Adopted: 29 May 2024

DOI: 10.2903/j.efsa.2024.8866

efsa JOURNAL

SCIENTIFIC OPINION

Scientific Committee guidance on appraising and integrating evidence from epidemiological studies for use in EFSA's scientific assessments

EFSA Scientific Committee | Simon More | Vasileios Bampidis | Diane Benford | Claude Bragard | Antonio Hernandez-Jerez | Susanne Hougaard Bennekou | Konstantinos Koutsoumanis | Claude Lambré | Kyriaki Machera | Wim Mennes | Ewen Mullins | Soren Saxmose Nielsen | Josef Schlatter | Dieter Schrenk | Dominique Turck | Maged Younes | Tony Fletcher | Matthias Greiner | Evangelia Ntzani | Neil Pearce | Marco Vinceti | Martine Vrijheid | Marios Georgiadis | Andrea Gervelmeyer | Thorhallur I. Halldorsson

Correspondence:
sc.secretariat@efsa.europa.eu

Abstract

EFSA requested its Scientific Committee to prepare a guidance document on appraising and integrating evidence from epidemiological studies for use in EFSA's scientific assessments. The guidance document provides an introduction to epidemiological studies and illustrates the typical biases, which may be present in different epidemiological study designs. It then describes key epidemiological concepts relevant for evidence appraisal. This includes brief explanations for measures of association, exposure assessment, statistical inference, systematic error and effect modification. The guidance then describes the concept of external validity and the principles of appraising epidemiological studies. The customisation of the study appraisal process is explained including tailoring of tools for assessing the risk of bias (RoB). Several examples of appraising experimental and observational studies using a RoB tool are annexed to the document to illustrate the application of the approach. The latter part of this guidance focuses on different steps of evidence integration, first within and then across different streams of evidence. With respect to risk characterisation, the guidance considers how evidence from human epidemiological studies can be used in dose–response modelling with several different options being presented. Finally, the guidance addresses the application of uncertainty factors in risk characterisation when using evidence from human epidemiological studies.

KEYWORDS

epidemiological studies, evidence integration, exposure assessment, hazard characterisation, risk assessment, risk of bias

CONTENTS

1 | INTRODUCTION

1.1 | Background and Terms of Reference as provided by the requestor

Epidemiology is the study of the distribution and determinants of health-related states or events in specified populations, and the application of this study to the control of health problems. Therefore, in the broadest sense, epidemiological studies examine determinants of health and disease conditions in defined populations, including humans, animals and plants. Epidemiological studies include both experimental and non-experimental studies, the latter is often referred to as 'observational' studies.

Within EFSA's remit there are well-established procedures and guidelines covering the use of controlled animal experiments for chemical risk assessment and the use of double blind randomised controlled trials. Other sources of evidence include non-experimental studies for assessing potential harm or benefits of different factors (chemicals, nutrients, biohazards) in humans, animals, including both analytical and descriptive monitoring studies, and nutritional intervention studies that deviate from randomised controlled trial (RCT) designs. For these sources of evidence, guidance on how to use these studies in EFSA's work is either more limited or lacking. This is particularly the case for epidemiological studies in humans, which are often characterised by high variability and uncertainties related to ethical constraints on what interventions can be made and how information can be collected. Therefore, the way human epidemiological studies are conducted and the information they provide do not always fit into existing frameworks for traditional chemical risk assessment or other established procedures within EFSA.

In light of the identified needs, it is important that clear guidance be developed on how evidence from epidemiological studies can be appraised, integrated and used in EFSA's scientific assessments. Such guidance would enable all areas in EFSA's remit to better exploit all sources of evidence, while correctly accounting for their potential limitations. The Scientific Committee has recommended in 2013 and in 2016 that a cross-cutting guidance be developed on the appraisal and use of epidemiological studies. This recommendation was based on the observation that limited use is made of evidence from non-experimental studies in chemical risk assessment.

TERMS OF REFERENCE

This project will:

A. deliver a Guidance addressing the following terms of reference (ToRs):

1. Set the basis for giving guidance on how to appraise and interpret findings from different types of epidemiological evidence and its application in EFSA scientific assessments.
2. Provide guidance on how to appraise and integrate evidence from epidemiological studies of humans or animals for specific scientific assessment questions of the different EFSA panels. Particular emphasis should be given to areas where guidance is lacking.
3. Provide guidance on how to use evidence from epidemiological studies in EFSA scientific assessments.

   Particularly,

- In relation to safety of chemicals for human health: Provide guidance on how to appraise and use epidemiological evidence from experimental and non-experimental human studies in scientific assessments of chemicals.
- In relation to efficacy assessment for human health, animal and plant health: Provide guidance on how to appraise and use epidemiological evidence from experimental and non-experimental human and animal studies in scientific assessment of efficacy for chemical and biological agents.

B. Facilitate the implementation of the Guidance in EFSA's scientific assessments by providing:

- Info sessions.
- Trainings.
- Assistance from a cross-cutting WG (to be agreed at the adoption of the guidance).

1.2 | Interpretation of the Terms of Reference

The use of epidemiological studies affects risk assessment in a broad range of areas that fall under EFSA's remit, e.g. nutrition, toxicology, animal and plant health as well as biological hazards. It requires a good understanding of the strengths and limitations of different study designs, and the ability to evaluate studies individually, in a structured manner, and to assimilate and interpret evidence from relevant epidemiological studies.

This guidance will provide a brief introduction to the different types of epidemiological studies (Section 4.1) and explain the key epidemiological concepts that are relevant for evidence appraisal (Section 4.2) to address the **first ToR**.

To address the **second ToR**, the guidance will explain (1) the three main types of bias, i.e. information bias, selection bias and confounding, in specific study designs; (2) how to make judgements about the direction and the magnitude of the bias; and (3) how to deal with bias when it is identified in a study. Further, different approaches to study appraisal will be described. Finally, the guidance will explain the integration of evidence across different types of epidemiological studies (Section 4.3) within the same population (e.g. summary across all human observational studies or across all human experimental studies) and highlight the respective value of the available study designs in the context of the research question and the population under study.

The use of epidemiological studies, e.g. for establishing reference values (e.g. health-based guidance values (HBGVs) such as acceptable daily intake or tolerable daily intake) varies among different panels and depends to a large extent on their different types of scientific assessments (nutrition, toxicology, animal and plant health). Therefore, Section 4.4, addressing the **third ToR**, will focus on the panel-specific needs regarding use of epidemiological studies. It will also provide guidance on specific issues regarding evidence integration.

In terms of scope, this guidance will cover the appraisal of experimental and observational epidemiological studies, giving particular emphasis on studies with humans as target populations. The particularities of experimental studies in animals (livestock, companion animals) and plants are briefly explained in Sections 4.1.3 and 4.1.4. The appraisal of evidence from studies of laboratory animals and in-vitro studies, as well as guidance on plant or animal disease epidemiology are outside the scope of this guidance.

2 | AUDIENCE AND DEGREE OF OBLIGATION

The aim of this guidance is to facilitate multidisciplinary and integrative scientific assessments, in particular to facilitate better integration of epidemiological and toxicological data. The guidance provides a harmonised, but flexible framework that is applicable to all areas of EFSA's work and all types of scientific assessment, including risk assessment. In line with improving transparency, the Scientific Committee considers the application of this guidance to be unconditional for EFSA. Assessors have the flexibility to choose appropriate methods and the degree of refinement in applying them.

3 | DATA AND METHODOLOGIES

The concepts used for the development of this guidance are covered in standard textbooks in human (Lash et al., 2021), animal (Dohoo et al., 2009) and plant disease (Cooke et al., 2006; Madden et al., 2007) epidemiology, as well as textbooks covering more focused topics such as nutritional epidemiology (Willett, 2012). Published papers, book chapters and reports from the biomedical literature are referred to where appropriate in support of arguments, statements and examples.

Concerning methodology, this guidance is drawing on basic concepts and methodologies within epidemiology, explains them and provides recommendations on how to use them in the context of EFSA's work.

ToR 3 demands that the guidance addresses the specific scientific assessment questions of the different EFSA panels. Therefore, the experience of the different scientific panels of EFSA in using epidemiological studies in their scientific assessments, and the specific questions for which guidance was needed, were collected via a questionnaire submitted to the EFSA coordinators of all 10 panels and their chairs. The responses were used to refine the guidance.

4 | GUIDANCE

EFSA has published in recent years a number of cross-cutting guidance documents with the aim of further increasing robustness, transparency and openness of its scientific assessments. Altogether the documents cover major approaches to the use and interpretation of data and scientific evidence in risk assessments.

In the PROMETHEUS project ('PROmoting METHods for Evidence Use in Scientific assessments'), EFSA defined a set of principles for the scientific assessment process and a 4-step approach (plan/carry out/verify/report) for their fulfilment (EFSA, 2015), which was piloted in 10 case studies, one from each EFSA panel. According to PROMETHEUS, the process of **validating or appraising evidence** must be planned for, conducted consistently, verified and thoroughly documented. To do so, pre-defined criteria must be applied to all individual studies of the same design included in the assessment. This is important as study appraisal both informs and influences the integration process, where all potentially relevant data are considered and weighted together. Based on the results from the pilot phase, several limitations were identified, and recommendations were made (EFSA, 2018). These included:

- the lack of guidance and of agreed in-house appraisal tools;
- the need for standardised templates that account for the diversity of the evidence;
- the lack of expertise in appraising studies using structured approaches;
- the need for multidisciplinary working groups (WGs) of experts (statisticians, epidemiologists, domain experts).

In the 'Guidance on the assessment of the biological relevance of data in scientific assessments' (EFSA Scientific Committee, 2017), biological relevance is considered at three main stages of the process of dealing with evidence. In that document, it is stated that 'For each effect, the first step is to determine whether it is causally related to the exposure or treatment, for instance according to the Bradford Hill viewpoints (Hill, 1965)'. Therefore, even if aspects related to the reliability of the various pieces of evidence used in the assessment are outside the scope of this guidance document, evidence appraisal is acknowledged as being a necessary step to reach conclusions about exposure–health associations.

The Weight of Evidence (WoE) guidance document (EFSA Scientific Committee, 2017) provides a general framework for considering and documenting the approach used to evaluate and weigh the assembled evidence when answering the main question of each scientific assessment. This includes assessing the relevance, reliability and consistency of the evidence.

Lastly, the guidance on protocol development for EFSA generic scientific assessment (EFSA Scientific Committee, 2023a) lays out a harmonised and flexible framework for developing protocols that consists of two main steps. In the first step, problem formulation, the APRIO (Agent, Pathway, Receptor, Intervention and Output) approach is introduced to translate the ToR into assessment questions. The APRIO approach aims to bridge the challenge in applying the PICO/PECO (Population, Intervention/Exposure, Comparator, Outcome) approach to the EFSA remit. In the second step, protocol development, the evidence needs and the methods that will be applied for answering the assessment questions, including uncertainty analysis, need to be specified.

In line with the concepts and approaches set out in these guidance documents, this document can be considered an addition that addresses needs for specific guidance on appraisal and integration of evidence from epidemiological studies (Sections 4.3 and 4.4).

4.1 | Introduction on epidemiological studies

In recent decades, principles and methodology of epidemiology have undergone rapid development. This development has partly been driven by advancements in methods for collecting and analysing large scale data. Several definitions of epidemiology have been proposed, with older definitions often being narrower in scope, stating for example that 'Epidemiology is concerned with the patterns of disease occurrence in human populations and the factors that influence these patterns' (Lilienfeld & Lilienfeld, 1980). Acknowledging the broader scope of epidemiological research today, Porta in the Dictionary of Epidemiology (2014) defined epidemiology as:

> The study of the occurrence and distribution of health-related events, states and processes in specified populations, including the study of the determinants influencing such processes, and the application of this knowledge to control relevant health problems.

By this definition, epidemiology is not just the study of disease[1] but also the study of any health-related endpoints/outcomes,[2] including risk factors, surrogate outcomes, and biomarkers of exposure/effect. Indeed, epidemiology provides a set of tools and methodologies to describe outcomes of interest in defined populations. Such outcomes can be of a variety of types (e.g. infection, disease, immunity to specific diseases, presence of certain conditions such as raised blood pressure or blood lipids, hard disease endpoints such as stroke, cancer occurrence). Epidemiological studies may also cover the occurrence of outcomes in members of a population where a direct link with health has not been well characterised so far. The growing number of studies in humans and animals linking environmental exposures to the composition of gut microbiota (Clemente et al., 2012; Lee & Hase, 2014; Snedeker & Hay, 2012; Wolter et al., 2021) is one example of such studies.

The Dictionary of Epidemiology definition is relevant for the study of '…health-related events, states and processes…' in any population, these being either humans, animals or plants. Regardless of the type of populations under study, these populations need to be defined explicitly. Although there can be important differences in design and conduct, it follows that general epidemiological principles and considerations also apply in settings, which have traditionally been viewed as non-epidemiological. As an example, potential biases that can occur in studies in laboratory animals are often the same as those encountered in experimental studies in humans. Such similarities in methodology generally exist across different fields of epidemiology (animals, humans).

Although various classifications exist (Lesko et al., 2020), epidemiological studies can be broadly classified as either descriptive or analytical. In descriptive studies, patterns of exposures or outcomes of interest are described across one or more factors, such as over time and place, while in analytical studies, relationships between identifiable factors and outcomes of interest are quantified. Analytical studies can be classified as either experimental or non-experimental studies, with the latter often referred to as observational studies.

[1]Disease is 'a pathological process, acute or chronic, inherited or acquired, of known or unknown origin, having a characteristic set of signs and symptoms, which are used for its diagnosis'. The diagnosis of a disease relies on widely accepted, well-defined criteria (i.e. the criteria used for diagnosis are widely accepted by the medical community and can be verified by a physician) (EFSA Scientific Committee, 2024).

[2]The terms endpoint and outcome are used interchangeably in this document, taking into consideration the evolution of the definitions over time (Ferreira & Patino, 2017; McLeod et al., 2019).

4.1.1 | Descriptive epidemiological studies

Descriptive epidemiological studies have the objective of describing and/or comparing the occurrence of exposure or outcome in a population (e.g. humans, livestock or companion animals, plants) over factors such as time and space. When all members of a defined population can be examined, i.e. when a census is possible, the characteristic(s) of interest can be determined directly. In practice, this is often not possible, for logistical or other reasons. In those cases, surveys need to be conducted, in which a sample is taken from the population of interest and their characteristics are then measured. These include prevalence or surveillance surveys of specific characteristics. The value of the estimates resulting from such surveys also depends on the representativeness of the sampling in relation to the scope of the survey. Examples of such studies of relevance for EFSA are The European Union One Health 2022 Zoonoses Report (EFSA and ECDC, 2023a); reports on antimicrobial resistance (EFSA and ECDC, 2023b); surveys for plant harmful organisms ('pests') relevant to the EU's plant health policy for which EFSA provides survey data sheets (EFSA, 2020) and reports on pesticide residues in food (EFSA, 2020) and dietary surveys (Ioannidou et al., 2021).

4.1.2 | Analytical epidemiological studies

A brief description of the most common designs of analytical epidemiological studies, both experimental and observational, is given in this section. Different designs exist due to their abilities to extract information under varying experimental or non-experimental (observational settings) for health outcomes of different frequency, severity and with varying latency periods.

4.1.2.1 | *Experimental studies*

Experimental studies (also named 'intervention studies' or 'trials') are primarily confined to experiments where the exposure conditions are controlled by the researcher to examine what effect an intervention may have on the population under study. In **Randomised Controlled Trials (RCTs)**, factors that may affect the outcome are (on average) balanced out by randomly allocating study participants to different treatments (two or more groups). At the end of the experiment, the groups are then compared with respect to the outcome of interest (parallel design). The unit of randomisation can either be the individual or a group of individuals within the study population (cluster randomisation). Examples of clusters are school units, families and neighbourhoods. Cluster randomisation may be chosen to serve convenience, to overcome ethical concerns raised by individual randomisation, to avoid departures from treatment or to assess group effects. Examples of experimental studies of randomised design include pharmaceutical trials, trials with foods, including dietary supplements, and changes in dietary patterns and toxicological studies in experimental animals.[3] The sample size needs to be sufficiently large to allow matched/stratified analyses and to better control for confounders, and to ensure a sufficient precision to the risk (or effect) estimates that are expected to be generated by the study (Rothman & Greenland, 2018). In humans, variability in lifestyle and genetic factors is generally high. As a result, larger sample sizes are needed, as compared to experimental studies in laboratory animals.

In RCTs, random treatment allocation alone is, however, not sufficient to achieve unbiased results. Blinding the investigators and caretakers (e.g. in the case of children and animals) to treatment assignment and outcome detection and assessment is essential to avoid bias resulting from unintended differences in co-intervention of the experimental groups or differences in the assessment of the outcome. For participants, being blinded to the treatment received is equally important to avoid bias from selective dropout, changes in behaviour or departures[4] from the assigned treatment (Dodd et al., 2011, 2012). When both investigators and study participants are blinded to treatment, these studies are traditionally referred to as **double blind RCTs**. If appropriately designed and conducted, they are expected to provide an unbiased measure of effect (gold standard).

Several variants of the RCT design exist. The simplest variant is when double blinding cannot be achieved. This is the case for many interventions, such as those based on nutritional and lifestyle changes, that test the efficacy of treatments when the exposure of interest cannot be masked. For example, for many foods and dietary components, such as fish oils or different types of sweeteners, participant blinding is difficult or impossible to achieve, while blinding of the investigators can still sometimes be ensured. Similar problems arise in many cognitive and physical activity interventions. Lack of blinding at the participant level may lead to selective dropout and differential departures from the assigned treatment, outcome detection bias, or confounding. Another design is the **crossover trial** (either randomised or not) where each participant receives both (or all) treatments in a sequential order, with a suitable 'wash out' period in between. This design has the advantage that each participant acts as its own control, which is more effective in balancing external factors than comparing different participants randomly allocated to two or more treatment groups. The limitation of this design is that it is only suitable for treatments where the anticipated effects are short term and fully reversible. That is, no carry-over effects between treatments are expected to occur, and response to treatment can be assumed to be independent of the order in

[3]Toxicological studies in experimental animals are usually not classified as epidemiological studies, but their design is similar to that of experimental studies in humans (i.e. RCTs).

[4]Departures here include participants'non-adhearance to treatment (not complying or simply doing something else) or changes in the prescribed treatment by the investigator.

which it is assigned. This design is often used when comparing the short-term effects of different treatments on clinical biomarkers, such as blood pressure or blood lipids. A variant of the crossover design in occupational settings is when the researcher changes workers' exposure by removing them temporarily (or permanently) from their workplace (or assigning them to other tasks), to see if their health conditions (such as asthma or allergies) improve (partial crossover design).

Other types of intervention studies that are relevant for the area of food safety are the so-called *Phase 0, I and II clinical trials*. These trials are conducted to assess safety, pharmacokinetics and -dynamics of pharmaceuticals in humans prior to conducting larger scale RCTs (Phase III trials). One characteristic of their design is that they may not include a well-defined control group for comparison and the study population can be quite different from the target population that the intended treatment is designed for. In **Phase 0 trials**, a group of healthy participants are given micro-doses of the test substance. Such trials are aimed at detecting potential adverse effect at low doses and/or provide relevant information on pharmacokinetics. General conclusions on the effect of the exposure are, however, hampered as this design does not include a control group. Further, healthy participants may be less likely to respond to treatment compared to more sensitive individuals. In **Phase I trials**, often called 'dose escalation trials', participants are dosed in small groups going from low to high doses to assess the safety or tolerability of the test substance. These studies may involve sensitive sub-groups (patients) and they provide valuable information on both pharmacokinetics and tolerability. However, in terms of evaluating the potential health effects across dose groups, the small number of participants per dose and possible dropout due to adverse events means that randomisation across dose groups is variably successful. As a result, bias (confounding) may occur. **Phase II trials** are designed to test therapeutic doses of the test substance often in sensitive individuals. These trials are of smaller scale (sample size) than Phase III trials and vary by design in terms of use of controls (currently preferred medication or placebo). Although Phase 0, I and II trials are mostly used for pharmaceuticals, these designs (or variants of these designs) may cover exposures falling under EFSA remit. An example of such studies includes a Phase 0 trial studying the pharmacokinetics of bisphenol A (Völkel et al., 2002), a Phase I dose escalation trial of caffeine (Altman et al., 2011) and advantame (Warrington et al., 2011) and in the area of novel foods a Phase II trial examining the possible therapeutic effects of flavanol-containing cocoa (Balzer et al., 2008).

Experimental studies involve a variety of **ethical considerations**. An extensive and rigid framework of ethical standards is in place, and it is constantly evolving based on new developments and their challenges. These ethical standards aim to safeguard the participants' safety, autonomy, and equal and respectful treatment within the experimental study (World Medical Association, 2013). Even when no apparent harm (side effect) is expected, such as in preventive interventions, the design of the study should ensure the best interest of the participants, including active surveillance for unexpected adverse events. In several cases, experimental studies aimed at testing beneficial effects of presumably non-harmful doses of nutritional substances at low doses, such as micronutrients and other food supplements, have shown unexpected harmful effects (Blumberg & Block, 1994[5]; Lippman et al., 2009[6]; Kristal et al., 2014[7]). These examples clearly highlight the importance of being cautious and maintaining high ethical standards when conducting experimental studies.

4.1.2.2 | *Non-experimental epidemiological studies*

In non-experimental (observational) epidemiological studies, the researcher does not control the circumstances or the amount of exposure. Instead, the researcher observes the outcome of interest in a given population, whose members may have been exposed to certain factors, inadvertently or by choice. The exposure of interest is observed (and quantified, where possible) and its relationship with the studied outcome assessed. The level and variation of the observed exposure reflect how participants have been exposed within their surroundings, which includes occupation and differences in dietary habits and other factors. Associations between exposures and outcomes are identified from such studies, but it needs to be ascertained whether the observed associations are attributed correctly to the exposure of interest. In fact, confounding may occur if other determinants of the outcome are not randomly associated with the exposure. For example, a study may find that elderly people with serum 25(OH)D (vitamin D) above 75 mmol/L perform better on physical function tests than those with vitamin D status below 50 mmol/L. Such an association may be confounded by the simple fact that those participants who are healthier (less frail) may spend more time outdoors, and therefore have higher serum vitamin D level, at least partly due to their exposure to the sun. If the observed association between physical function tests and level of vitamin D is attributed to vitamin D, confounding may occur. In practice, it is usually impossible to record and fully account for all factors that may influence the outcome. However, a possible replication of findings across different study populations with support from other experimental findings in vivo and/or in vitro, a low risk of bias (RoB) in such studies, and biological plausibility of the observed association between exposure and outcome(s) may support a stronger case for or against causality. Caution should be taken that the same biases may consistently exist in different studies across varying populations.

The main observational epidemiological study types are cohort, case–control, cross-sectional and ecological studies. These designs differ mainly in terms of selection of study participants, the timing between assessment of exposure and the outcome; and whether one or the other is assessed on an individual or group level.

[5]A prevention trial examining the effects of vitamin E and beta-carotene supplementation on reducing the incidence of lung cancers in male smokers.

[6]A prevention trial to examine the effects of selenium and Vitamin E on prostrate cancer and other cancers.

[7]A case–cohort study investigating effects of selenium and vitamin E supplementation on prostrate cancer risk conditional upon baseline selenium status.

In **cohort study designs,**[8] a source population is defined, and participants (the study population) are classified according to their exposure(s). Participants are then 'followed-up' for a specified period of time (the risk period), during which the outcome of interest is evaluated and compared across the exposure groups, while taking potential confounding factors into consideration. One advantage of this design is that it can be ascertained at the beginning of the study whether participants are free of the outcome of interest without the risk of differential misclassification depending on the outcome status. After a follow-up time considered sufficient to cover the known or assumed induction period of the outcome (or disease), it can then be examined if the exposure may have contributed to the development of the outcome. Frequently used variants of cohort studies are the case–cohort and cohort-nested case–control studies. These are generally more compact designs requiring smaller number of study participants and are often used for efficiency reasons, for example when chemical analyses or clinical assessment cannot be performed for all cohort participants (O'Brien et al., 2022).

In general, cohort studies are more resource demanding and difficult to conduct than other types of epidemiological studies, and the time it takes to generate results depends on the induction period of the outcome. Several large cohorts have been created to address many different exposure – outcome associations, including rare diseases, that are studied over time (e.g. the EPIC project,[9] the Danish National Birth Cohort,[10] the Avon Longitudinal Study of Parents and Children,[11] UK Biobank[12]).

Epidemiological studies frequently distinguish their findings among those related to the primary endpoint(s) or hypotheses. Different endpoints, related to additional objectives, are often added over time to the original study protocol, and this applies to both observational studies as well as experimental studies. In addition, studies presenting numerous disease outcomes may or may not adjust the presented *p*-values for multiple testing when different hypothesis are tested within the same study. Such adjustments are, however, generally not performed to account for different hypotheses presented in different studies. In general, a distinction between primary and secondary study hypotheses in terms of internal validity is useful in interpreting effect estimates (or published *p*-values). When primary and secondary endpoints are inter-related, they may re-enforce each other in terms of biological plausibility. Both types of endpoints are worth considering, particularly for assessing findings across studies. In such cases, the distinction between primary or secondary endpoints is less relevant.

Cohort studies can be **prospective, historical** (retrospective) **or a combination of both**. In prospective cohort studies, information on exposures is collected prior to assessment of the outcome while for historical studies the exposure and/or the outcome are assessed back in time (retrospectively). However, even a historical cohort study may involve exposure information that was recorded prospectively, e.g. a historical occupational cohort study may involve following participants over several years from recruitment, but the exposure information may be based on archived blood samples, clinical or other records collected prior to recruitment at the time that the relevant exposures occurred. In historical cohort studies, the exposure has already occurred before the study, but the outcome has yet to occur. This is a useful setup for assessing health outcomes with a long induction period and exposures that could trigger several outcomes of interest (Lazcano et al., 2019).

Case–control studies recruit participants based on the outcome of interest. That is, participants with a certain disease or health state (cases) and an appropriate group of participants that do not have such condition at the time of enrolment (controls) are recruited from the same source population. Thus, a case–control study involves studying cases (from a specific source population) and a sample of non-cases (ideally from the same source population). The distribution of past or current exposures among cases and non-cases (controls) is then compared, adjusting for confounding factors. Further details on the different types of case–control studies can be found in the paper of Knol et al. (2008).

It is important that selection of controls is conducted at random, i.e. that controls are a random sample of the source population over the risk period, with the qualification that controls may be matched to cases on some key factors such as age and gender. The strength of this design is its efficiency compared to the cohort design. In fact, case–control studies should be viewed in the context of a specific source population, in the sense that all cases from this population or a representative sample of these – over a defined period of time – are included in the study, and only a sample of non-cases is taken from the population. This sample of non-cases is used to estimate the distribution of exposures and confounders of interest in the source population from which the cases arose. The gain in efficiency of the case–control design derives from the fact that in a cohort study of the same source population, the entire population would have been studied. This gain is even more pronounced if the outcome under study is rare.

Similarly, case–control studies may be based on historical records or may involve interviews about historical or current exposures. The latter approach may result in problems if the health condition influences quantification of current or past exposures. For example, cancer cases may recall their past exposure differently than non-cases, even in situations when the exposure being assessed is not causally related to their disease condition (the same holds for many other health conditions). In addition, differences in the presence of certain health conditions among cases and controls, such as impaired kidney function or inflammation, can influence the measured concentrations for many biomarkers of exposure. In summary,

[8]We consider here typical cohort designs with two or more exposure groups, although most of the text equally applies to single-arm cohort studies, census studies or panel studies or any longitudinal studies (such as birth cohorts) in which exposure groups are not defined at the start of the study.

[9]https://pubmed.ncbi.nlm.nih.gov/9126529/.

[10]https://www.dnbc.dk/.

[11]http://www.bristol.ac.uk/alspac/.

[12]https://www.ukbiobank.ac.uk/.

the presence of certain health conditions when exposure is being assessed can create a spurious correlation between the quantified exposure and the health outcome under consideration, e.g. reverse causation. Assessing the exposure prospectively prior to the occurrence of the outcome should reduce the risk of such spurious correlation.

It is a common misunderstanding that case–control studies are always of lower value compared to cohort studies or randomised trials due to increased RoB. Past exposures can often be accurately assessed retrospectively through archived biomaterials stored in biobanks, or through access to high-quality health records or other similar sources, e.g. registries. If past exposures can be assessed in such manner, with appropriate temporal separation in relation to the outcome assessment, the RoB due to the exposure assessment should be comparable to that of a prospective design. Thus, for case–control studies the RoB is largely determined by (differential and non-differential) exposure misclassification, i.e. by how and when the exposure was assessed (retrospectively based on participant recall, cross-sectional or assessment of past exposures from high-quality records) and selection bias.

Other types of observational epidemiological study designs include the cross-sectional design and the ecological study design. In **cross-sectional studies**, a group of participants is recruited at one specific point in time, and information on both outcome and exposure is ascertained simultaneously. By design, it is often not possible to ascertain whether the exposure occurred before the outcome; therefore, the directionality of the observed association is often uncertain. That is, in some cases, the outcome (health state) itself, directly or through behavioural changes, may influence the parameter being assessed as exposure, as a result of reverse causation. The risk of such bias strongly depends on the time period that the measured exposure reflects and the health outcome under consideration. This has to be evaluated on a study-by-study basis. Still, a cross-sectional design may often be appropriate, such as in cases when exposure has short-term effects or for hypothesis generation. For example, for relatively rare exposures such as consumption of glycyrrhetinic acid from liquorice, which affects blood pressure (Sigurjónsdóttir et al., 2001), a simple cross-sectional study recording consumption for the past day and measuring blood pressure at the same time would be more appropriate than a cohort design that prospectively correlates exposures recorded in the previous year to current blood pressure. Additionally, no problems with reverse causation would exist in cross-sectional studies, for risk factors that do not change (e.g. blood type, genetic factors, etc.).

Finally, in **ecological studies,**[13] the units of observation are groups of participants defined, for example by region or community. Health-related states and exposures are measured, for example by rates in geographical areas, and their relation is examined. Limitations of these studies are lack of individual assessment and difficulty in accounting for confounders on a group level. Since the exact status of each member of the population (either in terms of exposure or in terms of outcome, or both) cannot be ascertained, the 'ecologic fallacy' may be produced. That is, the relationship between averages of population exposures and outcomes may not represent the relationship between exposure and outcomes at the individual level. However, despite this limitation, ecological studies sometimes have the advantage of achieving large exposure gradients, as exposure to certain nutrients or contaminants is generally greater across different units of observation than within individual units (Willett, 2012).

In summary, each of the observational study designs reviewed above has its strengths and limitations. Despite case–control and particularly cohort studies being generally considered to provide a higher certainty of evidence, for certain exposures and outcomes also cross-sectional and ecological studies can provide valuable information for risk assessment, complementing other lines of evidence.

4.1.3 | Epidemiological studies in animals

The basic principles of design and analysis of epidemiological studies are the same for human, animal and plant populations. Veterinary epidemiology, although based on the same methodological and study design principles as human epidemiology, often has different challenges to address, while some aspects of the execution of epidemiological studies may be simpler in livestock or companion animals rather than in human populations. For example, compared to humans, animal populations are sometimes easier to access, observe, control, test and follow-up. On the other hand, not all animals are individually identifiable, as is the case in intensively reared chicken, fish or wildlife. In those cases, probability sampling of populations and formation of study groups of individuals can prove very challenging or impossible. Additionally, exposures, outcomes and confounders may not be possible to assess at the individual level. In those cases, it may be necessary to use an entire group of animals (population of an entire fish tank, or an entire room of broilers, etc.) as the unit of the epidemiological study (in which case the exposures, outcomes and confounders are assessed at the group level). Sometimes, it may be feasible to introduce manipulations that make individual identification of animals possible, but this may not necessarily be part of the usual routine of animal rearing, and therefore it could affect the study findings.

As in any other branch of epidemiology, the definition of the target population, study population, enrolment process and sampling when dealing with animal populations is made with regard to the study objective, feasibility and bias minimisation. Studies on companion animals can have more similarities with human studies than studies on farm animals or wildlife. The hierarchical structure of farmed animal populations (e.g. different levels of organisation and possible social structures of such populations, clustering within production or housing units, litter, etc.) requires consideration in the design of the study and use of appropriate statistical methodology when analysing its results. Studies on wildlife are typically

[13] Ecological studies are generally not considered as descriptive studies as they consider the association between exposure and outcome.

restricted to descriptive or cross-sectional designs. Unique challenges exist on ascertainment of cases in wildlife studies when the entire population is not easily accessible. This is because observation of animals with the condition under study can, in those cases, be very challenging or dependent on other factors. For example, sick or dead wild animals may not be found, unless they are close to routes of human movements without ever being observed. The estimation of population sizes in these cases is a study objective in its own and requires specific methods (e.g. using capture–recapture). Information on population size is required as a denominator in measures of disease frequency. Population size, on the other hand, is usually not a challenge in farmed animal production systems (except sometimes when entire production units, or even entire farms, are the epidemiological unit). In all cases, daily operation of the system and the planning of the production need to be taken into consideration when conducting the study.

In veterinary epidemiology, obtaining exposure, disease and confounder information needs to focus on animal owners, breeders or farmers or on records or proxy measurements; therefore, the reliability of these sources of information always needs to be assessed. Distortions due to human behavioural or cognitive factors (compliance, non-response, recall and other intentional or non-intentional interferences with sampling, treatment or diagnosis) may still occur and, therefore, influence exposure or outcome assessments, treatment of study animals and other aspects of the epidemiological study.

4.1.4 | Epidemiological studies in plants

In plant health, an epidemic has been defined simply as 'the change in intensity of a disease in time and space' (Madden et al., 2007). Plant health focuses mainly on infectious disease (rather than non-communicable disease). A considerable number of plant health threats are caused by the invasion and spread of herbivorous insect populations in addition to pathogenic microorganisms, and the EFSA Plant Health (PLH) Panel thus operates at the intersection of epidemiology and population ecology. Consequently, fields of study relevant to the PLH Panel can be found in the study of infectious disease of humans and animals (e.g. Diekmann & Heesterbeek, 2000), and invasive species and entomology (e.g. Cock & Wittenberg, 2001). There is also a very strong focus on the environmental drivers of insect pest and pathogen populations in plant health, which are a major contributing factor to epidemics. Indeed, plant pest risk is usually viewed through the lens of the 'disease triangle' where there must be overlapping availability of host, pathogen and conducive environmental conditions for an epidemic to occur, with particular emphasis on the latter (Madden et al., 2007). In contrast to human and, to a lesser extent, animal disease epidemiology, plant health is concerned with a very large number of wild and domesticated host species. For example, *Xylella fastidiosa*, a current major plant health threat in the EU, is known to infect over 696 plant species (EFSA, 2023b). Despite this difference, the One Health concept, which has been used to unify human, animal, plant and environmental studies, has been identified as an opportunity to better integrate plant health (Boa et al., 2015) with a few examples of broadening the approaches commonly used in plant health (e.g. Rizzo et al., 2021). Integrated Pest Management (IPM) and especially the agroecological approach could perfectly fit into the One Health approach, allowing to achieve food safety and food security, while reducing the impact on the environment.

The EFSA PLH Panel uses a mechanistic population-based approach to capture the dynamics of insect pest and pathogen populations through the attributes of the disease triangle. This involves the definition of a conceptual model to compute changes in the population abundance and distribution across the different assessment steps (EFSA PLH Panel, 2018). For typical quantitative pest risk assessments (QPRA), questions are framed by the ISPM (International Standards for Phytosanitary Measures), in particular ISPM2 11 on entry, establishment, spread and impact of pest populations. These activities are supported by up-to-date panel guidance documents (EFSA PLH Panel, 2018, 2019). Problems encountered include the availability of data to parameterise pest risk models (but which can be supported by Expert Knowledge Elicitation (EKE)) as well as transferability of models in space and time, including the assessment of climate suitability and climate change. In general, this is exacerbated by the limited number of epidemiological studies in plant health from which to synthesise information. Though this uncertainty is in part offset, since small deviations in risk can in general be tolerated in plant health, which is often not the case in human disease.

4.1.5 | Cause and effect

In simple terms, causality is the process where one factor leads to the production of another process or state. Section 4.1.5.1 gives a short description of some of the existing theoretical frameworks on causality that have been developed within epidemiology. Considerations on how to make inferences of causality based on different study designs are then given in Sections 4.1.5.2 and 4.1.5.3. It should be noted that in general, the level that a study is aimed at (e.g. molecular, individual, population) needs to be considered when weighing the evidence for causation.

4.1.5.1 | *Existing frameworks on causality*

Much of the theoretical framework for causality in epidemiological studies has been developed in the 20th century, driven in part by studies on smoking and lung cancer (Lash et al., 2021; Vandenbroucke et al., 2016). Theoretical frameworks include the simple but much cited viewpoints formulated by Austin Bradford Hill (1965), and more elaborate theoretical frameworks such as the Sufficient-Component Cause Model (Lash et al., 2021). The subject of causality has also been

elaborated by Pearl (2009) and Pearl and Mackenzie (2018). Moreover, well-defined counterfactual conditionals can be used in causal reasoning as valuable tools for forming intermediate steps towards supporting causal claims (Hernán & Robins, 2020).

In general, finding a statistical association in epidemiological studies with observational design is not per se enough to assume an association is causal. The Bradford Hill (1965) paper has been very influential in the development of systematic assessment of evidence of causality. With his nine viewpoints, Hill laid a sound framework for assessing causality; some are specific to assessing an individual epidemiological paper, but most are directed at synthesising evidence across different types of studies. The features that he proposed were as follows: Strength of the observed association, consistency across repeated studies, specificity of the association, temporality – exposure preceding effect, a gradient of effect or dose–response relationship, biological plausibility – mechanistic evidence or support from animal studies, coherence between different types of epidemiological observations – the observed association should not contradict any previous knowledge available about the disease and/or exposure, experimental evidence, analogy with comparable causal associations with other exposures. He emphasised that his systematic approach serves to guide the assessment of the strength of evidence of causality and cannot be used mechanically to yield a yes/no decision.

> None of my nine viewpoints can bring indisputable evidence for or against the cause-and-effect hypothesis and none can be required as a sine qua non. What they can do, with greater or less strength, is to help us to make up our minds on the fundamental question – is there any other way of explaining the set of facts before us, is there any other answer equally, or more, likely than cause and effect? (Hill, 1965).

The nine viewpoints and the questions to answer when assessing them are listed in Appendix A of this document.

The original Bradford Hill viewpoints have been modified and adapted for toxicology (Adami et al., 2011), and applied within the Mode of Action framework for comparative analysis of the weight of evidence (WoE) (Meek et al., 2014). A useful tool for characterising biological plausibility for a toxicological exposure/disease association is the 'Adverse Outcome Pathway' (AOP) approach. An AOP is an analytical construct describing the sequential chain of causally linked events at different levels of biological organisation that lead to an adverse effect. These should be included, if available, in the hazard assessment for exposure. The Bradford Hill viewpoints can also be used to assess the WoE that supports an investigated AOP, and for making a judgement on how strong the evidence is to support a particular investigated mode of action (Gross et al., 2017). They have also been used to develop approaches to evaluate the confidence in a whole body of evidence when making inferences of causality (GRADE, (Grading of Recommendations, Assessment, Development, and Evaluations) and modified GRADE Approaches (Morgan et al., 2016)).

Another influential framework on causality is the *Sufficient-Component Cause Model*. This model is centred around the fact that disease causality is multifactorial, meaning that in most cases several component causes need to act together or sequentially in order to complete a sufficient disease cause (Rothman, 1976). Moreover, several different sufficient causes may lead to the same disease. The more component causes that are known, the more complete is the causal picture of the disease, which allows for more targeted and accurate interventions for prevention of the disease. Such component causes or risk factors are investigated in both experimental and observational epidemiological studies.

In recent years, the use of causal diagrams in epidemiology or directed acyclic graphs (DAGs) has increased and these are very helpful for careful planning of a study design and analysis. While not entirely consistently used, they help both study investigators and readers of their papers with guidance on the causal relationships of outcomes and risk factors (Tennant et al., 2021). DAGs thus provide the investigator with a simple and transparent way to identify and demonstrate their knowledge, theories and assumptions about the causal relationships between variables, and detailed guidance is available on how to prepare DAGs (Hernán & Robins, 2020; Lash et al., 2021).

4.1.5.2 | *Experimental studies and causality: Strengths and limitations*

Experimental studies, when they are feasible, for example for short-term effects of exposure, are better suited than non-experimental studies to determine if a certain exposure is causally related to a given outcome. When available and of good quality, these studies are generally considered the ideal design when making judgement on causality. Often the absence of an effect in such studies is considered a strong argument for 'no evidence for effect'. Such interpretations are however only valid in sufficiently powered studies where RoB is low, compliance is high, and dropout is low. These conditions can more easily be met in nutritional interventions with vitamins, minerals and other supplements where the assigned intervention requires modest commitment from the participants. However, when the assigned intervention requires substantial changes in habitual lifestyle, these conditions become more difficult to achieve. This is, for example, the case for some dietary intervention studies. Examples of such studies include interventions aimed at reducing risk of non-communicable diseases such as cardiovascular disease (CVD) (Howard et al., 2006) or individual CVD risk factors (Tang et al., 1998) through assignment to complex dietary regimes (in this example low-fat diets rich in whole grains, fruits and vegetables). In such studies, observed changes in dietary habits between intervention and controls have generally been modest and far from the goals set out for dietary changes. In such studies, compliance may decrease considerably over time, thus hampering the reliability of long-term intervention studies. That is the idea that one can randomise and ask people to change their lifestyle habits substantially over several months or years and see if they experience lower disease frequency is subject to substantial methodological challenges that may, if not overcome, provide limited evidence for or against causality.

4.1.5.3 | *Observational studies and causality: Strengths and limitations*

Compared to experimental studies which involve randomised allocation to exposure, observational studies are more prone to bias, particularly confounding. To make statements on causality based on their results, replication of findings in different study populations, where confounding factors may differ, and taking other lines of evidence into consideration are usually needed to build a strong case for causality (EFSA Scientific Committee, 2017). One point that is sometimes made is that a case for causality can only be made from observational epidemiology by relying on prospective cohort studies. This view, however, ignores the fact that different designs often complement each other, particularly when possible sources of bias differ. As an example, when studying diseases, which have a relatively long latency period, such as cancer, cohort studies may suffer from large dropout of participants during follow-up periods, which properly designed case–control studies can bypass. Another example is that cohort studies may not have information on potential confounders relevant for the outcome being examined (e.g. lifestyle factors), whereas case–control studies may have this information. Thus, if the case–control studies indicate that there is little or no confounding by a specific factor, or that such a confounder would have biased the study estimates towards the null, this suggests that any observed increased risks are unlikely to be due to confounding from this factor.

For studying long-term effects of exposure, there are examples where observational studies are more suitable than experimental studies and the only possible source to identify causal relationships, such as when assessing the safety of food supplements, food additives, or pesticides post-marketing. One famous example from the area of safety assessments of pharmaceuticals is the marketing of oral contraceptives in the 1960s. A few years later (in the 1970s), observational studies started to show a consistent association between oral contraceptives and venous thromboembolism, an outcome that previous clinical trials lacked power to detect. Based on these findings, the ethinylestradiol dosage in these pills was reduced substantially, which was associated with less side effects in subsequent studies (Dhont, 2010).

4.2 | Key epidemiological concepts relevant for evidence appraisal

Decision on how to use evidence from an epidemiological study in a scientific assessment should be supported by a rigorous appraisal. This includes assessment of individual studies in terms of their *internal validity*, which is the degree to which the observed findings from a given study or experiment are unbiased and accurate for the population studied. That is, a study of appropriate design conducted and analysed to minimise RoB and chance findings has high internal validity. In the section below, key concepts on how to assess and appraise epidemiological studies are introduced. This covers both practical issues relating to understanding and interpreting exposure and outcome measures and a brief description of the main sources of biases. A more practical application of these concepts is then introduced in Section 4.3.

4.2.1 | Study reliability

4.2.1.1 | *Use and interpretation of measures of frequency and measures of association*

Frequency measures refer to discrete variables that describe distributions of outcome, exposure or covariate measures such as, disease status, mortality, occupation and smoking. Although frequency measures are generally described as proportions or percentages, two key concepts for defining **categorical outcome measures** in epidemiology are prevalence and incidence:

- ***Prevalence*** *refers to the proportion of cases in a defined population at a given time.*
- ***Incidence rate*** *refers to the rate per unit of time at which new cases are occurring in a defined population.*

The prevalence and incidence rate are useful measures for describing how frequently a given outcome occurs (at a certain point in time) and the rate at which it is occurring (over time). In comparing exposure groups, the common approach is the comparison of ratios as measures of effect. The most common ratio measures are explained in **Box 1**. More complete descriptions can be found elsewhere (Dohoo et al., 2009; Lash et al., 2021).

Box 1 Measures of effect for frequency outcomes

Measures of effect are indexes that summarise the strength of the association between exposures and outcome. Effect measures can be expressed in both relative and absolute terms.
The relative effect measure comparing, for example, an exposed to a non-exposed population, can be called 'relative risk' and can be expressed as a ratio of incidence rates, ratio of prevalences, ratio of cumulative risks or can be estimated by the odds ratio.

Measures of effect from prevalence in cross sectional studies or cumulative risk in cohort studies: Let us assume we have two groups (1 and 2) that differ both in exposure and occurrence of a given outcome. The probability or prevalence (p) of the event occurring in Groups 1 and 2 is then:

$p_1 = \frac{a}{N_1}$, where a is the number of events and N_1 is the total number of subjects in Group 1.

$p_2 = \frac{b}{N_2}$, where b is the number of events and N_2 is the total number of subjects in Group 2.

