test a hypothesis, but just describe the health status or characteristics of a sample from a defined population. (Cochrane glossary, 2018).
Effect modification and effect modifier	Effect modification describes the situation where the magnitude of the effect of an exposure variable on an outcome variable differs depending on a third variable. In other words, the presence or absence of an effect modifier changes the association of an exposure with the outcome of interest.
Epidemiology	The study of the health status of populations and communities, not just particular individuals. (Cochrane glossary, 2018)
Evidence synthesis	Evidence synthesis involves the development of techniques to combine multiple sources of quantitative evidence. Synthesis techniques such as systematic reviews and meta-analysis, are increasingly being adapted and applied.

Experimental study	In this type of study, the investigators actively intervene to test a hypothesis. In a controlled trial, one type of experimental study, the subjects receiving the treatment being tested are said to be in the experimental group (or arm) of the trial. (Cochrane glossary, 2018)
Genotoxicity	Genotoxicity describes the property of chemical agents that damages the genetic information within a cell causing mutations, which may lead to cancer.
Good Laboratory Practise (GLP)	GLP is a quality system of management controls for research laboratories and organizations to ensure the uniformity, consistency, reliability, reproducibility, quality, and integrity of data in non-clinical safety tests, e.g. physio-chemical properties, acute to chronic toxicity tests.
Grading of Recommendations Assessment, Development and Evaluation (GRADE)	GRADE is a transparent framework for developing and presenting summaries of evidence and provides a systematic approach for making clinical practice recommendations.
Hazard	Set of inherent properties of a substance, mixture of substances or a process involving substances that make it capable of causing adverse effects to organisms or the environment.
Integrated Approaches to Testing and Assessment (IATA)	IATA are pragmatic, science-based approaches for chemical hazard characterization that rely on an integrated analysis of existing information coupled with the generation of new information using testing strategies.
Intervention studies	This type of study involves an intervention of people, groups, entities or objects in an experimental study. An intervention is sometimes used to describe the regimens in all comparison groups, including placebo and no-treatment arms in a controlled trial. (Cochrane glossary, 2018).
In silico	A term used to describe a computerized analysis of the structure of a chemical to assess its potential hazard.
In vitro	A Latin term used to describe studies of biological material outside the living animal or plant (literally “in glass”).
In vivo	A Latin term used to describe studies in living animals or plants (literally “in life”).
Key event (KE)	A measurable change in biological state that is essential, but not necessarily sufficient, for the progression from a defined biological perturbation toward a specific adverse outcome. KEs are represented as nodes in an AOP diagram or AOP network and provide verifiability to an AOP description.

Meta-analysis	A meta-analysis is the use of statistical techniques in a systematic review to integrate and quantify the results of included studies. This term is sometimes misused as a synonym for systematic reviews, where the review includes a meta-analysis. (Cochrane glossary, 2018).
Mode of Action	A biologically plausible sequence of key events leading to an observed effect supported by robust experimental observations and mechanistic data. It describes key cytological and biochemical events, i.e. those that are both measurable and necessary to the observed outcome, in a logical framework. It contrasts with mechanism of action.
Molecular initiating event (MIE)	The MIE is a key event that spans the disciplines of chemistry and biology, linking the chemical properties of a molecule to an interaction at a biological target.
Natural experiments	These are naturally occurring circumstances in which subsets of the population are exposed to different levels of a supposed causal factor, in a situation resembling an actual experiment where human subjects would be randomly allocated to groups, for example, accidental contamination of food or water with a substance. (International Epidemiological Association, 2008)
No observed adverse effect level (NOAEL)	The NOAEL denotes the level of exposure of an organism, found by experiment or observation, at which there is no biologically or statistically significant increase in the frequency or severity of any adverse effects of the tested protocol.
Observational studies	A non-experimental study - the investigators do not seek to intervene, and simply observe the course of events. Most epidemiological studies are observational. Changes or differences in one characteristic (e.g. whether or not people received the intervention of interest) are studied in relation to changes or differences in other characteristic(s), without action by the investigator. There is a greater risk of selection bias than in experimental studies. (Cochrane glossary, 2018).
Omics technologies	A scientific subdiscipline that combines the technologies of genomics and bioinformatics to identify and characterize mechanisms of action of known and suspected toxicants. The collective term ‘omics’ refers to the genomic (DNA sequence analysis) and post-genomic (e.g. transcriptomics, proteomics, metabolomics, epigenomics) technologies that are used for the characterization and quantitation of pools of biological molecules (e.g. DNA, mRNAs, proteins, metabolites), and the exploration of their roles, relationships and actions within an organism.

Physiologically based pharmacokinetic (PBPK) model	A mathematical model which is used to predict the absorption, distribution, metabolism and excretion of a chemical substance in humans.
Point of departure	In toxicology, the point of departure (POD) is defined as the point on a toxicological dose-response curve established from experimental data or observational data generally corresponding to an estimated low effect level or no effect level. The POD can then be used to calculate a toxicological reference dose (RfD). Points of departure include the BMD, BMDL, LOAEL, and carcinogenic potency estimates, such as the T25. (FAO/WHO, 2009a).
Protection goals	
(Q)SAR models	(Quantitative) structure-activity relationship models are mathematical models that can be used to predict the physicochemical, biological and environmental fate and properties of compounds from the knowledge of their chemical structure.
Randomised controlled trials (RCTs)	These are experiments in which two or more interventions, possibly including a control or no intervention, are compared through random allocation to study participants. Most trials assign one intervention to each individual but sometimes assignment is to defined groups of individuals (for example, in a household). Interventions may also be assigned within an individual (for example, in different orders or to different body parts). (Cochrane glossary, 2018).
Reference point	See “Point of Departure”
Risk assessment	Process of evaluating a potential hazard, likelihood of suffering, or any adverse effects from certain human activities. Comprised of the four aspects, hazard identification, hazard characterization, exposure assessment and risk characterization. Can be carried out retrospectively or prospectively.
Sensitivity analyses	An analysis used to determine how sensitive the results of a study or systematic review are to changes in parameters e.g. excluding earlier years, excluding studies with low quality scores from a meta-analysis, only including cohort studies.
Systematic review	A formalized review that has been prepared using a documented systematic approach to minimizing biases and random errors.
Threshold of toxicological concern (TTC)	A pragmatic, scientifically valid methodology to prioritise substances of unknown toxicity found in food for further evaluation. It is used when there are limited chemical-specific toxicity data and can be used for substances with or without

	structural alerts for genotoxicity and for cancer and non-cancer endpoints.
Triangulation	Triangulation is the practice of obtaining more reliable answers to research questions through integrating results from several different approaches, where each approach has different key sources of potential bias that are unrelated to each other (Lawlor et al., 2016).
Uncertainty intervals/estimates	The term uncertainty intervals is used to refer to confidence intervals. This is the measure of uncertainty around a statistical analysis result. There will be an upper and lower confidence limit. Most estimates use a 95% confidence interval which means that if a study were continually repeated the true value would be contained in 95% of the confidence intervals from those studies. (Cochrane glossary, 2018).
Uncertainty factors	Value used in extrapolation from a reference point (or POD), determined in experimental animals, to humans (assuming that humans may be more sensitive) or from a sub-population of individuals to the general population: for example, a value applied to the NOAEL to establish an ADI or TDI. The value depends on the size and type of population to be protected and the quality of the toxicological information available.
Weight of Evidence	This approach uses a combination of information from several independent sources of evidence (e.g. toxicological or genotoxicity data) to arrive at a conclusion regarding potential hazard (such as mutagenicity).

Organisational abbreviations

COC	Committee on Carcinogenicity of Chemicals in Food, Consumer Products and the Environment: Independent scientific committee that provides advice the government and government agencies on whether substances are likely to cause cancer.
COM	Committee on Mutagenicity of Chemicals in Food, Consumer Products and the Environment: Independent scientific committee that assesses and advises the government and government agencies on mutagenic risks to humans.
COMEAP	Committee on the Medical Effects of Air Pollutants: Provides independent advice to government departments and agencies on how air pollution impacts on health.
COT	Committee on Toxicity of Chemicals in Food, Consumer Products and the Environment: Independent scientific committee that provides advice to the government and