The risk ratio of an event occurring in Group 1 compared to Group 2 is then

Risk Ratio, $\mathbf{RR = \frac{p_1}{p_2}}$

When the event incidence takes into account time at risk, the effect measure becomes the **rate ratio (also for the case of Cox regression called the hazard ratio):** That is, the number of new cases (events) occurring divided by the number of person-years at risk (e.g. if 10 people are each followed for 10 years, this involves 100 person-years of follow-up)
Then the **rate ratio** is defined as

Rate Ratio $= \frac{\lambda_1}{\lambda_2}$, where λ_1 and λ_2 are the rates in Groups 1 and 2, respectively.

Relative effect measures are commonly used in epidemiological studies as they provide direct measure of the **strength of an association** between exposure and outcome.
On the other hand, **absolute difference measures** such as **the risk difference** ($p_1 - p_2$) or **the rate difference** ($\lambda_1 - \lambda_2$) provide a direct measure of **excess risk** of outcome (or disease) between two groups.

Measures of effect in case–control studies
Case–control studies compare exposures and other factors in cases in the source population (over the follow-up period) and a sample of the non-cases. In case–control studies the incidence of the outcome cannot usually be estimated, depending on how subjects are recruited. The outcome measure in a case–control study is the **odds ratio**, the ratio of odds of exposure in the cases to the odds in the referents. The odds of exposure in each group are the ratio of the proportion exposed (p) divided by the proportion of no ($1 - p$). The odds ratio is then the odds of an event in Group 1 divided by the odds of the event in Group 2:

Odds Ratio, $\mathbf{OR = \frac{odds_1}{odds_2} = \frac{\frac{p_1}{1-p_1}}{\frac{p_2}{1-p_2}}}$.

What this relative effect measure is estimating depends on how the controls were chosen. In most case–control studies, the odds ratio from the case–control study corresponds to the rate ratio from the corresponding cohort study. Sometimes the OR is used as an outcome measure in cross-sectional and cohort studies. In such cases, the OR generally overestimates of the ratio of prevalence or cumulative risk between exposure and outcome. However, for rare outcomes (< 10%), the value of the OR is not too different from the Risk Ratio.

Different views exist on whether measures of relative risk or absolute risk (see Box 1) are more appropriate for evaluating and interpreting effects or associations from epidemiological studies. However, the argument can be made that both are necessary to evaluate findings and 'one cannot be interpreted without the other' (Noordzij et al., 2017).

To give an example, let us say that in a well-defined community the prevalence of perinatal mortality has increased from 0.11% to 0.44% and one suspected cause is a dramatic increase in exposure to an environmental contaminant (e.g. contamination by accidental release of wastewater contaminated with mercury into a nearby aquatic environment). In terms of measures of effect, the absolute risk difference is 0.33%, which for the individual is quite small. At the community level, such an increase in perinatal mortality would also, perhaps, not be noticed in the absence of complete registration and publication of summary statistics from relevant authorities. However, the risk ratio (RR) is as large as 4.00 (OR is 4.01).

To take another example, let us say that in a RCT of a food supplement an unexpected side-effect is revealed. At baseline, the prevalence of hypertension among study participants is 28.7% in both intervention and control groups. However,

at the end of the study period, the prevalence in the intervention group was 34.5%, but 28.8% among controls (placebo). The risk difference here is 5.7%, which could be considered as relevant. The risk ratio here is only 1.20 (OR is 1.32).

To conclude, absolute risk measures are the most relevant measures when assessing the population impact of exposure. However, when quantifying effect size or strength of an association, relative risk estimates are more appropriate. A thorough evaluation of any association or effect reported in a study requires careful weighing of the actual effect size, the severity of the outcome and its implications for the individual and the community/population. Ideally, sufficient information allowing translating relative outcome measures to absolute measures should be reported in any publication, but the absence of the latter should not be used to downgrade studies, at the appraisal step (see Section 4.3).

The approach of modelling absolute risk is also used in one particular tool in risk assessment: benchmark dose (BMD) modelling (EFSA Scientific Committee, 2009, 2017, 2022). This approach was developed for toxicological studies with different groups of laboratory animals (e.g. rats or mice) exposed to several doses of a compound being tested. The absolute risk of developing disease (e.g. inflamed liver) increases from background rate at very low doses to very high or all of them at the highest doses. Based on fitting a smooth line through this data, the dose at which a fixed proportion being affected, say 5%. This can be used as a point of departure to set a protective level by taking into account the confidence interval of the estimate and adding safety factors. This methodology is now sometimes being adopted and applied to epidemiological data (WHO, 2010), with, for example, the absolute effect (e.g. IQ) related to the exposure level (e.g. lead in blood), and the BMD estimated for a fixed effect, in this case a shift of one IQ point (EFSA CONTAM Panel, 2010). BMD modelling is useful as a tool for establishing HBGVs but is not in itself a tool for assessing causality, which needs to be done based on integrating the strands of evidence from multiple studies and sources. Further reflections on this approach are presented in Section 4.4.3.

4.2.1.2 | *Exposure assessment*

In controlled experimental animal studies, the investigator usually has control over the exposure conditions and their changes for the whole duration of the experiment. In such cases, major exposure misclassifications are largely confined to lack of compliance by study participants or other deviation from intended treatment. In humans, similar control over exposure conditions may be achieved in highly controlled metabolic trials that can, for ethical and practical reasons, usually be conducted over a limited time period. For other experimental studies, including many RCTs, the investigator has less control, as exposure is only assigned and not always adequately monitored. As an example, in an RCT testing the effect of long-chain omega 3 fatty acids supplementation on blood pressure, the effect estimate, in strict terms, measures the average effect of administering the supplementation. That is, the average effect over those taking the supplement and those who did not (or did something else). Therefore, compliance with the treatment allocation should be carefully ensured and monitored, whenever possible, throughout the experiment. Exposure misclassification due to departures from the allocated treatment tends to distort the measured effects towards null, with some exceptions (Yland et al., 2022).

In observational epidemiological studies, the investigator does not control the exposure conditions. Therefore, the assessment of exposure must rely on laboratory measurements or other proxies of the exposure itself, such as questionnaires, historical records, geographical information systems, environmental modelling techniques and other tools. In such settings, the key challenge is not only to assess exposure in a reliable way, but also to do that in the appropriate time window, assuming that exposure duration and amount were consistent with a causal effect, and biologically plausible. What can be considered as 'acceptable' or 'valid' in exposure assessment depends, however, markedly on the exposure and outcome under study. For example, a single blood measure of a persistent substance such as dioxins that has an elimination half-life of several years could be considered a reliable marker of long-term exposure and of relevance for most long-term health outcomes, including chronic disease such as cancer or CVD. The same would not apply for a non-persistent compound such as caffeine, which has an elimination half-life of a few hours and whose body levels may markedly change over time. For caffeine, therefore, one or more objective measurements from blood samples would be enough to examine short-term effects on blood pressure, but repeated measurements in blood stretching over longer time period would be needed to reliably assess possible effects on disease such as stroke and other CVD. Despite blood measurements of a compound being an objective measure, substantial long-term exposure misclassification for single measurements may occur due to individual variation in uptake and excretion.

In a questionnaire, a simple question on behaviour, including habitual coffee, alcohol intake or smoking, can often be considered reasonably accurate measures of exposure, as such habits can be assumed (or have been shown) to stay rather constant over time for most individuals. However, self-reported exposures are often considered inferior to objective methods as, for example heavy smokers (or drinkers) are more likely to selectively underreport their habits. Objective methods are generally preferred, but when such methods do not exist or are not used, a RoB should not automatically be assumed. As an example, for smoking the use of urinary cotinine measurement as an objective biomarker can be useful to quantify exposure misclassifications, compared to relying on self-reported estimates only.

Exposure misclassification in epidemiological research may, however, also occur when 'objective' methods for assessing exposure are used. For example, providing subjects with a fitness watch to objectively measure physical activity may result in an activity higher than usual, simply because study participants have become motivated to use the instrument. It is also well known that use of dietary records can result in changes in dietary habits during the period of recording as some foods are more difficult to weigh and record than others. In addition, such records cannot generally assess rare or highly seasonal food consumption in a reliable way. Another example is use of 24-h urine sampling, which allows for accurate assessment of exposure to several substances over the past day. However, the burden of collecting all urine excreted during that period may lead to subjects becoming less mobile (or behaving differently), resulting in changes in exposures

that would not normally occur, or may decrease the number and completeness of participant recruitment due to lack of participation. Therefore, the simple act of trying to capture exposure with high precision may lead to biased estimates due to behavioural changes. In addition, even when an ideal biomarker of exposure, such as the determination of a substance in one or preferably multiple 24-h urine samples, is not available, determining such a substance in a less adequate matrix, such as in one or more random urine or morning samples, may still provide a useful estimate. By considering the strength and the limitations of the methods applied for exposure assessment, a more appropriate use of the available evidence can be made in the risk assessment process.

Based on the discussion above, a brief summary of strengths and weaknesses of different exposures measures commonly used in human studies is outlined in Table 1.

TABLE 1 Overview of the major strengths and limitations of different exposure measures frequently used in epidemiological studies.

	Limitations	Strengths
Self-reported measures	Can be prone to misclassification due to memory or selective reporting	May better capture long-term exposure than records or biomarkers
Records or monitoring data[a]	Records may also suffer from memory and selective reporting depending on how and when the recording is being done. Use of monitoring data (such as fitness watches) may influence behaviour of participants	Harmonised recording of exposure in a standardised manner. Correct use of monitoring devices gives an accurate measure of current status
Biomarkers	Often only capture short-term exposures, influenced by ADME.[14] May not be specific to the exposure under consideration	Objective and accurate measures of exposures relative to their half-life
Assigned exposure[b]	Uncertainty regarding participants' compliance or deviation from intended exposure is a limitation, particularly in long-term studies	Exposure is controlled and can be accurately quantified in terms of assigned exposure

[a]For example, clinical or other public health records containing information on past exposures (such as smoking or use of supplements or medication) or monitoring devices (such as fitness watches, air pollution monitors). Also includes occupational records (on past exposure).

[b]In randomised controlled trials and other experimental studies.

One common practice when examining continuous exposures in observational studies is to divide the exposure variables into categories, using a priori or data-driven (percentiles) cut points of exposure. The dose–response is then examined relative to one reference exposure category. One reason why this approach has historically been used is that the resulting effect estimates from quantile analyses provide simple representation of the underlying dose–response relationship that is easy to interpret in comparison to, for example, effect estimates obtained from non-linear regression. The use of quantiles does, however, lead to some loss of precision and other adverse consequences (Rothman, 2014) and the pros and cons of this approach are discussed in some detail in Appendix G.

4.2.1.3 | *Statistical inference for effect measures in epidemiological studies*

Effect measures, as estimated in epidemiological studies, represent an estimate of the underlying true parameter in the reference population. To make inferences about such parameters, uncertainties around the statistical (or central) estimate need to be considered. This is done by estimating the confidence interval which accounts for random errors[15] in the exposure and outcome. In general, the larger the sample size, the higher the precision, which is reflected by narrower confidence interval around the central estimate.

In terms of reporting effect measures, both the central estimate and its confidence interval should be reported. The *p*-value may provide useful supplementary information, but there is growing consensus that significance testing involving arbitrary cut-points (e.g. $p < 0.05$) may not be appropriate (Amrhein et al., 2019; EFSA Scientific Committee, 2011; Greenland et al., 2016; Wasserstein & Lazar, 2016). For further discussion on this issue, the reader is directed to Appendix B Hypothesis testing vs. estimation. Similarly, as for the effect measure from a single study, effect measures from several studies (or experiments) should, in the absence of systematic bias, follow a distribution affected only by random (study-specific) errors that are symmetric around the true estimate. It is, however, well known that publication bias can occur when the probability of publication of study results is correlated with the reported effect size (or statistical significance), i.e. when small effect sizes (or non-significant results) are systematically underrepresented in the available (published) body of evidence. As a result of publication bias, the body of available evidence may bias the summary of evidence away from the null in cases where there is truly no effect or skew the estimate from its actual value when an effect truly exists. A similar bias would result from a selection process of published studies for evidence integration. Thus, both the selection of results for publication and the selection of published studies in evidence integration should be independent of reported effect sizes. Mandatory pre-registration of clinical trials can mitigate publication bias.[16] A worldwide voluntary pre-registration of stud-

[14]ADME = absorption, distribution, metabolism and excretion.

[15]Random errors in measurements are caused by unknown and unpredictable changes in the experimental output. For continuous measurements, random errors follow a normal distribution. Random errors should not be confused with systematic errors (see Section 4.2.1.4).

[16]https://classic.clinicaltrials.gov; https://www.clinicaltrials.gov.

ies involving animals has been launched recently (Bert et al., 2019). A pre-registration and/or a publication of the protocol of observational epidemiological studies, as well as of systematic reviews and meta-analyses, can be assumed to have similar positive effects.

4.2.1.4 | *Systematic error (bias)*

Systematic errors differ from random errors in as far as the former would be present even in an infinitely large study, whereas random errors can be reduced by increasing the study size. Thus, systematic errors (or 'bias') occur if a systematic difference between the true value and the measured value exists (Pearce, 2005). Systematic errors are usually classified into three types of bias: information bias, confounding and selection bias.

4.2.1.4.1 | *Information bias*

Information bias concerns misclassification of the study participants with respect to exposure, outcome or confounder status. Usually, two types of misclassifications are considered: non-differential and differential misclassification.

Non-differential misclassification occurs when the probability of misclassification of exposure or health outcome is the same for cases and non-cases, i.e. exposed and non-exposed persons are equally likely to be misclassified according to disease outcome; or diseased and non-diseased persons are equally likely to be misclassified according to exposure. With some exceptions (Yland et al., 2022), non-differential misclassification of exposure biases the effect estimate towards the null and tends to reduce the size of the effect which is of particular concern in studies which find weak associations (Pearce, 2005).

Differential misclassification occurs when the probability of misclassification of exposure is different in cases and non-cases, or the probability of misclassification of disease is different in exposed and non-exposed persons. This can bias the observed effect estimate either towards or away from the null value (Pearce et al., 2007). For example, in a case–control study of lung cancer, the recall of past exposures, e.g. smoking, might differ in cases from that of the controls, leading to differential misclassification. This could bias the odds ratio towards or away from the null (value of 1.0).

4.2.1.4.2 | *Confounding*

While several detailed definitions of confounding exist (e.g. Lash et al., 2021), in this document, a confounder is referred to as a variable (or factor) that is associated with both the exposure and outcome, resulting in a spurious association between the two. Confounding is to be expected if the factor of interest is associated with a different factor (the 'confounder') which is a known or unknown risk factor for the outcome of interest. For example, assume that the exposure to substance X (risk factor of interest) is associated with co-exposure to cigarette smoke (confounding factor), i.e. individuals who are exposed to higher concentrations of substance X also smoke more cigarettes compared to the unexposed; and smoking is also causally related to the outcome of interest.

When confounding is not considered, the potential effect of the risk factor of interest may be mixed with the effect of the confounder or even entirely explained by that confounder. Consequently, the statistical effect estimate is biased with unknown magnitude and direction. It is a matter of subject expertise to identify potential confounders, to plan collection of confounder information at the design stage, to adjust for confounders in the analysis and to consider the possibility of residual confounding[17] in the interpretation of the study results. DAGs are an increasingly popular approach for identifying confounding variables that require conditioning when estimating causal effects (Tennant et al., 2021).

Confounding can be mitigated by the design of the study and through 'adjustment' in the statistical data analysis. An ideal study design to control confounding ensures that the expected variation of all potential confounders is identical across all levels of the main risk factor. RCT are epidemiological studies in which this is theoretically possible since allocation of intervention or treatment (main study factor) to the participants is at random. Thus, potential confounders should (on average) be evenly distributed in all treatment groups. Confounding can still occur in an RCT due to imbalanced distribution of confounding factors across treatment groups. In observational studies, where participants are not randomised, confounding is more likely to occur.

If potential confounders have been identified in the design of the study, and the respective information on all confounders is collected at the individual level, it is possible to statistically adjust for confounding. Several approaches for confounder control exist (Kestenbaum, 2019). One approach for this involves stratification by the confounding factor and construction of a weighted effect estimate (e.g. the Mantel–Haenszel odds ratio estimate). Multivariable models provide a similar adjustment (correction of confounding bias) and offer the additional flexibility to accommodate categorical as well as continuous risk factors. The fact that a risk model is adjusted for one or several confounding factors does not give a full guarantee against confounding bias. It requires a case-by-case expert judgement from a subject matter and statistical modelling viewpoint to decide whether potential confounding is adequately addressed.

While the technique of matching can be used to prevent confounding (from the matching factor) from occurring in cohort studies, in case–control studies, it may also lead to the opposite result, i.e. introduce a selection bias that behaves

[17]Residual confounding is the distortion that remains after incomplete confounder control in the design and/or analyses of the study.

like confounding, because it may violate the principle of selecting controls at random from the source population. In practice, matching may artificially bring the exposure distributions in the cases and controls closer together than they really are in the source population (overmatching) and therefore introduce bias. Therefore, in matched case–control studies, the matching factor will in most cases need to be controlled for in the analysis (Pearce, 2016). Other methods of controlling for confounding such as weighting and propensity scores can also be applied (Lash et al., 2021).

In observational epidemiological studies, usually more than one factor will differ between the compared groups, in which case they could all be potential confounders. For this reason, the results of such studies are always subjected to multiple regression analysis, which allows for the adjustment of the effect estimates for several factors simultaneously in the same statistical model. That means that the effect estimates obtained from such modelling are unconfounded by the effects of the other factors that are included in the same model (provided that these other factors have been defined appropriately and measured accurately). Residual confounding may still exist for several reasons, including (1) other confounders that have not been included in the model, (2) imprecise measurement of one or more confounders controlled for, or (3) inappropriate modelling of the confounder in the statistical analyses. Even though it is very important to evaluate the appropriateness of the statistical model used, the validity of the respective assumptions, and the model building strategy, etc., this is a very technical issue which is beyond the scope of this document. It is advised that for this task the assistance of a statistician or an epidemiologist be requested.

4.2.1.4.3 | *Selection bias*

Selection bias is an important systematic error in observational studies. It involves bias arising from how the study participants are selected (or select themselves) from the source population. It thus arises when the relation between exposure and disease in the study population (i.e. the actual study participants) differs from the relation in the source population from which study participants are drawn (Lash et al., 2021). In general, selection bias occurs as a result of the procedures used to select study participants (Pearce, 2005). Because usually only information from the recruited study population is known, selection bias must typically be evaluated indirectly or theoretically, and anticipated in the study design. It may be possible to 'correct' selection bias in a study, if the factors influencing selection can be controlled for in the analysis (in the same way that confounders can be) (Pearce, 2005). This requires, however, that additional information (on these factors) needs to be available for all study participants.

Selection bias could occur, for example, when people enrolled in a cohort study are self-referred. One such example would be if people self-referred to a study, knowing that they had the studied exposure and suspecting that they may also have the outcome (maybe experiencing relevant symptoms). Selection bias would occur if these people would indeed have a higher probability of the outcome compared to exposed people in general. Selection bias can be related not only to 'selection' to enter a study, but also to a 'selection' to exit a study. In this sense, the bias resulting from a loss to follow up (persons lost to the study investigators before the end of the study) that is differential between the two compared groups (e.g. exposed and non-exposed) is also a form of selection bias (Hernán et al., 2004).

Selection bias can also result from using an inappropriate control group in a case–control study. In these studies, the purpose of the control group is to provide an estimate of the distribution of exposure in the source population from which the cases originate. A control group may fail to provide this information, when, for example, the population from which the cases originate is not appropriately defined, or selection of controls is based on convenience rather than on specific criteria that need to be fulfilled. For a detailed discussion on selection of controls in case–control studies, the reader is referred to Wacholder, McLaughlin, et al. (1992); Wacholder, Silverman, et al. (1992b, 1992c).

Confounding generally involves biases that can occur even if everyone in the source population took part in the study as they are inherent in the source population. Selection bias, on the other hand, covers biases that stem from the procedures that are used to select the study participants from the source population. As a result, selection bias is not an issue in a cohort study with complete follow-up, as the study cohort composes the entire source population. Selection bias can, however, occur if participation in the study or follow-up is incomplete or if the response rate depends on the exposure and outcome (e.g. overrepresentation of heavily exposed persons who are more likely to be diagnosed with disease (Pearce, 2005)). Similarly, selection bias is not an issue if a case–control study involves all cases in the source population (and risk period) and the controls are a random sample of the source population, and the response rate is 100%. However, selection bias may occur if response varies by exposure and disease status.

4.2.1.5 | *Effect modification/interaction*

A key issue of epidemiological research is to identify and assess the extent to which the effect of an exposure may depend on the level of one or more other factors and whether or not such factors may have an independent causal effect on the endpoint under consideration. Such factors are described as effect modifiers, the underlying concept being the existence of interactions between two or more factors. For example, if an exposure vs no exposure has a rate ratio for the outcome of 3.0 in men and 1.5 in women, sex may be an effect modifier because the effect of the exposure seems different in men and women. Identification of effect modifiers has a key relevance in both scientific research and risk assessment, and also plays a crucial role when assessing the external validity of study findings. Therefore, assessment of interactions has become a key goal of scientific research, in order to identify higher susceptibilities to adverse or beneficial effects of a given exposure,

due to other exposures or endogenous factors (such as children, pregnant women, diseased persons, individuals with specific dietary/life-style habits or genetic backgrounds).

Effect modification is entirely different from confounding, since the former concerns the ability of one factor to modify the causal effect of another factor on a defined endpoint. Effect modification occurs when an exposure has a different effect among different subgroups; hence, it is associated with the outcome but not the exposure. Therefore, understanding of effect modification is necessary to characterise causal association, interactions and susceptibilities, which is important in risk assessment. Confounding. on the other hand. must be minimised when planning a study or controlled for at the analysis stage.

Effect modification may be assessed either as statistical interaction or biological interaction (Lash et al., 2021). Statistical interaction is just a departure from the basic form of a statistical model and is therefore dependent on the metrics used in the statistical model, e.g. multiplicative vs additive models. Biological interaction describes the mechanistic interaction between causal factors, assessing the departure over additive effects of the combination of single risk determinants. It amounts to an attempt to identify susceptibility factors, which are factors that modify the effect of an exposure on a specific health outcome. Unlike statistical interaction, it is a biological phenomenon. The assessment of biological interaction requires considerably more data than the assessment of the effect of a single factor and may involve the net effect of factors that are causes and preventives in varying combinations (see Section 4.1.5.1).

4.2.2 | Study relevance

Generally, epidemiological studies conducted in the target species have clear advantages in terms of relevance for risk assessments over studies conducted in non-target species, as uncertainties due to between-species extrapolation are eliminated. It can often be assumed that the exposure conditions in observational settings, if appropriately captured, are more similar to real conditions in terms of duration, concentration of exposure and other circumstances than in experimental studies where the exposure conditions are chosen by the investigator. A refined assessment of the relevance of the evidence from epidemiological studies for risk assessments requires that the choice and characteristics of the study population, the selection of study participants, the exposure conditions as well as the case definition and the measurement of the outcome be evaluated with respect to the specific research question.

4.2.2.1 | *External validity*

When assessing external validity several different concepts should be distinguished:

- *There is a target population to which we wish to draw inferences (e.g. all people in the EU, all people on the planet)*
- *There is a source population which is used as the source of participants for a particular study (e.g. everyone living in Parma, British doctors)*
- *There is a study population, i.e. the group of people who actually take part in the study, with some of the source population not taking part either due to selection by the investigators, or self-selection (i.e. non-response)*

External validity refers to whether the study findings can be generalised to the target population

Provided that disease outcome has not affected the choice of source population, and if 100% of the source population is included in the study, there can be no selection bias. Rather, any differences between the results in the source population, and what would have been obtained from studying the whole target population is a result of confounding (different confounding structures) and/or effect modification (see Section 4.2.1.5). In most studies, the 'target population' is left undefined, with the implication that the findings are intended to apply to the general population. In fact, there is no need to invoke some hypothetical target population to validly design and analyse a study. In more simple terms, generalisability is a matter of expert judgement, not statistical considerations (Lash et al., 2021). When studying exposure–health relationships in the target population, lack of representativeness may not be a problem when identifying risk factors (e.g. the original findings for smoking and lung cancer included a study in British doctors).

External validity is of particular relevance for descriptive studies and cross-sectional surveys aimed at determining disease frequency or other characteristics in a given population. If the study population recruited in these studies is representative of the source population, statistical inferences may be made about the characteristics (parameters) in the source population, based on the information from the study population. Representative samples can be obtained using specific probability sampling techniques. However, sometimes, it is not possible to obtain representative samples from a population. In those cases, it is very important to consider if the population has been sampled selectively, based on specific factors, and which 'sub-population' the sample may represent. For example, testing cattle at the slaughterhouse for bovine paratuberculosis, a chronic progressive infectious disease of cattle, may not provide a representative picture of the disease in the entire bovine population of the area served by the slaughterhouse. Animals at the slaughterhouse may have more advanced infections than in the 'general' populations of cattle in the area, because they might have been sent to the slaughterhouse due to their infection or due to old age. Or conversely, animals with very advanced cases might have already died or euthanised at the farm and never made it to the slaughterhouse (Nielsen et al., 2011).

4.2.2.2 | *External vs internal validity*

External validity and internal validity can be interlinked following various patterns (Steckler & McLeroy, 2007). For example, controlled experimental studies have, for reasons explained above, generally lower RoB than observational studies. This higher internal validity often comes at a price. For example, adverse effects of chemicals can only be studied under experimental conditions in animals, which then have to be extrapolated to humans where such experiments are not feasible. Conversely, when human data exist, addressing uncertainty in terms of external validity means relying on human observational studies, which generally have lower internal validity, compared to experimental studies in humans conducted in highly selected populations. Moreover, external validity is affected by the population characteristics of the studies comprising the best available evidence. As an example, RCTs testing the effect of pharmaceuticals or individual nutrients on health are usually conducted in specific populations that are most likely to benefit from treatment (pharmaceuticals) or demonstrate some beneficial effects (nutritional RCTs). Such a selection may, however, hamper extrapolations to the more general population. For example, results from a RCT showing modest increase in cancer risk as a result of beta carotene supplementation in male smokers (The Alpha-Tocopherol Beta Carotene Cancer Prevention Study Group, 1994) may provide a reasonable argument for not taking beta carotene as a food supplement for cancer prevention. On the other hand, it could be argued that these results would perhaps not be the same if conducted in healthy non-smokers who have a much lower cancer risk. In terms of extrapolating such exposures to more real-life setting, such increase in cancer risk due to use of beta carotene supplements is not comparable to exposure to beta carotene from the habitual diet. Similarly, findings of increased mortality in postmenopausal women with underlying CVD following supplementation with vitamins C and E (Waters et al., 2002) could be considered to have modest to low external validity for the general population. In terms of making conclusions on causality, both internal and external validity need to be considered.

4.2.3 | Summary and conclusions

In Section 4.1, a brief description of different experimental and non-experimental studies was given, highlighting their main strengths and limitations. In this section, different types of biases that may occur in each of these designs have been explained. When interpreting the findings of epidemiological studies and assessing the evidence generated by them, there is sometimes a preference towards ranking studies in terms of internal validity by their design (design hierarchy). This usually translates to emphasising the role of experimental studies (RCTs) and, among the non-experimental ones, that of cohort studies. However, such ranking is often not justified as the examples and discussions above have tried to highlight. In fact, all study designs are more (or less) prone to biases. In the absence of an empirical basis for the relative importance of biases in a given research area, it can be misleading to infer bias proneness from study design only. Some typical biases that may occur in different study designs are briefly summarised in Table **2**.

TABLE 2 Study designs and typical biases.

Bias	RCT	Cohort	Case–control	Cross-sectional	Ecological
Selection bias	a	b	b	b	c
Confounding	d	e	e	e	e
Information bias	f	g	h	h	i

General note on Table 2**:** a lack of external validity may be an issue with all study designs, e.g. patients included in a RCT may have severe disease and the findings may not be generalisable to mild disease, a study conducted in men may not be generalisable to women, in adults to children, etc. The letters are explained in the text below.

Selection bias

a. Selection bias at baseline is not usually a concern in RCTs, provided that the study is sufficiently large, if allocation is adequately randomised and concealed after the study participants are selected; selection bias may occur due to loss to follow-up if this differs by treatment group or outcome (Hernán et al., 2004).
b. Selection bias in cohort, case–control and cross-sectional studies may result from the way in which the study participants are selected (or select themselves) from the source population, leading to them being unrepresentative of the source population in terms of exposure or outcome. Selection bias can also occur due to loss to follow-up.
c. Selection bias is by definition rare in ecological studies provided that they cover an entire defined population.

Confounding

d. Confounding may occur in RCTs if allocation concealment and/or randomisation is not adequate, for example when the study group is small, thus making treatment groups not directly comparable at baseline.
e. Confounding can occur in all observational designs.

Information bias

f. Information bias on exposure (i.e. exposure misclassification) occurs if the treatment groups of RCTs are not maintained (i.e. participants stop or switch treatment); information bias on the outcome occurs if participants receiving the treatment may be subjected to more or less intensive diagnostics compared to the comparison group (lack of blinding, diagnostic bias).
g. Information bias may occur in cohort studies due to misclassification of exposure or the outcome, e.g. if exposed participants may receive more intensive diagnostics compared to non-exposed (lack of blinding, diagnostic bias). With some exceptions, non-differential (random) misclassification of exposure or disease will usually produce a bias towards the null (no effect) and cannot explain positive findings. Differential information bias (e.g. if classification of the outcome differs by exposure status) can produce bias in either direction.
h. Information bias may occur in case–control and cross-sectional studies due to misclassification of exposure or the outcome, particularly when the classification of exposure is based on participant recall (recall bias). Non-differential (random) misclassification of exposure or disease will usually produce a bias towards the null (no effect) – with some exceptions – and cannot explain by itself positive findings. Differential information bias (e.g. if recall is different in cases and controls) can produce bias in either direction. Exposure assessment may be affected by the disease condition (e.g. due to reverse causation).
i. Information bias is a major concern in ecological studies, since exposure and outcome information is only available on a population and not on the individual level – thus, even if there is an association between exposure and outcome at the population level, it may not be the case that the outcome was more common in the exposed individuals – i.e. there may be an association at the population level but not at the individual level, and vice versa. The assumption that the observed associations can be transferred from the population to the individual level is known as ecological fallacy (Hammer et al., 2009).

4.3 | Study appraisal frameworks

This chapter focusses on tools and processes for assessing characteristics of individual studies to enable their quality to be assessed in a thorough and consistent way.

4.3.1 | Background

Risk assessments undertaken by EFSA are an integral part of health-related regulatory decision making, a field characterised by large diversity in context, content, methods, information sources and implementation (Diefenbach et al., 2016). A common feature in any decision-making process is the efficient retrieval, organisation and integration of the available evidence on a specific question or term of reference (Langlois et al., 2018). For EFSA's risk assessments evidence should be retrieved from many and diverse sources. The information derived from each piece of evidence, however, is not necessarily equally relevant as different study design are prone to different sources of bias as described in the previous section. Taking into consideration the above, for a successful integration to be achieved, the selected evidence base must be organised in a way that assigns an appropriate role to each information piece.

The process of assigning such a role to each piece or body of evidence is complex and specific to each risk assessment question and context, and such decisions cannot be taken on the basis of study design only. Further guidance on this process can be found in Section 4.4 and in EFSA guidance documents (EFSA, 2015; EFSA Scientific Committee, 2017, 2023a).

Individual study appraisal is organised as follows (Agency for Healthcare Research and Quality, 2002):

a. identification of the key elements of the research/assessment question under study (EFSA Scientific Committee, 2023a)
b. assessment of internal validity (RoB)
c. summarisation of the study appraisal results.

Clarifying the **key elements of a research/assessment question** is the starting point of the study appraisal process. A clearly framed question 'creates the structure and delineates the approach to defining research objectives, conducting systematic reviews and developing health guidance' (Morgan et al., 2018). A formal strategy for identifying key elements is essential for:

- designing the literature search strategy,
- identifying the studies that by design and conduct best fit the risk assessment needs,
- clarifying important population characteristics and subgroups,
- expanding or narrowing the exposure spectrum and defining the different exposure strata,
- choosing the comparison that best fits the terms of reference among the usually large number of performed comparisons (i.e. combinations of exposure category and the various endpoints),
- organising and prioritising the relevant endpoints and follow-up timepoints thereof**.**

Before the publication of EFSA's GD on systematic review (EFSA, 2010), the PICO/PECO/PO/PIT[18] approach has been the framework most widely adopted in EFSA for defining the key elements of a question and thus structuring the problem formulation process. More recently, a new approach named APRIO[19] has been defined by EFSA's Scientific Committee in its guidance on protocol development (EFSA Scientific Committee, 2023a). Owing to its cross-cutting nature, the APRIO paradigm is broadly applicable within and across the various domains of EFSA and is considered to overcome the limited applicability of the existing frameworks in some of EFSA's domains. Therefore, although the PICO/PECO/PO/PIT represents a valid approach in some domains, the APRIO is currently EFSA's recommended framework for problem formulation and is considered preferable, to enhance harmonisation across domains.

After clarifying the key elements of a research/assessment question the next step is to **assess internal validity of different studies**. Internal validity is the extent to which a piece of evidence provides an unbiased estimate of the association between exposure and outcome, i.e. the extent to which the study results reflect the 'truth' among the study population. For a given study, assessment of internal validity refers to evaluation of its design and conduct, including reflections on the likelihood, degree and direction of possible biases. Such an assessment can be facilitated by organising the appraisal into various bias domains. Selection bias, information bias and confounding are key domains to be included and can be operationalised to specifically address e.g. classification of exposures, departures from intended exposures, missing data, outcome ascertainment**.**

Critically summarising the appraisal results of a study essentially pertains to the magnitude of the effects and the precision of the point estimates. However, while clarifying what are the main results of the study, various parameters are of considerable importance, such as the proportion of exposed and unexposed, the effect metric used and its appropriateness, the magnitude of the effect in absolute and relative association measures; the reporting of both crude and adjusted effect estimates; the confounders adjusted for; and the implementation of subgroup analysis.

The next sub-section provides a brief description on the development of appraisal and RoB tools and an overview of existing tools. Guidance on the use of RoB tools for assessing internal validity of individual studies and on summarising their results in a systematic manner (points b) and (c) above, respectively is given in Sections 4.4.1.5.1 and 4.4.1.5.2.

4.3.2 | Appraisal and RoB tools: Development and overview

In view of the challenges inherent in study appraisal, the practical need of a standardised process has led to the development of various appraisal instruments. Many reviews, inventories and annotated bibliographies of critical appraisal tools applicable to different study designs have been produced with different aims. Some of these exercises have been performed by groups of researchers, such as those developed by the Cochrane Collaboration.[20] Others have been the result of the efforts by risk assessment organisations or governmental bodies which were interested in implementing structured and harmonised approaches in their own assessment processes (BfR,[21] IARC,[22] ECETOC,[23] NIHS R&D HTA[24] Programme, AHRQ,[25] NTP-OHAT,[26] EPA-IRIS,[27] Navigation Guide, USDA-NESR[28]). Currently, there are no agreed gold standards and no standardised processes for developing such tools (see Appendix D).

Critical appraisal tools have been developed for different purposes and contexts such as (1) to appraise single studies; (2) to assess RoB in systematic reviews; and (3) to inform the weighing of the evidence in risk assessments. They cover one or more study designs and can have one of the following structures (Sanderson et al., 2007):

- summary checklist consisting of only a list of items (e.g. CASP),
- a checklist accompanied by a summary qualitative judgement (e.g. EPIQ),
- a scale with the list of items and scores attached, which result in a summary numerical score (e.g. Jadad; Newcastle-Ottawa),
- domain-based tools (e.g. Cochrane RoB 2.0; NTP-OHAT).

[18] *PICO/PECO* stands for Population, Intervention/Exposure, Comparator, Outcome; *PO* for Population, Outcome, *PIT* for Population, Index Test, Target Condition.

[19] The abbreviation APRIO stands for: *Agent* is anything that can cause an effect on a receptor; *Pathway* refers to any way in which an agent interacts with its receptor. It is the sequence of events leading the agent to cause an effect on the receptor. It can simply cover the route of exposure (typically dietary in EFSA assessments) or represent, for instance, the steps of introduction and spread when assessing a pathogen; *Receptor* refers to anything that experiences the effect of the agent. The receptor can also experience a secondary consequence to the exposure to the agent (e.g. farmers changing cropping practices as a consequence of the crops being affected by a pest); *Intervention* refers to any intentional measure aimed at changing directly or indirectly the exposure and/or the consequence of the exposure to the agent; *Output* is the form of the answer to the assessment question or sub-question, the result of an assessment process.

[20] Risk of bias Tools.

[21] Bundesinstitut für Risikobewertung.

[22] International Agency for Research on Cancer.

[23] European Centre for Ecotoxicology and Toxicology of Chemicals.

[24] National Institute for Health and Care Research Research & Development Health Technology Assessment.

[25] Agency for Healthcare Research and Quality.

[26] National Toxicology Program-Office of Health Assessment and Translation.

[27] US Environmental Protection Agency Integrated Risk Information System.

[28] US Department of Agriculture Nutrition Evidence Systematic Review.

EFSA has used the NTP-OHAT tool in several of its scientific assessments since 2015. The Office of Health Assessment and Translation (OHAT) from the National Toxicological Program (NTP) in the US has outlined operating procedures for systematic review and evidence integration for conducting literature-based evaluations in environmental health and toxicology (Rooney et al., 2014). They have developed a RoB Tool that applies a parallel approach to the evaluation of RoB for human and animal studies, facilitating consideration of potential bias across evidence streams with common terminology and domains (National Toxicology Program, 2019). This approach was developed drawing on several different sources including the most recent guidance from the Agency for Healthcare Research and Quality (Viswanathan et al., 2012), the Cochrane RoB tool for non-randomised studies of interventions (Sterne et al., 2016), Cochrane Handbook (Higgins & Green, 2011), SYRCLE's RoB tool for animal studies (Hooijmans et al., 2014) and the Navigation Guide (Woodruff & Sutton, 2014). The NTP-OHAT RoB tool is designed to evaluate, through different sets of questions, the internal validity of several of the most common study designs encountered in chemical and nutrient risk assessment. These questions are complemented by detailed criteria ('practices') that define aspects of the study design, conduct, and reporting required to reach each RoB rating. It can be applied to many research questions and tailored to the scope of the assessment.

In Appendix D Relevant inventories and reviews of critical appraisal tools, a table showing a selection of inventories and reviews on critical appraisal tools is provided. This includes a description of their context, objectives and study designs covered.

An overview of RoB tools for appraising systematic reviews (i.e. research synthesis) and for appraising individual primary research studies is provided for all types of EFSA assessments in Appendix E Overview of appraisal tools.

4.4 | Use of epidemiological evidence for human health risk assessment

4.4.1 | Evidence assessment and integration

Assembling and assessing the evidence from multiple epidemiological studies within a risk assessment framework serves three needs; first, the hazard assessment, i.e. contributing towards the assessment of causality in an association; second, characterising the exposure–response relationship; third, assessing the uncertainty underlying the two previous endeavours via characterising possible biases and consistency, or the lack thereof, across the appraised evidence.

The different steps of evidence assessment and evidence integration generally follow a systematic approach, and they are implemented first within a single evidence stream of either human or other studies (e.g. studies using laboratory animals or in vitro studies) (see Figure 1).

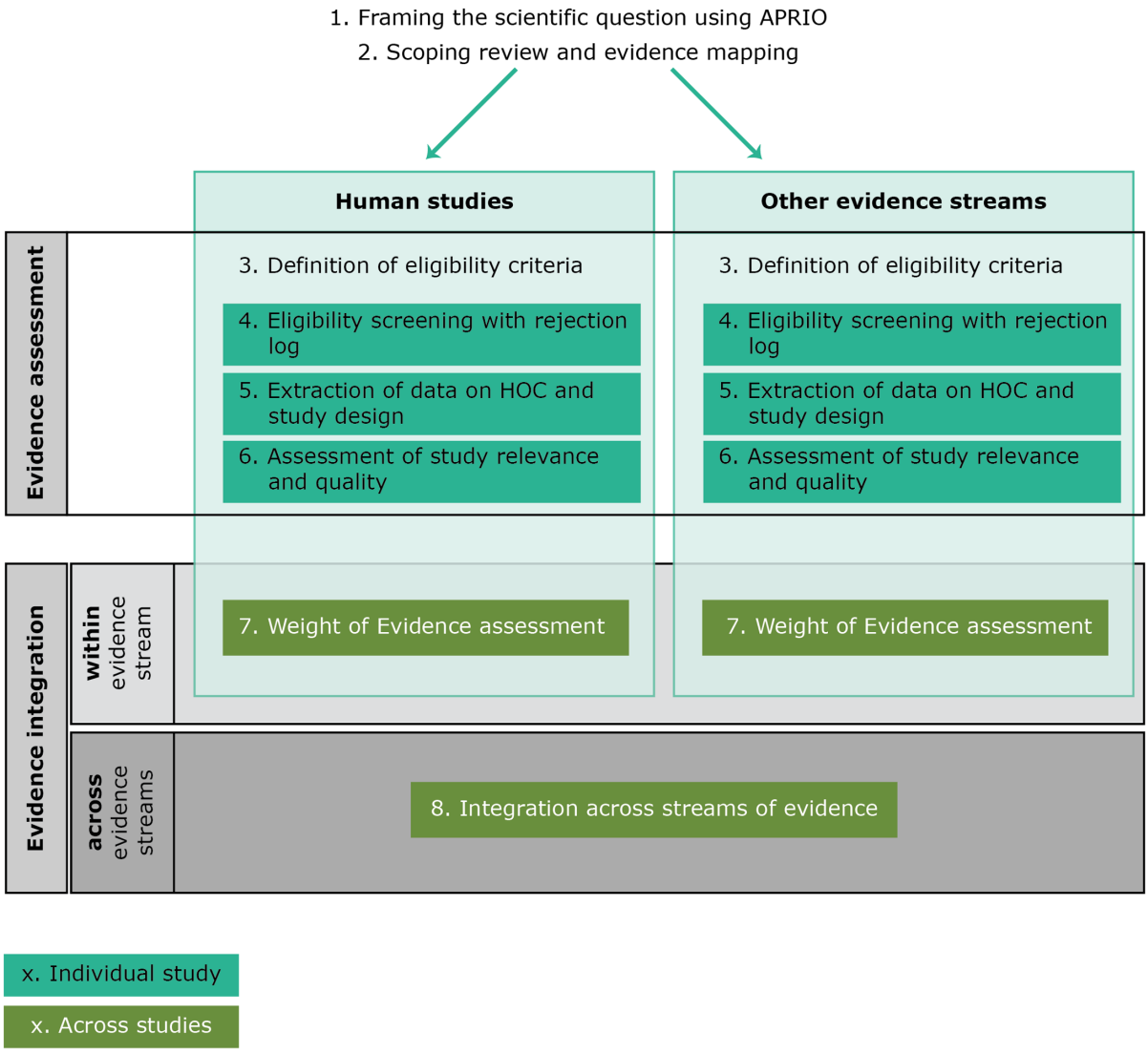

FIGURE 1 Steps of evidence assessment and evidence integration (HOC: health outcome category, i.e. a combination of similar/biologically related health outcomes into one group).

4.4.1.1 | *Planning by mapping the evidence base*

The process of gathering, assessing and integrating the pertinent epidemiological evidence is preceded by a customisation step that renders the process fit-for-purpose (see Section 4.4.1.2). During this step, evidence mapping is crucial to clarify the boundaries of a scientific assessment during both the mandate- and the assessment planning phase (Peters et al., 2020). Towards that end, scoping reviews and evidence maps[29] can be used and the right timing for this process is during the planning phase when developing the assessment protocol (EFSA Scientific Committee, 2023a). Scoping reviews are broad literature compendia that comprehensively examine the extent, range and nature of the relevant research activity, and identify gaps in an existing body of literature. Thus, they can inform a variety of decision-making settings, including but not limited to risk assessment and policy making, and can provide direction for future research priorities.

The use of evidence mapping and scoping reviews is particularly helpful in scientific assessments where the underlying evidence is characterised by great volume and considerable heterogeneity (both in terms of study design and endpoints under study). As the planning phase is an iterative process, the information from scoping reviews and evidence maps will help to specify the most appropriate methods for evidence synthesis by providing predictions on the amount and heterogeneity of studies that need to be assessed. The use of scoping reviews and evidence maps may also guard against ad hoc changes to the protocol at later stages, which are often made when it becomes clear that the approach originally chosen is not compatible with the available evidence, resources or time constraints.

While mapping the evidence related to a Scientific Assessment (SA) update, the previous SA could serve as a starting point and function as a basis for predictions on how much evidence may have accumulated since the last assessment. There are numerous examples where the difference between a specific SA or Opinion and its update is substantial. One

[29]'Evidence map is a systematic search of a broad field to identify gaps in knowledge and/or future research needs that presents results in a user-friendly format, often a visual figure or graph, or a searchable database' (Miyake-Lye et al., 2016).

such example is the update of EFSA's scientific opinion on polybrominated diphenyl ethers (PBDEs) in food (EFSA CONTAM Panel, 2011) in 2024 (EFSA CONTAM Panel, 2024), where more than 200 new human epidemiological studies had been published since the previous assessment and had to be assessed. The performed evidence mapping helped make important planning decisions related to the organisation and prioritisation of health outcome categories, the extent of multicongener exposure assessment, and the feasibility of quantitative evidence synthesis of the human data (and the required allocation of resources).

KEY POINTS

- Consider doing a scoping review/evidence map, especially in data-rich and/or heterogeneous topics.
- Frame the APRIO (Agent, Pathway, Receptor, Intervention and Output) based on the Terms of Reference.

4.4.1.2 | *Customisation of the study appraisal process*

The study appraisal process should be tailored to serve the specific purpose of the risk assessment, and this should be done at the stage of protocol development. Following the structure proposed before, the first level of customisation is implemented at the level of the APRIO elements of the research question (EFSA Scientific Committee, 2023a).