	government agencies on matters concerning the toxicity of chemicals.
EFSA	European Food Safety Authority: The agency of the European Union that provides independent scientific advice and communicates on existing and emerging risks associated with the food chain.
EURO ECVAM	European Union Reference Laboratory for alternatives to animal testing: An integral part of the Joint Research Centre (JRC), the science and knowledge service of the European Commission. Its mandate includes a number of duties to advance the Replacement, Reduction and Refinement (the Three Rs) of animal procedures.
FAO	Food and Agriculture Organization of the United Nations: A specialized agency of the United Nations that leads international efforts to defeat hunger and improve nutrition and food security.
FSA	Food Standards Agency: Independent government department working protecting public health and consumers' wider interests in relation to food in England, Wales and Northern Ireland.
IARC	International Agency for Research on Cancer: An intergovernmental agency forming part of the World Health Organization of the United Nations. Its role is to conduct and coordinate research into the causes of cancer. It also collects and publishes surveillance data regarding the occurrence of cancer worldwide.
ICCVAM	The Interagency Coordinating Committee on the Validation of Alternative Methods: The ICCVAM is composed of representatives from 17 U.S. federal regulatory and research agencies. Each of these regulatory and research agencies require, use, generate, or disseminate toxicological and safety testing information. • Increase the efficiency and effectiveness of U.S. federal agency test method review. • Eliminate unnecessary duplication of effort and share experience among U.S. federal regulatory agencies. • Optimize utilization of scientific expertise outside the U.S. federal government. • Ensure that new and revised test methods are validated to meet the needs of U.S. federal agencies. • Reduce, refine, or replace the use of animals in testing where feasible.

IPCS	International Programme on Chemical Safety: A collaboration between the World Health Organization, the International Labour Organization and the United Nations Environment Programme, to establish a scientific basis for safe use of chemicals and to strengthen national capabilities and capacities for chemical safety.
OECD	Organisation for Economic Cooperation and Development (OECD) Guidelines for the testing of chemicals: Tools for assessing the potential effects of chemicals on human health and the environment. Accepted internationally as standard methods for safety testing and assessment of chemicals.
PHE	Public Health England: An executive agency of the Department of Health and Social Care in the United Kingdom.
REACH	Registration, Evaluation, Authorisation and Restriction of Chemicals: A European Union regulation addressing the production and use of chemical substances, and their potential impacts on both human health and the environment.
UK Biobank	A large long-term biobank study in the United Kingdom which is investigating the respective contributions of genetic predisposition and environmental exposure to the development of disease. It began in 2006.
WHO	World Health Organization: Specialized agency of the United Nations responsible for international public health.

SETE membership list

Professor Alan Boobis (Chair)

Professor Alan Boobis is Emeritus Professor of Toxicology, National Heart & Lung Institute, Imperial College London. He was a member of Imperial College London (initially at the Royal Postgraduate Medical School, which merged with the College in 1997) for over 40 years, until his retiral in 2017. His main research interests lie in mechanistic toxicology, drug metabolism, toxicity pathway analysis and in the application of knowledge in these areas to risk assessment. He has published around 250 original research papers (H-index of 80~).

He is a member of a number of national and international advisory committees, including the Committee on the Medical Effects of Air Pollutants, the WHO Study Group on Tobacco Product Regulation (TobReg) and the WHO Chemical Risk Assessment Network Coordinating Group. He is a member/chair of JECFA (veterinary residues) and JMPR. He has been a member and deputy chair of the UK Advisory Committee on Pesticides, a member and deputy chair of the UK Committee on Toxicity (2003-2012), a member of the UK Committee on Carcinogenicity, the EFSA Panel on Contaminants in the Food Chain and a member and deputy chair of the EFSA Panel on Plant Protection Products. He has also served as a member of the HPA Board Sub-Committee for Radiation, Chemical and Environmental Hazards and the Veterinary Residues Committee.

He is a member (and a former chair) of the Board of Trustees of the International Life Sciences Institute (ILSI), a member (and a former vice-president) of the Board of Directors of ILSI Europe and a member (and a former chair) of the Board of Trustees of the Health and environmental Sciences Institute (HESI). He sits on several international scientific advisory boards, in both the public and private sectors. He is an honorary fellow of the British Toxicology Society, fellow of the British Pharmacological Society, recipient of the BTS John Barnes Prize Lectureship, honorary membership and Merit Award of EUROTOX, the Royal Society of Chemistry Toxicology Award, the Arnold J Lehman Award from the SOT, the Toxicology Forum Philippe Shubik Distinguished Scientist Award, and Officer of the British Empire (OBE).Professor Boobis was appointed chair of the Committee on Toxicity with effect from 1 April 2015 for 3 years and reappointed for a further 3 years from April 2018.

Professor Boobis was appointed chair of the Committee on Toxicity with effect from 1 April 2015 for 3 years and reappointed for a further 3 years from April 2021.

Dr Phil Botham

Dr Phil Botham is a Principal Science Advisor for Syngenta’s Product Safety Group. He is a toxicologist with over 40 years of experience in the safety assessment of chemicals. In addition to this part-time role, he is an independent consultant working

for Regulatory Science Associates. He is a Fellow of the British Toxicology Society and of the Royal College of Pathologists and has authored over 100 peer-reviewed publications.