Having framed the research question and related sub-questions that are relevant to the risk assessment, a decision needs to be made which study appraisal elements, during RoB assessments, should be generic and which need to be specific to the **study design**. The answer to this question largely depends on the capacity of certain study designs to contribute to the risk assessment and the volume of the accumulated evidence. This is specific to the hypothesis under consideration, and the context in which the studies are conducted.

Moving to the customisation of the RoB assessment, the inclusion, elaboration and decisions on how to assess various bias domains need to be decided and tailored. For example, specific concerns related to selection bias may have to be addressed as to whether there is bias arising from poor response rate or loss to follow-up. Other points to reflect on may include what issues should be considered when assessing confounder control. For interventional studies, a specific mention should be made on the feasibility of blinding and what should be considered as realistic compliance on case-by-case basis.

Exposure assessment is another domain where a priori customisation of the study appraisal process for all study designs is warranted for almost every risk assessment. Methodologically related to information bias, there are particular features within each type of exposure assessment that need to be addressed. For example, certain analytical techniques may serve as gold standards and thus a greater weight may be put on the studies that use them, while evolution of the exposure assessment methodology over time is also something to consider. Although giving priority to different methods for exposure assessment or favouring certain methods is both logical and common, it is important to remember that no method is perfect and different methods may provide complementary information that strengthen each other (see Table 1, Section 4.2.1.2). Similar to the exposure assessment, the endpoint/outcome ascertainment is a feature that should receive the same attention.

4.4.1.3 | *Study eligibility*

Criteria for inclusion/exclusion of human studies should be defined a priori based on either the APRIO (EFSA Scientific Committee, 2023a) or the PICO-PECO/PO/PIT approach. Exclusion of studies solely based on design, population characteristics, or endpoint attributes should not be done. These study characteristics reflect validity aspects (both internal and external) that should be systematically addressed later in the assessment and will finally inform the 'WoE exercise'. Conversely, exclusion criteria when assessing eligibility should be focused on the outcome assessment and on the intervention characteristics (for clinical trials) or the exposure assessment/characterisation (for observational studies). Rejections based on those considerations should be based on established scientific knowledge allowing a firm conclusion that the study methodology is not appropriate.

The approach described above can be modified in data-rich situations, where it is known a priori that for certain health outcome categories there is an abundance of large studies with methodological characteristics related to generally lower RoB (e.g. randomised controlled experimental studies, prospective cohort studies) that can directly address the assessment question. An example of such a case is the 2022 NDA opinion on tolerable upper intake for dietary sugars (EFSA NDA Panel, 2022), where it was already known at the planning stage that there was a large number (e.g. hundreds) of high-quality prospective cohort and intervention studies, that adequately addressed the key assessment questions. There, retrospective case–control studies, cross-sectional studies, ecological studies and case studies/series were captured but not prioritised first. Throughout the assessment, this volume of evidence was recorded and could have been used by the NDA Panel, should the need for further consideration had arisen.

When screening for eligibility, special attention should be given to studies suspected of violating basic ethical standards. These may include, for example, older studies performed in mental health or correctional institutions where there is an indication that the informed consent procedure was inappropriate (Newcomer et al., 1999; Sheppard et al., 2020). Inclusion of such studies should be carefully considered.

Eligibility screening of studies should be performed in a transparent manner and documented accordingly. It is recommended to be done in duplicate by two independent parties and to follow a tiered approach where titles/abstracts are to be assessed first followed by the full text scrutiny. Disagreements between the involved assessors can be resolved by consensus or by a third arbitrator. The use of relevant software with the capacity of providing a 'rejection log' along with brief reasons for exclusion is strongly recommended. Moreover, a summary of the excluded body of evidence is recommended pertaining to the cumulative size of the excluded evidence both in terms of the number of studies and the sample size. Based on this brief report, a proposal can be made about whether the exclusion of this part of the evidence may pose a threat to the relevance, validity and generalisability of the scientific assessment. The eligibility criteria, rejection log and summary of excluded evidence should be accessible throughout the entire risk assessment process. Subsequently, included evidence will be organised within different priority tiers as described in the following sections.

KEY POINTS

- Define inclusion and exclusion criteria a priori *based on Terms of Reference*.
- Record rejections.
- As a general rule, do not exclude evidence based on study design and sample size. Exceptional deviations from this approach should be well justified.

4.4.1.4 | *Evidence base organisation, reporting and data extraction*

Once the eligibility criteria are set, no further exclusions should be made during the assessment. It is recommended to extract data related to the study design and the pertinent health outcome categories for the eligible studies at the eligibility screening level. If the health outcome category is a disease, ideally the proposed health outcome categories should follow the International Statistical Classification of Diseases and Related Health Problems (ICD) classification[30] along with other related endpoints (e.g. biomarkers of effect or functional outcomes), with related endpoints included under the same health outcome category. Furthermore, alignment, to the extent possible, with comparable health outcomes from experimental animal data should be made to facilitate better integration later. For example, lung (function) studies in animals can be linked to human epidemiological studies assessing pulmonary diseases and markers thereof. This process can be particularly challenging as regards not only differences in dose/exposure but also the validity and the relevance of certain animal data to human morbidity. Similarly, a lack of comparable measures between human observations and animal data may complicate comparisons across human and animal health outcome categories, where detailed histopathology and organ weights may be available from animal studies that cannot be measured in human settings. In such cases, the construction of an inventory of health outcome categories at the very beginning of the process may be useful where human and animal health outcome categories are displayed, and their similarities and differences are discussed (EFSA CEP Panel, 2023).

Concerning description of studies, the details of the data extraction process need to be considered in the context of the relevance and number of the available studies, keeping in mind that the purpose of data extraction is to help assess and summarise the evidence in a systematic and concise manner so that the end-users of the Opinion can understand why certain decisions were reached. If the number of studies is small, more details can be justified. When the number of studies is large, decisions are made based on more extensive information and then concise reporting is needed that summarises the 'overall picture'.

If the data extraction process is outsourced, proper communication and preparation are needed between those doing the data extraction and the experts who will use the data. The final stage of this communication and preparation stage should include the list with the data extraction items, the data characteristics of the items to be extracted (e.g. string or numerical data), as well as the output of a piloting exercise performed by the WG/Panel members that will be used as a reference by the contractor. It is recommended that a literature review software is used for this purpose, allowing for comparative assessments across exposure categories and/or health outcome categories. Moreover, summary master files generated by standard systematic review software can facilitate a meta-analysis, if appropriate.

As an output of the data extraction, the recommended structure of the relevant section is as follows: (a) evidence base overview, (b) a short text description of the prioritised studies, (c) descriptive plots or summary tables including study characteristics and study results, (d) evidence synthesis (if appropriate), (e) hazard identification summary.

As regards the evidence overview, it is recommended that the overview starts with a small 'setting the scene' paragraph reporting on the number of studies, the cumulative sample size, the median and range of study sample sizes and interquartile range (IQR), the available study designs and proportions thereof, the countries/populations of origin of the studies and proportions thereof, followed by the APRIO attributes [the population characteristics of the included studies (based on gender, age groups, risk factors), the exposure assessment methods/matrices and the range of the exposure levels, the

[30] International Classification of Diseases 11th Revision.

endpoints under study and proportions thereof]. For example, in the recent update of the risk assessment of hexabromocyclododecanes (HBCDD) in food (EFSA CONTAM Panel, 2021), an evidence base consisting of 13 studies was formed across diverse endpoints. A short summary of the evidence base related to this risk assessment was as follows:

The evidence base includes one cohort study, one birth cohort study (reported in four publications) and 11 cross-sectional studies where the HBCDD exposure was assessed simultaneously or even later than the endpoint ascertainment. The sample size of the included observational studies ranged from 34 to 71,415 participants. All the evaluated populations came from European countries except for five cross-sectional studies in which populations from the USA (n = 2), China, South Korea and Tanzania were investigated. The populations under study were diverse. Four studies recruited younger children or adolescents, while the remaining studies assessed adult female (n = 6), male (n = 2) or mixed (n = 1) populations. HBCDD exposure was assessed via serum biomarkers (n = 9), biomarkers in breast milk (n = 1), biomarkers in adipose tissue (n = 1), HBCDD measurements in dust (n = 1), or through merging dietary patterns and the presence of HBCDDs in food samples (n = 1). Birth weight/length, neurodevelopment and thyroid dysfunction were the three endpoint categories assessed in children. Subfertility, type 2 diabetes, thyroid hormone levels, severe endometriosis and ovarian endometrioma and breast cancer metastasis were the endpoints assessed in the adult populations.

As far as the tables including study characteristics and study results are concerned and due to the complexity of the design and analysis of epidemiological studies, the full panel of results cannot always be tabulated. Keeping in mind that the full data extraction dataset can be available as a supplementary file, the emphasis in the main text should be put to the prioritised endpoints and exposure categories (Table 3).

TABLE 3 Example of study reporting table (EFSA CONTAM Panel, 2024).

Reference study population design	Outcome definition	Population size (*n*), age	Arsenic exposure	Results	Additional information/ confounders
Prenatal					
von Ehrenstein et al. (2007) India Prospective mother–child cohort study	Wechsler Intelligence Scale for Children (no edition provided), Raven Colored, Progressive Matrices test, Total Sentence Recall test, Purdue pegboard test	351 5–15 years	w-As (μg/L) Mean (SD) *Peak lifetime* 147 (322) *During pregnancy* 110 (243) Tertiles < 10 10–49 50–99 > 100 < 10 10–49 50–99 > 100	B (95% CI) Full scale IQ *Peak lifetime* Ref 0.006 (−0.031, 0.33) −0.16 (−0.56, 0.23) −0.06 (−0.30, 0.18) *During pregnancy* Ref −0.047 (−0.38, 0.28) −0.007 (−0.36, 0.34) −0.002 (−0.24, 0.24)	Adjusted for age, sex, BMI, maternal and paternal education, father's occupation, number of rooms in the house, type of house building material, BMI and mother's age

KEY POINTS

- Define health outcome categories based on the current ICD along with other related endpoints (e.g. biomarkers of effect or functional outcomes (e.g. impaired cognition/reduced IQ, impaired growth)).
- Alignment of health outcome categories with animal data and organising evidence by health outcome categories.
- Extract data related to health outcome categories and study design, using literature review software as needed.
- Create the evidence overview, using evidence mapping tools where needed.
- Report the evidence base in a structured manner, including a short summary of the evidence base, followed by a description of the eligible studies, summary tables, and evidence synthesis and hazard identification.

4.4.1.5 | *Assessment of study relevance and quality*

After asserting study eligibility, structuring the evidence base and data extraction, the next step would be the assessment of individual studies so that they can be integrated through the WoE approach. Before embarking on that step of the hazard identification, several aspects on the ability of each study to answer the assessment question need to be considered. They include:

- the **internal validity** of the available studies (i.e. RoB and its likely impact on the effect estimates).
- relevance of the **study design** (different interventions and observational designs) **and their specific design characteristics** for answering the research/assessment question.

- **other factors** such as characteristics of the recruited population including age and underlying health that may impact the external validity of the study. The relevance of these considerations is highly specific to the assessment question.

The **internal validity of a study**, or RoB assessed though structured appraisal tools (see Section 4.4.1.5.1), is one (albeit not the sole) of the core attributes that will determine how an individual study is integrated into the assessment of the totality of the evidence. When the RoB evaluation is used to assign studies to different tiers (e.g. 1, 2, or 3), such tiering can be used as a prior for assigning weights to individual studies. As the RoB assessment may not address the direction or magnitude of potential biases or the appropriateness, weighing the evidence cannot be performed relying only on the RoB assessment. For this reason, exclusion of studies based on their tiering during hazard identification is not recommended. Such a decision can only be justified if the RoB evaluation necessitates a reconsideration of the initial eligibility criteria or the assessment question (or both). This recommendation is in line with most guidance documents on evidence synthesis. For example, in the NTP-OHAT handbook on systematic reviews it is specifically stated that: 'The tiering approach outlined by OHAT favors inclusion of studies unless they are problematic **in multiple key aspects of study quality**, an approach that offsets concern about potentially excluding studies based on a single measure, which could seriously limit the evidence base available for an evaluation'. As formulated, the term 'multiple key aspects' would not justify the exclusion of a study due to strong concerns for bias for one or two bias domains, which commonly leads to 'Tier 3' allocation.

Moreover, as mentioned, individual studies will always have different uncertainties and biases. By evaluating the totality of the evidence during hazard identification, it can be determined whether the combination of individual studies can overcome possible biases identified in individual studies, thereby providing a more robust assessment (see also Section 4.4.1.7).

There are several **key considerations on specific design characteristics** that should be considered along with the RoB assessment. Here the assessor needs to assess to what degree the study design and conduct is appropriate for providing meaningful input for answering the assessment question. These considerations include:

- **Sample size.** A reasonable question to ask for each study is if the sample size is sufficient to allow for the detection of an association or a relevant effect size. Of note, post hoc power calculations may not be of further help in such assessment (Heinsberg & Weeks, 2022) as the uncertainty around a significant or non-significant effect estimate is already quantified by the confidence interval.
- **Study duration and temporal separation between exposure and outcome:** These issues are not explicitly included in many RoB tools (National Toxicology Program, 2019), although in some of EFSA's risk assessments the question on outcome has been customised for NTP-OHAT to include temporal separation (EFSA CEP Panel, 2023). Intervention studies of too short duration or observational studies with too short follow-up are likely to be biased towards the null (Falkingham et al., 2010), regardless of how well they are conducted.

To give a perspective on the points above, a study can be well conducted in terms of scoring low on RoB (relatively high internal validity) but can at the same time be close to meaningless for answering the question it aimed to address (low relevance). As an example, in EFSA's re-evaluation of the non-nutritive sweetener thaumatin, several intervention studies were identified examining if oral intake of this protein might lead to allergic responses, using the skin prick test (EFSA FAF Panel, 2021). Some of those studies recruited very few participants (e.g. indicatively, $n < 20$). With the prevalence of common food allergies generally being well below 10%, no meaningful conclusion (or even an effect estimate) could be derived from such small studies. Similarly, an intervention study of only 2-week duration aimed at examining the effect of a certain diet on blood lipids would provide limited information when taking into consideration that lipid-lowering drugs achieve maximal benefits only after several (> 4) weeks of treatment (Gencer & Giugliano, 2020). The main take home message here is that a study can be well conducted in terms of RoB but at the same time may be poorly suited by design to answer the assessment question.

The relevance of **study design** was largely covered in our discussion on experimental studies (Section 4.1.2.1), observational studies (Section 4.1.2.2) and their strengths and limitations (Sections 4.1.5.1 and 4.1.5.2). The different sources of bias of different study designs were also highlighted in Table **2**, which provides some background to their strengths and limitations, which need to be assessed relative to the exposure and health outcome being addressed.

EFSA's assessment of the tolerable upper intake level (abbreviation: UL (upper level)) for selenium provides an example of the potential relevance of cross-sectional studies in risk assessment, among available streams of evidence (i.e. mechanistic and animal data, and other experimental and observational studies in humans) (EFSA NDA Panel, 2023). Cross-sectional studies have been carried out in seleniferous areas, for instance of China, India, and South and North America, and in such cases, it is reasonable to assume that current measured exposure should reflect long-term exposure among those who have been living in such areas characterised by an exceptionally high selenium content in soil and drinking water, particularly if residents' diet largely depends on locally produced foods. This makes a stronger case for causality compared to cross sectional studies examining exposure that are less likely to be stable over time. For this reason, the cross-sectional design is generally considered valid when investigating settings such as the aforementioned ones for selenosis.

Conversely, there are several instances in which cross-sectional (as well as some case–control) studies are subject to substantial RoB, and therefore should be considered with caution, and their exclusion in the evidence integration may be justified. This is the case, for instance, for some nutrients, such as sodium and potassium, the consumption of which is modified by early disease symptoms following dietary advice and/or metabolic and nutritional alterations.

Similarly, RCTs focusing on complex dietary changes for outcomes with long latency periods like cancer are not necessarily more informative than well designed prospective studies (Hébert et al., 2016). Also, the assumption that ecological studies provide no or limited information on cause and effect can be questioned (Li et al., 2020). The reason for re-highlighting these examples from previous sections is to emphasise that different study designs can be variably useful in risk assessments and that rules of thumb on study designs serve educational purposes rather than actual risk assessment endeavours.

In summary, RoB assessment alone is not sufficient for determining the weight assigned to a given study prior to integrating the evidence. It is the combined assessment of the study type, its RoB and specific design attributes that should determine the weight given to a study when performing the WoE assessment.

KEY POINTS

- Risk of bias assessment is only one of several aspects that need to be considered when assessing relevance and reliability of individual studies.
- Key study characteristics such as timing between exposure and outcome and sample size must also be thoroughly assessed.
- The relevance of a given study design for a specific assessment needs to be considered on a case-by-case basis.
- Exclusion of studies based on their assigned tier is generally discouraged when assessing epidemiological evidence.

4.4.1.5.1 | *Use of RoB tools for assessing individual studies*

Critical appraisal tools provide a **structured** and **transparent** approach to assess the risk of systematic biases that may occur in individual studies (e.g. internal validity). In terms of use, it is important to take into consideration that many RoB tools were initially designed to assess RCTs. Some were later adapted, and new tools have been developed to address observational designs. Given the variety of observational designs and shorter history of use of RoB tools for such designs, some customisation (or tailoring) is often needed. When used on the basis of those principles, RoB tools provide a transparent structure for considering different types of bias, which is an important information for further evidence synthesis and assessment of uncertainty.

Another point to consider is that, for a particular hypothesis, the bias domains to focus on when using RoB tools may differ. For example, if environmental exposure to a chemical is hypothesised to cause lung cancer, then confounding by smoking may be of concern. Occupational cohort studies, for example using past employment records usually do not include smoking data, whereas case–control studies conducted in the general population usually include this – if smoking is not a strong confounder in the case–control studies, it is also not likely to be a confounder in cohort studies conducted in comparable populations. On the other hand, if chemical exposure occurs occupationally, then confounding is unlikely to be important (there may be little or no confounding in comparisons of different groups of manual workers[31]), but other potential biases (e.g. the healthy worker effect) may be of more concern. Thus, the outcome of the RoB assessment should focus on the bias domains that are expected to have the highest influence on the uncertainty related to the assessment question.

Regardless of study design, the domains covered by RoB tools include **selection bias**, **confounding** and **information bias** (see definitions in Sections 4.2.1.4 and 4.2.1.4.3).[32] When applying RoB tools to individual studies understanding the different processes that may lead to different biases, their complexity and knowing what to look for is fundamental for facilitating proper use. Below a short comparison for the three main types of biases that may occur in experimental and non-experimental designs are given:

- **Selection bias:** For **RCTs**, this relates to the appropriate randomisation of study subjects, including both the process of allocation and the procedures to conceal it. These factors are relatively straightforward to assess if the baseline characteristics across groups and the method of randomisation are properly described. If not, considerations around reported balance of the treatment groups at baseline provide important information on possible RoB.[33] For **observational studies,** selection bias relates to the procedures used to select study participants. This can be difficult or impossible to evaluate as information on whether study participants are systematically different from those who were eligible (but not

[31]Provided that other co-exposures can be neglected.

[32]Note that these biases can be divided further into specific types of biases and/or can in principle be grouped differently (Mansournia et al., 2017). However, the text here aims to explain differences in how such biases may occur depending on study design (experimental vs non-experimental) and what to look for when identifying risk of bias in practical terms.

[33]Selection bias as defined here for RCTs would be source of confounding, that is deviation from randomisation at baseline may result in imbalance in factors that can confound the results for the experiment.

recruited) is usually missing. Selection bias can also occur **in both experimental and observational studies** if participants are selectively lost to follow-up during the study period. In principle, the risk of such bias can be assessed by comparing the baseline characteristics of those lost to follow-up vs those who were not.

- **Confounding:** Assuming that the randomisation process is appropriate and study size is adequate, bias due to confounding in **RCTs** may still occur if there are **deviations from intended treatment**.[34] Such bias may occur if participants or investigators are not blinded to treatment allocation. The intention to blind participants and investigators can easily be evaluated by study reporting, but how influential blinding is in terms of avoiding differential treatment is more difficult to evaluate. For **observational studies** assessing confounding is even more complex as the exposure is rarely randomly allocated by nature. As a result, confounding has to be controlled for. Even if known confounders are accounted for, confounding due to unidentified factors or improper confounder control in the analysis ('residual confounding') can never be fully excluded – although it may be possible to estimate its likely strength and direction (and in some instances, residual confounding may be very small). Compared to RCTs, where some inferences on differential treatment can be made regardless of the research question, assessing bias due to confounding in the observational study setting is most often study specific and requires **expert judgement and experience** on a case-by-case basis (e.g. what are the likely sources of confounding for this particular setting).
- **Information bias:** Information bias involves misclassification of the study participants with respect to exposure, outcome or confounder status. For **RCTs and observational** studies, there are no differences in terms of how outcome misclassification may occur or how such biases are assessed. Problems with exposure and confounder misclassifications are however more specific to observational designs, as both exposure and relevant confounders need to be assessed (quantified) as opposed to being largely taken care of by design in RCTs. Reporting bias is also one form of information bias that is commonly assessed in RoB tools. Standards exist of both designs, but since RCTs are generally conducted to test one or very few hypotheses, selective reporting is often easier to identify compared to observational studies that can be of more explorative nature.

When using RoB tools it is important to note that all existing appraisal approaches have their strengths and limitations (Bero et al., 2018). Ideally, different instruments should lead to the same conclusion when applied to the same study. However, very comprehensive tools may indirectly lead to too much focus on minor issues. On the other hand, simple tools may lead to important aspects being overlooked in some cases. Different RoB tools may also use different formulations for questions aimed at assessing the same types of biases. To give an example of differences in formulations across RoB tools for **human studies** we used **selection bias** as an example. To demonstrate that, as an example, we compare the formulation in the NTP-OHAT[35] and USDA Nutrition Evidence Systematic Review[36] (**NESR**) RoB tools for observational designs:

- The **NTP-OHAT** tool asks a single question: 'Did selection of study participants result in appropriate comparison groups?'
- The **USDA-NESR** asks: 'Was selection of participants into the study (or into the analysis) based on participant characteristics observed after the start of exposure?' Based on the answer this question (yes/no), several sub-questions follow, including if post-exposure variables may have influenced exposure and outcome.

A similar formulation as used in the USDA-NESR is also used in the Robins-I tool for non-randomised interventions (Sterne et al., 2016). On the other hand, the Newcastle-Ottawa Scale for assessing the quality of non-randomised studies again uses a slightly different formulation and different scales for cohort and for case–controls studies (unlike NTP-OHAT and USDA-NESR). The purpose of this example is to highlight that formulations used to capture possible selection bias can be quite different. Without further consideration, this may lead to different conclusions if the focus of the assessment is on the exact wording of individual questions (the bias that should be captured is the same, independent of how the question is asked).

For human **RCTs** more consistent formulations are generally in use. As an example,

- **the Cochrane risk of bias tool** (RoB 2.0) asks: 'Was the allocation sequence (1) random (2) and concealed until participants were enrolled and assigned to interventions; and (3) did baseline differences between intervention groups suggest a problem with the randomisation process?'
- Similarly, the **NTP-OHAT** asks: '(1) was administered dose or exposure level adequately randomised, (2) was allocation to study groups adequately concealed; and (3) did selection of study participants result in appropriate comparison groups?'

Although slightly different, the two formulations are for all practical purposes identical. The similarity for this RCT example may be due to the longer history of RoB tools for RCTs which has perhaps resulted in better harmonisation. In contrast, there is currently no single standard or consensus about the best approach for assessing RoB in observational studies (Page et al., 2018; Viswanathan et al., 2013).

In summary, it is crucial that those using different RoB tools are aware of what types of biases are being captured and what to focus on and look for, being able to understand the issues queried by each of the individual

[34] This is often referred to as performance bias.

[35] https://ntp.niehs.nih.gov/sites/default/files/ntp/ohat/pubs/riskofbiastool_508.pdf.

[36] https://nesr.usda.gov/sites/default/files/2019-07/RiskOfBiasForNutritionObservationalStudies-RoB-NObs.pdf.

questions, which are formulated to guide the assessor. A simple checklist-type formulation can never cover all scenarios encountered when appraising different studies, but they do provide a structured approach for assessing biases, which needs to be tailored for each assessment.

Finally, to put the content of this section on use of RoB tools in perspective, Appendix F Appraisal of different studies using a RoB tool contains a series of appraisal examples aimed at demonstrating how appraisal of individual studies using a RoB tool could be performed. For this purpose, the examples cover appraisal of both double blind RCTs and randomised nutritional intervention studies, as well as observational designs (cohort and case control studies) relevant to chemical risk assessment. Each of these examples is aimed at highlighting the principles and considerations that need to be considered when assessing different types of biases for individual studies. The examples are chosen for illustrative purposes only and some of the points made could be subject to a different interpretation.

4.4.1.5.2 | *Summarising the outcome of a RoB assessment*

After assessing a study using RoB tools, the reviewers' judgements attached to each question in a given appraisal tool are documented and translated into an overall summary assessment. This could, for example, be in a form of a

- short text summary
- grouping of studies according to types of bias that may occur (see section below on evidence synthesis).
- ranking of studies into tiers (from low to high RoB) or
- numerical scoring.

Numerical scoring here refers to the approach of assigning a numerical value to each RoB question that is then summarised in some way (perhaps using different weights) into an overall score. The use of numerical scores (or scales) for assessing quality or RoB is currently explicitly discouraged by the Cochrane handbook (Higgins & Green, 2011). Despite their proffered convenience and simplicity, scaling systems rely on weight assignment on different items of the scale. Such an approach bears three major limitations: it is difficult to replicate, it is not transparent to the final user of the risk assessment, and it does not accurately reflect study validity (Emerson et al., 1990; Jüni et al., 1999; Schulz et al., 1995).

Ranking of quality of the evidence of a study into tiers is a better alternative to numerical scoring, as it is more transparent because it relies on fewer well-defined attributes (bias questions) that determine the overall summary assessment. Summary assessments by ranking (into tiers) allow not only for an overview of the evidence within each tier but also for a structured appraisal of the whole body of evidence. Such a summary can be done in the form of heat maps.

One potential problem when relying on summary assessments by ranking into tiers is that it may obscure the fact that judgements on the overall body of evidence should always consider the **type,** and possible **direction and magnitude of** potential **biases** identified across different studies. Even though it is often difficult to assess such parameters, it is important. Summarising RoB by tiers tends to hide these important attributes. In cases where most studies suffer from the same type of bias (including possible direction), assessing the overall body of evidence by looking at individual tiers from each study is more justified. In other cases, the type and direction of biases must be assessed in parallel.

For example, suppose that several studies of the same design have been rated as having either moderate or high RoB, but all the studies consistently show the same association of exposure and health outcome. Based on simple tiering (or scoring), some assessors may conclude, when weighing the evidence, that the quality is low and that limited conclusions on causality can be made due to the present RoB. However, a more careful inspection may reveal that there are different sources of biases across these studies, with some scoring low on selection bias, but high on other aspects, while other studies scoring low on confounding or information bias score high on other aspects. Further evaluation may then reveal that the direction of these potential biases across studies is likely to be different. In that case, it is highly unlikely that the consistently observed associations are due to these potential biases (since they would work in different directions). Such a scenario is not just hypothetical, and the approach of taking both type and possible direction of bias into account compared to just looking at the RoB scoring can lead to different conclusions. Of course, if the risk of same type of bias would have been present in most or all the studies evaluated and the expected direction is anticipated to be the same, then it is not surprising if the studies produce similar findings – they may all be wrong. A further discussion of these issues is the subject of Section 4.4.1.6.

In summary, the considerations above once again highlight the importance of assessing both the magnitude (where possible) and the direction of different biases when evaluating individual studies. This should allow for better evidence integration than simply focusing on individual study tiers.

4.4.1.6 | *Using causal inference by triangulation*

As discussed above, the outcome of a RoB assessment is just one of several key aspects that need to be considered in evidence integration. Parallel to wider use and experience gained of applying systematic reviews in evidence synthesis, the limitations of prioritising studies based on simple tiering through RoB assessments is increasingly being recognised (Steenland et al., 2020). This has led to some methodological development on how to make more thorough use of RoB assessment in evidence integration. One approach that has been suggested is causal inference by triangulation. That approach is more in line with the Bradford Hill viewpoints that were intended to aid integration of all available evidence but

not to 'judge' individual studies. The concept of 'triangulation' extends the approach of Bradford Hill, in that it explicitly seeks to consider evidence from different types of studies and/or studies in different contexts, so that the strength and direction of various possible biases can be assessed.

To give an example, Pearce et al. (1986) conducted a case–control study of pesticide exposure and non-Hodgkin lymphoma, which involved two control groups: (i) a general population control group and (ii) an 'other cancers' control group. It was hypothesised that the former control group would produce an upward bias in the estimated odds ratio (differential recall bias if healthy general population controls are less likely to remember previous exposure than the cancer cases), whereas the 'other cancers' control group could produce a downward bias in the estimated odds ratio (if any of the other cancers were also caused by the pesticide exposure under study). Both groups yielded similar findings, indicating that neither bias was occurring to any discernible degree. This provided strong evidence that little recall bias (a type of information bias) or selection bias was occurring. However, a simple tiering of this study as a result of bias evaluation likely would have led to low prioritisation as both components of the study might have been considered high RoB (albeit in opposite directions).

Triangulation is also consistent with the approach advocated by Savitz et al. (2019) who argue that RoB assessments should focus on identifying a small number of the most likely influential sources of bias, classifying each study on how effectively it has addressed each of these potential biases (or was likely to have the bias) and determining whether results differ across studies in relation to susceptibility to each hypothesised source of bias. For example, information bias is unlikely to explain positive findings of studies with non-differential exposure and/or outcome misclassification if stronger findings are found among studies with more accurate assessment. A good example of triangulation by assessing exposure quality can be found in Lenters et al. (2011) who evaluated the association between asbestos and lung cancer. In this analysis, stratification by exposure assessment characteristics revealed that studies with a well-documented exposure assessment, larger contrast in exposures, greater coverage of the exposure history by the exposure measurement data, and more complete job histories had higher risk estimates per unit dose than did studies without these characteristics. Similar observations have been made for other important environmental and occupational exposures (Vlaanderen et al., 2011).

KEY POINTS

- An alternative to study tiering is to use the outcome of the risk of bias evaluation for causal inference by triangulation.
- This may allow for more thorough evaluation of biases and their possible consequences for the risk assessment rather than just focusing on the presence of possible biases in individual studies.

4.4.1.7 | *Weight of evidence assessment*

The aim of this section is to provide specific guidance on how to assemble and integrate evidence from human studies in line with EFSA's WoE guidance (EFSA Scientific Committee, 2017).

EFSA's guidance on WoE provides a flexible framework for assembling available information into lines of evidence and weighing them at each step of the integration process. The NTP-OHAT guidance on systematic reviews (National Toxicology Program, 2019) and GRADE (Morgan et al., 2019) can be considered one of several possible implementations under that framework. A modified illustration from the WoE guidance is shown in Figure 2. The figure shows how all evidence (human, animal, in vitro and in silico) can be assembled and integrated.

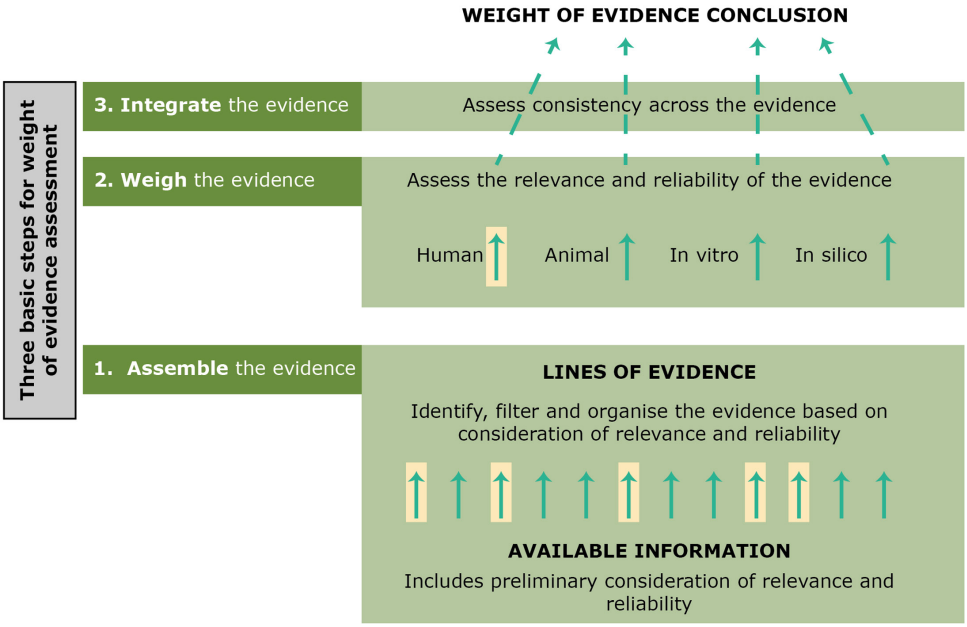

FIGURE 2 General framework for hazard assessment, assembling and integrating different lines of evidence. Adapted from the EFSA guidance on WoE (EFSA Scientific Committee, 2017). The yellow bars indicate human lines of evidence.

Although human studies are shown as one line of evidence in Figure 2, this line would reflect integration of several lines of human evidence covering all studies meeting the inclusion criteria (of varying design addressing different exposure and outcome measures in different populations). In some cases, e.g. within the area of nutrition, the evidence base can be quite large with many observational and intervention studies being available for each health outcome category (EFSA NDA Panel, 2019, 2022). In other areas, such as for certain contaminants or food additives, much fewer studies are usually available (EFSA FAF Panel, 2021; EFSA Scientific Committee, 2023b).

Ideally, the evidence should be assembled around a specific health outcome category. Within each health outcome category, separate lines of evidence can be established, taking into consideration:

- Different study population subgroups, such as children, pregnant and lactating women, adults, and the elderly
- Timing of exposure
- Representativeness of the study population for the population of concern

The reason for considering these factors when assigning studies to different lines of evidence is usually expected differences in sensitivity or susceptibility with respect to exposure. This could, for example be due to differences in age, health and underlying nutritional status and, when relevant, genetic background (see Figure 3).

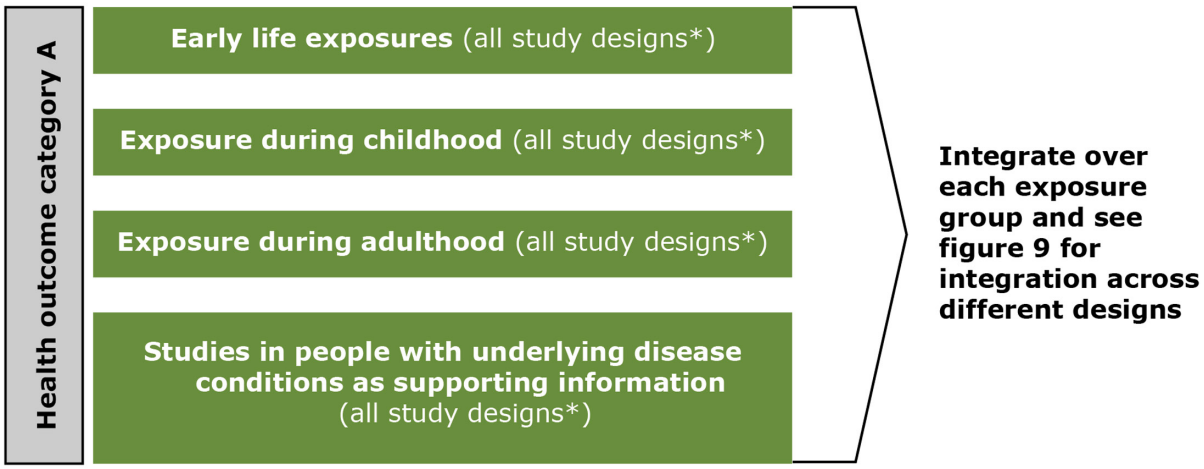

FIGURE 3 Different sub-categories within a given health outcome category constructed based on timing of exposure or population under study (*within each sub-category, different study designs can be assembled as shown in Figure 4).

Within each sub-category available studies can be grouped according to their design (see Figure 4), as this allows for integration across studies with similar sources of design-related attributes and potential biases (see Table **2**). Deviations

from this approach can be taken on a case-by-case basis and this may, for example, be appropriate when using the method of triangulation discussed above.

Concerning Figure 4, more weight should be given to studies that by design are better suited to answer the assessment question. A priori this does not mean that RCTs should receive the highest weight according to some pre-defined evidence pyramid (Pandis, 2011). For example, if the outcome under consideration has a long latency period such as CVD, well conducted cohort studies could be prioritised. RCTs assessing intermediate CVD risk factors could then be used as supportive line of evidence when making judgement on causality. In other cases, if several well-conducted and sufficiently powered RCTs are available that can address the assessment question, such evidence would take precedence over observational studies that would then provide supporting evidence. In short, the study design (different interventional or observational) should not by default determine the weight assigned to each study without careful consideration on its suitability to answer the assessment question.

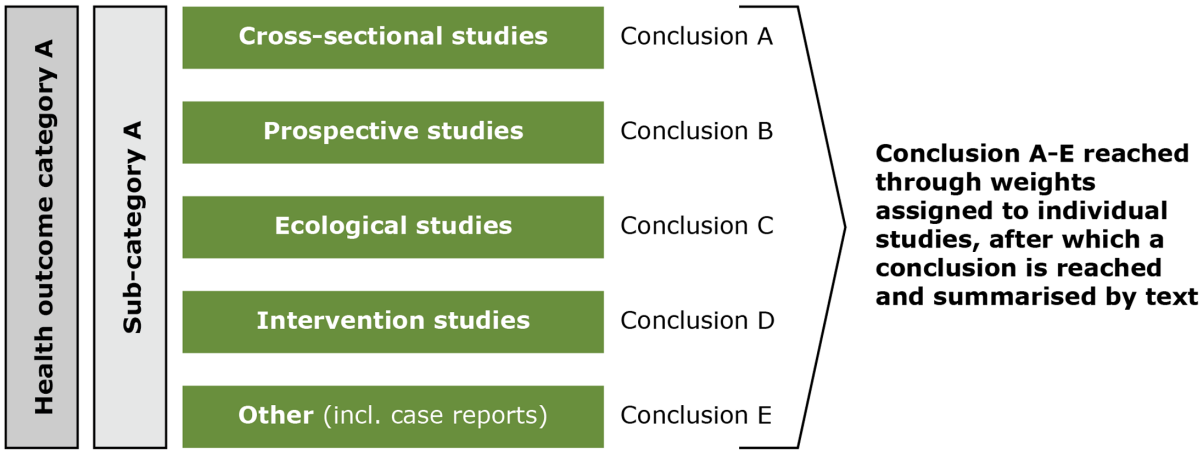

FIGURE 4 General framework for assembling and weighing epidemiological evidence for a specific health outcome category.

Based on the assembling studies as suggested in Figures 3, 4, integration across studies can be performed by assigning weights or confidence levels to individual studies based on the study design, other design specific considerations, the RoB evaluation and other relevant factors identified. Then the evidence can be integrated across all studies (Figure 4) within each sub-category (Figure 3). Although in principle the weights or confidence levels assigned to individual studies could be both qualitative or quantitative, a common approach is to provide a short-written argument explaining how weight/confidence is assigned, taking all the above-mentioned factors into consideration. In other cases, such as the NTP-OHAT guidance on systematic reviews,[37] confidence levels assigned to individual studies are already predefined based on study design (very low to high initial confidence). Using those levels as starting point, the evidence integration then takes into consideration the RoB evaluation and other factors that may result in upgrading or downgrading of the body of evidence (see the NTP-OHAT guidance[38] and, for a practical example, the re-evaluation of erythritol (E 968) as a food additive (EFSA FAF Panel, 2023)).

It is important to note that the grouping of studies as suggested in in Figures 3, 4 may not be feasible when few studies are available. Formal procedures for WoE (Higgins et al., 2023; National Toxicology Program, 2019), that tend to be time consuming, may also be less relevant when the evidence is small (although in principle they can be applied). When the evidence base is small, a simple narrative description reflecting on the strengths and limitations of the evidence can be more appropriate than a structured approach designed around a large evidence base. The Scientific Committee's opinion on copper provides a good example of how the WoE guidance (EFSA Scientific Committee, 2017) can be applied in a structured but simple manner to an assessment with few available studies (EFSA Scientific Committee, 2023b). Consistency or inconsistency across studies may be highlighted in the summary of all relevant studies, which may be done narratively or graphically, or more formally through a meta-analysis.

[37] https://ntp.niehs.nih.gov/sites/default/files/ntp/ohat/pubs/handbookmarch2019_508.pdf.

[38] National Toxicology Program-Office of Health Assessment and Translation - Handbook for Conducting Systematic Reviews for Health Effects Evaluations.

KEY POINTS

- Evidence should preferably be assembled around specific health outcome categories.
- Within each health outcome category, separate lines of evidence can be established, taking into consideration different study population subgroups with known (or assumed) differences in sensitivity or susceptibility with respect to exposure.
- If the number of available studies allows, grouping studies by design is one option for structured evidence integration. This allows for weighing the combined evidence taking the complimentary strength and weaknesses of each design into consideration.

4.4.1.8 | *Integrating several lines of human evidence for related health outcomes*

Although each health outcome category is usually constructed around a range of related health outcomes, how broadly each health outcome category is defined depends on the size of the evidence base (i.e. how many studies). For example, in the EFSA opinion on BPA (EFSA CEP Panel, 2023), a range of partly unrelated outcomes such as sex ratio, live birth rate, follicular phase length and endometrial wall thickness were all included in one health outcome category named 'female fertility'. The few studies addressing each health outcome justified such grouping. The obvious limitation, however, is that integrating the evidence across a few weakly related outcomes is challenging and may not be substantiated in terms of suspected mode of action. An alternative approach may be equally challenging as the conclusions drawn from single or very few studies are usually not very robust, irrespective of how well they have been conducted.

The NDA opinion on sodium provides an example on how to construct different lines of evidence in situations where the evidence base is large (EFSA NDA Panel, 2019). In that opinion, a few hundred epidemiological studies addressing cardiovascular health outcomes were identified. In such cases, it is possible to construct different lines of evidence within each health outcome category around related health outcomes. Here the lines of evidence for sodium excretion and (1) raised blood pressure, (2) hypertension and (3) stroke were constructed. As expected, the number of both experimental and observational studies for raised blood pressure was large. Fewer studies were available for hypertension. The number of RCTs and non-experimental studies on the effects of sodium intake on blood pressure was quite large, which allowed to conduct a dose–response meta-analysis between 24-h sodium urinary excretion and both systolic and diastolic blood pressure. Therefore, only interventions examining the effect of sodium reduction on blood pressure were included. For hypertension, fewer intervention studies were available than observational studies. For stroke or coronary heart disease (CHD), only three observational cohort studies were available, and the same was true for the risk of overall CVD. In such cases, conclusion from clinical markers, which are intermediate steps or risk factors (e.g. changes in blood pressure), can provide key supporting evidence for the disease endpoint of concern (e.g. stroke[39]).

Similar grouping of related health outcomes into different lines of evidence, each re-enforcing each other, is strongly encouraged in risk assessment. The same approach could, for example, be made for other related health outcomes such as grouping blood lipids, hypercholesterolaemia and CHD into three related but separate lines of evidence (see Figures 4 and 5), provided that they influence the health outcome being assessed. For each health outcome, all study designs should preferably be included because the findings of interventions conducted in controlled conditions in healthy volunteers or selected populations may not be very representative compared to the general population or older persons more at risk.

The key here is to use the available evidence to the extent possible, building a case for or against causality for a disease endpoint with supporting studies examining related clinical markers (or risk factors). Thus, the assembled lines of evidence should also include health outcomes that may ultimately not necessarily be considered for establishing a HBGV.

[39]Few individual studies on measures of 24-h sodium excretion do not provide robust evidence for lowering of blood pressure because of sodium reduction making the independent conclusion for that line of evidence less plausible. The combined evidence from intervention studies for a positive association of sodium intake with blood pressure and a possible negative association with stroke was used to identify 2 g of daily sodium intake as the 'safe and adequate intake for the general EU population of adults'.

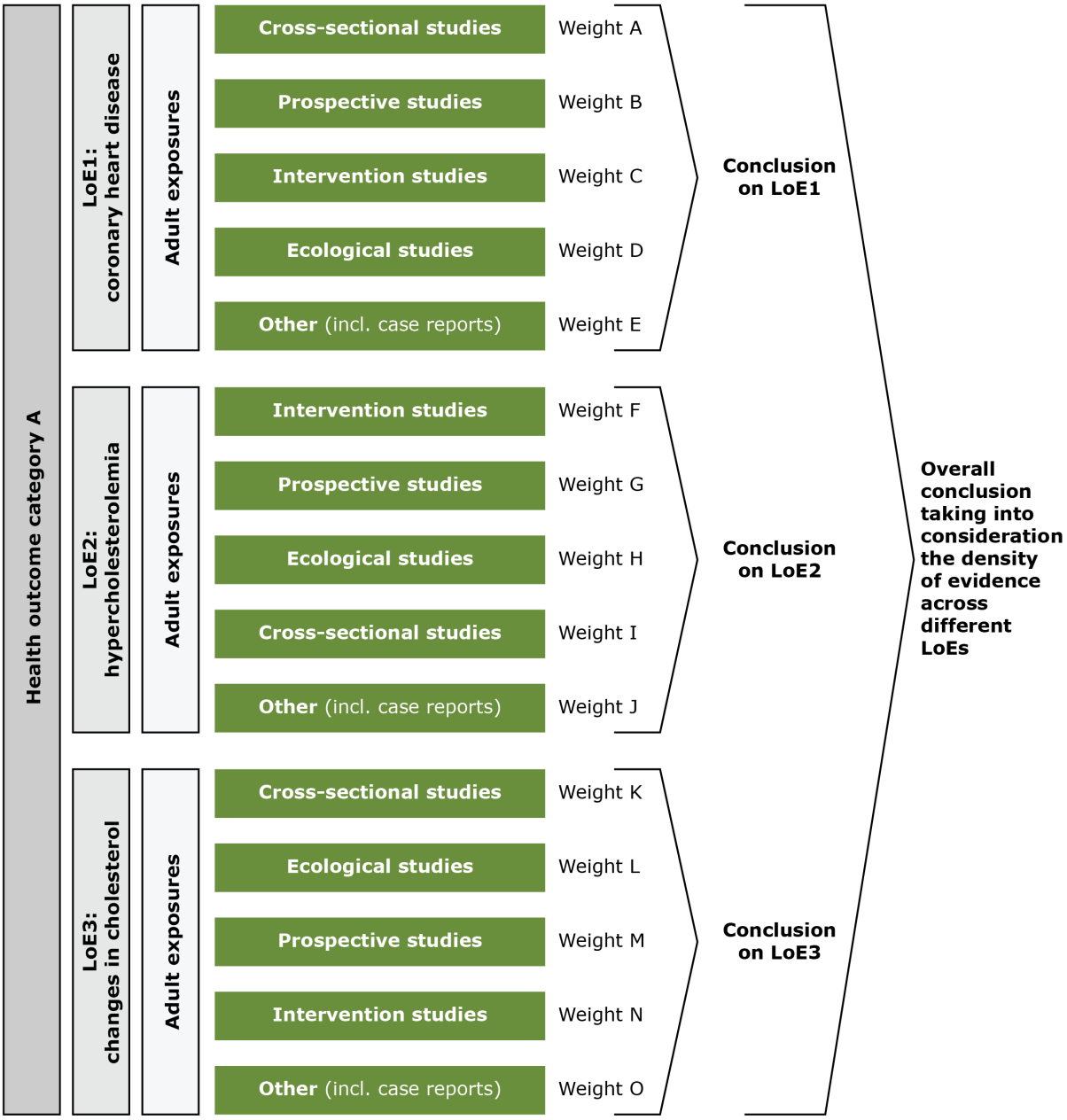

FIGURE 5 Aligning the available human evidence within a given health outcome category into different lines of evidence (LoE) based on individual health outcomes.

KEY POINTS

- It is recommended to assemble different lines of evidence within each health outcome category around related health outcomes.
- Conclusions reached from biomarkers of effect which are intermediate steps or risk factors for the disease endpoint of concern could be used as evidence for or against causality for the disease endpoint of concern, especially where there is limited density of evidence on the endpoint.

4.4.2 | Integrating the evidence from human epidemiological studies with other streams of evidence

Simultaneously with the retrieval, appraisal and WoE for human studies, a similar process is carried out for other evidence streams (i.e. toxicological data), see Figure 1. It is not within the scope here to provide guidance on this process for other evidence streams but it is recommended to align the endpoints with comparable human health outcomes.

When integrating human evidence with other evidence streams, information and assessment of mode of action (MoA) provides a structured biologically driven way of integration (see also Section 4.1.5 on cause and effect). Key considerations

for a MoA assessment have been described in the EFSA/ECHA guidance for the identification of endocrine disruptors (ECHA and EFSA, 2018) and in a joint report by the Committee on Toxicity and Committee on Carcinogenicity in the UK (2021). These considerations encompass for example:

- Have substance-related adverse effects been observed in experimental studies in laboratory animals?
- Is there sufficient information from those studies to establish a MoA for each Key Event[40]?
- Is the relationship between the Key Events in the MoA biologically plausible? This should be assessed based on a broader knowledge of biology.
- Is it plausible that the effect can occur in humans based on toxicokinetic and toxicodynamic considerations?
- Does the available evidence support the biological plausibility for the MoA? Here, the evidence must be assessed for dose and temporal concordance.

Use of AOPs (either existing or postulated for the purpose of the risk assessment) can be particularly useful, as these provide links between data generated from in vitro/in silico methods, animal models and humans. Furthermore, depending on the maturity of the AOP, it may facilitate quantitative assessment of the relationships between the key events, thereby allowing for predictions on adversity based on in vitro methods measuring key events, including how to integrate exposure considerations by using physiologically based pharmacokinetic (PBPK) models. The guidance for identification of endocrine disruptors also provides guidance on the reporting and examples and how the evidence is mapped corresponding to the level of biological organisation, thus aligning to MoA/AOP frameworks (ECHA and EFSA, 2018). Furthermore, capturing the strength of the evidence is also recommended when conducting the WoE assessment. For more detail, see Section 4.4.2.1 on AOP-based integrated approach to testing and assessment (IATA).

Considerations for the integration of MoA can be found in several assessments of pesticide active substances on assessment of identification of endocrine disruptive properties. One relevant example is the assessment of metribuzitan, where it was concluded that the substance had endocrine disruptive properties regarding the thyroid modality, but not for oestrogen, androgen and steroid modalities (EFSA, 2023a). An example on MoA considerations related to species differences and human relevance can be found in the Opinion on inorganic arsenic (EFSA, 2009). Here it was concluded that data from experimental animals could not be used for risk characterisation, because of toxicokinetic differences between humans and animals in their ability to methylate inorganic arsenic and differences in excretion of the metabolites (humans excrete more).

In certain situations, experimental animal data are considered but not fully integrated into the risk assessment, e.g. when the human data are abundant and robust, as is often the case for nutrition. In such cases only certain data from laboratory animals may be integrated, such as data on absorption, distribution, metabolism and excretion (ADME). An example of this is the 2015 opinion on caffeine, where rodent studies were not considered in the hazard characterisation (EFSA NDA Panel, 2015). Another example is the scientific advice on a tolerable upper intake level for dietary sugars (EFSA NDA Panel, 2022). In this case, the large availability of human data about both ADME of dietary sugars and their potential adverse effects, arising from both experimental and observational epidemiological studies, allowed the assessment to be based almost entirely on human evidence.

For regulated products and non-regulated chemicals, differences exist in the availability and nature of data. The initial approval process (pre-marketing authorisation) of regulated products, such as food additives and pesticides, is based on toxicological experiments in animals and in vitro/in silico data, as determined by the respective data requirements, and, if it exists, on human experimental data (e.g. food additives). However, for later post-marketing assessments, human observational studies might be available, and these should be taken into consideration. The latter are often not designed to support the authorisation process as they consider post-marketing observations where the population is not only exposed to the chemical under assessment. For non-regulated chemicals, all evidence is taken account of, and the fact that available human observational studies as well as laboratory animal studies are rarely designed to directly address the risk assessment questions needs to be considered.

4.4.2.1 | *Approach for systematic integration of epidemiological data with other streams of evidence*

One generic approach for the integration of evidence from human studies with other toxicological data for regulated product and non-regulated chemicals is proposed based on an approach initially developed for regulated products, in this case pesticides (EFSA PPR Panel, 2017), see Figure 6.

[40]Key event = A measurable change in biological state that is essential, but not necessarily sufficient, for the progression from a defined biological perturbation towards a specific adverse outcome. KEs are represented as nodes in an AOP or MOA diagram or AOP network and provide verifiability to an AOP or MOA description (Committee on Toxicity and Committee on Carcinogenicity, 2021).

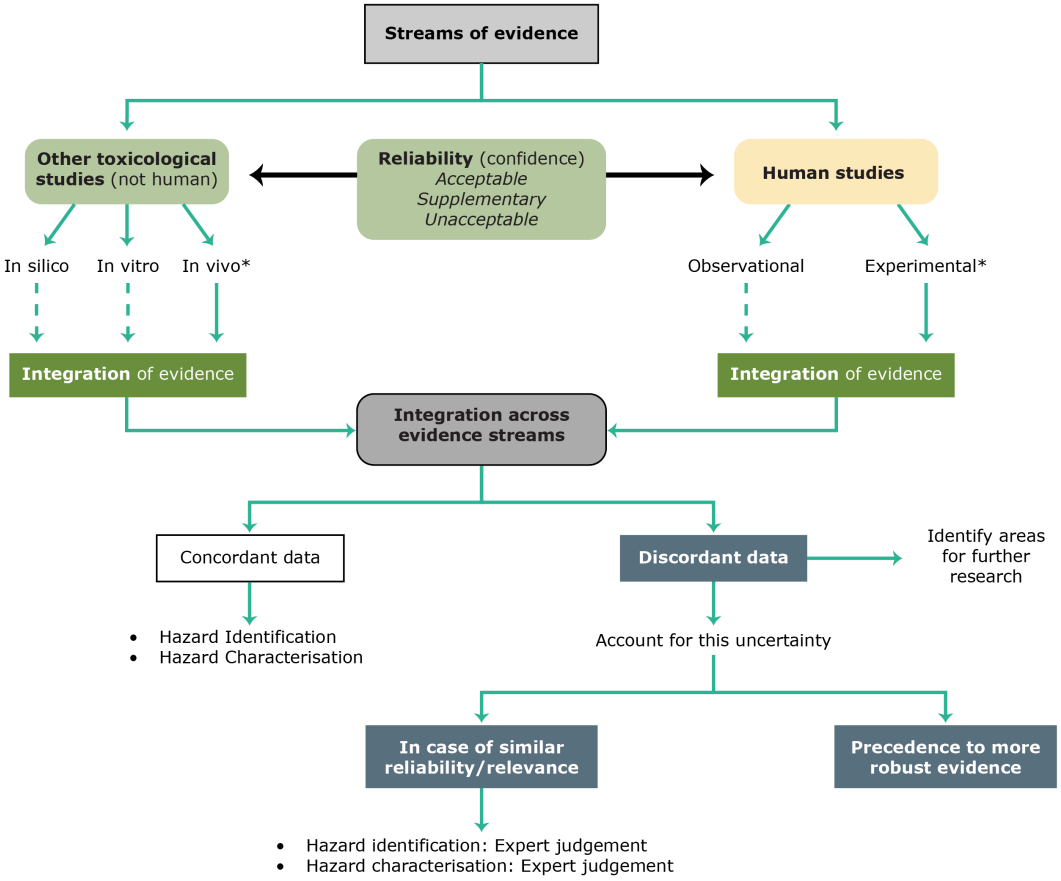

FIGURE 6 Approach for systematic integration of epidemiological data with other streams of evidence, see supporting explanations below (*generally, precedence will be given to these studies).

After all relevant and reliable epidemiological evidence has been identified, the separate lines of evidence need to be integrated with the other relevant and reliable lines of evidence (laboratory animal, in vitro and in silico data). Available reliable and relevant laboratory animal studies would generally take precedence over in vitro/in silico data unless there is evidence that the laboratory animal model studies do not capture the effect in question. Furthermore, studies compliant with OECD[41] guidelines are by default considered to have a high reliability, unless there is evidence of the contrary.

The toxicological data might be corroborated by mechanistic in silico/in vitro data and thus strengthened, whereas a meta-analysis of the human studies may provide more precise effect estimate for possible health effects. Studies that are found to be more relevant for the health outcome in question are to be given more weight, regardless of whether the data come from human or laboratory animal studies.

Where human observational/experimental data are of highest relevance, and they are supported by a mechanistic scientific foundation (see above on MoA considerations), they should take precedence over the experimental animal data. When human and toxicological data are judged to be of similar relevance, it is important to assess their concordance (consistency across the lines of evidence) in order to determine which data set may be given precedence.

Concerning human observational data, it is important to stress that a single study in isolation, no matter how well-conducted, generally only provides part of the evidence, as there are variations in study settings and conduct and possible influence of biases (that can rarely be fully eliminated). A basis for WoE and assessment of concordance ideally requires several studies that vary by design and/or study population/setting so that consistency can be assessed, as pointed out in Section 4.4.1.7. As discussed above, a robust human experimental database may in some cases be used as a stand-alone, without a full integration of other streams of evidence.

In case of **concordance** between human and other toxicological data, the risk assessment should use all the data, as both yield similar results in either hazard identification (e.g. both indicate the same hazard) or hazard characterisation (e.g. both suggest similar levels). Thus, both can reinforce each other, and similar mechanisms may be assumed in both cases.

In case of **non-concordance** and **similar reliability/relevance**, one needs to take account of this uncertainty. For any non-concordance, the reason behind the difference should be considered. It is a matter of expert judgement to select the most appropriate study/studies depending on the situation, for example on the data density, dose-spacing/exposure in different studies, mechanistic understanding (toxicodynamics and toxicokinetics), possible species differences, specific biases in the evidence streams.

[41] Organisation for Economic Co-operation and Development guidelines.

Regardless of the consistency between experimental animal and human observational data, an assessment of biological plausibility is warranted by including other lines of evidence (mechanistic data and experimental animal data). Challenges are present when results related to specific endpoints are markedly inconsistent between humans and animals (e.g. the association between pesticide exposure and Parkinsonian disorder (EFSA PPR Panel, 2017)), or when the health outcome associated to exposure has no animal correlate (e.g. the association between exposure to the pyrethroid pesticide, deltamethrin, and autism spectrum disorder and attention deficit hyperactivity disorder (EFSA PPR Panel, 2021). In both cases, biological plausibility for the associations found in epidemiological studies was strengthened by means of developing AOPs.

Three different approaches for using AOPs may be applied that improve the interpretation of human data by providing a plausible mechanistic link to adverse outcomes supporting the contextualisation in the risk assessment process:

1. Look for an already developed AOP, see OECD AOP wiki.[42] This platform is both open access and is continuously being updated. If a relevant AOP(s) exist and in silico/in vitro/in vivo data exist that supports that the chemical under assessment triggers the AOP at relevant dose/concentration levels, then there is support for biological plausibility for causality. The assessment of endocrine disrupting properties of pesticides and biocides is in practice applying such an approach on a routine basis.
2. Develop an AOP based on retrieved data on a known stressor compound (not the compound under assessment). For example, data (in vitro as well as in vivo) for the well-known neurotoxic compound rotenone was used to develop the AOP on inhibition of the mitochondrial complex I leading to Parkinsonian motor deficiencies which has been endorsed by OECD (EFSA PPR Panel, 2017). Once the AOP has been developed, it can be concluded whether it is biologically plausible that chemicals (in this case certain pesticides) triggering the AOP are plausible risk factors for Parkinson's Disease. Such an approach is resource demanding in terms of time and expertise. However, it may be helpful in the situation where there is human observational evidence of an association while the data from animal experiments does not show an effect, because the animal model is not capturing the effect.
3. Carry out a systematic literature review and critical appraisal of all the evidence on the compound in question (human observational studies, in vivo rodent studies, in vitro data), a quantitative uncertainty analysis of all the evidence using expert knowledge elicitation (EKE) and a probabilistic approach, and finally integrate the data using the AOP conceptual framework. The AOP triggered by a specific compound (stressor-based AOP) can then support an IATA as exemplified for the developmental neurotoxic effects of deltamethrin (EFSA PPR Panel, 2021).

A full integration into a risk assessment also requires a careful assessment of the exposure, including detailed information on ADME. This was, for instance, done in the deltamethrin case mentioned above (EFSA PPR Panel, 2021). First, the evidence was integrated in an AOP conceptual framework, where a probabilistic quantification of the WoE was conducted to assess and quantify the uncertainty of the evidence. This ultimately allowed the assessment of whether the concentration of the specific compound, which targets the molecular initiating events (MIE) of the AOP, will be relevant or not for its activation.

This assessment of the biological plausibility for causality of outcomes observed in epidemiological studies for regulated products could generally be adopted for any risk assessment. However, as for all frameworks, exceptions can arise. In the case of certain food additives, short-term experimental studies in humans assessing tolerability are available, but long-term studies assessing chronic disease risk have received much less attention (compared to the pesticide area). Therefore, the approach shown in Figure 6 may apply to food additives in most cases.

KEY POINTS

- Evidence from human epidemiological data should be well-integrated with findings from other streams of evidence.
- Structured, biologically driven ways for integrating the data from human observational studies with other data are (1) relating the data to already existing AOP(s) or postulating a new AOP or (2) assessment of Mode of Action.
- The approach for integrating information from different studies outlined in this guidance provides a structured framework for integrating and for deciding on precedence of different data.

4.4.3 | Risk characterisation – considerations on dose–response modelling

In this section, options for dose–response modelling for epidemiological data are addressed. Following a summary of the experience of modelling animal dose–response data using BMD modelling, the overall objectives for identifying reference points (RPs) from which HBGVs can be established are described. Then possible approaches to deriving human benchmark

[42] https://aopwiki.org/.

values are discussed. This includes reflections on the choice of benchmark response (BMR), and the application of benchmark approaches to different epidemiological datasets. An overview of alternatives to BMD modelling is also provided: (1) the use of meta-analysis to derive pooled estimates of dose–response slopes from multiple studies; (2) the specific case of estimating a change in dose–response relationships indicative of a risk threshold; and (3) simple modelling of risk. Finally, the application of uncertainty factors is discussed.

4.4.3.1 | *Current approach to dose–response modelling using animal studies*

To prevent harm from a dietary component, the ideal goal is to establish a HBGV, which is defined as an amount that can be ingested over a defined time-period without appreciable health risk. Formerly, this was done by extrapolating findings from laboratory animal data using the NOAEL as the RP and dividing by uncertainty factors to account for uncertainty in differences between and within species (see EFSA Scientific Committee, 2012). Alternatively, the RP may be divided by the estimated dietary exposure to set a margin of exposure[43] (MOE), (EFSA, 2005).

For laboratory animal data, EFSA prefers the BMD approach to the NOAEL approach for identifying a RP, partly because it makes better use of the entire data of the dose–response curve compared to simple pairwise comparisons, and the use of the BMDL as a RP includes an adjustment for the uncertainty around the BMD. The methods for applying the BMD approach are described in detail in the EFSA guidance on the use of benchmark dose approach in risk assessment (EFSA Scientific Committee, 2022) and a software platform for modelling of controlled animal experiments in line with this guidance has been developed (Hasselt University, 2022). The principle involves modelling a number of plausible dose–response curves (e.g. Hill, gamma, exponential, probit, etc.) across the exposure groups, and calculating an exposure (BMD) corresponding to a defined response or BMR. The RP is then identified by taking the lower bound of the model average 95% credible interval of the BMD (the BMDL), which is built considering the posterior weights across all models. Such BMDLs are then calculated for each endpoint measured in the relevant studies. The various BMDLs are assessed in relation to the relevance of the outcome and the quality of the study, to lead to selection of a BMDL to be used as the RP for establishing HBGVs. In this process, both robustness of the underlying study and adversity of the outcome are taken into consideration (EFSA Scientific Committee, 2012).

EFSA considers that it is not appropriate to establish HBGVs for compounds for which a threshold cannot be assumed, such as substances that are directly acting genotoxic carcinogens. In these circumstances, an MOE is estimated (EFSA, 2005). The Scientific Committee has concluded that an MOE of 10,000 or higher, if it is based on the BMDL10 from a laboratory animal study, would be of low concern from a public health point of view (EFSA, 2005).

4.4.3.2 | *Dose–response modelling using human observational studies*

The goal of dose–response modelling is to identify an exposure RP, associated with one of three degrees of response:

- A non-minimal adverse change[44] over background incidence specified as absolute incidence or relative risk for rare outcomes; and for a continuous endpoint a certain adverse change in the outcome.
- A threshold of exposure below which there is taken to be no appreciable adverse effect. Such a threshold should be taken into consideration, for example, if there is no increase in incidence of a quantal response, or no change in a continuous measure.
- Minimisation of risk when there is a U-shaped relationship between exposure and health effects, as is frequently the case for nutrients for which both deficiency and excess exposure tend to have adverse effects.

These RP values can then be used to establish a HBGV or MoE. This may be done by applying different adjustment factors to the RP, which are related to, for example, sensitive subgroups.

KEY POINTS

- Human dose–response modelling in food safety assessments requires a broad range of modelling tools in addition to BMD modelling.
- Modelling of U-shaped relationships for nutrients (dual-risk) and identification of thresholds for health effects or other biological responses are some examples of the methods frequently applied for human data.

[43]The margin of exposure (MOE) is the ratio between a defined point on the dose–response curve for the adverse effect and the human intake.

[44]See Section 4.4.3.4.1 on BMR for more detail.

4.4.3.3 | *BMD modelling of experimental vs observational data*

As discussed in Section 4.4.3.1, the BMD modelling approach has been the preferred approach in EFSA for modelling data from controlled animal experiments (EFSA Scientific Committee, 2022). For such data, the background response should, in the absence of cross-contamination, be well defined by the un-exposed controls. Furthermore, as the laboratory animals are randomised into exposure groups, there should not be a problem of confounding variables varying between exposure groups. Lastly, with proper dose selection the whole sigmoidal dose–response curve from background to maximum response should be captured. However, in studies assigning only few animals confounding, such as by sex, may occur that needs to be accounted for. The maximum response may also not be accurately captured if the number of dose groups is small, or the highest dose is not sufficiently high.

In principle, the current EFSA Guidance for BMD modelling would work well for human experimental data, but for obvious ethical reasons such data rarely exist. There are rare exceptions such as in EFSA's 2014 opinion on perchlorate (EFSA CONTAM Panel, 2014). Similar data availability may also occur in nutrition where experimental data on food supplements is frequently available (revealing side effects or unexpected adverse events). However, in most cases one can expect that available experimental data would not be compatible with multi-dose RCT design that the current BMD guidance addresses. The reason being that prior to conducting such experiments (phase III), the safety and tolerability of a substance is usually tested in smaller phase 0, I or II trials and any sign of harm would not justify further experiments (see Section 4.1.2.1). These studies usually suffer from correlated observations (dose escalation trials)/or small dose groups which may introduce confounding (e.g. imbalance in sex, smoking, age, or other factors).

However, the above-mentioned challenges with experimental studies in humans could, perhaps, be more easily addressed compared to trying to model human observational data. That is, the modelling of human observational data needs special considerations and deviations from existing BMD guidance (EFSA Scientific Committee, 2022) for the following reasons:

- There are likely differences between exposure groups for several confounders, which need to be adjusted for.
- For quantile data, results are usually presented in terms of adjusted risk measures relative to a reference group, and the baseline risk level of 1 does not have a standard error associated with it.
- For continuous outcome data, there may be methodological challenges in establishing the baseline level at zero exposure
- With individual or grouped exposure levels, the baseline risk level is likely not zero, so the BMR would not be relative to zero exposure.

These considerations are not addressed in the 2022 guidance (EFSA Scientific Committee, 2022), but other research groups have published on these topics (Budtz-Jørgensen et al., 2001; Whitney & Ryan, 2013).

The reference group in most observational studies with exposure grouping is the lowest exposure group, not a true zero exposure group. So, the BMR may need to be defined as the incremental increase relative to that of non-zero exposure. This may vary between studies, affecting the calculated BMD. To bypass this, assumptions on background response at zero concentrations need to be made. In many cases, no health risk would be expected at very low exposure levels, and therefore, the non-zero exposure group could be defined as a reference. If the exposure–response relationship is very flat/shallow at the lowest exposure group relative to higher exposure groups, the error introduced when estimating the BMD by using the lower exposure group (because the response at zero dose is not known) should be minimal. This can be checked by sensitivity analysis.

Unadjusted observational epidemiological data can be modelled ignoring risk factors other than the exposure of interest, but such modelling is vulnerable to the presence of confounding, so this is not advised. Current BMD modelling platforms developed by EFSA, RIVM and US EPA do not allow inclusion of multiple covariates in the models nor direct modelling of relative risk. While not yielding adjusted BMD values, the EFSA platform allows some evaluation of categorical covariates (Hasselt, 2022). In this approach, the data are stratified by the covariate categories and BMDs calculated for each stratum. For multiple covariates, a single stratifying variable comprising all combinations of covariates (e.g. sex and age groups) would need to be created. In many cases, this will have the disadvantage of generating one or more strata with sparse numbers. Such analyses may potentially highlight the most sensitive stratum, which could then be selected as RP. This approach has been used to calculate BMDLs in simple subgrouping cases, e.g. sex-specific BMDLs. In cases where the BMD results are the same across these strata, the overall BMD and BMDL can be calculated in the normal way for the whole data and be considered unconfounded. However, in the presence of confounding, this stratification approach does not allow the direct estimate of BMDLs adjusted for these confounders. It is frequently the case that several covariates need to be adjusted for. In such situations, covariate adjustment should ideally be made using multivariable regression. Furthermore, to fully implement such approaches and to comply with the legal constraints of sharing individual participant data, the software would need to be downloadable so full data could be modelled by data owners instead of relying on summary data.[45]

[45]The R version of PROAST is downloadable online (https://www.rivm.nl/en/proast) and can be used locally.

Looking ahead, a solution would be to develop new versions of BMD software programs better suited to epidemiological data, which would allow adjustment for multiple covariates and modelling of RR in line with how human observational data are traditionally modelled. In the absence of such software, there are ways to use existing BMD software and still address the three concerns listed above. This has been done to varying degree in previous EFSA opinions primarily for continuous outcomes (see examples provided in Section 4.4.3.4.5 below). Modelling of risk ratios, however, had not been performed until EFSA's 2024 opinion on inorganic arsenic (iAs) (EFSA CONTAM Panel, 2024) where the following indirect approach to modelling adjusted relative risk was used:

> Since the current BMD approach is not designed to model relative risk estimates such as IRR, HR, or OR, it was necessary to transform the relative risks to natural numbers/integers. This was based on the approach used by JECFA (FAO and WHO, 2011). For cohort studies, the incidence rate or the cumulative incidence in the reference category was calculated. This incidence was multiplied by the adjusted risk estimate to obtain an adjusted incidence estimate, which was then used to calculate the "adjusted number of cases" (as integers). Having obtained the "adjusted number of cases" and the population size (provided in the papers or calculated from number of cases and incidence rates) in each exposure category, these data could be used as input into the EFSA BMD webtool.

This approach has the advantage of being able to use the existing BMD modelling platform to covariate adjusted data. In some cases, such use of data may introduce errors, for example if the numbers of observed cases in exposure groups are small, as the rounding to integers may mean that the RRs are somewhat changed.

While BMD modelling is well characterised for laboratory animal studies, it is novel for epidemiological data. As epidemiological data have the advantage of avoiding the uncertainty of animal to human extrapolation, there is a need for guidance on BMD modelling of human data so that it can be consistently used for risk characterisation in EFSA's assessments. Although developing such guidance is outside of the scope of this document, main principles and considerations for dose–response modelling using data from human observational studies are outlined below. These considerations represent both the current status and suggestions for further empirical and/or modelling studies.

KEY POINTS

- Although many of the principles laid out in the EFSA guidance on BMD modelling for controlled animal experiments may also apply for human data, a direct application is rarely feasible due to differences in study designs and the nature of human data.
- This may equally apply to modelling of human experimental data as such data, relevant for chemical risk assessment, are often not compatible with multidose randomised controlled experiments.
- Although modelling of human data has been performed by EFSA on several occasions, those efforts have had to overcome limitations in existing software platform that are designed to model specific type of experimental data.
- Adjustment for multiple covariates and modelling of relative risks are key specific aspects that BMD modelling software have to address to allow for appropriate modelling of human observational data.

4.4.3.4 | *BMD modelling using human epidemiological studies*

For BMD modelling of an epidemiological study, there are certain minimum conditions. There need to be multiple exposure groups, including one group with little or no exposure. If data are not grouped, the exposure needs to cover a wide range, including little or no exposure. In many cases, there will be several outcomes of interest, and for each outcome, there will be multiple studies, each of which may provide a BMD. Before carrying out BMD modelling, the BMR needs to be defined (in terms of additional absolute risk or relative risk).

4.4.3.4.1 | *BMR selection*

The selection of the BMR from human data shares common principles with the selection of BMRs from laboratory animal data. 'The BMR is a degree of change that defines a level of response in a specific endpoint that is measurable, considered relevant to humans or to the model species, and that is used for estimating the associated dose (the "true" BMD)' (EFSA Scientific Committee, 2022). For continuous outcomes, the BMR is, ideally, the smallest measurable change that reflects adversity. In human studies, the link between some clinical biomarkers and disease endpoints is well established, and that could be considered when selecting a BMR. For quantal outcomes with relatively low absolute risk, the relative risk approach is more appropriate than modelling absolute risk (see examples in Section 4.2.1.1).

Given the BMR is a direct estimate of effects in the underlying study population, further adjustment factors may be applied to the BMDL to establish a HBGV. No clear guidance on when such factors should be applied exist for human data, but some considerations and past examples are highlighted in Section 4.4.3.4.5 below.

4.4.3.4.2 | *Cohort studies*

Epidemiology data which fit most readily into the BMD approach established for laboratory animal studies are quantal data from prospective cohorts with a common outcome (e.g. CVD). Ideally such studies should have sufficient follow-up time. The data to be entered reflect the adjusted incremental risk per group and are based on the size of each exposure group and the number of cases in each group. However, the crude (unadjusted) observational data cannot be used as this is likely subject to confounding. Similar to the approach used by the CONTAM panel for inorganic arsenic described above (EFSA CONTAM Panel, 2024), the number of cases in each group above reference can be estimated from the adjusted relative risk, rounded to the nearest integer and entered into the BMD modelling platform. The error of rounding is relatively minor if groups are not too small.[46] The BMDL can then be calculated at the BMR.

However, when studies with low incidence of the outcome are modelled, a BMR based on absolute low risk (say 1%) is usually not going to be reached. In such cases, modelling the relative increase in risk would make more sense. The same calculations would be done but the target BMR would be defined as a relative risk increase of, e.g. 5% relative to the reference group rate.[47] Another reason for focusing on relative risk is due to limited follow-up in many studies. That is, the observed absolute risk during insufficient follow-up may underestimate the lifetime risk given sufficient follow-up, but the relative risk more appropriately captures the effect of exposure.

Another concern when applying the BMD approach to human data is whether the 'low exposure' reference category in epidemiology can properly be considered as equivalent to the zero-dose referent category in experiments. With individual or grouped exposure levels, the baseline risk level is likely not zero, so the BMR would not be relative to zero exposure. If there is a threshold exposure–response relationship and the lowest exposure category is below that threshold, then this category still provides a reasonable baseline equivalent to the risk level at a true zero exposure. If the baseline category is close to the general background population exposure level, then it can be assumed to reflect the real-world contrast between additional exposure and unavoidable exposure. However, if the lowest exposure group is, on average, substantially exposed, then this is a weakness that needs to be acknowledged in the review of study-specific BMDLs.[48]

4.4.3.4.3 | *Case–control studies*

The BMD modelling approach with grouped exposure data and quantal outcome data suggested for cohort studies can be extended to case–control data but this requires some additional assumptions. If we assume that all cases are collected from a given base population, and the controls provide a sufficient estimate of the exposure distribution of the base population, then there is data on the total population in each exposure group. The numbers of cases above the reference population can be adjusted to conform with the odds ratio in a comparable manner to the preceding case for cohort data. The data can be processed to calculate BMDLs to a BMR by considering the target relative risk or odds ratio (EFSA CONTAM Panel, 2024).

4.4.3.4.4 | *Continuous outcomes*

BMD approaches can also be applied to continuous outcomes (e.g. cholesterol, glucose or antibodies in serum, IQ score, lung function tests). The BMR should reflect a minimal change which is adverse and, therefore, it will depend on the nature of the endpoint selected. No default values exist and the BMR should be based on health considerations, but some examples based on previous EFSA assessments are given in next section.

It would be desirable for future iterations of the current EFSA BMD software platform (Hasselt University, 2022) to allow for direct input of continuous exposure data (i.e. at individual level) that are not divided into subgroups. Such data may be in individual studies or may be combined by well-established rules for meta-analysis, leading to a pooled, more precise slope for the exposure–response.

[46]The numerical value of a group that is too small depends on the outcome.

[47]The number of new cases of a disease divided by the number of persons at risk for the disease.

[48]If the lowest exposure group is, on average, substantially exposed, this leads to a less protective BMDL.

KEY POINTS

- Selection of benchmark response for human benchmark dose modelling follows the same principle as laid out in the EFSA guidance for benchmark dose modelling for controlled studies in experimental animals.
- Based on biological and modelling considerations, it needs to be assessed on a case-by-case basis whether 'low exposure' reference category in human observational studies can be considered as equivalent to the zero-dose referent category in experiments. In many cases, the associated uncertainties are marginal and easily dealt with.
- Modelling of adjusted continuous outcomes where exposure is categorised is relatively straight forward with existing benchmark dose modelling software. For quantal health outcomes, modelling of adjusted incident data is more challenging with existing software, but the approach used for inorganic arsenic by EFSA is one example of how such data can be modelled based on several assumptions.
- Modelling of unadjusted observational data is strongly discouraged.
- The principles of human benchmark dose modelling should be used as starting point for developing guidance for human benchmark dose modelling. Such guidance may require some modification of the existing benchmark dose modelling framework to make it compatible with modelling of observational data.

4.4.3.4.5 | *Examples of BMD modelling in EFSA opinions*

There has been some experience of using BMD modelling for continuous human data in EFSA assessments. Those assessments provide some examples on how decisions on BMRs vary depending on the outcome.

- For cadmium, urinary cadmium was related to the BMR of 5% of having beta-2-microglobulin (B2M) levels exceeding 300 μg/g creatinine (EFSA, 2009).
- In the EFSA opinion on perchlorate BMD modelling of thyroidal radioiodine uptake, a BMR of 5, 10 and 20% was based on modelling a human intervention study (dose escalation trial) (EFSA CONTAM Panel, 2014).
- For lead (Pb), a 1% BMR was selected for two continuous outcomes – decrease in IQ score and increase in systolic blood pressure corresponding to an absolute change of 1-IQ point and 1.2 mmHg, respectively (EFSA CONTAM Panel, 2010). A 10% BMR for one quantal outcome, for the change in the prevalence of chronic kidney disease (CKD) was also quantified.
- For the opinion of perfluorooctane sulfonate (PFOS) and perfluorooctanoic acid (PFOA) in 2018, a BMR of 5% relative increase in serum cholesterol was used for both compounds; and in a later revised opinion on perfluoroalkyl substances (PFAS) from 2020, a BMR of 10% absolute reduction in antibody response (antibody titres found using serological tests) was used for the sum of four PFAS (EFSA CONTAM Panel, 2018, 2020).

Besides continuous data, there are also some examples of BMR being used for quantal outcomes:

- In the EFSA opinion on perchlorate BMD modelling of cutaneous effects (quantal outcome data) a BMR of 1% and 10% was used. This was done by merging data from three independent human interventions (EFSA CONTAM Panel, 2014).
- BMD modelling of human data based on quantal outcomes (e.g. cancer) was performed in the update of EFSA's risk assessment of inorganic arsenic in food (EFSA CONTAM Panel, 2024). The observed cumulative incidence (the ratio between the number of cases and the size of the source population over the observation time) in the assessed studies was estimated to be around 0.02%. Moreover, for the assessed cancer endpoints, a BMR of 1%–5%, expressed as relative increase of the background incidence after adjustment for confounders, was regarded to be relevant for public health. Thus, a BMR of 5% was used of 0.06 μg iAs/kg bw per day obtained from a study on skin cancer as a RP.

4.4.3.5 | *Other modelling approaches for human dose–response assessment*

Modelling approaches other than the BMD have been applied when assessing dose–response in human observational data. These include dose–response meta-analyses for establishing dietary reference values (DRV)/HBGV/UL for nutrients, modelling approaches to detect changes in risk or incidence (e.g. in cancer research); and assessing excess risk for chemical exposure in the occupational setting. Although there has been somewhat less focus on the use of these approaches in chemical risk assessment compared to the BMD, this may be more related to the frequent use of data from experimental animals for establishing HBGV rather than the appropriateness of these modelling approaches. Below, a short description of alternative methods, their suitability and pros and cons are provided. The choice of method depends both on the nature of the data and specific objective of the modelling. Therefore, in this document no general prescription on which method to use can be provided; the choice needs to be made on a case-by-case basis using expert judgement.

4.4.3.5.1 | *Dose–response meta-analysis*

A fundamental statistical tool that is playing a key role in risk assessment is the implementation of dose–response meta-analysis, which is based on a flexible modelling framework that can incorporate epidemiological studies encompassing different levels and categories (2 or more) of exposure. Until recently, dose–response meta-analyses have been based on forest plots, which have the limitation of ignoring the heterogeneity of exposure categories across studies and the shape of the dose–response relation (Vinceti et al., 2020).

Recently, the use of the so-called one-stage or mixed-effects framework, allowing the synthesis of tables of empirical contrasts, has become common, allowing the estimation of heterogeneous and frequently curvilinear dose–response relations. This method is generally based on the described random-effects dose–response restricted cubic spline modelling using a one-stage mixed effect meta-analytic model for aggregated data (Crippa et al., 2018; Orsini et al., 2012; Orsini & Spiegelman, 2021; Sera et al., 2019; Vinceti et al., 2020).

In a meta-analysis, parameter estimates are generally obtained with the restricted maximum likelihood method, and statistical inference typically focuses on the summary dose–response relation. Use of these methods to carry out dose–response meta-analysis allows the identification and characterisation of complex relations between exposure and endpoints, such as chronic disease risk associated with intake of nutrients like potassium, sodium, manganese and selenium, and of contaminants like acrylamide and cadmium (Adani et al., 2020; Filippini et al., 2021; Filippini et al., 2022; Vinceti et al., 2016). In such cases, adverse health effects may arise at too low and/or too high exposure and characterising the shape of such patterns of association is of paramount value in risk assessment. In these instances, the use of linear functions, such as linear regression analysis, would likely lead to wrong statistical inferences and conclusions.

In nutritional risk assessment, non-linear dose–response meta-analytic modelling is of key importance not only in shaping the exposure–disease relations, but also when characterising the relation between intake and biomarkers of exposure, or more generally quantitative variables. Conversely, when some studies do not report findings in a way that is suitable for dose–response meta-analyses, their omission may reduce the body of evidence available for assessment. The availability of a large number of studies for the dose–response meta-analysis is also very important not only to yield more precise risk/effect estimates, but also to: (1) broaden the range of exposure for which the risk assessment is conducted, (2) identify population characteristics that may act as effect modifiers in the exposure–endpoint relation, (3) carry out sensitivity analyses according to the RoB of the studies; (4) assess publication bias or small study effect.

Of relevance are also methodological issues such as the general preference in using the most adjusted estimates from the studies to be included in the meta-analyses, the selection of the RP in plotting the curves of RR/effect estimates in relation with the Y-axis, the selection of the knots, i.e. the fixed points among which the curves are smoothly interpolated in the most commonly used flexible method to model non-linear functions, i.e. the restricted cubic spline function (Crippa & Orsini, 2016; Orsini et al., 2012). Dose–response meta-analyses based on cubic spline modelling are largely used for both continuous (e.g. blood glucose, blood pressure) and dichotomous (disease occurrence) endpoints (Vinceti et al., 2020).

KEY POINTS

- Dose–response-meta-analysis is a more robust way of integrating data from several studies compared to traditional forest plots used on meta-analyses that generally ignore the heterogeneity of exposure.
- In data rich cases, the use of dose–response meta-analyses may also provide a more realistic and more comprehensive assessment of an underlying exposure health relationship than relying on the results from a single or a few studies.

4.4.3.5.2 | *Use of piecewise linear regression for identifying a change in risk*

For certain disease outcomes, such as cancer or CVD, the primary interest when quantifying risk may not necessarily be assessment of the whole dose–response but rather to estimate at what exposure level a statistically significant change occurs in the slope for risk (or incidence). This could, for example, include assessing the level of exposure associated with significant changes in incidence or risk from background. Such modelling can both be done using individual participant data or summary data.

Although this type of modelling has not been commonly applied in risk assessment within food safety, it has been extensively used in cancer research for monitoring changes in cancer incidence over time in different populations. The National Institute of Cancer, has for example, specifically developed a software, called *JoinPoint* (Kim et al., 2000), to estimate changes in cancer incidence over time. The method is based on fitting a piecewise linear model to the dose–response data. The presence of change-points are then assessed by fitting where on the dose–response curve the slope of the linear segment changes. The resulting change-point is estimated along with associated uncertainty (based on the standard error).

Although this method has primarily been used to assess changes in cancer incidence over time, this methodology is increasingly being applied to model epidemiological studies, such as assessing the level of compliance needed for

antihypertensive medication to significantly reduce later risk of CVD (Yang et al., 2017); or assessing where selenoprotein P concentrations plateau in relation to blood selenium (Hurst et al., 2010).

In addition to the JoinPoint package, other software is available. The package 'Segmented' in R[49] allows for the estimation of change-points through regression analyses where adjustment for covariates can also be performed. This package in R was, for example, used in a recent publication to assess the change-point (Figure 7A) at which antibody titres started to decrease significantly with higher serum PFAS concentrations in 1-year old children (Source: Abraham et al., 2020). The 'Segmented' package in R can also be used for quantile data. Although no example within the area of food safety is available, the use of this method is well illustrated in the modelling of the relationship between maternal age and Down syndrome (Figure 7B); Source: Muggeo, 2008 (from Davidson & Hinkley, 1997)).

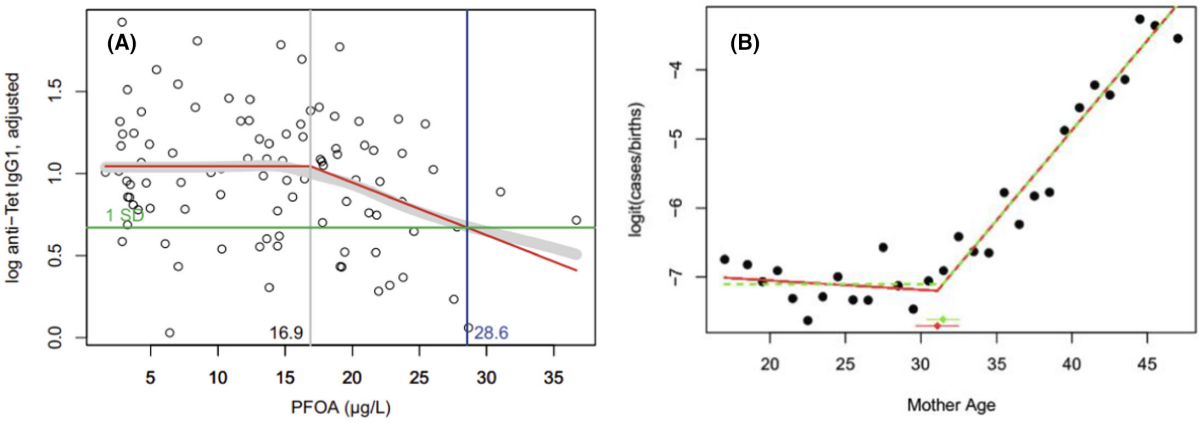

FIGURE 7 Examples of applications of segmented regression. (A) The association between serum plasma concentrations of PFOA in 1-year-old children in relation to adjusted antibody concentrations to tetanus (Source: Abraham et al., 2020). The red line is the fitted piece-wise linear function while the broader grey line is a moving average. The vertical grey line shows the PFOA concentration where the change-point is identified (16.9 ng/mL). (B) How a piece-wise linear function is fitted through a data for the association between maternal age and percentage of children being born with Down syndrome (Source: Muggeo, 2008 (from Davidson & Hinkley, 1997)). The red line identifies a change-point at 31.1 y (Std error of 0.7 y). However, as the first segment of the line was not significantly different from null the fitted green line was fitted with the constraint the first slope being $\beta_1 = 0$. In that case the change-point was estimated as 31.5 y (Std. error of 0.6 y).

Primarily biological considerations and the nature of the endpoint should be guiding the selection of change-points. Assessment of change-points may be relevant when the underlying dose–response data in the study of interest appears to show a threshold for effect. The uncertainty around the change-point can be quantified objectively, based on the standard errors.

This modelling approach may be of relevance for certain disease endpoints such as cancer and CVD, where assessment of when a significant change in risk (or incidence) from background occurs may be of more interest than assessing where a relative or absolute BMR occurs. The identified change-point can then be used for further risk characterisation.

KEY POINTS

- The use of statistical methods assesses the presence of change-points, where a significant change in risk or response occurs relative to baseline is one option for characterising risk.
- Although use of such methods has been traditionally confined to cancer research, this methodology is increasingly being used for other health related outcomes.
- Available software packages make the use of this methodology relatively straight forward and allow for quantification of uncertainty.

4.4.3.5.3 | *Other Modelling approaches*

Other modelling approaches have, compared to the BMD approach, been applied when assessing dose–response in human observational data. These include dose–response meta-analyses discussed in the previous section, and various approaches to establish the shape of the exposure–response relationship including simple linear relationships.

[49] https://cran.r-project.org/web/packages/segmented/segmented.pdf.

The most common approach has been to estimate the slope of the dose–response curve (in terms of a disease rate, risk, hazards ratio or odds, usually log transformed) with an extrapolation to zero exposure. Based on that extrapolation, the exposure equivalent to an excess risk of say 1/1000 or 1/10,000 can be estimated (Steenland & Deddens, 2004). This approach can be extended to continuous variables. Splines or linear functions are commonly used for this purpose, but other models can be applied as well. The RP would be identified based on the excess risk which forms the basis for establishing the HBGV with or without the use of an uncertainty factor (UF) (see Section 4.4.3).

Different approaches and conventions exist for choosing a function for such modelling. Today the use of splines is common and well accepted, frequently in the form of restricted cubic splines. The use of splines is well suited to capture any deviation from non-linearity and generally provides better flexibility than polynomials.

The use of linear models is also common in epidemiology and may be justified in cases where strong indication of linearity exists, depending on whether or not variables are on a log or linear scale. Approximate linear relationships may occur in human epidemiology when the exposure range is small (narrow). That is, the full non-linearity of a relationship is not captured. Furthermore, if variability of the exposure and/or the outcome is high, as it is often the case in human studies, it may be difficult to distinguish between a linear and non-linear response. Often the assumption on linearity beyond the observed exposure range is simpler and easier to interpret than extrapolation based on non-linear functions. One way to justify the use of linear functions is to test if a significantly better fit is obtained (e.g. based on the residual sum of squares) when using a non-linear function such as splines. If the non-linear function does not provide a significantly better fit, the use of a linear function is partly justified.