Dr Gill Clare

Dr Gill Clare is currently an independent consultant and has over 30 years of experience in genetic toxicology, working across the university, health and private sectors. She specialises in the identification and characterisation of genotoxic hazards of substances to human health and has experience is performing risk assessment for substances, including those found in food. She is a member of the COC, a member of COM until recently and has served on VRC and HSAC (formerly ACHS).

Alison Gowers

Alison Gowers works in the Air Quality and Public Health (AQPH) team at Public Health England (PHE), advising on the health risks of air pollutants. She leads the Scientific Secretariat which supports the work of the independent expert advisory Committee on the Medical Effects of Air Pollutants (COMEAP). Since obtaining her MSc (Distinction) in Toxicology from Surrey University in 1996, Alison has worked on the hazard characterisation and risk assessment of chemical contaminants in the environment within both consultancy and government departments and agencies.

Dr Valentina Guercio

Dr Valentina Guercio works at Public Health England as a senior environmental scientist. After obtaining her PhD she worked for 6 years as a research fellow at the University of Milan and Mario Negri Institute in Milano. Her research interests are in the epidemiology of cancer and other chronic diseases and the identification of the major risk factors, including air pollution and environmental chemicals. This has been done by carrying out observational studies and systematic reviews and meta-analyses. She was also involved in national and international projects that aimed to combine the epidemiological and toxicological evidence in order to establish a causal relationship between exposures and outcomes.

Professor Gunter Kuhnle

Professor Gunter Kuhnle is a Professor of Nutrition and Food Science at the University of Reading. His research interest is the development of objective measures of exposure and dietary intake using a range of different analytical techniques. Further interests are the link between diet and health, in particular the health effect of polyphenols and the link between meat and cancer.

George Loizou

Dr George Loizou is a computational toxicologist with over 36 years' experience in quantitative, mechanistic, chemical toxicology. For the past 23 years, George has been engaged in strategic research for the Health & Safety Executive (HSE) and external customers investigating whether computational tools can be designed to exploit new technologies and mathematical modelling to provide a biologically based, quantitative chemical risk assessment. This work had focused on the use of physiologically-based pharmacokinetic (PBPK) modelling to analyse, quantify and explain toxicological data with the ultimate aim of replacing the current slow, inefficient and expensive animal-based chemical risk assessment paradigm. For the past 4 years, George had also been investigating developments in personalised medicine where data obtained in people may potentially be appropriate for occupational and environmental toxicology. The use of gene expression (transcriptomics), metabolite (metabolomics) data and bioinformatics could lead to the development of a 'next generation' approach to chemical risk assessment based on human data.

Dr David Lovell

Dr David Lovell is Emeritus Reader in Medical Statistics at St. Georges Medical School, University of London. He was previously Associate Director and Head of Biostatistics Support to Clinical Pharmacogenomics at Pfizer Global Research and Development in Kent, where he provided data management and statistical support to pharmacogenetics and genomics.

David has conducted and managed research programmes on genetics, statistics and quantitative risk assessment. Dr Lovell has been a member of COM since 2006 and the Chair of COM since 2012. He was a Member of COC from 2009 until 2012 and is now an ex officio member of COC. He has been a member of the Scientific Committee of EFSA and a member of the Independent Scientific Advisory Committee, an expert committee of the Medicines and Healthcare Products Regulatory Agency. He is a Board Member of UK NC3Rs. He is also currently a Member of the COT.

Professor Neil Pearce

Professor Neil Pearce joined the London School of Hygiene and Tropical Medicine (LSHTM) at the beginning of 2011, after working in New Zealand for the last 30 years. He originally trained in biostatistics, before moving over to do a PhD in epidemiological methods. Since the completion of his PhD in epidemiology in 1985 he has been engaged in a wide range of public health research activities. In 1988 he co-founded the Wellington Asthma Research Group (WARG) at the Wellington School of Medicine. In 2000 he established the Massey University Centre for Public Health Research. He is a Fellow of the Royal Society of New Zealand (FRSNZ) and the Academy of Medical Sciences (FMedSCi) and is currently Past-President of the

International Epidemiological Association (IEA). He is also currently a Member of the COC.

Professor Neil Pearce currently teaches epidemiology, biostatistics and public health courses at the LSHTM. He also teaches at the annual European Educational Programme in Epidemiology (EEPE) summer course, and on various IEA courses in developing countries.

He has a broad range of research interests with a common theme of applied epidemiological and biostatistical methods, particularly methods of study design and data analysis for non-communicable diseases (NCDs). In terms of substantive research, during 1980-1988 his main research interest was in occupational epidemiology, and during this time he co-authored the leading textbook of occupational epidemiology, published by Oxford University Press in 1989. During the 1990s, at the Wellington Asthma Research group, he conducted a wide range of research projects including the identification of the role of the asthma drug fenoterol in the New Zealand asthma mortality epidemic, studies of the management of asthma in the community, and more recently studies of the causes of the increases in asthma prevalence in New Zealand and worldwide. He co-authored a textbook of asthma epidemiology which was published by Oxford University Press in 1998. During his ten years at the Massey University Centre for Public Health Research, they conducted a wide range of public health research including respiratory disease, cancer, diabetes, Maori health, Pacific health and occupational and environmental health research. His current research interests focus on epidemiological and biostatistical methods, and their application to studies of neurological disease, occupational and environmental health, asthma, cancer, and health inequalities.

Dr Lesley Rushton

Dr Lesley Rushton is an epidemiologist/statistician with extensive research experience into occupational and environmental causes of ill health. She has worked in several UK academic institutions and is currently a Reader in Occupational Epidemiology at Imperial College London.

Dr Rushton has specialised in health studies in various industries, including leukaemia and related diseases from benzene in the petrochemical industry, studies of lung cancer and silicosis in the silica sand industry, and dermatitis in the printing industry. She has also carried out several studies in the area of indoor and outdoor air pollution, particularly in relation to children’s health. She led the study to estimate the current and future burden of cancer due to occupation in Britain. A major new project involves the design and application of an occupational module for UK Biobank. Dr Rushton’s methodological research includes systematic review and meta-analysis in the areas of risk assessment and cross-design synthesis. She has been a member of several UK government committees including the COT, the Industrial Injuries Advisory Council, and the EU Scientific Committee on Emerging

and Newly Identified Health Risks. She is currently a member of COC and the HSE Scientific and Engineering Assurance Committee.

Prof Mireille Toledano

Professor Mireille Toledano was appointed to the Committee on 1st April 2018. She is Professor of Perinatal and Paediatric Epidemiology at Imperial College London and is an investigator of the MRC-PHE Centre for Environment and Health, specialising in environmental epidemiology. As an academic public health scientist working in the UK's top ranked School of Public Health (in the latest RAE/REF exercise), she has devoted her professional life to conducting epidemiological research and risk assessment focusing on environmental chemicals and sources of pollution, with the goal of improving public health. She is a member of a number of other UK and international advisory committees. Professor Toledano is a reader in epidemiology at Imperial College London and an investigator of the MRC-PHE Centre for Environment and Health specialising in environmental epidemiology. As an academic public health scientist working in the UK's top ranked School of Public Health (in the latest RAE/REF exercise) for more than a decade, she has devoted her professional life to conducting objective and impartial epidemiological research with the goal of improving public health.

Professor Heather Wallace

Professor Heather Wallace is Professor of Biochemical Pharmacology and Toxicology at the University of Aberdeen. Her research interests are in carcinogenesis, cancer biology, cancer therapeutics and prevention, selective drug delivery and the use of biomarkers for diagnosis and monitoring efficacy of anticancer drug therapy. She co-ordinates postgraduate and undergraduate teaching at the University of Aberdeen in toxicology, pharmacology, drug discovery and cancer biology.

Heather is President of EUROTOX, which is a federation of the National Societies of Toxicology across Europe and is a strong advocate for raising the profile of European toxicology in the wider world and in the recognition of the European Registration and training process for toxicologists, the European Registered Toxicologist (ERT) status.

Externally, Heather works with the Medicines and Healthcare Regulatory Products Agency (MHRA) on the Paediatric Medicines Expert Advisory Group and on the Herbal Medicine Advisory Committee. She is a member of the Committee on Carcinogenicity of Chemicals in Food, Consumer Products and the Environment (COC) and works with the European Food Safety Authority (EFSA) where she is vice chair of the CONTAM Panel working in the risk assessment environment indicating commitment to health protection at human, animal and environmental level.

Heather is a member of the Scientific Programme Committee of the Society of Toxicology (USA), an advisor to IUTOX Scientific Programme Committee and a Past President of the British Toxicology Society. She is a Fellow of the Royal College of Pathologists and a Specialty Advisor for Toxicology at the College. She is also a Fellow of four other learned Societies and is a member of the UK Register of Toxicologists and a European Registered Toxicologist. Heather is Editor-in-Chief of Toxicology Research and a Trustee and Vice Chair of the Board of Trustees for Medical Research Scotland.