KEY POINTS

- Modelling of risk using simple linear or non-linear function to quantify the excess risk relative to a reference category or by extrapolating to zero exposure is one alternative to benchmark dose modelling.
- Such an approach may be considered when conditions for benchmark dose modelling are, for varying reasons, not met, or when a simpler modelling approach is considered more appropriate.
- The corresponding reference point based on the excess risk would form the basis of the HBGV with or without the use of an UF.

4.4.4 | Use of uncertainty factors for risk characterisation using evidence from human epidemiological studies

The situations in which use of uncertainty factors is considered relate primarily to the uncertainty around the identified RP for establishing HBGVs and the level of protection that is aimed at. The uncertainty can for example relate to the external validity of the study used to identify a RP. For example, a RP identified in a population of healthy adults may not necessarily be protective in the case of more vulnerable population subgroups, such as pregnant women or the elderly. Similarly, the use of uncertainty factors may be appropriate when there is limited information on dose–response. This is quite common in the area of nutrition, where adverse effects are observed as a result of high intake in an observational setting, or where adverse effects are observed in food supplementation trials. In these situations, it is often not possible to identify the specific intake level at which adversity occurs, and limited information exists on the level of risk associated with slightly lower intake. In such cases, uncertainty factors have often been applied (EFSA, 2005).

As no specific guidance exists on applying UFs when the RP is identified from human data besides the application of a 10-fold UF for accounting for human variability, case-by-case assessments relying on expert judgement need to be made. If available, an a priori standard UF may be a starting point for the process of selecting a final UF in the risk assessment. However, such UFs should always be assessed, tailored to the real data and overall evidence available, and eventually adapted based on expert judgement. Factors to consider for applying or not an UF could, for example, be:

- whether the pivotal studies appropriately represent the general population or the population relevant for the risk assessment,
- to account for uncertainties of the methodology used for assessing the exposure or the health outcome,
- whether the health outcome investigated is a primary or surrogate measure for the health outcome,
- whether physiological requirements need to be considered, such as for establishing upper levels for nutrients,
- uncertainties associated with deriving intake from biomarkers of exposure.

A few examples on how these considerations have been applied previously are provided below.

In the re-evaluation of the existing HBGVs for copper (EFSA Scientific Committee, 2023b), the established HBGV was based on retention of copper (a predictor of future toxicity), on the basis of measurements of excretion in healthy individuals. Although the pivotal study was conducted in healthy individuals that may not represent the general population, the fact that copper retention was a surrogate measure for the health outcome (liver toxicity), the RP identified

was considered sufficiently protective for most consumers over long-term. Use of UF was, therefore, not considered necessary.

In case of selenium, an UF of 1.3 was applied to account for the uncertainties associated with extrapolating findings from a large RCT carried out in the US to the general risk assessment for the European population. The scientific justification for the UF of 1.3 was based on expert judgement taking into account various considerations, among them the uncertainty around the dose–response due to the use of single-dose trials and current dietary intakes in the EU. These are explained in detail in the Scientific Opinion (EFSA NDA Panel, 2023).

An example of adjusting for uncertainty due to the use of a biomarker of exposure in establishing a HBGV, can be found in the CONTAM Panel assessment of mercury (EFSA CONTAM Panel, 2012). The exposure estimation was based on the mercury concentration in hair samples from mothers and converted to maternal blood concentrations. Here a data-driven chemical specific adjustment factor of 2 was applied, in addition to the standard factor for interindividual toxicokinetic variability of 3.2., to adjust for variability in the hair to blood mercury ratio. The resulting UF was thus 6.4.

KEY POINTS

- The recommended approach for application of uncertainty factors for deriving health-based guidance values should be based on the case at hand, tailored to the available data and based on the uncertainty analysis of the data.
- Currently, no specific guidance exists for using uncertainty factors when risk characterisation is based on human studies.

RECOMMENDATIONS

Recommendations for EFSA risk assessments

1. Evidence from epidemiological studies in humans should be used in risk assessments to the extent possible.
2. The overall assessment should consider the entire body of evidence.
3. Judgements on the overall body of evidence should always be made by considering the type, and, if possible, direction and magnitude of potential biases identified across different studies, for example by using a triangulation approach.
4. To facilitate more structured and time efficient risk assessment, the use of evidence maps and scoping reviews during the planning phase of an Opinion is recommended.
5. RoB tools provide a structured way to identify different biases that may occur to varying degrees in different studies. The key elements to capture within each study are the source, magnitude and direction of possible biases. That complexity cannot be accurately captured by assigning a numerical score of study quality, which is therefore discouraged.
6. The type of dose–response modelling for risk or benefit characterisation should be selected based on the type and nature of the available data and the objective aimed for (minimising risk, maximising benefits or balancing the two).

Recommendations for further developments

7. RoB tools have a long history of use for RCTs in humans. There is room for further development of these tools to capture the differences of different observational designs and use for other populations (e.g. livestock or companion animals and plants). It is recommended that EFSA collaborates at the European and international levels with relevant organisations and initiatives to harmonise developments in this area.
8. Based on the principles outlined in this document, a guidance on human BMD modelling specifically addressed to modelling of human observational data should be developed. This requires adaptation to the existing methodological framework designed around controlled animal experiments. It would need to be accompanied by changes to existing BMD software platforms to allow for adjustments of multiple covariates and modelling of relative risk. This would allow for more rigorous and consistent use of human data.
9. The use of multivariable regression analysis is recommended to account for covariates/confounders for BMD modelling of data from epidemiological studies.
10. Efforts are needed to provide guidance on the use of UFs and in particular on the MoE approach when using human epidemiological data.
11. The use of other modelling approaches frequently applied in epidemiology such as dose–response meta-analyses or estimation of thresholds should be explored and developed further for the area of chemical risk assessment.
12. Although design and conduct of epidemiological studies in humans, animals and plants often differ, many similarities exist. Better understanding of those similarities and differences and the terminology used is essential to address cross-cutting challenges that EFSA will face in the future. This requires training and closer collaboration among experts and staff across panels.

GLOSSARY

Accuracy	The extent to which systematic error (bias) is minimised. Risk of bias addresses also aspects like the sensitivity and specificity of the detection method used in an assessment (also referred to as 'Internal Validity').
Aggregated data	Information resulting from the combination of individual data (e.g. mean exposure in a treatment group, standard deviation of the observations in a group, etc.). See Individual data.
Assembling the Evidence	The first of three basic steps of weight of evidence assessment, as proposed in this guidance. Includes identification of potentially relevant evidence, selection of evidence to include in the weight of evidence assessment and grouping the evidence into lines of evidence.
Assessment	The term refers to all types of scientific assessments produced in the EFSA context, and for referring to both assessments based on data generated ex novo, assessments based on already existing data or assessments conducted by eliciting expert knowledge. Also referred to as 'scientific assessment'.
Best professional judgement	A category of weight of evidence assessment methods involving qualitative listing and qualitative integration of multiple pieces or lines of evidence.
Case-specific assessment	Case-specific assessments, where there is no pre-specified procedure and assessors need to choose and apply weight of evidence approaches on a case-by-case basis.
Causal criteria	A category of weight of evidence assessment methods based on criteria for determining cause and effect relationships.
Complementary line of evidence	A line of evidence which can only answer a question or sub-question when it is combined with other line(s) of evidence.
Conceptual framework	The context of the assessment; all sub-question(s) that must be answered; and how they combine in the overall assessment.
Consistency	The extent to which the contributions of different pieces or lines of evidence to answering the specified question are compatible.
Critical Appraisal Tool (CAT)	Tool for appraising study methodological quality (see definition). A CAT contains a comprehensive list of elements to consider for appraising study methodological quality and detailed guidance for performing the appraisal. CATs are tailored for the specific study designs. For instance, the items to be considered when appraising a randomised controlled trial are different from those considered in an observational study. Within the same study design CATs should be applied by outcome or endpoint. This is because the same study can be of different methodological quality depending on the outcomes that are reported. CATs should be applied to each individual study included in the assessment so to allow a consistent classification of studies according to their methodological quality (which is then considered when assessing the reliability of the evidence they provide).
Data	A piece of information. See also Individual data and Aggregated data.
Ecological studies	Studies in which the unit of analysis are populations or groups of people rather than individuals. Conclusions of ecological studies may not apply to individuals, but ecological studies can reach valid inferences on causal relationships at the aggregate/ group (ecological) level. Ecological studies have a role when implementing or evaluating policies that affect entire groups or regions.
Emergency assessment	Emergency procedures, where the choice of approach is constrained by unusually severe limitations on time and resources.
Estimate	A calculation or judgement of the approximate value, number, quantity, or extent of something. Some weight of evidence questions refer to estimates, while others refer to hypotheses.
Evidence	Information that is relevant for assessing the answer to a specified question. In PROMETHEUS, a piece of evidence for an assessment is defined as data (information) that is deemed relevant for the specific objectives of the assessment (EFSA, 2015). In this Guidance, this is expanded to all potentially relevant information, i.e. all evidence identified by the initial search process, to recognise that the assessment of relevance in the search process is necessarily a preliminary one (e.g. based on keywords and titles alone). 'Evidence' can refer to a single piece of potentially relevant information or to multiple pieces.
Ex novo data generation	The process of generating new data as it occurs when designing and conducting an experiment or an observational study (e.g. a survey). Sometimes also referred to as 'primary research study' as opposite to a 'secondary research study' based on existing data (i.e. a review). In the EFSA context, studies generating data ex novo

	are designed and conducted for instance by the applicants submitting a dossier to EFSA in support of an application or by EFSA, when e.g. performing surveys (e.g. baseline surveys).
Expert judgement	An expert judgement is a judgement made by an expert about a question or consideration in the domain in which they are expert. Such judgements may be qualitative or quantitative, but should always be careful, reasoned, evidence-based and transparently documented.
Extensive Literature Search (ELS)	A literature search process structured in a way to identify as many studies relevant to a review question as needed. It is tailored in order to address the trade-off between sensitivity and specificity depending on the context of the review question. The fundamental characteristics of an ELS are: (1) use of tailored search strings, and (2) extensive use of literature sources (i.e. bibliographic databases and other sources accessed via electronic or hand-searching – for example, websites, journal tables of content, theses repositories, etc.).
External Validity	The validity of the inferences as they pertain to participants outside the source population which is either a target or can be argued to experience effects similar to the targets.
GRADE	An approach for grading the quality of evidence and the strength of recommendations in environmental and occupational health, proposed and developed by the Grading of Recommendations, Assessment, Development and Evaluation (GRADE) Working Group (see Morgan et al., 2016).
Hypothesis	One type of framing for weight of evidence questions. Defined as a proposition proposed to be a potential explanation of a phenomenon or a potential outcome of a phenomenon. Some weight of evidence questions refer to hypotheses, while others refer to estimates.
Individual data	Information collected at the level of the finest unit on which variables are measured (e.g. exposure observed on each individual belonging to a study). By definition, they cannot be further 'disaggregated'.
Influence analysis	A study of possible change in the assessment output resulting not just from uncertainties about inputs to the assessment but also from uncertainties about choices made in the assessment.
Integrating the evidence	The third of three basic steps of weight of evidence assessment, as proposed in this guidance. Includes developing a conceptual model for integration, assessing the consistency of the evidence, applying the method chosen for integration and developing the weight of evidence conclusion.
Internal Validity	See accuracy.
Line of evidence	A set of evidence of similar type
Meta-analysis	A statistical analysis that combines the results of multiple scientific studies
OHAT	An approach to systematic review and evidence integration for literature based environmental health science assessments, developed by the NTP Office of Health Assessment and Translation (OHAT) (see Rooney et al., 2014).
Piece of evidence	A broad term used to refer to distinct elements of evidence that may be combined to form a line of evidence, e.g. a single study, expert judgement or experience, a model, or even a single observation.
Precision	The extent to which random error is minimised and the outcome of the approach, method, process or assessment is reproducible over time.
Probability	Defined depending on philosophical perspective: (1) the frequency with which samples arise within a specified range or for a specified category; (2) quantification of uncertainty as degree of belief regarding the likelihood of a particular range or category. The latter perspective is implied when probability is used in a weight of evidence assessment to express relative support for possible answers
Problem formulation	In the present guidance, problem formulation refers to the process of clarifying the questions posed by the Terms of Reference, deciding whether and how to subdivide them, and deciding whether they require weight of evidence assessment
Qualitative assessment	An assessment performed or expressed using words, categories or labels
Quantification	A category of weight of evidence assessment methods defined as comprising formal decision analysis and statistical methods. Would also include probabilistic reasoning.
Quantitative assessment	An assessment performed or expressed using a numerical scale (see
Rating	A category of weight of evidence assessment methods for weighing and/or integration of evidence based on qualitative logic models, ranks, scores and empirical models.

Refinement	One or more changes to an initial assessment, made with the aim of reducing uncertainty in the answer to a question. Sometimes done as part of a 'tiered approach' to risk or benefit assessment.
Relative support	An expression of the extent to which evidence supports one possible answer to a weight of evidence question, relative to other possible answers. Can be expressed qualitatively or quantitatively. Quantitative expression can be in terms of probability
Relevance	The contribution a piece or line of evidence would make to answer a specified question, if the information comprising the line of evidence was fully reliable. In other words, how close is the quantity, characteristic or event that the evidence represents to the quantity, characteristic or event that is required in the assessment. This includes biological relevance as well as relevance based on other considerations, e.g. temporal, spatial, chemical, etc.
Reliability	Reliability of a piece of evidence refers to: (i) precision (see definition); and (ii) accuracy (see definition). It is influenced by the methodological quality of the process for producing such evidence.
Representativeness	Ability of a subset of a population (e.g. a sample of individuals) to reflect accurately specific characteristics of the population of origin.
Scientific assessment	See Assessment.
Scope of the assessment	What is to be evaluated in the assessment.
Sensitivity analysis	A study of how the variation in the outputs of a model can be attributed to, qualitatively or quantitatively, different sources of uncertainty or variability. Implemented by observing how model output changes when model inputs are changed in a structured way.
Standalone line of evidence	A line of evidence which offers an answer to a question or sub-question without needing to be combined with other lines of evidence.
Standardised assessment procedures	Assessments where the approach to integrating evidence is fully specified in a standardised assessment procedure. They generally include standardised elements that are assumed to provide adequate cover for uncertainty.
Sub-question	A scientific question that does not need to be further broken down to be answered and is formulated in a way that is directly answerable in an experiment or observational study (or as a single question in an expert elicitation study).
Uncertainty	A general term referring to all types of limitations in available knowledge that affect the range and probability of possible answers to an assessment question.
Uncertainty analysis	A collective term for the processes used to identify, characterise, explain and account for sources of uncertainty.
Variability	Heterogeneity of values over time, space or different members of a population, including stochastic variability and controllable variability.
Weighing	In this guidance, weighing refers to the process of assessing the contribution of evidence to answering a weight of evidence question. The basic considerations to be weighed are identified in this guidance as reliability, relevance and consistency of the evidence.
Weighing the evidence	The second of three basic steps of weight of evidence assessment, as proposed in this guidance. Includes deciding what considerations are relevant for weighing the evidence, deciding on the methods to be used, and applying those methods to weigh the evidence.
Weight of evidence	The extent to which evidence supports one or more possible answers to a scientific question. Hence 'weight of evidence methods' and 'weight of evidence approach' refer to ways of assessing relative support for possible answers.
Weight of Evidence	A function of relevance and reliability.
Weight of evidence assessment	A process in which evidence is integrated to determine the relative support for possible answers to a scientific question.
Weight of evidence conclusion	The outcome of a weight of evidence assessment, expressed in terms of relative support for possible answers to the weight of evidence question.
Weight of evidence question	A question addressed by a weight of evidence assessment. This may be the overall scientific question for an assessment, or a sub-question that contributes to answering the overall question. Weight of evidence questions may be framed in terms of hypotheses (which are often qualitative) or estimates (quantitative).

ABBREVIATIONS

ADME	absorption, distribution, metabolism and excretion
AOP	adverse outcome pathway
APRIO	Agent, Pathway, Receptor, Intervention and Output

BMD	benchmark dose
BMDL	benchmark dose lower confidence limit
BMI	body mass index
BMR	benchmark response
BPA	bisphenol A
CASP	Critical Appraisal Skills Programme
CHD	coronary heart disease
CI	confidence interval
CVD	cardiovascular disease
DAGs	directed acyclic graphs
DRV	dietary reference value
EKE	Expert Knowledge Elicitation
EPIQ	Evidence-based Practice for Improving Quality
GRADE	Grading of Recommendations, Assessment, Development, and Evaluations
HBCDD	hexabromocyclododecane
HBGV	health-based guidance value
HDL	high-density lipoprotein
HOC	health outcome category
HR	hazard ratio
IATA	Integrated Approach to Testing and Assessment
ICD	International Statistical Classification of Diseases and Related Health Problems
IPM	Integrated Pest Management
IQ	Intelligence Quotient
IRR	incidence rate ratio
ISPM	International Standards for Phytosanitary Measures
LoE	line of evidence
MIE	molecular initiating events
MoA	mode of action
MOE	margin of exposure
NOAEL	no observed adverse effect level
OR	odds ratio
PBPK	physiologically based pharmacokinetic
PFAS	perfluoroalkyl substances
PFOS	perfluorooctane sulfonate
PFOA	perfluorooctanoic acid
PECO	Population, Exposure, Comparator, Outcome
PICO	Population, Intervention, Comparator, Outcome
PIT	Population, Index Test, Target Condition
PLH	Plant Health
PO	Population, Outcome
QPRA	Quantitative Pest Risk Assessments
RCT	randomised controlled trials
RD	risk difference
RoB	risk of bias
RP	reference point
RR	risk ratio
SD	standard deviation
SYRCLE	Systematic Review Center for Laboratory animal Experimentation
ToR	Terms of Reference
WoE	Weight of Evidence
UF	uncertainty factor

ACKNOWLEDGEMENTS

The Panel wishes to thank the following for the support provided to this scientific output: the hearing experts Henning Thole and Marco Zeilmaker, the interim Tuuri Tauriainen and the EFSA staff members Laura Ciccolallo and Hans Steinkellner.

CONFLICT OF INTEREST

If you wish to access the declaration of interests of any expert contributing to an EFSA scientific assessment, please contact interestmanagement@efsa.europa.eu.

REQUESTOR

EFSA

QUESTION NUMBER

EFSA-Q-2019-00200

COPYRIGHT FOR NON-EFSA CONTENT

PANEL MEMBERS

Vasileios Bampidis, Diane Benford, Claude Bragard, Thorhallur I. Halldorsson, Antonio Hernandez-Jerez, Susanne Hougaard Bennekou, Konstantinos Koutsoumanis, Claude Lambré, Kyriaki Machera, Wim Mennes, Ewen Mullins, Simon More, Soren Saxmose Nielsen, Josef Schlatter, Dieter Schrenk, Dominique Turck and Maged Younes.

REFERENCES

Abraham, K., Mielke, H., Fromme, H., Völkel, W., Menzel, J., Peiser, M., Zepp, F., Willich, S. N., & Weikert, C. (2020). Internal exposure to perfluoroalkyl substances (PFASs) and biological markers in 101 healthy 1-year-old children: Associations between levels of perfluorooctanoic acid (PFOA) and vaccine response. *Archives of Toxicology*, *94*(6), 2131–2147. https://doi.org/10.1007/s00204-020-02715-4

Adami, H. O., Berry, C. L., Breckenridge, C. B., Smith, L. L., Swenberg, J. A., Trichopoulos, D., Weiss, N. S., & Pastoor, T. P. (2011). Toxicology and epidemiology: Improving the science with a framework for combining toxicological and epidemiological evidence to establish causal inference. *Toxicological Sciences*, *122*(2), 223–234. https://doi.org/10.1093/toxsci/kfr113

Adani, G., Filippini, T., Wise, L. A., Halldorsson, T. I., Blaha, L., & Vinceti, M. (2020). Dietary intake of acrylamide and risk of breast, endometrial, and ovarian cancers: A systematic review and dose-response meta-analysis. *Cancer Epidemiology, Biomarkers & Prevention*, *29*, 1095–1106. https://doi.org/10.1158/1055-9965.EPI-19-1628

Agency for Healthcare Research and Quality. (2002). Systems to rate the strength of scientific evidence. Evidence Report/Technology Assessment No 47, Publication No 02-E019 Rockville: Agency for Healthcare Research and Quality.

Altman, D. G., Gore, S. M., Gardner, M. J., & Pocock, S. J. (1983). Statistical guidelines for contributors to medical journals. *British Medical Journal*, *286*(6376), 1489–1493. https://doi.org/10.1136/bmj.286.6376.1489

Altman, R. D., Lang, A. E., & Postuma, R. B. (2011). Caffeine in Parkinson's disease: A pilot open-label, Dose-Escalation Study. *Movement Disorders*, *26*(13), 2427–2431. https://doi.org/10.1002/mds.23873

Amrhein, V., Greenland, S., & McShane, B. (2019). Scientists rise up against statistical significance. *Nature*, *567*(7748), 305–307. https://doi.org/10.1038/d41586-019-00857-9

Apelberg, B. J., Witter, F. R., Herbstman, J. B., Calafat, A. M., Halden, R. U., Needham, L. L., & Goldman, L. R. (2007). Cord serum concentrations of Perfluorooctane sulfonate (PFOS) and Perfluorooctanoate (PFOA) in relation to weight and size at birth. *Environmental Health Perspectives*, *115*(11), 1670–1676. https://doi.org/10.1289/ehp.10334

Armitage, J. M., Macleod, M., & Cousins, I. T. (2009). Comparative assessment of the global fate and transport pathways of long-chain perfluorocarboxylic acids (PFCAs) and perfluorocarboxylates (PFCs) emitted from direct sources. *Environmental Science & Technology*, *43*(15), 5830–5836. https://doi.org/10.1021/es900753y

Balzer, J., Rassaf, T., Heiss, C., Kleinbongard, P., Lauer, T., Merx, M., Heussen, N., Gross, H. B., Keen, C. L., Schroeter, H., & Kelm, M. (2008). Sustained benefits in vascular function through Flavanol-containing cocoa in medicated diabetic patients a double-masked, randomized, controlled trial. *Journal of the American College of Cardiology*, *51*(22), 2141–2149. https://doi.org/10.1016/j.jacc.2008.01.059

Bennette, C., & Vickers, A. (2012). Against quantiles: Categorization of continuous variables in epidemiologic research, and its discontents. *BMC Medical Research Methodology*, *12*, 21. https://doi.org/10.1186/1471-2288-12-21

Bero, L., Chartres, N., Diong, J., Fabbri, A., Ghersi, D., Lam, J., Lau, A., McDonald, S., Mintzes, B., Sutton, P., Turton, J. L., & Woodruff, T. J. (2018). The risk of bias in observational studies of exposures (ROBINS-E) tool: Concerns arising from application to observational studies of exposures. *Systematic Reviews*, *7*(1), 242. https://doi.org/10.1186/s13643-018-0915-2

Bert, B., Heinl, C., Chmielewska, J., Schwarz, F., Grune, B., Hensel, A., Greiner, M., & Schönfelder, G. (2019). Refining animal research: The animal study registry. *PLoS Biology*, *17*(10), e3000463. https://doi.org/10.1371/journal.pbio.3000463

Blumberg, J., & Block, G. (1994). The alpha-tocopherol, Beta-carotene cancer prevention study in Finland. *Nutrition Reviews*, *52*(7), 242–245. https://doi.org/10.1111/j.1753-4887.1994.tb01430.x

Boa, E., Danielsen, S., & Haesen, S. (2015). *Better together – Identifying the benefits of a closer integration between plant health, agriculture and one health . Published in one health: The added value of integrated health approaches. Zinsstag et al. (eds)*. CABI Publishing.

Budtz-Jørgensen, E., Keiding, N., & Grandjean, P. (2001). Benchmark dose calculation from epidemiological data. *Biometrics*, *57*(3), 698–706. https://doi.org/10.1111/j.0006-341x.2001.00698.x

Clemente, J. C., Ursell, L. K., Wegener Parfrey, L., & Knight, R. (2012). The impact of the gut microbiota on human health: An integrative view. *Cell*, *148*(6), 1258–1270. https://doi.org/10.1016/j.cell.2012.01.035

Cock, M. J. W., & Wittenberg, R. (2001). *Invasive alien species: A toolkit of best prevention and management practices*. CAB International.

Committee on Toxicity and Committee on Carcinogenicity. (2021). Report of the Synthesis and Integration of Epidemiological and Toxicological Evidence Subgroup (SETE). https://doi.org/10.46756/sci.fsa.sjm598

Cooke, B. M., Jones, D. G., & Kayle, B. (2006). *The epidemiology of plant diseases* (2nd ed.). Springer.

Crippa, A., Discacciati, A., Bottai, M., Spiegelman, D., & Orsini, N. (2018). One-stage dose-response meta-analysis for aggregated data. *Statistical Methods in Medical Research*, *28*(5), 1579–1596. https://doi.org/10.1177/0962280218773122

Crippa, A., & Orsini, N. (2016). Dose-response meta-analysis of differences in means. *BMC Medical Research Methodology*, *16*, 91. https://doi.org/10.1186/s12874-016-0189-0

Deeks, J. J., Dinnes, J., D'Amico, R., Sowden, A. J., Sakarovitch, C., & Song, F. (2003). Evaluating non-randomised intervention studies. *Health Technology Assessment*, *7*(27). 185 pp. https://doi.org/10.3310/hta7270

Dhont, M. (2010). History of oral contraception. *The European Journal of Contraception & Reproductive Health Care*, *15*(2), S12–S18. https://doi.org/10.3109/13625187.2010.513071

Diefenbach, M. A., Miller-Halegoua, S., & Bowen, D. J. (Eds.). (2016). *Handbook of health decision science* (1st ed.). Springer Science+Business Media LLC.

Diekmann, O., & Heesterbeek, J. A. P. (2000). Mathematical epidemiology of infectious diseases: Model building. In *Analysis and interpretation*. Wiley.

Dodd, S., White, I., & Williamson, P. (2011). Departure from treatment protocol in published randomised controlled trials: A review. *Trials*, *12*(A129). 1 p. https://doi.org/10.1186/1745-6215-12-S1-A129

Dodd, S., White, I., & Williamson, P. (2012). Nonadherence to treatment protocol in published randomised controlled trials: A review. *Trials*, *13*, 84. https://doi.org/10.1186/1745-6215-13-84

Dohoo, I., Martin, W., & Stryhn, H. (2009). *Veterinary epidemiological research* (2nd ed.). VER Inc. .

Duffield-Lillico, A. J., Dalkin, B. L., Reid, M. E., Turnbull, B. W., Slate, E. H., Jacobs, E. T., Marshall, J. R., Clark, L. C., & Nutritional Prevention of Cancer Study Group. (2003). Selenium supplementation, baseline plasma selenium status and incidence of prostate cancer: an analysis of the complete treatment period of the Nutritional Prevention of Cancer Trial. *BJU International*, *91*(7), 608–612. https://doi.org/10.1046/j.1464-410x.2003.04167.x

ECHA and EFSA (European Chemicals Agency and European Food Safety Authority) with the technical support of the Joint Research Centre (JRC), Andersson, N., Arena, M., Auteri, D., Barmaz, S., Grignard, E., Kienzler, A., Lepper, P., Lostia, A. M., Munn, S., Parra Morte, J. M., Pellizzato, F., Tarazona, J., Terron, A., & Van der Linden, S. (2018). Guidance for the identification of endocrine disruptors in the context of regulations (EU) No 528/2012 and (EC) No 1107/2009. *EFSA Journal*, *16*(6), 5311. https://doi.org/10.2903/j.efsa.2018.5311

EFSA (European Food Safety Authority). (2005). Opinion of the scientific committee on a request from EFSA related to a harmonised approach for risk assessment of substances which are both genotoxic and carcinogenic. *EFSA Journal*, *3*(10), 282. https://doi.org/10.2903/j.efsa.2005.282

EFSA (European Food Safety Authority). (2009). Scientific Opinion of the Panel on Contaminants in the Food Chain on a request from the European Commission on cadmium in food. *EFSA Journal*, *7*(3), 980. https://doi.org/10.2903/j.efsa.2009.980

EFSA (European Food Safety Authority). (2010). Application of systematic review methodology to food and feed safety assessments to support decision making. *EFSA Journal*, *8*(6), 1637. https://doi.org/10.2903/j.efsa.2010.1637

EFSA (European Food Safety Authority). (2015). Scientific report on principles and process for dealing with data and evidence in scientific assessments. *EFSA Journal*, *13*(5), 4121. https://doi.org/10.2903/j.efsa.2015.4121

EFSA (European Food Safety Authority), Aiassa, E., Martino, L., Barizzone, F., Ciccolallo, L., Garcia, A., Georgiadis, M., Muñoz Guajardo, I., Tomcikova, D., Alexander, J., Calistri, P., Gundert-Remy, U., Hart, H. D., Hoogenboom, R. L., Messean, A., Naska, A., Navajas Navarro, M., Noerrung, B., Ockleford, C., … Verloo, D. (2018). Implementation of PROMETHEUS 4-step approach for evidence use in EFSA scientific assessments: Benefits, issues, needs and solutions. *EFSA Supporting Publications*, EN-1395. https://doi.org/10.2903/sp.efsa.2018.EN-1395

EFSA (European Food Safety Authority), Medina-Pastor, P., & Triacchini, G. (2020a). The 2018 European Union report on pesticide residues in food. *EFSA Journal*, *18*(4), 6057. https://doi.org/10.2903/j.efsa.2020.6057

EFSA (European Food Safety Authority), Schrader, G., Kinkar, M., & Vos, S. (2020b). Pest survey card on Agrilus anxius. *EFSA Supporting Publications*, EN-1777. https://doi.org/10.2903/sp.efsa.2020.EN-1777

EFSA (European Food Safety Authority), Gibin, D., Gutierrez Linares, A., Fasanelli, E., Pasinato, L., & Delbianco, A. (2023a). Update of the Xylella spp. host plant database – Systematic literature search up to 30 June 2023. *EFSA Journal*, *21*(12), e8477. https://doi.org/10.2903/j.efsa.2023.8477

EFSA (European Food Safety Authority), Álvarez, F., Arena, M., Auteri, D., Leite, S. B., Binaglia, M., Castoldi, A. F., Chiusolo, A., Cioca, A.-A., Colagiorgi, A., Colas, M., Crivellente, F., De Lentdecker, C., De Magistris, I., Egsmose, M., Fait, G., Ferilli, F., Gouliarmou, V., Halling, K., … Villamar-Bouza, L. (2023b). Peer review of the pesticide risk assessment of the active substance metribuzin. *EFSA Journal*, *21*(8), 8140. https://doi.org/10.2903/j.efsa.2023.8140

EFSA and ECDC (European Food Safety Authority and European Centre for Disease Prevention and Control). (2023a). The European Union one health 2022 Zoonoses report. *EFSA Journal*, *21*(12), e8442. https://doi.org/10.2903/j.efsa.2023.8442

EFSA and ECDC (European Food Safety Authority and European Centre for Disease Prevention and Control). (2023b). The European Union summary report on antimicrobial resistance in zoonotic and indicator bacteria from humans, animals and food in 2020/2021. *EFSA Journal*, *21*(3), 7867. https://doi.org/10.2903/j.efsa.2023.7867

EFSA CEP Panel (EFSA Panel on Food Contact Materials, Enzymes and Processing Aids), Lambré, C., Barat Baviera, J. M., Bolognesi, C., Chesson, A., Cocconcelli, P. S., Crebelli, R., Gott, D. M., Grob, K., Lampi, E., Mengelers, M., Mortensen, A., Rivière, G., Silano, V., Steffensen, I.-L., Tlustos, C., Vernis, L., Zorn, H., Batke, M., … Van Loveren, H. (2023). Scientific Opinion on the re-evaluation of the risks to public health related to the presence of bisphenol a (BPA) in foodstuffs. *EFSA Journal*, *21*(4), 6857. https://doi.org/10.2903/j.efsa.2023.6857

EFSA CONTAM Panel (EFSA Panel on Contaminants in the Food Chain). (2010). Scientific opinion on Lead in food. *EFSA Journal*, *8*(4), 1570. https://doi.org/10.2903/j.efsa.2010.1570

EFSA CONTAM Panel (EFSA Panel on Contaminants in the Food Chain). (2011). Scientific opinion on Polybrominated diphenyl ethers (PBDEs) in food. *EFSA Journal*, *9*(5), 2156. https://doi.org/10.2903/j.efsa.2011.2156

EFSA CONTAM Panel (EFSA Panel on Contaminants in the Food Chain). (2012). Scientific opinion on the risk for public health related to the presence of mercury and methylmercury in food. *EFSA Journal*, *10*(12), 2985. https://doi.org/10.2903/j.efsa.2012.2985

EFSA CONTAM Panel (EFSA Panel on Contaminants in the Food Chain). (2014). Scientific opinion on the risks to public health related to the presence of perchlorate in food, in particular fruits and vegetables. *EFSA Journal*, *12*(10), 3869. https://doi.org/10.2903/j.efsa.2014.3869

EFSA CONTAM Panel (EFSA Panel on Contaminants in the Food Chain), Knutsen, H. K., Alexander, J., Barregard, L., Bignami, M., Bruschweiler, B., Ceccatelli, S., Cottrill, B., Dinovi, M., Edler, L., Grasl-Kraupp, B., Hogstrand, C., Hoogenboom, L. R., Nebbia, C. S., Oswald, I. P., Petersen, A., Rose, M., Roudot, A.-C., Vleminckx, C., … Schwerdtle, T. (2018). Scientific opinion on the risk to human health related to the presence of perfluorooctanesulfonic acid and perfluorooctanoic acid in food. *EFSA Journal*, *16*(12), 5194. https://doi.org/10.2903/j.efsa.2018.5194

EFSA CONTAM Panel (EFSA Panel on Contaminants in the Food Chain), Schrenk, D., Bignami, M., Bodin, L., Chipman, J. K., del Mazo, J., Grasl-Kraupp, B., Hogstrand, C., Hoogenboom, L., Leblanc, J.-C., Nebbia, C. S., Nielsen, E., Ntzani, E., Petersen, A., Sand, S., Schwerdtle, T., Wallace, H., Benford, D., Fürst, P., … Vleminckx, C. (2021). Scientific Opinion on the update of the risk assessment of hexabromocyclododecanes (HBCDDs) in food. *EFSA Journal*, *19*(3), 6421. https://doi.org/10.2903/j.efsa.2021.6421

EFSA CONTAM Panel (EFSA Panel on Contaminants in the Food Chain), Schrenk, D., Bignami, M., Bodin, L., Chipman, J. K., del Mazo, J., Grasl-Kraupp, B., Hogstrand, C., Hoogenboom, L. R., Leblanc, J.-C., Nebbia, C. S., Nielsen, E., Ntzani, E., Petersen, A., Sand, S., Vleminckx, C., Wallace, H., Barregard, L., Ceccatelli, S., … Schwerdtle, T. (2020). Scientific opinion on the risk to human health related to the presence of perfluoroalkyl substances in food. *EFSA Journal*, *18*(9), 6223. https://doi.org/10.2903/j.efsa.2020.6223

EFSA CONTAM Panel (EFSA Panel on Contaminants in the Food Chain), Schrenk, D., Bignami, M., Bodin, L., Chipman, J. K., del Mazo, J., Grasl-Kraupp, B., Hogstrand, C., Hoogenboom, L. R., Leblanc, J.-C., Nebbia, C. S., Nielsen, E., Ntzani, E., Petersen, A., Sand, S., Schwerdtle, T., Wallace, H., Benford, D., Fürst, P., … Vleminckx, C. (2024a). Update of the risk assessment of polybrominated diphenyl ethers (PBDEs) in food. *EFSA Journal*, *22*(1), e8497. https://doi.org/10.2903/j.efsa.2024.8497

EFSA CONTAM Panel (EFSA Panel on Contaminants in the Food Chain), Schrenk, D., Bignami, M., Bodin, L., Chipman, J. K., del Mazo, J., Grasl-Kraupp, B., Hogstrand, C., Hoogenboom, L. R., Leblanc, J.-C., Nebbia, C. S., Nielsen, E., Ntzani, E., Petersen, A., Sand, S., Vleminckx, C., Wallace, H., Barregård, L., Benford, D., … Schwerdtle, T. (2024b). Update of the risk assessment of inorganic arsenic in food. *EFSA Journal*, *22*(1), e8488. https://doi.org/10.2903/j.efsa.2024.8488

EFSA FAF Panel (EFSA Panel on Food Additives and Flavouring), Younes, M., Aquilina, G., Castle, L., Engel, K.-H., Fowler, P., Frutos Fernandez, M. J., Fürst, P., Gürtler, R., Gundert-Remy, U., Husøy, T., Manco, M., Mennes, W., Passamonti, S., Moldeus, P., Shah, R., Waalkens-Berendsen, I., Wölfle, D., Wright, M., … Vianello, G. (2021). Scientific opinion on the re-evaluation of thaumatin (E 957) as food additive. *EFSA Journal*, *19*(11), 6884. https://doi.org/10.2903/j.efsa.2021.6884

EFSA FAF Panel (EFSA Panel on Food Additives and Flavourings), Younes, M., Aquilina, G., Castle, L., Degen, G., Engel, K.-H., Fowler, P. J., Frutos Fernandez, M. J., Fürst, P., Gundert-Remy, U., Gürtler, R., Husøy, T., Manco, M., Mennes, W., Moldeus, P., Passamonti, S., Shah, R., Waalkens-Berendsen, I., Wright, M., … Tard, A. (2023). Re-evaluation of erythritol (E 968) as a food additive. *EFSA Journal*, *21*(12), e8430. https://doi.org/10.2903/j.efsa.2023.8430

EFSA NDA Panel (EFSA Panel on Dietetic Products, Nutrition and Allergies). (2015). Scientific opinion on the safety of caffeine. *EFSA Journal*, *13*(5), 4102. https://doi.org/10.2903/j.efsa.2015.4102

EFSA NDA Panel (EFSA Panel on Nutrition, Novel Foods and Food Allergens), Turck, D., Castenmiller, J., de Henauw, S., Hirsch-Ernst, K.-I., Kearney, J., Knutsen, H. K., Maciuk, A., Mangelsdorf, I., McArdle, H. J., Pelaez, C., Pentieva, K., Siani, A., Thies, F., Tsabouri, S., Vinceti, M., Aggett, P., Fairweather-Tait, S., Martin, A., … Naska, A. (2019). Dietary reference values for sodium. *EFSA Journal*, *17*(9), 5778. https://doi.org/10.2903/j.efsa.2019.5778

EFSA NDA Panel (EFSA Panel on Nutrition, Novel Foods and Food Allergens), Turck, D., Bohn, T., Castenmiller, J., de Henauw, S., Hirsch-Ernst, K. I., Knutsen, H. K., Maciuk, A., Mangelsdorf, I., McArdle, H. J., Naska, A., Peláez, C., Pentieva, K., Siani, A., Thies, F., Tsabouri, S., Adan, R., Emmett, P., Galli, C., … Vinceti, M. (2022). Tolerable upper intake level for dietary sugars. *EFSA Journal*, *20*(2), e07074. https://doi.org/10.2903/j.efsa.2022.7074

EFSA NDA Panel (EFSA Panel on Nutrition, Novel Foods and Food Allergens), Turck, D., Bohn, T., Castenmiller, J., de Henauw, S., Hirsch-Ernst, K.-I., Knutsen, H. K., Maciuk, A., Mangelsdorf, I., McArdle, H. J., Peláez, C., Pentieva, K., Siani, A., Thies, F., Tsabouri, S., Vinceti, M., Aggett, P., Crous Bou, M., Cubadda, F., … Naska, A. (2023). Scientific opinion on the tolerable upper intake level for selenium. *EFSA Journal*, *21*(1), 7704. https://doi.org/10.2903/j.efsa.2023.7704

EFSA PLH Panel (EFSA Panel on Plant Health), Jeger, M., Bragard, C., Caffier, D., Candresse, T., Chatzivassiliou, E., Dehnen-Schmutz, K., Gregoire, J.-C., Jaques Miret, J. A., MacLeod, A., Navajas Navarro, M., Niere, B., Parnell, S., Potting, R., Rafoss, T., Rossi, V., Urek, G., Van Bruggen, A., Van Der Werf, W., … Gilioli, G. (2018). Guidance on quantitative pest risk assessment. *EFSA Journal*, *16*(8), 5350. https://doi.org/10.2903/j.efsa.2018.5350

EFSA PLH Panel (EFSA Panel on Plant Health), Bragard, C., Dehnen-Schmutz, K., Di Serio, F., Gonthier, P., Jacques, M.-A., Jaques Miret, J. A., Fejer Justesen, A., MacLeod, A., Magnusson, C. S., Milonas, P., Navas-Cortes, J. A., Parnell, S., Reignault, P. L., Thulke, H.-H., Van der Werf, W., Vicent Civera, A., Yuen, J., Zappala, L., … Potting, R. (2019). Guidance on commodity risk assessment for the evaluation of high risk plants dossiers. *EFSA Journal*, *17*(4), 5668. https://doi.org/10.2903/j.efsa.2019.5668

EFSA PPR Panel (EFSA Panel on Plant Protection Products and their Residues), Ockleford, C., Adriaanse, P., Berny, P., Brock, T., Duquesne, S., Grilli, S., Hougaard, S., Klein, M., Kuhl, T., Laskowski, R., Machera, K., Pelkonen, O., Pieper, S., Smith, R., Stemmer, M., Sundh, I., Teodorovic, I., Tiktak, A., … Hernandez-Jerez, A. F. (2017a). 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://doi.org/10.2903/j.efsa.2017.5007

EFSA PPR Panel (EFSA Panel on Plant Protection Products and their Residues), Ockleford, C., Adriaanse, P., Berny, P., Brock, T., Duquesne, S., Grilli, S., Hernandez-Jerez, A. F., Bennekou, S. H., Klein, M., Kuhl, T., Laskowski, R., Machera, K., Pelkonen, O., Pieper, S., Smith, R., Stemmer, M., Sundh, I., Teodorovic, I., … Bennekou, S. H. (2017b). Scientific opinion on the investigation into experimental toxicological properties of plant protection products having a potential link to Parkinson's disease and childhood leukaemia. *EFSA Journal*, *15*(3), 4691. https://doi.org/10.2903/j.efsa.2017.4691

EFSA PPR Panel (EFSA Panel on Plant Protection Products and their Residues), Hernandez-Jerez, A., Adriaanse, P., Aldrich, A., Berny, P., Coja, T., Duquesne, S., Focks, A., Marinovich, M., Millet, M., Pelkonen, O., Pieper, S., Tiktak, A., Topping, C., Widenfalk, A., Wilks, M., Wolterink, G., Crofton, K., Hougaard Bennekou, S., … Tzoulaki, I. (2021). Scientific Opinion on the development of integrated approaches to testing and assessment (IATA) case studies on developmental neurotoxicity (DNT) risk assessment. *EFSA Journal*, *19*(6), 6599. https://doi.org/10.2903/j.efsa.2021.6599

EFSA Scientific Committee. (2009). Guidance of the scientific committee on a request from EFSA on the use of the benchmark dose approach in risk assessment. *EFSA Journal*, *7*(6), 1150. https://doi.org/10.2903/j.efsa.2009.1150

EFSA Scientific Committee. (2011). Statistical Significance and Biological Relevance. *EFSA Journal*, *9*(9), 2372. https://doi.org/10.2903/j.efsa.2011.2372

EFSA Scientific Committee. (2012). Guidance on selected default values to be used by the EFSA scientific committee, scientific panels and units in the absence of actual measured data. *EFSA Journal*, *10*(3), 2579. https://doi.org/10.2903/j.efsa.2012.2579

EFSA Scientific Committee, Hardy, A., Benford, D., Halldorsson, T., Jeger, M. J., Knutsen, H. K., More, S., Younes, M., Naegeli, H., Noteborn, H., Ockleford, C., Ricci, A., Rychen, G., Schlatter, J. R., Silano, V., Solecki, R., Turck, D., Bresson, J.-L., Griffin, J., … Alexander, J. (2017a). Guidance on biological relevance. *EFSA Journal*, *15*(8), 4970. https://doi.org/10.2903/j.efsa.2017.4970

EFSA Scientific Committee, Hardy, A., Benford, D., Halldorsson, T., Jeger, M. J., Knutsen, H. K., More, S., Naegeli, H., Noteborn, H., Ockleford, C., Ricci, A., Rychen, G., Schlatter, J. R., Silano, V., Solecki, R., Turck, D., Benfenati, E., Chaudhry, Q. M., Craig, P., … Younes, M. (2017b). Scientific opinion on the guidance on the use of the weight of evidence approach in scientific assessments. *EFSA Journal*, *15*(8), 4971. https://doi.org/10.2903/j.efsa.2017.4971

EFSA Scientific Committee, Hardy, A., Benford, D., Halldorsson, T., Jeger, M. J., Knutsen, K. H., More, S., Mortense, N. A., Naegeli, H., Noteborn, H., Ockleford, C., Ricci, A., Rychen, G., Silano, V., Solecki, R., Turck, D., Aerts, M., Bodin, L., Davis, A., … Schlatter, J. R. (2017c). Update: Guidance on the use of the benchmark dose approach in risk assessment. *EFSA Journal*, *15*(1), 4658. https://doi.org/10.2903/j.efsa.2017.4658

EFSA Scientific Committee, More, S. J., Bampidis, V., Benford, D., Bragard, C., Halldorsson, T. I., Hernandez-Jerez, A. F., Bennekou, S. H., Koutsoumanis, K., Lambre, C., Machera, K., Mennes, W., Mullins, E., Nielsen, S. S., Schrenk, D., Turck, D., Younes, M., Aerts, M., Edler, L., … Schlatter, J. (2022). Guidance on the use of the benchmark dose approach in risk assessment. *EFSA Journal*, *20*(10), 7584. https://doi.org/10.2903/j.efsa.2022.7584

EFSA Scientific Committee, More, S., Bampidis, V., Benford, D., Bragard, C., Hernández-Jerez, A. F., Bennekou, S. H., Koutsoumanis, K., Lambré, C., Machera, K., Mullins, E., Nielsen, S. S., Schrenk, D., Turck, D., Younes, M., Kraft, A., Naegeli, H., Tsaioun, K., Aiassa, E., … Halldorsson, T. I. (2023a). Guidance on protocol development for EFSA generic scientific assessments. *EFSA Journal*, *21*(10), 8312. https://doi.org/10.2903/j.efsa.2023.8312

EFSA Scientific Committee, More, S. J., Bampidis, V., Benford, D., Bragard, C., Halldorsson, T. I., Hernández-Jerez, A. F., Bennekou, S. H., Koutsoumanis, K., Lambré, C., Machera, K., Mullins, E., Nielsen, S. S., Schlatter, J. R., Schrenk, D., Turck, D., Younes, M., Boon, P., Ferns, G. A. A., … Leblanc, J.-C. (2023b). Scientific Opinion on the re-evaluation of the existing health-based guidance values for copper and exposure assessment from all sources. *EFSA Journal*, *21*(1), 7728. https://doi.org/10.2903/j.efsa.2023.7728

EFSA Scientific Committee, More, S. J., Benford, D., Bennekou, S. H., Bampidis, V., Bragard, C., Halldorsson, T. I., Hernández-Jerez, A. F., Koutsoumanis, K., Lambré, C., Machera, K., Mullins, E., Nielsen, S. S., Schlatter, J., Schrenk, D., Turck, D., Naska, A., Poulsen, M., Ranta, J., … Younes, M. (2024). Guidance on risk-benefit assessment of foods. *EFSA Journal*, *22*(7), e8875. https://doi.org/10.2903/j.efsa.2024.8875

Emerson, J. D., Burdick, E., Hoaglin, D. C., Mosteller, F., & Chalmers, T. C. (1990). An empirical study of the possible relation of treatment differences to quality scores in controlled randomized clinical trials. *Controlled Clinical Trials*, *11*(5), 339–352. https://doi.org/10.1016/0197-2456(90)90175-2

Falkingham, M., Abdelhamid, A., Curtis, P., Fairweather-Tait, S., Dye, L., & Hooper, L. (2010). The effects of oral iron supplementation on cognition in older children and adults: A systematic review and meta-analysis. *Nutrition Journal*, *2010*(9), 4. https://doi.org/10.1186/1475-2891-9-4

FAO and WHO (Food and Agricultural Organisation and World Health Organisation). (2011). *Safety evaluation of certain contaminants in food/prepared by the seventy-second meeting of the joint FAO/WHO expert committee on food additives (JECFA)* (Vol. *8*, p. 799). FAO JECFA MONOGRAPHS. https://apps.who.int/iris/bitstream/handle/10665/44520/9789241660631_eng.pdf?sequence=1&isAllowed=y

Fei, C., McLaughlin, J. K., Tarone, R. E., & Olsen, J. (2007). Perfluorinated chemicals and fetal growth: A study within the Danish national birth cohort. *Environmental Health Perspectives*, *115*(11), 1677–1682. https://doi.org/10.1289/ehp.10506

Ferreira, J. C., & Patino, C. M. (2017). Types of outcomes in clinical research. *Jornal Brasileiro de Pneumologia*, *43*(1), 5. https://doi.org/10.1590/S1806-37562017000000021

Filippini, T., Halldorsson, T. I., Capitão, C., Martins, R., Giannakou, K., Hogervorst, J., Vinceti, M., Åkesson, A., Leander, K., Katsonouri, A., Santos, O., Virgolino, A., & Laguzzi, F. (2022). Dietary acrylamide exposure and risk of site-specific cancer: A systematic review and dose-response meta-analysis of epidemiological studies. *Frontiers in Nutrition*, *9*, 875607. https://doi.org/10.3389/fnut.2022.875607

Filippini, T., Malavolti, M., Whelton, P. K., Naska, A., Orsini, N., & Vinceti, M. (2021). Blood pressure effects of sodium reduction. *Circulation*, *143*, 1542–1567. https://doi.org/10.1161/CIRCULATIONAHA.120.050371

Genaidy, A. M., LeMasters, G. K., Lockey, J., Succop, P., Deddens, J., Sobeih, T., & Dunning, K. (2007). An epidemiological appraisal instrument – a tool for evaluation of epidemiological studies. *Ergonomics*, *50*(6), 920–960. https://doi.org/10.1080/00140130701237667

Gencer, B., & Giugliano, R. P. (2020). Management of LDL-cholesterol after an acute coronary syndrome: Key comparisons of the American and European clinical guidelines to the attention of the healthcare providers. *Clinical Cardiology*, *43*, 684–690. https://doi.org/10.1002/clc.23410

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, P values, confidence intervals, and power: A guide to misinterpretations. *European Journal of Epidemiology*, *31*(4), 337–350. https://doi.org/10.1007/s10654-016-0149-3

Gross, M., Green, R. M., Weltje, L., & Wheeler, J. R. (2017). Weight of evidence approaches for the identification of endocrine disrupting properties of chemicals: Review and recommendations for EU regulatory application. *Regulatory Toxicology and Pharmacology*, *91*, 20–28. https://doi.org/10.1016/j.yrtph.2017.10.004

Halldorsson, T. I., Fei, C., Olsen, J., Lipworth, L., McLaughlin, J. K., & Olsen, S. F. (2008). Dietary predictors of perfluorinated chemicals: A study from the Danish National Birth Cohort. *Environmental Science & Technology*, *42*(23), 8971–8977. https://doi.org/10.1021/es801907r

Hammer, G. P., du Prel, J. B., & Blettner, M. (2009). Avoiding bias in observational studies: Part 8 in a series of articles on evaluation of scientific publications. *Deutsches Ärzteblatt*, *106*(41), 664–668. https://doi.org/10.3238/arztebl.2009.0664

Hartley, L., May, M. D., Loveman, E., Colquitt, J. L., & Rees, K. (2016). Dietary fibre for the primary prevention of cardiovascular disease. *Cochrane Database of Systematic Reviews*, *7*, CD011472. https://doi.org/10.1002/14651858.CD011472.pub2

Hasselt University. (2022). EFSA platform for Bayesian benchmark dose analysis. *EFSA Supporting Publication*, *19*(12), EN-7740. https://doi.org/10.2903/sp.efsa.2022.EN-7740

Hébert, J. R., Frongillo, E. A., Adams, S. A., Turner-McGrievy, G. M., Hurley, T. G., Miller, D. R., & Ockene, I. S. (2016). Perspective: Randomized controlled trials are not a panacea for diet-related research. *Advances in Nutrition*, *7*(3), 423–432. https://doi.org/10.3945/an.115.011023

Heinsberg, L. W., & Weeks, D. E. (2022). Post hoc power is not informative. *Genetic Epidemiology*, *2022*(46), 390–394. https://doi.org/10.1002/gepi.22464

Hernán, M. A., Hernández-Díaz, S., & Robins, J. M. (2004). A structural approach to selection bias. *Epidemiology*, *15*(5), 615–625. https://doi.org/10.1097/01.ede.0000135174.63482.43

Hernán, M. A., & Robins, J. M. (2020). *Causal inference: What if*. Chapman & Hall/CRC.

Higgins JPT and Green S (eds). 2011. *Cochrane handbook for systematic reviews of interventions version 5.1.0 [updated march 2011]. The Cochrane collaboration, 2011*. www.handbook.cochrane.org

Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ and Welch VA (editors). *Cochrane handbook for systematic reviews of interventions version 6.4 (updated August 2023), Cochrane*, 2023. Available from www.training.cochrane.org/handbook

Hill, A. B. (1965). The environment and disease: Association or causation? *Proceedings of the Royal Society of Medicine*, *58*, 295–300. https://doi.org/10.1177/003591576505800503

Hooijmans, C. R., Rovers, M. M., de Vries, R. B., Leenaars, M., Ritskes-Hoitinga, M., & Langendam, M. W. (2014). SYRCLE's risk of bias tool for animal studies. *BMC Medical Research Methodology*, *14*, 43. https://doi.org/10.1186/1471-2288-14-43

Howard, B. V., Van Horn, L., Hsia, J., Manson, J. E., Stefanick, M. L., Wassertheil-Smoller, S., Kuller, L. H., LaCroix, A. Z., Langer, R. D., Lasser, N. L., Lewis, C. E., Limacher, M. C., Margolis, K. L., Mysiw, W. J., Ockene, J. K., Parker, L. M., Perri, M. G., Phillips, L., Prentice, R. L., … Kotchen, J. M. (2006). Low-fat dietary pattern and risk of cardiovascular disease: The Women's Health Initiative randomized controlled dietary modification trial. *Journal of the American Medical Association*, *295*(6), 655–666. https://doi.org/10.1001/jama.295.6.655

Hurst, R., Armah, C. N., Dainty, J. R., Hart, D. J., Teucher, B., Goldson, A. J., Broadley, M. R., Motley, A. K., & Fairweather-Tait, S. J. (2010). Establishing optimal selenium status: Results of a randomized, double-blind, placebo-controlled trial. *American Journal of Clinical Nutrition*, *91*(4), 923–931. https://doi.org/10.3945/ajcn.2009.28169

Ingelsson, E., Schaefer, E. J., Contois, J. H., McNamara, J. R., Sullivan, L., Keyes, M. J., Pencina, M. J., Schoonmaker, C., Wilson, P. W., D'Agostino, R. B., & Vasan, R. S. (2007). Clinical utility of different lipid measures for prediction of coronary heart disease in men and women. *Journal of the American Medical Association*, *298*(7), 776–785. https://doi.org/10.1001/jama.298.7.776

Ioannidou, S., Cascio, C., & Gilsenan, M. B. (2021). European food safety authority open access tools to estimate dietary exposure to food chemicals. *Environment International*, *149*, 106357. https://doi.org/10.1016/j.envint.2020.106357

Johnson, P. I., Sutton, P., Atchley, D. S., Koustas, E., Lam, J., Sen, S., Robinson, K. A., Axelrad, D. A., & Woodruff, T. J. (2014). The navigation guide - evidence-based medicine meets environmental health: Systematic review of human evidence for PFOA effects on fetal growth. *Environmental Health Perspectives*, *122*(10), 1028–1039. https://doi.org/10.1289/ehp.1307893

Jüni, P., Witschi, A., Bloch, R., & Egger, M. (1999). The hazards of scoring the quality of clinical trials for meta-analysis. *Journal of the American Medical Association*, *282*(11), 1054–1060. https://doi.org/10.1001/jama.282.11.1054

Kestenbaum, B. (2019). *Epidemiology and biostatistics - an introduction to clinical research*. Springer International Publishing. https://doi.org/10.1007/978-3-319-96644-1

Kim, H. J., Fay, M. P., Feuer, E. J., & Midthune, D. N. (2000). Permutation tests for joinpoint regression with applications to cancer rates. *Statistics in Medicine*, *19*, 335–351. https://doi.org/10.1002/(sici)1097-0258(20000215)19:3<335::aid-sim336>3.0.co;2-z Erratum in: Statistics in Medicine 2001 Feb 28;20(4):655.

Kilkenny, C., Browne, W. J., Cuthill, I. C., Emerson, M., & Altman, D. G. (2010). Improving bioscience research reporting: the ARRIVE guidelines for reporting animal research. *PLoS Biology*, *8*(6), e1000412.

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

Knol, M. J., Vandenbroucke, J. P., Scott, P., & Egger, M. (2008). What do case-control studies estimate? Survey of methods and assumptions in published case-control research. *American Journal of Epidemiology*, *168*, 1073–1081. https://doi.org/10.1093/aje/kwn217

Koustas, E., Lam, J., Sutton, P., Johnson, P. I., Atchley, D. S., Sen, S., Robinson, K. A., Axelrad, D. A., & Woodruff, T. J., (2013). Applying the Navigation Guide: Case Study #1 – the Impact of Developmental Exposure to Perfluorooctanoic Acid (PFOA) on Fetal Growth a Systematic Review of the Non-human Evidence (Final Protocol).

Kristal, A. R., Darke, A. K., Morris, J. S., Tangen, C. M., Goodman, P. J., Thompson, I. M., Meyskens, F. L., Goodman, G. E., Minasian, L. M., Parnes, H. L., Lippman, S. M., & Klein, E. A. (2014). Baseline selenium status and effects of selenium and vitamin E supplementation on prostate cancer risk. *Journal of the National Cancer Institute*, *106*(3), djt456. https://doi.org/10.1093/jnci/djt456

Lam, J., Koustas, E., Sutton, P., Johnson, P. I., Atchley, D. S., Sen, S., Robinson, K. A., Axelrad, D. A., & Woodruff, T. J. (2014). The Navigation Guide – evidence-based medicine meets environmental health: integration of animal and human evidence for PFOA effects on fetal growth. *Environmental Health Perspectives*, *122*(10), 1040–1051. https://doi.org/10.1289/ehp.1307923

Langlois, E. V., Daniels, K., & Akl, E. A. (2018). *Evidence synthesis for health policy and systems: A methods guide*. World Health Organization, . Licence: CC BY-NC-SA 3.0 IGO. https://apps.who.int/iris/bitstream/handle/10665/275367/9789241514552-eng.pdf?ua=1

Lash, T. L., VanderWeele, T. J., Haneuse, S., & Rothman, K. J. (2021). *Modern Epidemiology* (p. 1250). Wolters Kluwer.

Lazcano, G., Papuzinski, C., Madrid, E., & Arancibia, M. (2019). General concepts in biostatistics and clinical epidemiology: Observational studies with cohort design. *Medwave*, *19*(11), e7748. https://doi.org/10.5867/medwave.2019.11.7748

Lee, W. J., & Hase, K. (2014). Gut microbiota–generated metabolites in animal health and disease. *Nature Chemical Biology*, *10*, 416–424. https://doi.org/10.1038/nchembio.1535

Lenters, V., Vermeulen, R., Dogger, S., Stayner, L., Portengen, L., Burdorf, A., & Heederik, D. (2011). A meta-analysis of asbestos and lung cancer: is better quality exposure assessment associated with steeper slopes of the exposure–response relationships? *Environmental Health Perspectives*, *119*, 1547–1555. https://doi.org/10.1289/ehp.1002879

Lesko, C. R., Keil, A. P., & Edwards, J. K. (2020). The epidemiologic toolbox: Identifying, honing, and using the right tools for the job. *American Journal of Epidemiology*, *189*, 511–517. https://doi.org/10.1093/aje/kwaa030

Li, Y., Barregard, L., Xu, Y., Scott, K., Pineda, D., Lindh, C. H., Jakobsson, K., & Fletcher, T. (2020). Associations between perfluoroalkyl substances and serum lipids in a Swedish adult population with contaminated drinking water. *Environmental Health*, *2020*(19), 33. https://doi.org/10.1186/s12940-020-00588-9

Li, Y., Fletcher, T., Mucs, D., Scott, K., Lindh, C. H., Tallving, P., & Jakobsson, K. (2018). Half-lives of PFOS, PFHxS and PFOA after end of exposure to contaminated drinking water. *Occupational and Environmental Medicine*, *75*(1), 46–51. https://doi.org/10.1136/oemed-2017-104651

Lilienfeld, A. M., & Lilienfeld, D. E. (1980). *Foundations of epidemiology* (2d ed.). Oxford University Press.

Lippman, S. M., Klein, E. A., Goodman, P. J., Lucia, M. S., Thompson, I. M., Ford, L. G., Parnes, H. L., Minasian, L. M., Gaziano, J. M., Hartline, J. A., Parsons, J. K., Bearden, J. D., Crawford, E. D., Goodman, E. D., Claudio, J., Winquist, E., Cook, E. D., Karp, D. D., Walther, P., … Coltman, C. A. (2009). Effect of selenium and vitamin E on risk of prostate cancer and other cancers: The selenium and vitamin E cancer prevention trial (SELECT). *Journal of the American Medical Association*, *301*(1), 39–51. https://doi.org/10.1001/jama.2008.864

Longnecker, M. P., Smith, C. S., Kissling, G. E., Hoppin, J. A., Butenhoff, J. L., Decker, E., Ehresman, D. J., Ellefson, M. E., Flaherty, J., Gardner, M. S., Langlois, E., Leblanc, A., Lindstrom, A. B., Reagen, W. R., Strynar, M. J., & Studabaker, W. B. (2008). An Interlaboratory study of Perfluorinated alkyl compound levels in human plasma. *Environmental Research*, *107*(2), 152–159. https://doi.org/10.1016/j.envres.2008.01.005

Lynch, H. N., Goodman, J. E., Tabony, J. A., & Rhomberg, L. R. (2016). Systematic comparison of study quality criteria. *Regulatory Toxicology and Pharmacology*, *76*, 187–198. https://doi.org/10.1016/j.yrtph.2015.12.017

Madden, L. V., Hughes, G., & van den Bosch, F. (2007). *The study of plant disease epidemics*. The American Phytopathological Society.

Mansournia, M. A., Higgins, J. P. T., Sterne, J. A. C., & Hernán, M. A. (2017). Biases in randomized trials: A conversation between Trialists and epidemiologists. *Epidemiology*, *28*(1), 54–59. https://doi.org/10.1097/EDE.0000000000000564

McLeod, C., Norman, R., Litton, E., Saville, B. R., Webb, S., & Snelling, T. L. (2019). Choosing primary endpoints for clinical trials of health care interventions. *Contemporary Clinical Trials Communications*, *12*(16), 100486. https://doi.org/10.1016/j.conctc.2019.100486

Meek, M. E., Palermo, C. M., Bachmann, A. N., North, C. M., & Lewis, R. J. (2014). Mode of action human relevance (species concordance) framework: Evolution of the Bradford Hill considerations and comparative analysis of weight of evidence. *Journal of Applied Toxicology*, *34*, 595–606. https://doi.org/10.1002/jat.2984

Miyake-Lye, I. M., Hempel, S., Shanman, R., & Shekelle, P. G. (2016). What is an evidence map? A systematic review of published evidence maps and their definitions, methods, and products. *Systematic Reviews*, *5*, 28. https://doi.org/10.1186/s13643-016-0204-x

Money, C. D., Tomenson, J. A., Penman, M. G., Boogaard, P. J., & Jeffrey Lewis, R. (2013). A systematic approach for evaluating and scoring human data. *Regulatory Toxicology and Pharmacology*, *66*(2), 241–247. https://doi.org/10.1016/j.yrtph.2013.03.011

Morgan, R. L., Thayer, K. A., Bero, L., Bruce, N., Falck-Ytter, Y., Ghersi, D., Guyatt, G., Hooijmans, C., Langendam, M., Mandrioli, D., Mustafa, R. A., Rehfuess, E. A., Rooney, A. A., Shea, B., Silbergeld, E. K., Sutton, P., Wolfe, M. S., Woodruff, T. J., Verbeek, J. H., … Schünemann, H. J. (2016). GRADE: Assessing the quality of evidence in environmental and occupational health. *Environment International*, *92-93*, 611–616. https://doi.org/10.1016/j.envint.2016.01.004

Morgan, R. L., Thayer, K. A., Santesso, N., Holloway, A. C., Blain, R., Eftim, S. E., Goldstone, A. E., Ross, P., Ansari, M., Akl, E. A., Filippini, T., Hansell, A., Meerpohl, J., Mustafa, R. A., Verbeek, J., Vinceti, M., Whaley, P., & Schünemann HJ and the GRADE Working Group. (2019). A risk of bias instrument for non-randomized studies of exposures: A users' guide to its application in the context of GRADE. *Environment International*, *122*, 168–184. https://doi.org/10.1016/j.envint.2018.11.004

Morgan, R. L., Whaley, P., Thayer, K. A., & Schünemann, H. J. (2018). Identifying the PECO: A framework for formulating good questions to explore the association of environmental and other exposures with health outcomes. *Environment International*, *121*(Pt 1), 1027–1031. https://doi.org/10.1016/j.envint.2018.07.015

Muggeo, V. (2008). Segmented: An R package to fit regression models with broken-line relationships. *R News*, *8*, 20–25.

Newcomer, J. W., Craft, S., Fucetola, R., Moldin, S. O., Selke, G., Paras, L., & Miller, R. (1999). Glucose-induced increase in memory performance in patients with schizophrenia. *Schizophrenia Bulletin*, *25*(2), 321–335. https://doi.org/10.1093/oxfordjournals.schbul.a033381

Nielsen, S. S., Toft, N., & Gardner, I. A. (2011). Structured approach to design of diagnostic test evaluation studies for chronic progressive infections in animals. *Veterinary Microbiology*, *150*, 115–125. https://doi.org/10.1016/j.vetmic.2011.01.019

Noordzij, M., van Diepen, M., Caskey, F. C., & Jager, K. J. (2017). Relative risk versus absolute risk: One cannot be interpreted without the other. *Nephrology, Dialysis, Transplantation*, *32*(suppl_2), ii13–ii18. https://doi.org/10.1093/ndt/gfw465

NTP (National Toxicology Program). (2019). Handbook for conducting a literature-based health assessment using OHAT approach for systematic review and evidence integration. Office of Health Assessment and Translation, Division of the National Toxicology Program, National Institute of Environmental Health Sciences. https://ntp.niehs.nih.gov/whatwestudy/assessments/noncancer/handbook/index.html

Organisation for Economic Co-operation and Development (OECD). (2005). Guidance document on the validation and international acceptance of new or updated test methods for hazard assessment. ENV/JM/MONO(2005), 14, 96 p.

O'Brien, K. M., Lawrence, K. G., & Keil, A. P. (2022). The case for case-cohort: An applied Epidemiologist's guide to reframing case-cohort studies to improve usability and flexibility. *Epidemiology*, *33*, 354–361. https://doi.org/10.1097/EDE.0000000000001469

Orsini, N., Li, R., Wolk, A., Khudyakov, P., & Spiegelman, D. (2012). Meta-analysis for linear and nonlinear dose-response relations: Examples, an evaluation of approximations, and software. *American Journal of Epidemiology*, *175*(1), 66–73. https://doi.org/10.1093/aje/kwr265

Orsini, N., & Spiegelman, D. (2021). Meta-analysis of dose–response relationships. In C. H. Schmid, T. Stijnen, & I. White (Eds.), *Handbook of meta-analysis* (1st ed., p. 2021). Chapman and Hall/CRC. https://doi.org/10.1201/9781315119403

Page, M. J., McKenzie, J. E., & Higgins, J. P. T. (2018). Tools for assessing risk of reporting biases in studies and syntheses of studies: A systematic review. *BMJ Open*, *8*, e019703. https://doi.org/10.1136/bmjopen-2017-019703

Pandis, N. (2011). The evidence pyramid and introduction to randomized controlled trials. *American Journal of Orthodontics and Dentofacial Orthopedics*, *140*(3), 446–447. https://doi.org/10.1016/j.ajodo.2011.04.016

Pearce, N. (2005). *A short introduction to epidemiology* (second ed.. Occasional Report Series No 2,). Centre for Public Health Research, Massey University.

Pearce, N. (2016). Analysis of matched case-control studies. *BMJ*, *2016*, 352. https://doi.org/10.1136/bmj.i969

Pearce, N., Checkoway, H., & Kriebel, D. (2007). Bias in occupational epidemiology studies. *Occupational and Environmental Medicine*, *64*(8), 562–568. https://doi.org/10.1136/oem.2006.026690

Pearce, N., Smith, A. H., Howard, J. K., Sheppard, R. A., Giles, H. J., & Teague, C. A. (1986). Non-Hodgkin's lymphoma and exposure to phenoxyherbicides, chlorophenols, fencing work, and meat works employment: A case-control study. *British Journal for Industrial Medicine*, *43*(2), 75–83. https://doi.org/10.1136/oem.43.2.75

Pearl, J. (2009). Causal inference in statistics: An overview. *Statistics Survey*, *3*, 96–146. https://doi.org/10.1214/09-SS057

Pearl, J., & Mackenzie, D. (2018). *The book of why: The new science of cause and effect*. Basic Books.

Peters, M. D. J., Marnie, C., Tricco, A. C., Pollock, D., Munn, Z., Alexander, L., McInerney, P., Godfrey, C. M., & Khalil, H. (2020). Updated methodological guidance for the conduct of scoping reviews. *JBI Evidence Synthesis*, *18*, 2119–2126. https://doi.org/10.11124/JBIES-20-00167

Porta. (2014). *Dictionary of epidemiology* (6th ed.). Oxford University Press.

Poston, L., Bell, R., Croker, H., Flynn, A. C., Godfrey, K. M., Goff, L., Hayes, L., Khazaezadeh, N., Nelson, S. M., Oteng-Ntim, E., Pasupathy, D., Patel, N., Robson, S. C., Sandall, J., Sanders, T. A. B., Sattar, N., Seed, P. T., Wardle, J., Whitworth, M. K., … the UPBEAT Trial Consortium. (2015). Effect of a Behavioural intervention in obese pregnant women (the UPBEAT study): A multicentre. *Randomised Controlled Trial. Lancet Diabetes Endocrinology*, *3*(10), 767–777. https://doi.org/10.1016/S2213-8587(15)00227-2

Quigley, J. M., Thompson, J. C., Halfpenny, N. J., & Scott, D. A. (2019). Critical appraisal of nonrandomized studies - a review of recommended and commonly used tools. *Journal of Evaluation in Clinical Practice*, *25*(1), 44–52. https://doi.org/10.1111/jep.12889

Rizzo, D. M., Lichtveld, M., Mazet, J. A. K., Togami, E., & Miller, S. A. (2021). Plant health and its effects on food safety and security in a one health framework: Four case studies. *One Health Outlook*, *3*(2021), 1–9. https://doi.org/10.1186/s42522-021-00038-7

Rooney, A. A., Boyles, A. L., Wolfe, M. S., Bucher, J. R., & Thayer, K. A. (2014). Systematic review and evidence integration for literature-based environmental health science assessments. *Environmental Health Perspectives*, *122*(7), 711–718. https://doi.org/10.1289/ehp.1307972

Rothman, K. J. (1976). Causes. *American Journal of Epidemiology*, *104*(6), 587–592. https://doi.org/10.1093/oxfordjournals.aje.a112335

Rothman, K. J. (2014). Six persistent research misconceptions. *Journal of General Internal Medicine*, *29*(7), 1060–1064. https://doi.org/10.1007/s11606-013-2755-z

Rothman, K. J. (2016). Disengaging from statistical significance. *European Journal of Epidemiology*, *31*(5), 443–444. https://doi.org/10.1007/s10654-016-0158-2

Rothman, K. J., & Greenland, S. (2018). Planning study size based on precision rather than power. *Epidemiology*, *29*, 599–603. https://doi.org/10.1097/EDE.0000000000000876

Samuel, G. O., Hoffmann, S., Wright, R. A., Lalu, M. M., Patlewicz, G., Becker, R. A., DeGeorge, G. L., Fergusson, D., Hartung, T., Lewis, R. J., & Stephens, M. L. (2016). Guidance on assessing the methodological and reporting quality of toxicologically relevant studies: A scoping review. *Environment International*, *92–93*, 630–646. https://doi.org/10.1016/j.envint.2016.03.010

Sanderson, S., Tatt, I. D., & Higgins, J. P. (2007). Tools for assessing quality and susceptibility to bias in observational studies in epidemiology: A systematic review and annotated bibliography. *International Journal of Epidemiology*, *36*(3), 666–676. https://doi.org/10.1093/ije/dym018

Savitz, D. A. (2007). Guest editorial: Biomarkers of Perfluorinated chemicals and birth weight. *Environmental Health Perspectives*, *115*(11), A528–A529. https://doi.org/10.1289/ehp.10923

Savitz, D. A., Stein, C. R., Elston, B., Wellenius, G. A., Bartell, S. M., Shin, H. M., Vieira, V. M., & Fletcher, T. (2012). Relationship of perfluorooctanoic acid exposure to pregnancy outcome based on birth records in the mid-Ohio Valley. *Environmental Health Perspectives*, *120*(8), 1201–1207. https://doi.org/10.1289/ehp.1104752

Savitz, D. A., Wellenius, G. A., & Trikalinos, T. A. (2019). The problem with mechanistic risk of bias assessments in evidence synthesis of observational studies and a practical alternative: Assessing the impact of specific sources of potential bias. *American Journal of Epidemiology*, *188, 9*, 1581–1585. https://doi.org/10.1093/aje/kwz131

Schulz, K. F., Chalmers, I., Hayes, R. J., & Altman, D. G. (1995). Empirical evidence of bias-dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *Journal of the American Medical Association*, *273*(5), 408–412. https://doi.org/10.1001/jama.273.5.408

Sera, F., Armstrong, B., Blangiardo, M., & Gasparrini, A. (2019). An extended mixed effects framework for meta-analysis. *Statistics in Medicine*, *38*, 5429–5444. https://doi.org/10.1002/sim.8362

Shamliyan, T., Kane, R. L., & Dickinson, S. (2010). A systematic review of tools used to assess the quality of observational studies that examine incidence or prevalence and risk factors for diseases. *Journal of Clinical Epidemiology*, *63*(10), 1061–1070. https://doi.org/10.1016/j.jclinepi.2010.04.014

Shea, B. J., Grimshaw, J. M., Wells, G. A., Boers, M., Andersson, N., Hamel, C., Porter, A. C., Tugwell, P., Moher, D., & Bouter, L. M. (2007). Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews. *BMC Medical Research Methodology*, *7*, 10. https://doi.org/10.1186/1471-2288-7-10

Sheppard, L., McGrew, S., & Fenske, R. A. (2020). Flawed analysis of an intentional human dosing study and its impact on chlorpyrifos risk assessments. *Environment International*, *143*, 105905. https://doi.org/10.1016/j.envint.2020.105905

Shin, H. M., Vieira, V. M., Ryan, P. B., Steenland, K., & Bartell, S. M. (2011). Retrospective exposure estimation and predicted versus observed serum perfluorooctanoic acid concentrations for participants in the C8 health project. *Environmental Health Perspectives*, *119*(12), 1760–1765. https://doi.org/10.1289/ehp.1103729

Sigurjónsdóttir, H. A., Franzson, L., Manhem, K., Ragnarsson, J., Sigurdsson, G., & Wallerstedt, S. (2001). Liquorice-induced rise in blood pressure: A linear dose-response relationship. *Journal of Human Hypertension*, *15*(8), 549–552. https://doi.org/10.1038/sj.jhh.1001215

Snedeker, S. M., & Hay, A. G. (2012). Do interactions between gut ecology and environmental chemicals contribute to obesity and diabetes? *Environmental Health Perspectives*, *120*(3), 332–339. https://doi.org/10.1289/ehp.1104204

Sommar, J. N., Pettersson-Kymmer, U., Lundh, T., Svensson, O., Hallmans, G., & Bergdahl, I. A. (2013). Hip fracture risk and cadmium in erythrocytes: A nested case–control study with prospectively collected samples. *Calcified Tissue International*, *94*, 183–190. https://doi.org/10.1007/s00223-013-9796-5

Steckler, A., & McLeroy, K. R. (2007). The importance of external validity. *American Journal of Public Health*, *98*(1), 9–10. https://doi.org/10.2105/ajph.2007.126847

Steenland, K., Barry, V., & Savitz, D. (2018). Serum Perfluorooctanoic acid and birthweight: An updated meta-analysis with bias analysis. *Epidemiology*, *29*(6), 765–776. https://doi.org/10.1097/EDE.0000000000000903

Steenland, K., & Deddens, J. A. (2004). A practical guide to dose-response analyses and risk assessment in occupational epidemiology. *Epidemiology*, *15*(1), 63–70. https://doi.org/10.1097/01.ede.0000100287.45004.e7

Steenland, K., Schubauer-Berigan, M. K., Vermeulen, R., Lunn, R. M., Straif, K., Zahm, S., Stewart, P., Arroyave, W. D., Mehta, S. S., & 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*, *128*, 9. https://doi.org/10.1289/EHP6980

Sterne, J. A. C., Hernán, M. A., Reeves, B. C., Savović, J., Berkman, N. D., Viswanathan, M., Henry, D., Altman, D. G., Ansari, M. T., Boutron, I., Carpenter, J. R., Chan, A. W., Churchill, R., Deeks, J. J., Hróbjartsson, A., Kirkham, J., Jüni, P., Loke, Y. K., Pigott, T. D., … Higgins, J. P. T. (2016). ROBINS-I: A tool for assessing risk of bias in non-randomized studies of interventions. *British Medical Journal*, *355*, i4919. https://doi.org/10.1136/bmj.i4919

Tang, J. L., Armitage, J. M., Lancaster, T., Silagy, C. A., Fowler, G. H., & Neil, H. A. (1998). Systematic review of dietary intervention trials to lower blood total cholesterol in free-living subjects. *British Medical Journal*, *316*(7139), 1213–1220. https://doi.org/10.1136/bmj.316.7139.1213

Tennant, P. W. G., Murray, E. J., Arnold, K. F., Berrie, L., Fox, M. P., Gadd, S. C., Harrison, W. J., Keeble, C., Ranker, L. R., Textor, J., Tomova, G. D., Gilthorpe, M. S., & Ellison, G. T. H. (2021). Use of directed acyclic graphs (DAGs) to identify confounders in applied health research: Review and recommendations. *International Journal of Epidemiology*, *50*(2), 620–632. https://doi.org/10.1093/ije/dyaa213

The Alpha-Tocopherol Beta Carotene Cancer Prevention Study Group. (1994). The effect of vitamin E and beta carotene on the incidence of lung cancer and other cancers in male smokers. *New England Journal of Medicine*, *330*(15), 1029–1035. https://doi.org/10.1056/NEJM199404143301501

US EPA. (2013). Materials Submitted to the National Research Council Part I: Status of Implementation of Recommendations Integrated Risk Information System Program. Submitted to National Research Council, p. 142.

Vandenbroucke, J. P., Broadbent, A., & Pearce, N. (2016). Causality and causal inference in epidemiology: The need for a pluralistic approach. *International Journal of Epidemiology*, *45*(6), 1776–1786. https://doi.org/10.1093/ije/dyv341

Verner, M. A., Loccisano, A. E., Morken, N. H., Yoon, M., Wu, H., McDougall, R., Maisonet, M., Marcus, M., Kishi, R., Miyashita, C., Chen, M. H., Hsieh, W. S., Andersen, M. E., Clewell, H. J., 3rd, & Longnecker, M. P. (2015). Associations of Perfluoroalkyl substances (PFAS) with lower birth weight: An evaluation of potential confounding by glomerular filtration rate using a physiologically based pharmacokinetic model (PBPK). *Environmental Health Perspectives*, *123*(12), 1317–1324. https://doi.org/10.1289/ehp.1408837

Vinceti, M., Filippini, T., Crippa, A., de Sesmaison, A., Wise, L. A., & Orsini, N. (2016). Meta-analysis of potassium intake and the risk of stroke. *Journal of the American Heart Association*, *5*, e004210. https://doi.org/10.1161/JAHA.116.004210

Vinceti, M., Filippini, T., Malavolti, M., Naska, A., Kasdagli, M. I., Torres, D., Lopes, C., Carvalho, C., Moreira, P., & Orsini, N. (2020). *Dose-response relationships in health risk assessment of nutritional and toxicological factors in foods: Development and application of novel biostatistical methods* (p. EN-1899). EFSA supporting publication. https://doi.org/10.2903/sp.efsa.2020.EN-1899

Viswanathan, M., Ansari, M., Berkman, N. D., Chang, S., Hartling, L., McPheeters, L. M., Santaguida, P. L., Shamliyan, T., Singh, K., Tsertsvadze, A., & Treadwell, J. R. (2012). Assessing the risk of bias of individual studies when comparing medical interventions. Agency for Healthcare Research and Quality, Methods Guide for Comparative Effectiveness Reviews. AHRQ Publication No. 12-EHC047-EF. www.effectivehealthcare.ahrq.gov/

Viswanathan, M., Berkman, N. D., Dryden, D. M., & Hartling, L. (2013). Assessing risk of bias and confounding in observational studies of interventions or exposures: Further development of the RTI item Bank. Agency for Healthcare Research and Quality, Methods for Effective Health Care. AHRQ Report No. 13-EHC106-EF. www.effectivehealthcare.ahrq.gov/

Vlaanderen, J., Lan, Q., Kromhout, H., Rothman, N., & Vermeulen, R. (2011). Occupational benzene exposure and the risk of lymphoma subtypes: A meta-analysis of cohort studies incorporating three study quality dimensions. *Environmental Health Perspectives*, *119*(2), 159–167. https://doi.org/10.1289/ehp.1002318

Völkel, W., Colnot, T., Csanády, G. A., Filser, J. G., & Dekant, W. (2002). Metabolism and kinetics of Bisphenol a in humans at low doses following Oral administration. *Chemical Research in Toxicology*, *15*(10), 1281–1287. https://doi.org/10.1021/tx025548t

Wacholder, S., McLaughlin, J. K., Silverman, D. T., & Mandel, J. S. (1992a). Selection of controls in case-control studies. I. Principles. *American Journal of Epidemiology*, *135*(9), 1019–1028. https://doi.org/10.1093/oxfordjournals.aje.a116396

Wacholder, S., Silverman, D. T., McLaughlin, J. K., & Mandel, J. S. (1992b). Selection of controls in case-control studies. II. Types of controls. *American Journal of Epidemiology*, *135*(9), 1029–1041. https://doi.org/10.1093/oxfordjournals.aje.a116397

Wacholder, S., Silverman, D. T., McLaughlin, J. K., & Mandel, J. S. (1992c). Selection of controls in case-control studies. III. Design options. *American Journal of Epidemiology*, *135*(9), 1042–1050. https://doi.org/10.1093/oxfordjournals.aje.a116398

Wang, Z., Taylor, K., Allman-Farinelli, M., Armstrong, B., Askie, L., Ghersi, D., McKenzie, J., Norris, S., Page, M., Rooney, A., Woodruff, T., & Bero, L. (2019). A systematic review: Tools for assessing methodological quality of human observational studies. NHMRC. https://nhmrc.gov.au/guidelinesforguidelines/develop/assessing-risk-bias

Warrington, S., Lee, C., Otabe, A., Narita, T., Polnjak, O., Pirags, V., & Krievins, D. (2011). Acute and multiple-dose studies to determine the safety, tolerability, and pharmacokinetic profile of Advantame in healthy volunteers. *Food Chemistry and Toxicology*, *49*(Suppl. 1), S77–S83. https://doi.org/10.1016/j.fct.2011.06.043

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA statement on p-values: Context, process, and purpose. *The American Statistician*, *70*(2), 129–133. https://doi.org/10.1080/00031305.2016.1154108

Waters, D. D., Alderman, E. L., Hsia, J., Howard, B. V., Cobb, F. R., Rogers, W. J., Ouyang, P., Thompson, P., Tardif, J. C., Higginson, L., Bittner, V., Steffes, M., Gordon, D. J., Proschan, M., Younes, N., & Verter, J. I. (2002). Effects of hormone replacement therapy and antioxidant vitamin supplements on coronary atherosclerosis in postmenopausal women: A randomized controlled trial. *Journal of the American Medical Association*, *288*(19), 2432–2440. https://doi.org/10.1001/jama.288.19.2432

Whitney, M., & Ryan, L. (2013). Uncertainty due to low-dose extrapolation: Modified BMD methodology for epidemiological data. *Environmetrics*, *24*(5), 289–297. https://doi.org/10.1002/env.2217

WHO (World Health Organization). (2010). Environmental health criteria 240. Principles and methods for the risk assessment of Chemicals in Food. In *Chapter 5: Dose-response assessment and derivation of health-based guidance values* (p. 34). WHO. https://www.who.int/publications/i/item/principles-and-methods-for-the-risk-assessment-of-chemicals-in-food

Willett, W. (2012). *Nutritional epidemiology* (3rd ed., p. 2012). Monographs in Epidemiology and Biostatistics.

Wolter, M., Grant, E. T., Boudaud, M., Steinle, A., Pereira, G. V., Martens, E. C., & Desai, M. S. (2021). Leveraging diet to engineer the gut microbiome. *National Reviews Gastroenterology and Hepatology*, *18*, 885–902. https://doi.org/10.1038/s41575-021-00512-7

Woodruff, T. J., & Sutton, P. (2014). The navigation guide systematic review methodology: A rigorous and transparent method for translating environmental health science into better health outcomes. *Environmental Health Perspectives*, *122*(10), 1007–1014. https://doi.org/10.1289/ehp.1307175

World Medical Association. (2013). Declaration of Helsinki: Ethical principles for medical research involving human participants. *Journal of the American Medical Association*, *310*(20), 2191–2194. https://doi.org/10.1001/jama.2013.281053

Yang, Q., Chang, A., Ritchey, M. D., & Loustalot, F. (2017). Antihypertensive medication adherence and risk of cardiovascular disease among older adults: A population-based cohort study. *Journal of the American Heart Association*, *6*, 6. https://doi.org/10.1161/JAHA.117.006056

Yland, J. J., Wesselink, A. K., Lash, T. L., & Fox, M. P. (2022). Misconceptions about the direction of bias from nondifferential misclassification. *American Journal of Epidemiology*, *191*(8), 1485–1495. https://doi.org/10.1093/aje/kwac035 Erratum in: American Journal of Epidemiology, 2022, 19;191(12):2123.

How to cite this article: EFSA Scientific Committee, More, S., Bampidis, V., Benford, D., Bragard, C., Hernandez-Jerez, A., Bennekou, S. H., Koutsoumanis, K., Lambré, C., Machera, K., Mennes, W., Mullins, E., Nielsen, S. S., Schlatter, J., Schrenk, D., Turck, D., Younes, M., Fletcher, T., Greiner, M., … Halldorsson, T. I. (2024). Scientific Committee guidance on appraising and integrating evidence from epidemiological studies for use in EFSA's scientific assessments. *EFSA Journal*, *22*(7), e8866. https://doi.org/10.2903/j.efsa.2024.8866

ANNEX A

Report of the Public consultation on the draft guidance of EFSA's Scientific Committee on appraising and integrating evidence from epidemiological studies for use in EFSA's scientific assessments

Annex A can be found in the online version of this output (in the 'Supporting information' section): https://doi.org/10.2903/j.efsa.2024.8866

APPENDIX A

Bradford Hill viewpoints

In order to assess the strength of evidence that exposure "A" can cause a health outcome, Austin Bradford Hill proposed a systematic review based on considering the following nine viewpoints as he described them. Some are specific to assessing an individual epidemiological paper, but most are directed at synthesizing evidence across different types of study. Some are simple yes/no assessments, such as 'does the exposure precede the health outcome', but most viewpoints make reference to characteristics that could range in their related evidence from weak to strong, for example the strength of association. They were not intended as a checklist leading to indisputable evidence of causality if all are ticked, but rather as guidelines that could be used when weighing the entire set of evidence. Here the viewpoints are explained in terms of easily understood questions. For further details, the Guidance on the use of the weight of evidence approach in scientific assessments (EFSA Scientific Committee, 2017) should be consulted.

Strength of the association	Does exposure A increase the rate of the health outcome? To what extent is the rate of the health outcome increased following exposure A, with an increased risk of the health outcome in the exposed relative to the unexposed?
Consistency	Has the association between the health outcome and exposure A been repeatedly observed across multiple independent studies, particularly those conducted with different designs, in different populations, under different circumstances?
Specificity	Is the association with exposure A limited to one health outcome and not to numerous other the health outcomes?
Temporality	Does exposure A precede onset of the health outcome?
Biological gradient	Does the association between exposure A and the health outcome follow a biological gradient, i.e., is there a dose-response relationship of increasing risk of the health outcome with increasing exposure? Are increased effects associated with greater exposures or duration of exposures?
Plausibility	Is the association between exposure A and the health outcome biologically plausible considering current understanding of biological mechanisms?
Coherence	Is the association between exposure A and the health outcome in line with the generally known facts about the natural history and biology of the health outcome?
Experiment	Has the frequency of this health outcome changed due to either starting or removing exposure A?
Analogy	Does exposure to chemically or biologically similar hazards lead to the same or comparable health outcome?

Adapted from: Hill (1965).

APPENDIX B

Hypothesis testing vs estimation

In statistical hypothesis testing, a test statistic (e.g. risk difference (RD)) is estimated from observed data and, based on assumptions and the given sample size, used to compute a p-value on which the decision is based whether the null hypothesis of 'no effect' (e.g. RD being 0) can be rejected or not. The null hypothesis of 'no effect' is appropriate when the research interest is to actually demonstrate the effect. In contrast, the null hypothesis is 'non-equality' (i.e. there is an effect) when evidence of absence of an effect is the study objective. This is the so-called equivalence setting. The following descriptions apply in both settings with the respective adaptation of the null hypothesis.

The properties of the statistical test are described using the probability of type I error (α), i.e. drawing a false positive conclusion and rejecting the null hypothesis (concluding there is an effect) when in fact the null hypothesis is true (observed effect size is due to chance alone) and type II error (β), i.e. drawing a false negative conclusion and not rejecting the null hypothesis (concluding there is no effect) when in fact the null hypothesis is not true (study fails to demonstrate that there is a true effect). Type I and type II error are inversely related, i.e. increasing one will decrease the other and vice versa. The probability of type I error (α) is also referred to as 'significance level' or 'threshold of significance' and conventionally set to a level of 5%. The quantity ($1-\beta$) is also referred to as the power of the study which is often set to a level of 80%. Although these levels for statistical significance and power are frequently encountered, other choices can be made to control type I and II errors. The levels for significance and power have to be set during the design phase of the study. The required sample size for the study is obtained as a function of significance, power and anticipated or biologically relevant effect size along with other input quantities that may be required for the given test statistic. The power of the study (for a given effect size) can be increased by increasing the sample size. Sub-optimal study designs may lead to 'under-powered' or 'over-powered' studies. An under-powered study is too small to demonstrate the anticipated effect and can be seen as a waste of resources, which can only be remedied using meta-analysis. Likewise, an over-powered, too big study can be criticised for wasting excess resources that are not required to demonstrate a biologically relevant effect size. Over-powered studies are prone to misinterpretation (confusing statistical and biological relevance). Increased type I error rates should be anticipated when the interpretation of a biological relevant effect size is corrected based on the findings of an over-powered study.

A *p*-value of a hypothesis test is affected both by the size of the effect and the size of the sample studied. This means, that for the same effect, the analysis of the results of a study with a larger sample size would yield a lower *p*-value, compared to a study with smaller sample size. As a result, the *p*-value cannot be used as a measure of effect size and vice versa. The above-mentioned problem of under-powered or over-powered studies is another consequence of this characteristic. Sample size calculations based on the desired precision of the estimate preclude this problem at least for the primary analysis goals specified in the design phase. A better understanding and appreciation of study results can be reached when assessing separately the estimated size of the effect and the precision around that estimate, rather than performing the respective hypothesis test. In this way, the biological relevance of the effect can also be judged, while also the precision around that estimate (stemming from the sample size of the study and the statistical properties of the estimator) can be evaluated. This is possible using the point estimate of the statistic in question and its confidence interval obtained for a single sample or the comparison groups. As an example, let us consider two hypothetical epidemiological studies, the first of which yielding a 95% confidence interval (95% CI) for a RD of 0.01 to 0.03, while the second CI would be −0.01 to 0.70. In a hypothesis test using a significance threshold of 0.05, the first study would indicate that the RD was statistically significantly different than 0 (i.e. there was an effect) while the second would not. It needs to be noted, however, that the 95% CI for the first study shows a relatively small effect coming from a large (over-powered) study, while the second one shows a potentially large effect coming from a small (under-powered) study. Given that increasing the sample size of the second study would narrow the CI around the (central) point estimate, it would be expected that with a larger sample size the lower limit of the CI would come to exclude zero (which would be analogous to a statistically significant result for the null hypothesis that the RD equals zero). Three numbers (point estimate and limits of CI) on the scale of the statistic of interest (RD in our example) provide much more relevant information than the single *p*-value.

Indeed, emphasis on precision of effect measures with confidence intervals is gradually replacing the approach that focuses on hypothesis testing (Amrhein et al., 2019; Greenland et al., 2016; Rothman, 2016). As Altman and colleagues stated in 1983 (Altman et al., 1983) 'The confidence interval conveys more information because it indicates the lowest and highest true effect likely to be compatible with the sample observations'. Among the recommendations of the scientific opinion on Statistical Significance and Biological Relevance (EFSA Scientific Committee, 2011), there is a plea for not using hypothesis testing as the sole tool for decision making and the level of statistical significance as the main driver to derive conclusions; also, less emphasis should be placed on the reporting of statistical significance, and more on statistical point estimation and associated confidence intervals. If the calculation of both point and interval estimates as well as a *p*-value is possible, they should all be reported. It is also noted that meta-analysis uses the estimation rather than the testing framework. A combination of individual studies based on tallying 'significant' and 'non-significant' results without accounting for individual effect sizes and precisions would be a serious methodological flaw.

APPENDIX C

Random error and statistical precision

As mentioned in Section 4.2.1.1, inferences about measures of effect or association in epidemiological studies can be based either on statistical hypothesis testing or statistical estimation. Random error considerations play an important role in both these approaches.

Random error will occur in any epidemiological study. An example by Pearce (2005) is given to illustrate this: 'Suppose that 50 lung cancer deaths occurred among 10,000 people aged 35–39 exposed to a particular factor during one year. Then, if each person had exactly the same cumulative exposure, we might expect two subgroups of 5,000 people each to experience 25 deaths during the one-year period. However, just as 50 tosses of a coin will not usually produce exactly 25 heads and 25 tails, neither will there be exactly 25 deaths in each group'. Even in an experimental study, that randomises participants into 'exposed' and 'non-exposed' groups, there will be 'random' differences in background risk (before the intervention is assigned) between the compared groups. These will diminish in importance as the study size grows (i.e. the random differences will tend to 'even out'). In observational epidemiological studies, because of the lack of randomisation, the baseline (background) risk may be different among the compared groups. Therefore, the compared groups cannot be considered comparable by default in terms of risk by all factors other than the exposure of interest. Estimating how much of the observed difference in the outcome is due to random error and how much due to a real systematic difference is important in the statistical analysis of the results of epidemiological studies. Increasing the study size will (on the average) reduce difference due to random error but will not reduce systematic differences.

Random error affects both statistical hypothesis testing and interval estimation. In the former case, it explicitly enters the calculation of the value of the test statistic and, therefore, of the p-value, while in the latter case, it affects the width of the Confidence Interval (CI) (at a given confidence level). This width represents the precision of the estimation, showing thus how much information we have about the specific parameter. For example, a 95% CI for a risk ratio from 2.1 to 2.2 would show much higher precision than an interval from 2.1 to 9. Since the CI indicates with 95% confidence what values the real (unobserved) population parameter may have, it is obvious that the former interval is much more informative than the latter.

APPENDIX D

Relevant inventories and reviews of critical appraisal tools

The following table reports on a selection of recent inventories and reviews on critical appraisal tools. It includes a description of their context, the study designs, the specific objectives of the review and some of the authors' conclusions as deemed relevant. The purpose of this table is to give an overview of other available tools than those mentioned in the Guidance document, including relevant frameworks and guidance on evaluating study quality.

Publication	Context	Study designs covered	Purpose of the review	Further details	Authors' conclusions
Quigley et al. (2019)	Health Technology Assessment (HTA)	Randomised controlled trials, Non-randomised intervention studies (NRIS)	• **Identify tools** commonly used to assess bias in NRIS • Determine those **recommended** by HTA bodies	48 critical appraisal tools identified, among them those from Cochrane (**RoB, ROBINS-I**), the Centre for Reviews and Dissemination (**CRD**), and the Scottish Intercollegiate Guidelines Network (**SIGN**)	There is no consensus between HTA groups on the preferred appraisal tool. Reviewers should select from a suite of tools based on the design of studies included in their review
Wang et al. (2019)	Public and Environmental Health	Human observational designs (cohort, case–control, cross-sectional)	• Identify, describe, categorise **key elements for appraisal** into domains • Develop **guidance on selecting risk of bias tools** for public health decision makers	62 tools identified (with 17 categories of similar or overlapping items), full list available here (https://ntp.niehs.nih.gov/go/ohat_tools)	Need for a common tool for assessing risk of bias in human observational studies of exposures. Absent that common tool, a selection should be based on the following: (1) the tool should have clear definitions for each item and be transparent regarding the empirical or theoretical basis for each domain, (2) tools should include questions addressing 9 domains: Selection, Exposure, Outcome assessment, Confounding, Loss to follow-up, Analysis, Selective reporting, Conflicts of interest and Other, (3) the ratings for each domain should be reported, rather than an overall score, (4) the tool should be rigorously and independently tested for usability and reliability

(Continued)

Publication	Context	Study designs covered	Purpose of the review	Further details	Authors' conclusions
Lynch et al. (2016)	Chemical risk assessment	Human observational designs (cohort, case–control, cross-sectional), in vivo studies, in vitro studies, Systematic reviews	• Critically evaluate several **available frameworks** for evaluating study quality • Assess the criteria separately for human, animal, and in vitro studies as well as for systematic reviews and **evaluate commonalities** across disciplines.	10 systems for evaluating the quality of studies or systematic reviews were assessed: **the Klimisch system** (Klimisch et al., 1997); the **OECD Guidance Document (GD) 34** (OECD, 2005); the **Toxicological Data Reliability Assessment Tool** (ToxRTool) (European Commission, Undated); the approaches that have been used in recent **IRIS** systematic review documents (US EPA, 2013); the framework being developed by **NTP's OHAT** (NTP, 2013, 2015); Animal Research: Reporting of In Vivo Experiments (**ARRIVE**) guidelines for animal research (Kilkenny et al., 2010); the **Navigation Guide** for systematic reviews (Koustas et al., 2013, 2014; Woodruff & Sutton, 2014; Johnson et al., 2014; Lam et al., 2014); the 'assessment of multiple systematic reviews' (**AMSTAR**) system (Shea et al., 2007); the 'strengthening the reporting of observational studies in epidemiology' (**STROBE**) system (von Elm et al., 2007); the **Systematic Approach for Scoring Human Data**, as developed by Money et al. (2013)	Although study quality evaluation systems vary in the specifics of good study design, conduct, and reporting that are examined, there are several elements that are nearly universally named as essential to recognising study results as robust and reliable. For **human studies** (especially observational epidemiological ones): aspects of care in identifying and choosing study populations, investigating and adequately addressing potential confounding factors, and avoiding selectively reporting results. For **animal studies:** careful and documented control of animal provenance, environmental conditions, food and water (focus on aspects that might introduce variations in outcomes that are not attributable to the test agent); use of appropriate control groups, the randomisation of animals among treatments, documentation of procedures and thorough reporting; blinding of endpoint evaluators to the dosing status of individual animals
Samuel et al. (2016)	Toxicology	Human observational designs (cohort, case–control, cross-sectional), in vivo studies, in vitro studies, QSAR, physico-chemical properties studies	• **Scoping the available guidance to assess the methodological or reporting quality** of studies relevant to toxicology • Distill the **common elements** of these documents for each of the four study types	23 guidance documents for in vitro and in vivo studies included, 7 addressing methodological quality, 2 addressing reporting quality and 14 addressing both methodological and reporting quality 3 guidance documents for QSAR studies included 3 guidance documents included for studies of physico-chemical properties 12 publications in total identified as providing guidance on the assessment of human studies, including 10 on methodological quality, 1 on reporting quality and 1 on mixed guidance	There is considerable overlap in the proposed criteria by study type, despite some difference across guidance documents. This is reassuring, as quality appraisals should ideally be based on consensus criteria in order to facilitate broad understanding, buy-in and comparison across assessments, as well as to facilitate the conduct of the appraisals themselves. The results also illustrate that the proposed criteria differ somewhat across study types, suggesting that appraisal tools may need to be tailored to particular study types

(Continues)

(Continued)

Publication	Context	Study designs covered	Purpose of the review	Further details	Authors' conclusions
Sanderson et al. (2007)	Systematic reviews	Observational designs (cohort, case–control, cross-sectional – OBS), systematic reviews	• Provide an annotated **bibliography of tools** specifically designed to assess quality or susceptibility to bias in OBS epidemiological studies • Identify whether there is an **existing tool that could be recommended** for widespread use	86 tools reviewed, comprising 41 simple checklists, 12 checklists with additional summary judgements and 33 scales	Tools should be rigorously developed, evidence-based, valid, reliable and easy to use Tool components should, where possible, be based on empirical evidence of bias, although this may be difficult to obtain, and there is a need for more empirical research on relationships between specific quality items and findings from epidemiological studies Most tools included items to assess methods for selecting study participants (92%) and to assess methods for measuring study variable and design-specific sources of bias (both 86%). Over three-quarters of tools assessed the appropriate use of statistics, and the control of confounding (both 78%) but conflict of interest was only included in 4% of tools
Deeks et al. (2003)	Health Technology Assessment	Non-randomised intervention studies (NRIS)	• Review **empirical evidence of bias** associated with NRIS • Review the **content** of quality assessment tools for non-randomised studies • Review the **use** of quality assessment in systematic reviews of non-randomised studies	182 tools identified; 60 of them were selected as 'top' tools, covering at least five of six internal validity domains as characterised in the review. Of these, 14 met the criteria for 'best tools', covering at least three of four core items	Although many quality assessment tools exist and have been used for appraising non-randomised studies, most omit key quality domains. Six tools were considered potentially suitable for use in systematic reviews, but each requires revision to cover all relevant quality domains

APPENDIX E

Overview of appraisal tools

The tools for the appraisal of primary studies are presented stratified by study question, study population, and study design, allowing the readers to identify the tools available for their specific appraisal task. Where no specific tool is available,[50] tools from other contexts can be adapted to cover the needs of a specific assessment (e.g. if no tool is available for observational studies on livestock, one could adapt NTP-OHAT to the animal population). The figures reflect the tools available in 2019; in the meantime, considerable developments have occurred (e.g. the Robins-E tool[51]).

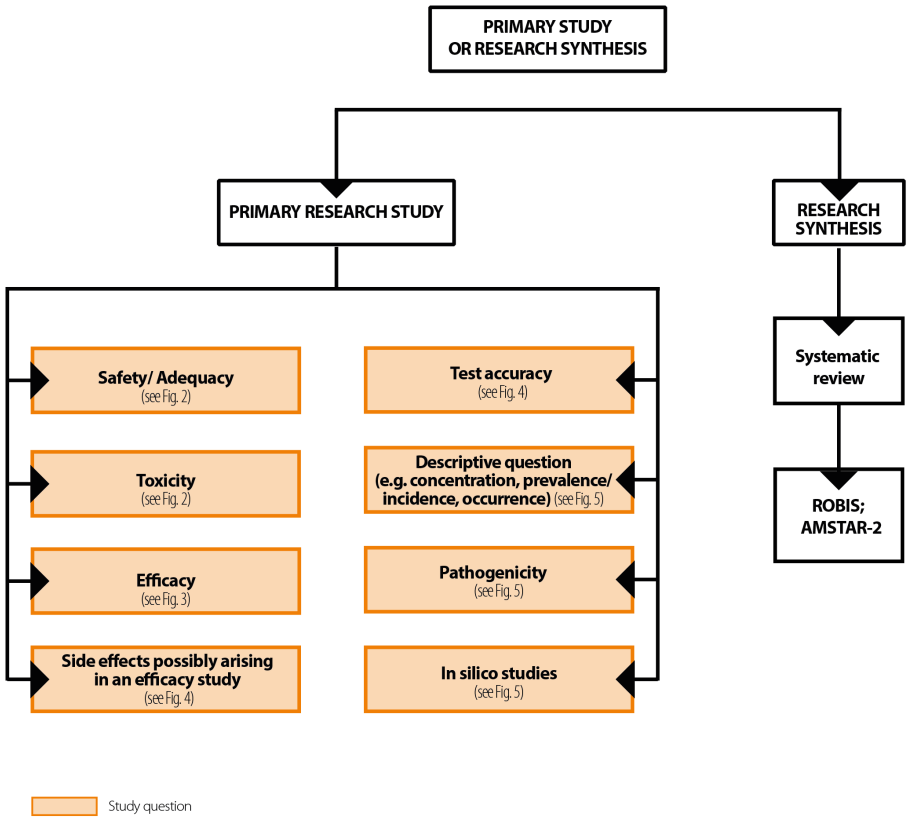

FIGURE E.1 Risk of bias (RoB) tools for appraisal of research syntheses or primary research studies (the RoB tools for these are shown in Figures E.1–E.5, G.1).[52]

[50]'Not available' in Figure E.1 intends 'not yet available'; note that no comprehensive review has been carried out.

[51]https://www.riskofbias.info/welcome/robins-e-tool.

[52]AMSTAR-2, ROBIS.

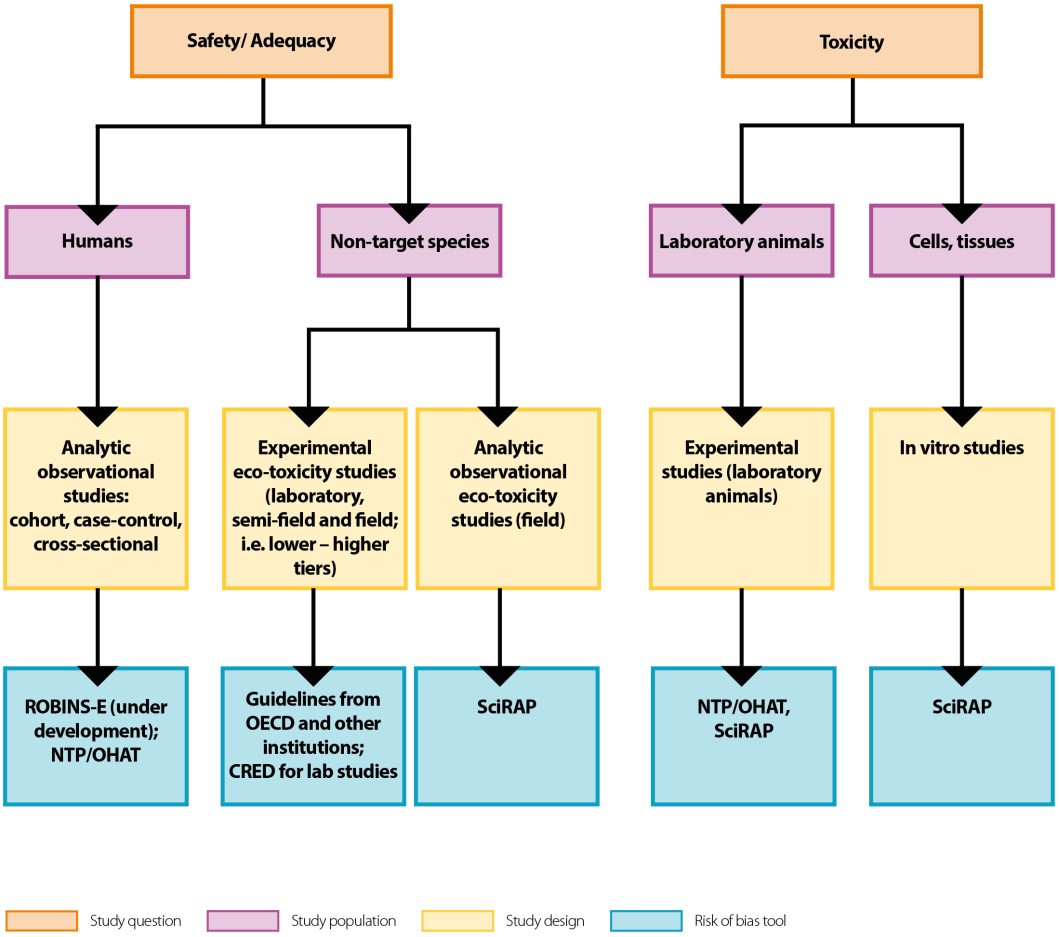

FIGURE E.2 Risk of bias tools for appraisal of primary research studies assessing safety/adequacy or toxicity (the term 'toxicity 'is intended to harm/induce toxicity).[53]

[53]ROBINS-E, CRED for lab studies, SciRAP, NTP-OHAT.

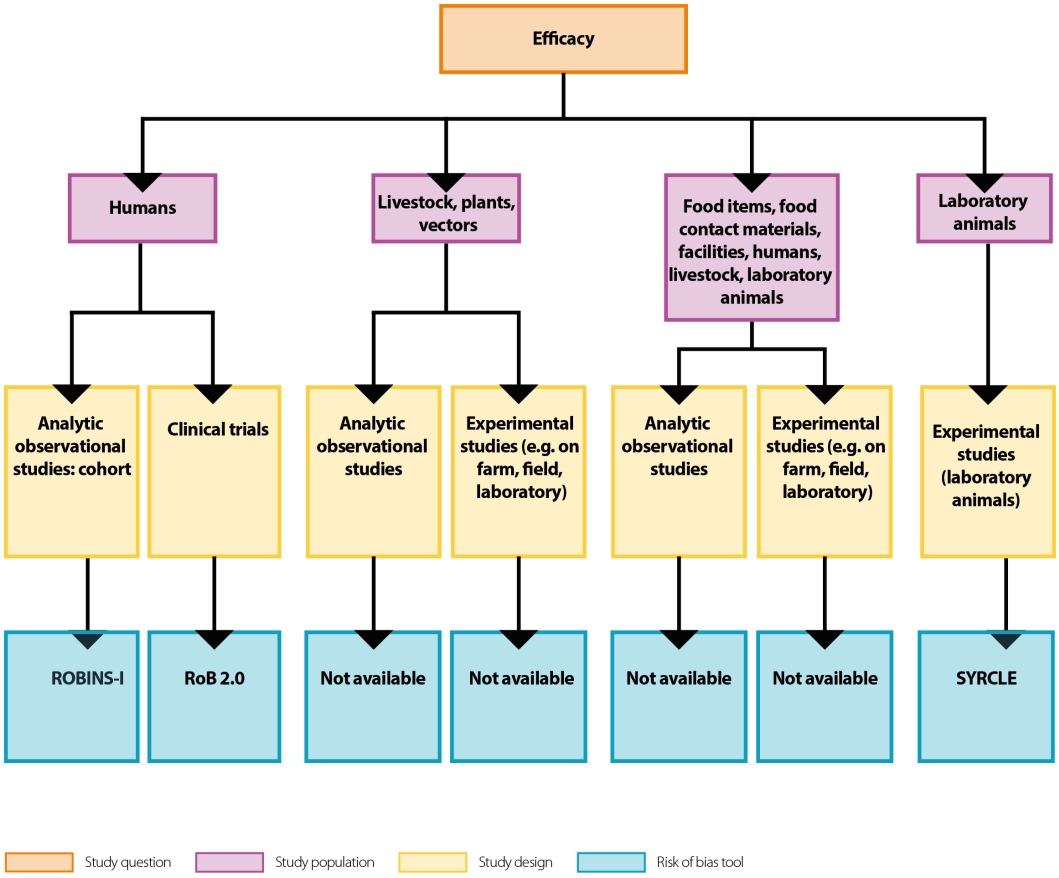

FIGURE E.3 Risk of bias tools for appraisal of primary research studies assessing efficacy.[54]

[54] ROBINS-I, RoB 2.0, Syrcle.

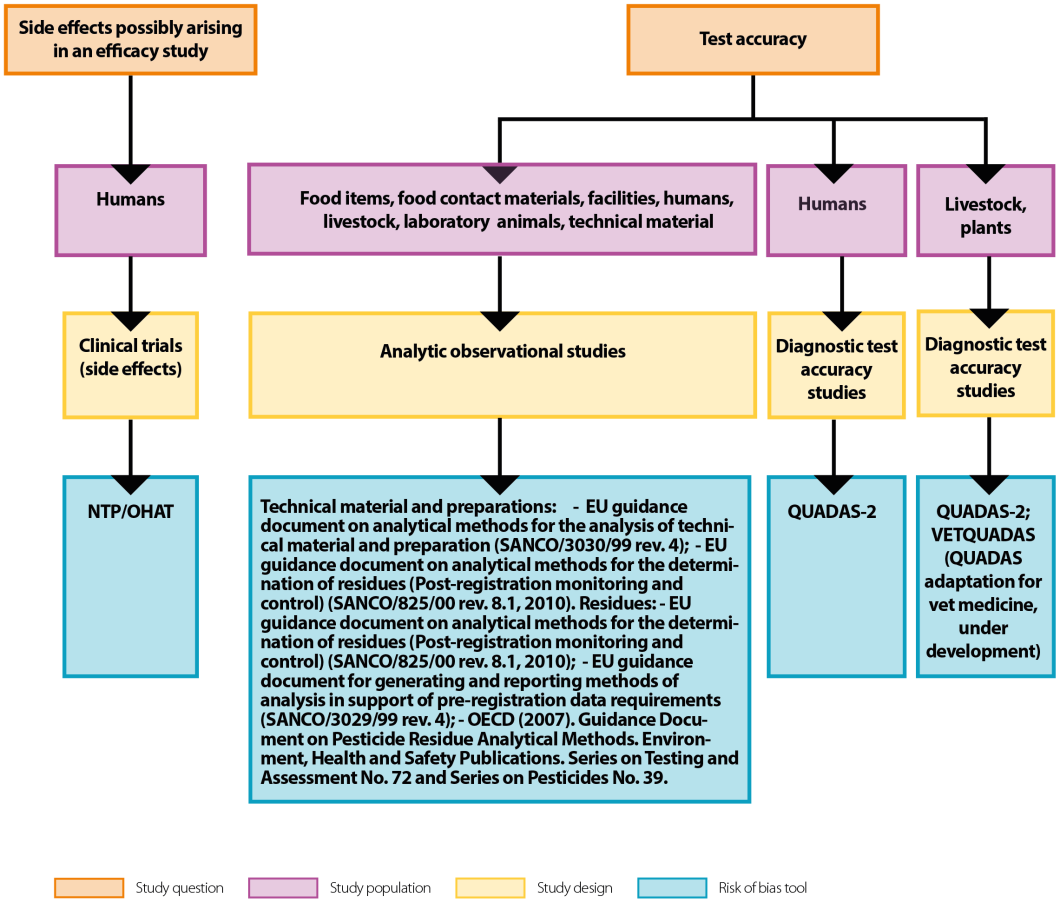

FIGURE E.4 Risk of bias tools for appraisal of primary research studies assessing side effects possibly arising in efficacy studies or test accuracy.[55]

[55] NTP-OHAT, QUADAS-2, VETQUADAS.

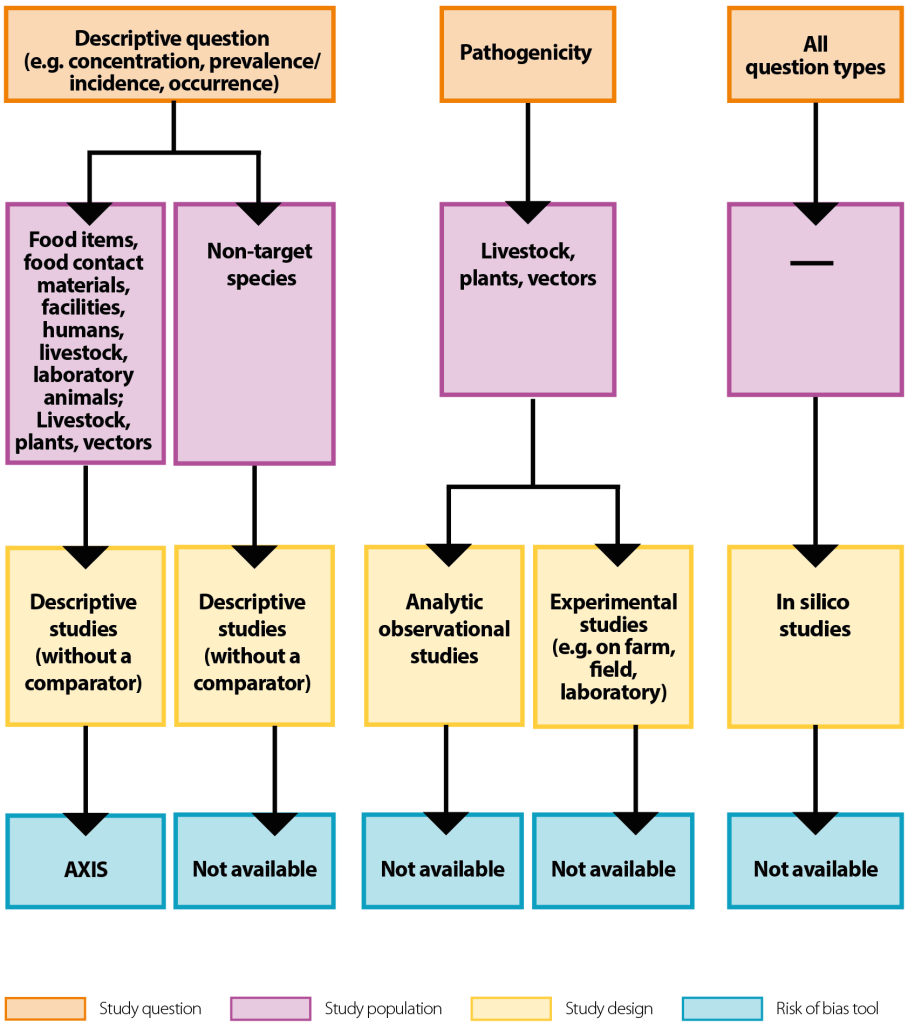

FIGURE E.5 Risk of bias tools for appraisal of primary research studies assessing descriptive questions, pathogenicity or in silico studies.

APPENDIX F

Appraisal of different studies using a risk of bias tool

Background

The selected risk of bias (RoB) tool used for all examples in this appendix is the NTP-OHAT Risk of Bias Rating Tool for Human and Animal Studies. The reason for this selection is that this tool has been used in many recent EFSA opinions.

The NTP-OHAT tool is quite comprehensive and complex. Below we use the instructions as given in the NTP-OHAT manual and, for illustrative purposes, have copied the original text, were useful. Following the manual in more detail, the specific questions and instructions for each RoB question should be tailored prior to conducting an assessment. This is done for each study design(s) and by reflecting on how individual questions should be evaluated. The use of the general instructions from the NTP-OHAT manual in the examples shown here should illustrate the importance of tailoring. For more detailed clarifications on the use of the NTP-OHAT RoB tool, we refer to the manual[56] (National Toxicology Program, 2019).

Using the NTP-OHAT RoB tool

This tool contains 11 risk-of-bias questions (or 'domains') that cover six types of bias (selection, confounding, performance, attrition/exclusion, detection and selective reporting bias). The six bias types used in the NTP-OHAT RoB tool reflect a finer categorisation of biases that fall under either selection bias, information bias and confounding as defined in Section 4.2. The correspondence between them and the three bias definitions provided in Section 4.2 is given in Table F.5.

For each of the 11 bias-questions, the RoB present in the study is then rated by selecting among four possible answer formats:

Answer Format:

++ ***Definitely Low*** *risk of bias*:

There is direct evidence of low risk-of-bias practices

(May include specific examples of relevant low risk-of-bias practices)

\+ ***Probably Low*** *risk of bias*:

There is indirect evidence of low risk-of-bias practices **OR** it is deemed that deviations from low risk-of-bias practices for these criteria during the study would not appreciably bias results, including consideration of direction and magnitude of bias

\- NR ***Probably High*** *risk of bias*:

There is indirect evidence of high risk-of-bias practices **OR** there is insufficient information (e.g. not reported or 'NR') provided about relevant risk-of-bias practices

-- ***Definitely High*** *risk of bias*:

There is direct evidence of high risk-of-bias practices (May include specific examples of relevant high risk-of-bias practices)

Source: NTP-OHAT.

By weighing the ratings of the individual bias questions, an overall RoB tier for each study is identified. How such weighing is performed should be decided on a priori. In the NTP/OHAT approach, this is usually done by identifying certain questions that reflect the most important biases for the specific design as key questions. Their ratings have more weight in the overall set of rules that combines the ratings from all questions to characterise (or rank) studies into tiers of reliability (this tool proposes three tiers: low, intermediate and high RoB).

There are no fixed sets of rules on how to weigh different questions when assessing the overall RoB. Any decision of weighing is always subject to expert judgement, but if done in a structured and transparent way it provides a framework for condensing the judgement assigned to individual bias questions. Such a summary not only can be helpful but it also has its pitfalls, as within each tier (low, medium and high RoB) the source and possible direction of the biases may differ between studies. This information may get lost when referring only to individual studies by their summary (or ranking) into tiers.

In the examples given below the weighing of the ratings for the different questions is not considered. The purpose of the example is first and foremost to give an example of the considerations needed when assessing the RoB for individual questions. Also, the rating scale (++, +, -, --) for each bias example is not always reported in its complete form to keep the text more concise.

When reading the example below, it is important to keep in mind that study appraisal is a question of judgement (this is not an automatic process) and that two different assessors might have reached different conclusions. The key issue is transparency in justifying and documenting the choices and discussions (among reviewers) in each appraisal made.

[56] https://ntp.niehs.nih.gov/ntp/ohat/pubs/handbookmarch2019_508.pdf.

Randomised Controlled Trials – risk of bias assessment

Example 1: Selenium and cancer, performance bias:

The RCT by Duffield-Lillico et al. (2003) is an example of a study whose findings may have been influenced by performance bias (in particular) and selection bias. This study was a randomised, double-blind, placebo-controlled trial conducted in 1312 participants recruited between 1983 and 1991, from seven dermatology clinics in low-selenium areas of the USA. This trial was originally designed to test the ability of selenium supplements to prevent non-melanoma skin cancer incidence.[57] In this study, a protective effect of selenium supplementation on prostate cancer was observed.

The 1312 participants were recruited in seven different centres and then randomised to receive placebo or selenium supplementation. The randomisation was 'blocked by time and stratified by clinic' as reported by the authors. The resulting baseline characteristics of study participants are shown in Table 4 (SS: selenium supplementation).

Table F.1 Baseline characteristics of study participants (© Duffield-Lillico et al., 2003).

TABLE F.1 The baseline characteristics of the men participating in the NPC, by treatment group.

Characteristic	SS	Placebo
Participants randomized, *n*	457	470
Mean (SD)		
Age, years	64.9 (8.8)	63.7 (9.4)*
BMI, kg/m^2	26.0 (3.6)	25.9 (3.7)
Smoking status, %		
Never	25	21
Former	47	46
Current	28	33
Plasma selenium, ng/mL		
Mean (SD)	115.1 (22.1)	115.1 (22.0)
33rd centile	106.8	106.0
50th centile	113.6	114.0
66th centile	123.2	122.4
PSA, %†		
Mean (SD), ng/mL	2.0 (3.4)	1.9 (3.3)
< 4	90.3	89.5
4–7	5.5	6.6
7.1–10	1.4	2.7
> 10	2.8	1.2

*Two-sample *t*-test, $p < 0.05$.

†PSA test within 6 months of randomization (694 men).

Question on selection bias:

Was administered dose or exposure level adequately randomised?

As can be seen in Table F.1, the mean age at baseline was significantly higher in the selenium supplemental group (~ 1.4 years), and there were some notable (non-significant) differences in never and current smokers. This may have occurred due to chance, but since prostate cancer is strongly dependent on age, and smoking is a risk factor, one could expect these differences to affect the results, in particular as the mean age difference is not small compared to the mean follow-up time of the study (~ 7 years). Higher mean age among those receiving placebo may well result in more cases being detected in the placebo group. Considering that the method of randomisation is poorly described (here it matters) there is quite clearly some RoB, and selecting either the '-' or '--' rating seems justifiable:

[57] This is an open access publication: https://pubmed.ncbi.nlm.nih.gov/12699469/.

'-' There is indirect evidence that subjects were allocated to study groups using a method with a non-random component, OR there is insufficient information provided about how subjects were allocated to study groups (record 'NR' as basis for answer).

'--' There is direct evidence that subjects were allocated to study groups using a non-random method including judgment of the clinician, preference of the participant, the results of a laboratory test or a series of tests, or availability of the intervention (Higgins & Green, 2011).

As there is no direct mentioning of the method for random allocation, and clear indications that the randomisation may have not been perfect, the appropriate judgement here would be '-'.

Question on performance bias:

Were the research personnel and human subjects blinded to the study group during the study?

Even though this study was a double-blind RCT, there are several indications provided in the manuscript that lead to the suspicion that blinding was not successful. First, the mean reported follow-up time in the selenium supplement group and placebo group was 7.3 and 7.6 years, respectively. This difference, although small, could reflect higher motivation in the selenium group to get screening. An alternative consequence is that shorter follow-up time in the selenium group (for whatever reasons) may lead to fever number of cases being diagnosed, if not corrected in the analysis. Second, and most importantly, the authors noted that

the follow-up, as per the current clinical standard for a man with an abnormal PSA level, differed significantly between treatment groups; 35% of men with an abnormal PSA in the placebo group underwent biopsy at some point throughout the trial, compared with only 14% in the selenium group ($p < 0.05$; Table 3). This observed difference in biopsy rates could not be accounted for by PSA concentration, age at which the abnormal PSA was detected, nor alternative diagnostic procedures including TURP or TRUS. This discrepancy suggests a potential bias against the detection of prostate cancer in the SS group.

Differences in rate of testing among men with abnormal PSA strongly suggest differential treatment, either by chance, but more likely because blinding of either participants or research personal could not be achieved. Lower rate of detection of prostate cancer in the selenium supplemental group would bias the effect estimate in the direction of showing a protective effect of the supplementation (which was the reported study finding). In terms of rating performance bias, here the two following options seem most logical:

'-' There is indirect evidence that it was possible for research personnel or subjects to infer the study group, OR there is insufficient information provided about blinding to study group during the study (record 'NR' as basis for answer).

'--' There is direct evidence for lack of adequate blinding of the study group including no blinding or incomplete blinding of research personnel and subjects. For some treatments, such as behavioural interventions, allocation to study groups cannot be concealed.

'The rating here would be '-', as the evidence is indirect based on study results and reported data.

Example 2: Dietary intervention in pregnant women. Performance bias and lack of blinding.

Dietary intervention studies are a good example of studies where performance bias due to lack of blinding may occur. That is, when dietary regimens are assigned as treatment and participant concealment is not made, this may result in either differential allocation to treatment by the investigator, or deviation from the assigned intervention by the participants. As a possible example of such a study we look at a dietary intervention study in pregnant women by Poston et al. (2015).[58]

In this study, the authors aimed to examine 'whether a complex intervention addressing diet and physical activity could reduce the incidence of gestational diabetes and large-for-gestational-age infants'. For that purpose, the 1555 women were randomly assigned to either standard antenatal ($n = 772$ of which 651 (84%) completed the follow-up) or behavioural intervention ($N = 783$ of which 629 (80%) completed the follow-up). The intervention as described by the authors 'was informed by control theory and elements of social cognitive theory, consisted of eight further health trainer-led group or individual sessions of 1 h duration once a week for 8 weeks'…. 'The intention of the intervention was to improve glucose tolerance through dietary and physical activity behaviour change'. In short, no effect of the intervention was observed for the primary outcome of this trial.

[58]This is an open access publication: https://pubmed.ncbi.nlm.nih.gov/26165396/.

Question on Performance Bias:

Were the research personnel and human subjects blinded to the study group during the study?

In terms of bias, it is clear from the description that participant blinding was not possible. Blinding of those assigning the intervention was also not possible but blinding of those performing the outcome assessment could have been (if the participants do not reveal their treatment status). If evaluated against double blind randomised controlled trials, the rating in terms of performance bias would have to be '--' according to the instructions:

> '--': *There is direct evidence for lack of adequate blinding of the study group including no blinding or incomplete blinding of research personnel and subjects. For some treatments, such as behavioural interventions, allocation to study groups cannot be concealed.*

Let us first reflect on the possible bias and its magnitude and direction that may occur due to lack of participant blinding. Participants are being encouraged to shift their dietary habits in a healthier direction in terms of carbohydrate and fat quality, and they are asked to increase their physical activity. Changing such habits on the investigators' request is generally difficult and one likely outcome in these types of studies is low or poor compliance. It is less likely that participants in the intervention group decide to follow a different diet which they were not assigned to. On the other hand, the controls who receive standard care are not assigned to any specific treatment, but they are aware that their health and lifestyle is being monitored. This may motivate them to act healthier than they would have done otherwise. Lack of compliance in the intervention group and possible changes in lifestyle habits in a healthier direction among controls would lead to smaller contrast in exposure than intended or to some confounding, biasing the effect estimate towards the NULL.

The exposure contrast in both dietary habits in this well-conducted trial was quite small in absolute terms, despite being significantly different (see Table F.2). For example, mean changes in dietary fibres are less than 1 g/day and the measured differences in glycaemic load and saturated fat are only a fraction of the between-subject variation (as measured by the standard deviations). This can be interpreted as an indication of performance bias (possibly due to poor compliance).

TABLE F.2 Maternal nutritional and physical activity outcomes, by period of gestation (© Poston et al., 2015).

	Standard care	Intervention	Mean difference (95% CI)	*p*
Nutrition				
Total energy (MJ/day)				
15–18 weeks +6 days	7.8 (2.6)	7.6 (2.5)		
27–28 weeks +6 days	7.5 (2.3)	6.8 (1.9)	−0.70 (−0.96 to −0.45)	< 0.0001
Glycaemic index (0–100)				
15–18 weeks +6 days	56.9 (4.1)	56.8 (3.9)		
27–28 weeks +6 days	57.0 (3.9)	54.3 (3.9)	−2.6 (−3.0 to −2.1)	< 0.0001
Glycaemic load per day				
15–18 weeks +6 days	141 (56)	135 (51)		
27–28 weeks +6 days	133 (47)	112 (38)	−21 (−26 to −16)	< 0.0001
Carbohydrate (% energy)				
15–18 weeks +6 days	49.4 (7.4)	49.0 (7.4)		
27–28 weeks +6 days	48.6 (6.6)	47.2 (7.2)	−1.4 (−2.2 to −0.58)	0.0011
Protein (% energy)				
15–18 weeks +6 days	19.7 (4.4)	20.1 (4.5)		
27–28 weeks +6 days	20.1 (4.0)	22.3 (4.6)	2.05 (1.5 to 2.5)	< 0.0001
Total fat (% energy)				
15–18 weeks +6 days	31.0 (5.5)	31.0 (5.3)		
27–28 weeks +6 days	31.5 (5.1)	30.5 (5.2)	−0.88 (−1.49 to −0.26)	0.0011
Saturated fat (g/day)				
15–18 weeks +6 days	26.5 (11.5)	25.4 (11.0)		
27–28 weeks +6 days	26.4 (10.9)	22.0 (8.3)	−4.3 (−5.4 to −3.1)	< 0.0001
Saturated fat (% energy)				
15–18 weeks +6 days	12.7 (3.0)	12.5 (2.9)		
27–28 weeks +6 days	13.1 (3.0)	12.1 (2.8)	−0.85 (−1.2 to −0.51)	< 0.0001

(Continues)

TABLE F.2 (Continued)

	Standard care	Intervention	Mean difference (95% CI)	*p*
Fibre (g/day)				
15–18 weeks +6 days	13.6 (6.0)	13.1 (5.3)		
27–28 weeks +6 days	12.6 (5.3)	13.4 (5.3)	0.83 (0.17 to 1.48)	0.013
Physical activity				
MET (min/week)				
15–18 weeks +6 days	1386 (660–3052)	1386 (594–2982)		
27–28 weeks +6 days	1386 (639–3363)	1836 (792–4158)	295 (105 to 485)	0.0015
Moderate or vigorous activity (min/week)				
15–18 weeks +6 days	0 (0–180)	0 (0–180)		
27–28 weeks +6 days	0 (0–240)	30 (0–240)	0 (–18 to 18)*	> 0.99
Walking (min/week)				
15–18 weeks +6 days	280 (140–600)	280 (140–540)		
27–28 weeks +6 days	300 (132–630)	420 (180–840)	77 (28 to 126)*	0.0018

Notes: Data are mean (SD) or median (IQR). Women with reported total energy ≤ 4.5 MJ/day or ≥ 20 MJ/day at 15–18 weeks +6 days of gestation were excluded from analyses of diet. Thus, in the standard care group, 571 women were assessed at 15–18 weeks +6 days of gestation and 511 were assessed at 27–28 weeks +6 days of gestation; corresponding figures in the intervention group were 574 and 435. Dietary intervention estimates were calculated by multiple regression and adjusted for pretrial values. For analyses of physical activity, in the standard care group, 678 women were included at 15–18 weeks +6 days of gestation and 588 were assessed at 27–28 weeks +6 days of gestation; in the intervention group, 683 and 559 women, respectively, were analysed. Physical activity estimates were calculated by bootstrapped (1000 replications) median regression, adjusting for pretrial values. MET is defined as the energy expenditure ratio of activity to rest; one MET is roughly equal to an individual's resting energy expenditure. MET, vigorous activity, moderate or vigorous activity, and walking were not prespecified endpoints.

Abbreviation: MET, metabolic equivalent of task.

*Median difference (95% CI).

Regarding blinding of the research personnel, the primary outcomes were gestational diabetes (GDM) and infants born large for gestational age. It is theoretically possible that lack of blinding may result in differential monitoring depending on intervention status, but as these outcomes were assessed as a part of the standard antenatal care, risk of bias appears low (but cannot be excluded for GDM which is assessed based on suspected symptoms and is not a standard measure applied to all). There is no statement on blinding of the outcome assessors in the manuscript.

Final Comment:

Despite being a high RoB study when assessed in relation to a double-blind RCT (as the gold standard), the study by Poston et al. (2015) could not have been performed to a higher standard when it comes to performance bias, suggesting that while some degree of bias is found in all studies, some study designs may be exposed to a higher and unavoidable amount of bias. In this instance, the nature of the assigned treatment simply means that blinding cannot be ensured. Still, this trial has the advantage over observational studies that the randomisation at baseline should minimise confounding by other factors possibly related to the treatment and the outcome. The disadvantage compared to the observational setting is that the exposure contrast in these types of trials is often (as here) much lower compared to what can be investigated in observational setting or in less labour-intensive interventions like the selenium trials above, and therefore may not be enough to elicit a biologically relevant effect.

Observational studies – risk of bias assessment

Example 3: Per- and polyfluoroalkyl substances (PFAS) and fetal growth – risk of bias assessment in a cohort study

In this example, three studies with different designs are presented, all addressing the relationship between PFAS in the mother and birthweight. They use different approaches to assessing exposure, each of which may be vulnerable to bias when judged separately but considering them together they can provide a more complete understanding of the epidemiological evidence. Specifically, it relates to the use of biomarkers of exposure which have the potential advantage of providing a precise measurement of individual integrated exposure, but may introduce confounding if some of the determinants of the biomarker level such as metabolism or excretion are related to the outcome of interest. The alternative to biomarkers of exposure could be modelled exposure which introduces a different potential disadvantage, model uncertainty. Thus, as these examples seek to illustrate it is not straightforward to rank these alternative exposure assessment approaches and different assessors may well disagree as to which is 'better'. However, acknowledging these differences and assessing the differing results from different designs in an integrated assessment helps to both reach a conclusion on the evidence of hazard, and shed light on the relative importance of these different potential sources of bias arising from exposure assessment.

First, we selected a study on '*Perfluorinated Chemicals and Fetal Growth: A Study within the Danish National Birth Cohort*' by Fei et al. (2007).[59]

[59]This is an open access publication: https://pubmed.ncbi.nlm.nih.gov/18008003/.

Unlike in the example for the RCTs above, we put slightly more emphasis on the study results in this section. The focus is also on drawing conclusions on the overall body of evidence from the examples given. In this study, the authors examined the association between perfluoro-octane-sulfonate (PFOS) and perfluoro-octanoate (PFOA) measured in maternal serum drawn in early gestation (weeks 4–14) and birth weight. The participants were 1400 subjects randomly selected form the Danish National Birth Cohort (***n*** ~ 100.000). During the recruitment period (1996–2002), around 30% of all births occurring in Denmark were recruited. In this study, a modest but significant inverse association was observed between maternal concentrations of PFOA and birth weight, while a non-significant inverse association was observed for PFOS (the regression coefficient in the table below from the manuscript reflects decrease in birth weight per 1-ng/mL increase in maternal concentrations of PFOS or PFOA).

TABLE F.3 Adjusted regression coefficients (ß (95% CI)) between PFOS and PFOA (ng/mL) in first maternal blood during pregnancy and birth weight (g) (Adapted from: Fei et al., 2007).

Strata	PFOS	PFOA
All[a]	−0.46 (−2.34 to 1.41)	−10.63 (−20.79 to −0.47)

[a]Adjusted for gestational age, quadratic gestational age, infant sex, maternal age, socio-occupational status, parity, cigarette smoking, prepregnancy BMI, gestational weeks at blood drawing. The category definitions of the covariates were the same as shown in Table 1.

Question on Selection bias

Did selection of study participants result in appropriate comparison groups?

The authors provide the following description in their method section:

Among all participants who gave birth to a single live-born child without a reported congenital malformation (n = 87,752), who provided the first blood sample between gestational weeks 4 and 14 (n = 80,678), and who had responded to all four telephone interviews (n = 43,045), we randomly selected 1,400 mothers.

To answer this question, we have the following rating options:

*'++': There is direct evidence that subjects (both exposed and non-exposed) were similar (e.g. recruited from the same eligible population, recruited with the same method of ascertainment using the same inclusion and exclusion criteria, and were of similar age and health status), recruited within the same time frame, and had the similar participation/response rates***.**

'+': There is indirect evidence that subjects (both exposed and non-exposed) were similar (e.g. recruited from the same eligible population, recruited with the same method of ascertainment using the same inclusion and exclusion criteria, and were of similar age and health status), recruited within the same time frame, and had the similar participation/response rates, OR differences between groups would not appreciably bias results.

'-': There is indirect evidence that subjects (both exposed and non-exposed) were not similar, recruited within very different time frames, or had the very different participation/response rates, OR there is insufficient information provided about the comparison group including a different rate of non-response without an explanation (record 'NR' as basis for answer).

'--': There is direct evidence that subjects (both exposed and non-exposed) were not similar, recruited within very different time frames, or had the very different participation/response rates.

The word 'unexposed' is here perhaps not a good term to use when the substance being investigated is an environmental contaminant that can be detected in humans and wildlife world-wide. When subjects were selected at random, the exposure status was unknown and random sampling should ensure that there is no selection with respect to exposure or outcome. No direct evidence is provided (it is unclear how that could be achieved).

Both '++' and '+' could be justified here but following the strict formulation of the text the rating here would be '+'.

Question on confounding

Did the study design or analysis account for important confounding and modifying variables?

The authors adjusted for the following-set of variables in their statistical analyses:

Gestational age, infant sex, maternal age, socio-occupational status, parity, cigarette smoking, pre-pregnancy BMI, gestational weeks at blood drawing.

Gestational age, infant sex, maternal age and weeks of blood drawing were extracted from clinical records. Other variables were based on self-reports. Misclassification due to self-reported parity should be low (simple to answer). Some misclassification in reporting of smoking, BMI and socio-occupational status could be expected, but since neither of these

characteristics are strongly associated with exposure to PFOS and PFOA, and self-report has been shown to be quite reliable in several other validation studies. The method for covariate assessment here is therefore judged to be valid.

The selected variables are standard for analyses aimed at examining the relationship between pregnancy-exposure and birth weight. All these variables are important predictors of birth weight and some of them are predictors of serum PFO and PFOA concentration (see Table F.4 from the manuscript (Source: Fei et al., 2007) below) (Table F.4):

TABLE F.4 Plasma concentrations of PFOS, PFOA, and birth weight (mean ± SD) by characteristics of study subjects ($n = 1400$).

Characteristic[a]	No. (%)	PFOS (ng/mL)	PFOA (ng/mL)	Birth weight (g)
Maternal age at delivery (years)				
< 25	118 (8.4)	38.6 ± 12.0	6.2 ± 2.1	3573 ± 492
25–29	547 (39.1)	36.8 ± 12.8	6.0 ± 2.8	3590 ± 521
30–34	504 (36.0)	33.9 ± 13.2	5.2 ± 2.2	3676 ± 519
≥ 35	230 (16.4)	33.0 ± 12.7	5.1 ± 2.4	3629 ± 581
Parity				
0	626 (44.7)	37.7 ± 13.0	6.6 ± 2.7	3524 ± 514
1	508 (36.3)	33.2 ± 12.7	4.7 ± 1.9	3689 ± 501
2	225 (16.1)	34.0 ± 12.6	4.8 ± 2.3	3723 ± 580
≥ 3	41 (2.9)	30.5 ± 11.7	3.7 ± 1.6	3862 ± 540
Socio-occupational status				
High	709 (50.8)	34.0 ± 12.7	5.6 ± 2.3	3648 ± 527
Middle	566 (40.5)	36.6 ± 12.9	5.6 ± 2.8	3603 ± 531
Low	121 (8.7)	36.5 ± 14.1	5.6 ± 2.3	3606 ± 536
Prepregnancy BMI (kg/m^2)				
<18.5	58 (4.2)	33.1 ± 14.3	5.2 ± 2.2	3396 ± 539
18.5–24.9	905 (66.2)	34.6 ± 12.9	5.5 ± 2.6	3620 ± 506
25.0–29.9	299 (21.9)	36.3 ± 12.0	5.6 ± 2.3	3638 ± 547
≥30.0	105 (7.7)	39.3 ± 14.4	6.1 ± 2.7	3770 ± 542
Smoking during the pregnancy				
Nonsmoker	1052 (75.1)	35.7 ± 13.3	5.6 ± 2.6	3661 ± 514
Quit smoking	131 (9.4)	33.9 ± 11.6	5.8 ± 2.2	3700 ± 592
1–9 cigarettes/day	109 (7.8)	35.5 ± 12.7	5.8 ± 2.6	3434 ± 469
≥10 cigarettes/day	108 (7.7)	32.5 ± 11.9	4.9 ± 1.9	3384 ± 560
Sex				
Female	690 (49.3)	35.3 ± 13.0	5.5 ± 2.4	3582 ± 538
Male	710 (50.7)	35.2 ± 12.9	5.6 ± 2.7	3668 ± 518

[a]Missing data: maternal age (1), socio-occupational status (4), prepregnancy BMI (33), birth weight (12).

Some assessors could ask the question why other factors such as dietary habits or other co-exposures were not taken into consideration. One answer to that might be that in well-nourished populations diet is generally not a strong predictor of birth weight. In this (and other) population, the relationship between self-reported diet and PFOS and PFOA measured in maternal serum was very modest (Halldorsson et al., 2008). This study was conducted in a time period when environmental release and use was at peak levels (Armitage et al., 2009) and these substances were found in common household products (carpets, clothing). These substances may therefore have had different exposure profiles compared to many other legacy contaminants. These arguments are, however, speculative but that is usually the case when assessing risk of confounding bias in observational studies.

One suspected confounder not taken into consideration in this study is possible confounding by physiological changes in pregnancy (Savitz, 2007; Verner et al., 2015). During pregnancy, blood volume increases which would lead to lower circulating concentrations of PFOS and PFOA, and the blood volume expansion is partly proportional to fetal growth. In addition, increase in glomerular filtration rate could result in more rapid excretion of PFOS and PFOA and the filtration rate is again partly driven by fetal growth (Verner et al., 2015). These changes would however be of more influence in late gestation when the fetus grows rapidly, but not in samples drawn in early pregnancy as is the case here (weeks 4–14).

Based on these considerations the judgement here would be a selection between a '+' or a '++':

> **'++'**: *There is direct evidence that appropriate adjustments or explicit considerations were made for primary covariates and confounders in the final analyses through the use of statistical models to reduce research-specific bias including standardisation, matching, adjustment in multivariate model, stratification, propensity scoring, or other methods that were appropriately justified. Acceptable consideration of appropriate adjustment factors includes cases when the factor is not included in the final adjustment model because the author conducted analyses that indicated it did not need to be included, AND there is direct evidence that primary covariates and confounders were assessed using valid and reliable measurements, AND there is direct evidence that other exposures anticipated to bias results were not present or were appropriately measured and adjusted for. In occupational studies or studies of contaminated sites, other chemical exposures known to be associated with those settings were appropriately considered***.***
>
> *'+': There is indirect evidence that appropriate adjustments were made, OR it is deemed that not considering or only considering a partial list of covariates or confounders in the final analyses would not appreciably bias results. AND there is evidence (direct or indirect) that primary covariates and confounders were assessed using valid and reliable measurements, OR it is deemed that the measures used would not appreciably bias results (i.e. the authors justified the validity of the measures from previously published research), AND there is evidence (direct or indirect) that other co-exposures anticipated to bias results were not present or were appropriately adjusted for, OR it is deemed that co-exposures present would not appreciably bias results*.

Since we cannot answer 'yes' to the condition on 'direct evidence', but it is included that there is sufficient indirect evidence that appropriate adjustments were made, the evaluation here would be a '+'. Note, however, that this evaluation depends on several assumptions and considerations, and another judgement could have been argued for. Assessing risk of confounding bias for observational studies requires expert judgement and rating is often largely determined by the experts' own views.

Question on attrition/exclusion bias.

Were outcome data complete without attrition or exclusion from analysis?

When selecting the 1400 maternal samples, the only inclusion criteria were singleton live born infants. Otherwise, data was complete (maternal concentration of PFOS and PFOA and birth weight was available for all samples). Based on that the rating '++' seems justified:

> *'++': There is direct evidence that loss of subjects (i.e. incomplete outcome data) was adequately addressed and reasons were documented when human subjects were removed from a study. Acceptable handling of subject attrition includes: very little missing outcome data; reasons for missing subjects unlikely to be related to outcome (for survival data, censoring unlikely to be introducing bias); missing outcome data balanced in numbers across study groups, with similar reasons for missing data across groups, OR missing data have been imputed using appropriate methods and characteristics of subjects lost to follow-up or with unavailable records are described in identical way and are not significantly different from those of the study participants*.

Question on detection bias.

1. Can we be confident in the exposure characterisation?

Both serum PFOS and PFOA have elimination half-life in humans that is measured in years. Since these are persistent compounds and serum concentrations are considered appropriate biomarker, bias due to the exposure assessment can be considered as low.

> *'++': There is direct evidence that exposure was consistently assessed (i.e. under the same method and time-frame) using well-established methods that directly measure exposure (e.g. measurement of the chemical in air or measurement of the chemical in blood, plasma, urine, etc.), OR exposure was assessed using less-established methods that directly measure exposure and are validated against well-established methods***.***
>
> *'+': There is indirect evidence that the exposure was consistently assessed using well-established methods that directly measure exposure, OR exposure was assessed using indirect measures (e.g. questionnaire or occupational exposure assessment by a certified industrial hygienist) that have been validated or empirically shown to be consistent with methods that directly measure exposure (i.e. inter-methods validation: one method vs. another).*

The method of analysis also appears appropriate ('liquid chromatography- tandem mass spectrometry with laboratory personnel being blinded to the birth outcomes and types of blood drawn'). When the samples were measured (2006–2007) analytical methods for PFOS and PFOA were in the early stage and different laboratories did not always produce consistent results in inter-calibration exercises (Longnecker et al., 2008). For this study, it was the 3M laboratory that did these analyses and it is reasonable to assume that their methods were of high standards. Still, given the uncertainty of analytical methods around this time period, a strict but reasonable judgement would be a '+'.

2. Can we be confident in the outcome assessment?
All birth outcomes were extracted from clinical records, which can be considered as gold standard. All subjects were followed-up for the same length of time (until birth). Blinding is ensured by design in this case. Based on that the rating of '++' seems justified.

> *'++': There is direct evidence that the outcome was assessed using well-established methods (e.g. the 'gold standard' with validity and reliability >0.70; Genaidy et al., 2007), AND subjects had been followed for the same length of time in all study groups. Acceptable assessment methods will depend on the outcome, but examples of such methods may include: objectively measured with diagnostic methods, measured by trained interviewers, obtained from registries (Shamliyan et al., 2010), AND there is direct evidence that the outcome assessors (including study subjects, if outcomes were self-reported) were adequately blinded to the study group, and it is unlikely that they could have broken the blinding prior to reporting outcomes.*

Question on selective reporting bias

Were all measured outcomes reported?
All measured outcomes relevant to assess fetal growth, including birth weight and gestational length, were reported. The relationship between birth outcomes and the covariates accounted for in the statistical analyses were also clearly reported (see Table above). All effect estimates were also clearly reported with confidence intervals (see manuscript).

Based on that the judgement '++' seems appropriate.

> *'++': There is direct evidence that all of the study's measured outcomes (primary and secondary) outlined in the protocol, methods, abstract, and/or introduction (that are relevant for the evaluation) have been reported. This would include outcomes reported with sufficient detail to be included in meta-analysis or fully tabulated during data extraction and analyses had been planned in advance.*

Other Biases

The use of statistical methods seems appropriate, i.e. adjusting for covariates using multivariate regression analyses is a standard method for confounder control. No other issues detected.
To conclude, a summary of the proposed rating for the individual bias questions is given in Table F.5.

TABLE F.5 Summary of rating for Fei et al. (2007).

Type of bias as defined in Section 4.2	OHAT formulation of bias[60]	Question	Rating	Remarks
Selection bias	Selection bias	Did selection of study participants result in appropriate comparison groups?	+	Random selection of 1400 subjects from the cohort that had participated in four different data collections until 18 months postpartum Random selection should guard against biased selection in relation to exposure or outcome. No direct evidence for that is, however, provided
Confounding bias	Confounding bias	Did the study design or analysis account for important confounding and modifying variables?	+	Use of early pregnancy sample should minimise risk of confounding (but confounding is still possible). Important covariates are considered and accounted for
Selection bias	Attrition/Exclusion bias	Were outcome data complete without attrition or exclusion from analysis?	++	No attrition, information on fetal growth is available for all 1400 subjects
Information bias	*Detection bias*	Can we be confident in the exposure characterisation?	+	PFOS and PFOA have long elimination half-life and serum concentrations are considered an optimal biomarker. The authors evaluated stability of a single serum measurement by measuring concentrations in subset of participants in late gestation and in cord blood. Analytical methods appear valid (but no results from an inter-calibration exercises are reported)
Information bias	*Detection bias*	Can we be confident in the outcome assessment?	++	Birth outcomes were extracted from clinical record, which can be considered as gold standard
Information bias	*Selective reporting bias*	Were all measured outcomes reported?	++	All outcomes relating to fetal growth were clearly reported
Information bias	Other bias	….such as appropriateness of the statistical methods applied	No rating	Statistical methods judged appropriate; no other potential sources of bias detected

As Table F.5 shows each bias question was in all cases either rated as '+' or a '++' (no '--' or '--'). Compared to the example for the RCTs above some decisions, such as the question on confounder control, are more subjective with no clear yes/no answer, as all possible confounders can never be accounted for. For that question, there is no right or wrong answer, only expert judgement that should be clearly documented.

Example 4: Per- and polyfluoroalkyl substances (PFAS) and fetal growth, partial confounding? – risk of bias assessment in cross-sectional studies

In another paper on PFOS/PFOA and fetal growth, Apelberg et al. (2007)[61] examined the same relationship in 293 mother–child pairs but using cord blood drawn at delivery. Since cord blood is drawn at a similar time as birth weight is recorded, this is in fact a cross-sectional study. However, since PFOS and PFOA have long elimination half-life and it is well documented that serum samples drawn at different time points in pregnancy are strongly correlated (see Table 3 from the manuscript by Apelberg et al. (2007), below), the cross-sectional label is perhaps not important. Similar to the study above, a significant inverse association was observed between cord blood concentrations of PFOS and PFOA with birth weight adjusted for gestational age. Similar but non-nonsignificant inverse associations were observed after adjusting for several other potential confounders. When interpreting the effect estimates in the table below, it is relevant to consider that this is much smaller study than the study by Fei et al. (2007) (***n*** = 293 vs 1400). As study size directly influences the standard errors of the effect estimate, the magnitude of the observed effect is perhaps more appropriate to focus on when comparing the two studies (not only formal statistical significance).

[60]The six bias types used in the NTP-OHAT risk of bias tool reflect a finer categorization of biases that fall under either selection bias, information bias and confounding, as defined in Section 4.2.

[61]This is an open access publication: https://pubmed.ncbi.nlm.nih.gov/18008002/.

TABLE F.6 Estimated change in mean gestational age, birth weight, and birth size parameters with a change in PFOS or PFOA concentrations equal to one ln-unit or from the 25th to 75th percentile.

	PFOS		PFOA	
Model	**Change in end point (95% CI) per ln-unit[a]**	**Per increase from the 25th to 75th percentile (IQR)[b]**	**Change in end point (95% CI) per ln-unit[a]**	**Per increase from the 25th to 75th percentile (IQR)[b]**
Birth weight (g)				
Univariate	−37 (−139 to 64)	−31 (−117 to 54)	−97 (−234 to 40)	−54 (−131 to 23)
Adjusted for GA	−89 (−170 to −8)*	−75 (−143 to −7)*	−161 (−270 to −52)*	−90 (−151 to −29)*
Fully adjusted[d]	−69 (−149 to 10)	−58 (−125 to 9)	−104 (−213 to 5)	−58 (−119 to 3)
Head circumference (cm)				

Abbreviations: GA, gestational age; IQR, interquartile range; ln-unit, natural log unit.

[a]Represents the change in the end point associated with a unit increase in ln(PFOS) or ln(PFOA), which is equivalent to a 2.7-fold increase in PFOS or PFOA.

[b]Interquartile range is 3.4–7.9 ng/mL for PFOS and 1.2–2.1 ng/mL for PFOA.

[c]Adjusted for maternal age, BMI, race, previous preterm birth, smoking, diabetes, and hypertension.

[d]Adjusted for gestational age, maternal age, BMI, race, parity, smoking, baby sex, height, net weight gain, diabetes, and hypertension. For head circumference, adjusted model also includes delivery mode (C-section/vaginal).

*Statistically significant ($p < 0.05$).

Question on confounding

Did the study design or analysis account for important confounding and modifying variables?

The same covariates are accounted for in the statistical analyses by Apelberg et al. (2007) as in the study by Fei et al. (2007). Same clarity in terms of relationship between covariates and concentrations of PFOS and PFOA is provided (see Table 1 in the manuscript). It is, however, important that samples are drawn in late gestation. In that case confounding due to increased blood volume expansion and/or higher glomerular filtration rate (more rapid excretion) among women carrying larger fetuses may occur (Savitz, 2007; Verner et al., 2015). Under that assumption, selecting one of the two rating options therefore seems appropriate:

> *'-': There is indirect evidence that the distribution of primary covariates and known confounders differed between the groups and was not appropriately adjusted for in the final analyses, OR there is insufficient information provided about the distribution of known confounders (record 'NR' as basis for answer), OR there is indirect evidence that primary covariates and confounders were assessed using measurements of unknown validity, OR there is insufficient information provided about the measurement techniques used to assess primary covariates and confounders (record 'NR' as basis for answer), OR there is indirect evidence that there was an unbalanced provision of additional co-exposures across the primary study groups, which were not appropriately adjusted for, OR there is insufficient information provided about co-exposures in occupational studies or studies of contaminated sites where high exposures to other chemical exposures would have been reasonably anticipated (record 'NR' as basis for answer).*
>
> *'--': There is direct evidence that the distribution of primary covariates and known confounders differed between the groups, confounding was demonstrated, and was not appropriately adjusted for in the final analyses, OR there is direct evidence that primary covariates and confounders were assessed using non valid measurements, OR there is direct evidence that there was an unbalanced provision of additional co-exposures across the primary study groups, which were not appropriately adjusted for.*

In this case, it seems that we have 'indirect evidence that the distribution of primary covariates and known confounders differed between the groups and was not appropriately adjusted'. The appropriate rating here would be '-'.

In terms of magnitude and direction, this type of confounding may not account for the full associations, and that conclusion is partly supported by PBPK modelling (Verner et al., 2015). Partial confounding here would be in the direction of creating a stronger inverse association with birth weight. Although the results from Apelberg et al. (2007) and Fei et al. (2007) are not reported in an identical manner, the two studies appear to be in line with that pattern.

Example 5: Per- and polyfluoroalkyl substances (PFAS) and fetal growth, detection bias? – risk of bias assessment of a cohort study.

As a final example, we take another study on PFOA and fetal growth. The study population are subjects from the C8 study that were exposed to high levels of PFOA due to contaminated drinking water around duPont facilities in Little Hocking, Ohio. This study examined the association between exposure during pregnancy to PFOA and term birth weight among 4534 mother–child pairs from the C8 cohort (Savitz et al., 2012).[62] The C8 cohort was established in 2005 after the

[62]This is an open access publication: https://ehp.niehs.nih.gov/doi/10.1289/ehp.1104752.

water contamination was known. In order to examine associations with birth weight, the authors extracted obstetric outcomes from birth records among cohort participants that had occurred in the area between 1990 and 2004. There were serum measurements in 2005–2006 but not at that time of pregnancy and so serum PFOA relied on a pharmacokinetic model linked to residential history and estimated historical drinking water concentrations identified from a fate and transport model (Shin et al., 2011). Comparison of the predicted serum concentrations to measured concentrations in 2005–2006 showed good agreement overall, with Spearman's correlation coefficient of 0,67 (Shin et al., 2011). One strength of this study compared to the two examples above is that the exposure gradient was much larger due to large exposure contrast in drinking water among subjects living in different water district areas. Several different model assumptions were examined. In short, no significant association was observed between modelled PFOA exposure and birth weight at term. Most effect estimates were negative but much weaker than in the studies mentioned above, for example in the linear model the birthweight change for 100 ng/mL of PFOA was -9g (CI –20 to +2), equivalent to 0.1 g per ng/ml PFOA, essentially null (see Table F.7 from Savitz et al., 2012).

TABLE F.7 Study II: PFOA and pregnancy outcome based on birth records linked to the C8 Health Project: association of PFOA with indicators of fetal growth, Mid-Ohio Valley, 1990–2004.

	Term low birth weight				Term SGA				Change in term birth weight (g)		
Estimated PFOA	**Term births ≥2500 g (*n*)**	**Cases (*n*)**	**Crude OR**	**Adjusted[a] OR (95% CI)**	**Term, AGE (*n*)**	**Cases (*n*)**	**Crude OR**	**Adjusted[a] OR (95% CI)**	**Cases (*n*)**	**Crude difference**	**Adjusted[a] difference (95% CI)**
Uncalibrated											
IQR(lnPFOA)[b] increase	4043	99	0.87	1.04 (0.75, 1.44)	3375	362	1.02	1.18 (0.97, 1.43)	4142	0.97	−21.89 (−45.91, 2.13)
100-ng/mL increase	4043	99	0.93	1.00 (0.82, 1.21)	3375	362	1.01	1.07 (0.98, 1.17)	4142	4.27	−9.14 (−20.30, 2.02)
<40th percentile (3.9 to <8.9 ng/mL)	1629	40	1.0	1.0	1356	144	1.0	1.0	1669	0	0 (referent)
40th to <60th percentile (8.9 to <21.8 ng/mL)	791	19	0.9	0.9 (0.5, 1.7)	659	72	1.0	1.0 (0.7, 1.4)	810	−19.9	−3.8 (−40.4, 32.8)
60th to <80th percentile (21.8 to 83.3 ng/mL)	814	27	1.4	1.6 (1.0, 2.8)	689	76	1.0	1.1 (0.8, 1.6)	841	−30.4	−25.4 (−63.7, 12.9)
≥80th percentile (83.3 to 921.3 ng/mL)	809	13	0.7	0.9 (0.5, 1.7)	671	70	1.0	1.3 (0.9, 1.7)	822	4.9	−33.3 (−73.1, 6.5)
Bayesian calibration											
IQR(lnPFOA)[c] increase	4043	99	0.97	1.16 (0.86, 1.58)	3375	362	1.03	1.19 (1.00, 1.43)	4142	7.78	−21.51 (−43.62, 0.61)
100-ng/mL increase	4043	99	0.93	1.04 (0.85, 1.27)	3375	362	0.98	1.06 (0.97, 1.16)	4142	3.73	−18.55 (−31.31, −5.80)
<40th percentile (3.9 to <8.9 ng/mL)	1624	42	1.0	1.0	1358	148	1.0	1.0	1666	0	0 (referent)
40th to <60th percentile (8.9 to <19.6 ng/mL)	803	17	0.8	0.8 (0.4, 1.5)	676	68	0.9	0.9 (0.7, 1.3)	820	−15.8	10.8 (−24.9, 46.5)
60th to <80th percentile (19.6 to 53.1 ng/mL)	803	24	1.1	1.3 (0.7, 2.2)	664	74	1.0	1.1 (0.8, 1.5)	827	−9.0	−11.0 (−49.8, 27.8)
≥80th percentile (53.1 to 1897.0 ng/mL)	813	16	0.8	1.0 (0.6, 1.9)	677	72	1.0	1.3 (0.9, 1.7)	829	8.7	−32.3 (−71.5, 6.8)

(Continues)

TABLE F.7 (Continued)

	Term low birth weight				Term SGA				Change in term birth weight (g)		
Estimated PFOA	**Term births ≥2500 g (*n*)**	**Cases (*n*)**	**Crude OR**	**Adjusted[a] OR (95% CI)**	**Term, AGE (*n*)**	**Cases (*n*)**	**Crude OR**	**Adjusted[a] OR (95% CI)**	**Cases (*n*)**	**Crude difference**	**Adjusted[a] difference (95% CI)**
Traditional calibration											
IQR(lnPFOA)[d] increase	4043	99	1.16	1.33 (1.04, 1.69)	3375	362	1.05	1.17 (1.00, 1.35)	4142	2.45	−16.90 (−34.89, 1.08)
100-ng/mL increase	4043	99	1.00	1.07 (0.96, 1.18)	3375	362	1.03	1.08 (1.01, 1.16)	4142	4.54	−12.76 (−26.08, 0.57)
<40th percentile (0.05 to <11.4 ng/mL)	1614	35	1.0	1.0	1351	147	1.0	1.0	1649	0	0 (referent)
40th to <60th percentile (11.4 to <21.0 ng/mL)	817	14	0.8	0.8 (0.4, 1.5)	686	70	0.9	1.0 (0.7, 1.3)	831	−9.2	4.2 (−31.2, 39.6)
60th to <80th percentile (21.0 to 49.0 ng/mL)	793	32	1.8	2.2 (1.3, 3.6)	659	72	1.0	1.1 (0.8, 1.5)	825	1.1	1.8 (−37.7, 41.4)
≥80th percentile (49.0 to 2468.4 ng/mL)	819	18	1.1	1.4 (0.8, 2.5)	679	73	1.0	1.2 (0.9, 1.7)	837	22.7	−21.2 (−59.6, 17.2)

Abbreviation: AGA, appropriate for gestational age.

[a]Adjusted for maternal age, education, parity, smoking status, exposure year, state of residence, gestational age (term birth weight analysis only).

[b]Effect estimates represent the change in outcome for a shift from the 25th percentile to the 75th percentile in estimated PFOA serum levels [IQR(lnPFOA) = 2.39].

[c]Effect estimates represent the change in outcome for a shift from the 25th percentile to the 75th percentile in estimated PFOA serum levels [IQR(lnPFOA) = 1.92].

[d]Effect estimates represent the change in outcome for a shift from the 25th percentile to the 75th percentile in estimated PFOA serum levels [IQR(lnPFOA) = 1.6].

Question on detection bias in relation to the exposure assessment

Can we be confident in the exposure characterisation?

This study relied on modelled, estimated serum levels for the year when the pregnancy occurred, and although the model predictions correlated well overall there are individual uncertainties in the prediction, leading to concerns about exposure misclassification. Indeed, in the risk of bias approach in the Navigation Guide system, this was scored as a high risk of bias in one systematic review (Johnson et al., 2014) and dismissed from the overall summary of 'better studies', because a model was considered inherently more at risk of bias than a measurement. On the other hand, the model was immune to the potential biases related to excretion described in the studies above. Thus although Johnson et al. (2014) evaluated the possible bias from exposure as '-' there are alternative arguments that were not considered suggesting that modelled exposure may be less prone to some of the biases associated with the use of biomarker. Those argument would support the use of '+'. In short, different arguments relating to risk of bias and expert views may result in different scoring for individual bias questions. In case such differences occur, the use of RoB tools allows for a transparent way of identifying such differences.

Based on that assumption, we would choose from the two following options:

'-': There is indirect evidence that the exposure was assessed using poorly validated methods that directly measure exposure, OR there is direct evidence that the exposure was assessed using indirect measures that have not been validated or empirically shown to be consistent with methods that directly measure exposure (e.g. a job–exposure matrix or self-report without validation) (record 'NR' as basis for answer), OR there is insufficient information provided about the exposure assessment, including validity and reliability, but no evidence for concern about the method used (record "NR" as basis for answer).

'+': There is indirect evidence that the exposure was consistently assessed using well-established methods that directly measure exposure, OR exposure was assessed using indirect measures (e.g. questionnaire or occupational exposure assessment by a certified industrial hygienist) that have been validated or empirically shown to be consistent with methods that directly measure exposure (i.e. inter-methods validation: one method vs. another).

Final remarks

Based on the three studies assessed here, one likely conclusion would have been that the study by Fei et al. (2007) is a low risk of bias study. The study by Apelberg et al. (2007) would have bene considered at a higher risk of bias, but the finding of that study could be considered supportive. The study by Savitz et al. (2012) could also have been ranked as high risk of bias and focusing on that summary assessment alone it would have been easy to reject its findings. The body of evidence from the three studies might therefore have been assessed in favour of strong evidence for an association between PFOA and birth weight.

One possible mechanism of confounding has been highlighted, that during the progress of pregnancy PFAS levels may change in a manner correlated with the degree of fetal growth, leading to a confounded relationship between maternal serum PFAS and birthweight. This would be expected to be more evident in associations between PFAS measured later in pregnancy and such a pattern was evident in a metanalysis stratifying birthweight studies by whether the measurement was done early (no overall association) or late in pregnancy (a significant effect) (Steenland et al., 2018).

Biomarkers such as serum PFAS, stable and with a long half-life are very attractive exposure indicators for assessing exposure to the fetus, as they reflect exactly the individual body burden. Studies of PFAS half-life also have shown that there is wide individual variability in excretion rates (Li et al., 2018), which in turn affects body burden as well as intake. However, the determinants of variation in excretion rate are poorly understood, and if those determinants were also linked to determinants of fetal growth, this could introduce confounding.

The study of Savitz et al. (2012) is an example with exposure being based on external exposure – converted to predicted serum levels, not taking any account of individual intake or excretion. This was scored as a high risk of bias in one systematic review (Johnson et al., 2014) and dismissed from the overall summary of 'better studies', because a model was considered inherently more at risk of bias than a measurement. However, it might be better to treat these different observational studies as trading off different potential biases.

Studies which classify exposure by degree of intake only, miss the individual variation of excretion rates, but also miss this potential unmeasured confounding. If results were consistent between studies assessing contrasts in exposure by serum measurements and by intake or other external exposure measure, then the consistent result (positive or absent) is more persuasive. This is an example of triangulation where results are compared between studies with different potential biases. The difference between a study with measurements early in pregnancy vs late in pregnancy, can reveal a bias due to pregnancy affecting PFAS body burden. The difference between a modelled vs measured serum level can reveal trade-off between excretion-related confounding vs model imprecision. The task of synthesising the evidence should include an expert assessment of the relevant importance of these various potential biases.

Given the larger exposure contrast in the C8 study, the modelled exposure is likely, despite some misclassification, to accurately rank those with high vs low PFOA exposures. Despite large exposure contrast and large study size, no significant difference in birth weight is detected. In addition, the modelled exposure is not influenced by physiological changes in pregnancy that we suspect may act as a confounder. This argument casts some doubts over the associations observed in the studies by Fei et al. (2007) and Apelberg et al. (2007), despite the fact that these two studies are considered to have a lower risk of bias. The resulting conclusion could therefore have been rather weak evidence for an association between PFOA and birth weight.

In summary, different conclusions may be reached if type, magnitude and direction of bias of individual studies are not considered when assessing the body of evidence. Ranking studies into tiers may be helpful, but this categorisation alone should not be used to guide the assessment of the body of evidence. For example, it is very useful to make use of the full mapping of the different ratings by type of bias and by individual studies.

Example 6: Cadmium and Osteoporosis– risk of bias assessment of a case–control study.

Sommar et al. (2013)[63] carried out a nested case–control study to evaluate the association between exposure to cadmium (Ery-Cd) and low-trauma hip fracture risk.

[63]This is an open access publication: https://link.springer.com/article/10.1007/s00223-013-9796-5.

Question on Selection bias

Did selection of study participants result in appropriate comparison groups?

There is direct evidence that cases and controls were similar (nested case–control: recruited prospectively from the same 'general' population including being of similar age, gender, ethnicity and eligibility criteria), recruited within the same time frame, and controls are described as having no history of the outcome. The observed statistically significant difference for smoking at baseline is considered as a potential confounder.

Question on Confounding

Did the study design or analysis account for important confounding and modifying variables?

These are biobank data and a 'candidate compound' approach was implemented. It is deemed that co-exposures present would not appreciably bias results. Adjustments were made for height, BMI, smoking (traditional fracture risk confounders).

Question on Selection bias

Were outcome data complete without attrition or exclusion from analysis?

The exposure data is incomplete (109/4900). No comparison with the whole sample is reported. There is indirect evidence that exclusion of subjects from analyses was not adequately addressed.

Question on Information bias

Can we be confident in the exposure characterisation?

There is direct evidence that exposure was consistently assessed (under the same method and timeframe) using well-established methods that directly measure exposure (Cd measurement in erythrocytes).

Question on Information bias

Can we be confident in the outcome assessment?

Control status deferred by exclusion. There is indirect evidence that the outcome was assessed in cases (i.e. case definition) and controls using acceptable methods, and subjects had been followed for the same length of time in all study groups.

Question on Information bias

Were all measured outcomes reported?

There is indirect evidence that all of the studies measured outcomes (primary and secondary) outlined in the methods, abstract and/or introduction (that are relevant for the evaluation) have been reported. This includes fracture data that are reported with sufficient detail to be included in meta-analysis or fully tabulated during data extraction. However, no protocol has been described.

Question on other sources of bias

Were there no other potential threats to internal validity (e.g. statistical methods were appropriate, and researchers adhered to the study protocol)?

Probably Low risk of bias. Appropriate statistical methods have been used. A summary of the ratings is given in Table F.8.

TABLE F.8 Summary of rating for Sommar et al. (2013).

Type of bias as defined in Section 4.2	OHAT formulation of bias*	Question	Rating	Remarks
Selection bias	Selection bias	Did selection of study participants result in appropriate comparison groups?	++ or +	Nested case–control; 2 population-based sub-cohorts (repeated mammography screening, general health examinations service); fracture cases identified from a 12-year prospective injury-fracture database and cross-matched within the sub-cohorts data; controls: one or two, selected from the same NSHDS cohort (VIP or V-MSP), matched for sex, age at recruitment (within 1 year), and date of blood sampling (within 1 week); cases and controls identical on most background factors (ss for smoking); Figure 1, Table 1

TABLE F.8 (Continued)

Type of bias as defined in Section 4.2	OHAT formulation of bias*	Question	Rating	Remarks
Confounding	Confounding	Did the study design or analysis account for important confounding and modifying variables?	+	Biobank; Univariate and multivariate analyses also performed for body mass index (BMI), height and smoking
Selection bias	Attrition/Exclusion Bias	Were outcome data complete without attrition or exclusion from analysis?	–	About 17% of all fracture cases (4900) were represented in the biobank, and 85% of them had left their sample before the time of the fracture. Finally, out of 158 identified cases, Ery-Cd was analysed in 111 cases and 109 of these were included in the analysis (Figure 1)
Information bias	*Detection Bias*	Can we be confident in the exposure characterisation?	++ or +	Ery-Cd (established but urine/blood>ery); sampling and measurement protocol adequately described; single measurement; FU not reported
Information bias	*Detection Bias*	Can we be confident in the outcome assessment?	+	Registry-based; a 12-year prospective injury-fracture academic database at the Umea° University Hospital; fractures verified by X-rays; trauma type: records; database merged with the NSHDS register
Information bias	*Selective Reporting Bias*	Were all measured outcomes reported?	+	Methods mirror Results. No protocol.
Other sources of bias	*Other sources of bias*	Were there no other potential threats to internal validity (e.g. statistical methods were appropriate, and researchers adhered to the study protocol)?	+	Appropriate statistical methods

*The six bias types used in the NTP-OHAT risk of bias tool reflect a finer categorization of biases that fall under either selection bias, information bias and confounding, as defined in Section 4.2.

APPENDIX G

Categorisation of continuous exposures: Analysis and interpretation

In epidemiological research, particularly for human studies, continuous exposure variables are often divided into categories, using a priori or data-driven (percentiles) cut points of exposure. One reason for such categorisation is that interpreting and conveying the results of such analyses in terms of public health messages is often simpler compared to effect estimates from analyses based on continuous measures, generally generated by linear or non-linear regression analyses.

For this example, we look at the association between dietary fibre intake and the ratio of total cholesterol to HDL-cholesterol (total-cholesterol:HDL) in a small cross-sectional study of 178 overweight and obese women aged between 21 and 44 years. These women were enrolled at three different recruitment centers and were asked to record their diet by weighted 2-day food records prior to having their blood samples drawn. The blood samples were then analysed for serum lipid profile and other biomarkers of cardiovascular health. The outcome considered here, the total-cholesterol:HDL ratio, is considered to be a reliable predictor of later coronary heart disease in both men and women (Hartley et al., 2016; Ingelsson et al., 2007).

NOTE: The example is chosen for illustrative purposes and discussions of study quality, risk of bias and causality are beyond the scope of this example. The example is based on real data, but the results have not been published in a peer-reviewed journal (the primary aim of this study was not to examine this cross-sectional association). Even though the example is within the nutritional domain, the same considerations as reflected on here would apply if the example would have addressed exposures more relevant to toxicological risk assessment.

Continuous exposure: The scatter plot for the association between fibre intake and the (total-cholesterol:HDL ratio is shown in Figure G.1). The distribution of exposure and outcome variables are shown in Table G.1.

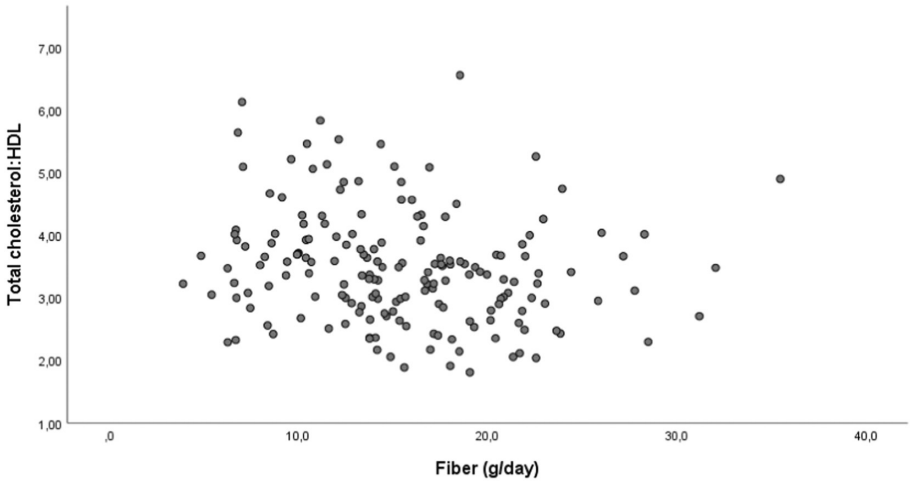

FIGURE G.1 Scatterplot of the cross-sectional association between dietary fibre intake and the total-cholesterol:HDL ratio (based on the continuous values of the variables) in 178 overweight and obese women aged between 21 and 44 year.

TABLE G.1 Distribution of the exposure and outcome variables.

	Mean (SD)	Range
Total-cholesterol:HDL	3.5 (0.90)	1.8–6,6
Total cholesterol (mg/dL)	193 (33)	108–288
HDL cholesterol (mg/dL)	58 (14)	33–126
Dietary fibre (g/day)	15 (6)	4–35

When looking at the scatter plot there appears to be a modest decrease in the lipid ratio with increasing fibre intake. Due to the high between-person variability it is, however, difficult to evaluate this association visually. Furthermore, the scatter plot only shows the crude association where potential confounders, such as age and body mass index (BMI), have not been accounted for.

To evaluate this association statistically we assume a linear relationship using linear regression analyses. For the unadjusted association shown in Table G.2 below the regression model would be written as:

$$Y(x) = \alpha + \beta x,$$

where x is the fibre intake and β is the slope for the association between the fibre intake and the total-cholesterol:HDL ratio; and α is the value of that ratio at the zero fibre intake. For this particular example, the intercept α gives the total-cholesterol:HDL ratio at 0 fibre intake, which does not exist in this population (the intercept α is therefore of limited relevance). Table G.2 shows the slope for both the unadjusted association and the slope after adjusting for covariates (the adjusted model has more terms added to the model above).

TABLE G.2 Association between dietary fibre intake entered as continuous variable and the total-cholesterol:HDL ratio. The regression coefficient gives the change in the outcome per 10-g increase in dietary fibre intake.

Unadjusted β (95% CI)	P for trend[1]	Adjusted[2] β (95% CI)	p for trend[1]
−0.28 (−0.51, −0.06)	0.02	−0.35 (−0.60, −0.11)	0.005

[1]T-test with fibre intake entered as continuous variable in the crude regression model.
[2]Adjusted for age, body mass index, total energy intake and recruitment center.

Based on the unadjusted regression coefficient, the total-cholesterol:HDL ratio decreases by 0.28 per 10-g increase in fibre intake. After adjustment the estimated decrease is slightly stronger (0.35). The high variability seen in Figure G.1 is partly reflected by the relatively wide confidence intervals with the upper limit being strictly below but close to zero.

When interpreting the association in Table G.2, some assumptions have to be made. One interpretation would be that the maximum possible increase in fibre intake in this population (see Table G.1) is about 30 grams. Based on the adjusted slope, the maximum expected decrease in the total-cholesterol:HDL ratio would be around 1.05, which is s around 30% of the mean value for that ratio. The above interpretation rests on two assumptions. First, we assume a linear relationship across the full range of fibre intake, which may not be the case. Therefore, fitting a non-linear function might give a better estimate of the underlying dose–response, though interpreting the resulting coefficients and the associated uncertainty (confidence intervals) would be more complex. Second, assuming a change in fibre intake of 30 g/day seems quite large as such an increase is quite extreme based on the intake distribution in this population.

Categorical exposure

Examining the association between dietary fibres and the total-cholesterol:HDL ratio by breaking the continuous fibre variable into categories offers a different way of examining and interpreting the data. In the example below the fibre variable has been a priori divided into four intake groups:

- Group 1 (4 to < 10 g/day),
- Group 2 (10 to < 15 g/day),
- Group 3 (15 to < 20 g/day) and
- Group 4 (20 to 35 g/day).

The corresponding scatter plot is shown in Figure G.2 where the median intake in each fibre group is used. Figure G.2 shows the same data as in Figure G.1 but now participants in each group have been assigned the median value in their respective exposure category. As for Figure G.1 the high between-person variability in the outcome combined with assigning the same value to all participants within the same group makes it difficult to evaluate the association by visual inspection.

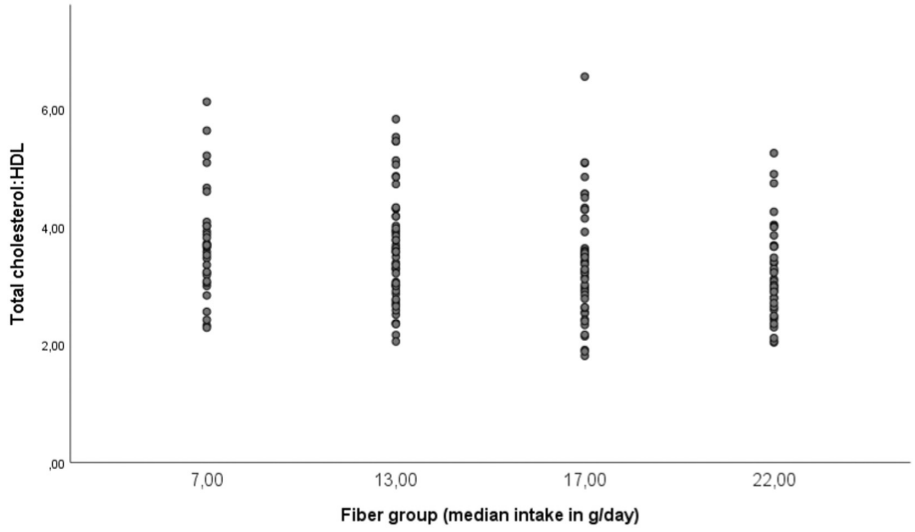

FIGURE G.2 Scatterplot of the cross-sectional association between dietary fibre intake categorised into 4 groups e and the total-cholesterol:HDL ratio in 178 overweight and obese women aged between 21 and 44 years.

To evaluate this association statistically we again use linear regression analysis based on the following model for the unadjusted association:

$$Y(x) = \alpha + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4.$$

Here $\boldsymbol{x}_2$ is a quantal variable for exposure group 2 and is set to zero except when fibre intake is within the range of that group (10 to < 15 g/day). The variables x_3 and x_4 are coded in the same way. Based on modelling, the intercept α represents the mean total-cholesterol:HDL ratio in group 1. The slopes $\boldsymbol{\beta}_2$, $\boldsymbol{\beta}_3$ and $\boldsymbol{\beta}_4$ represent the mean change in the total-cholesterol:HDL ratio relative to group 1. The estimated slopes for this unadjusted association and the slope after adjusting for covariates are shown in Table G.3 (for the adjusted model more terms have been added to the model above).

TABLE G.3 Association between dietary fibre intake and the total cholesterol:HDL ratio. The dietary fibre variable has been divided into four categories and regression coefficients reflect the mean change relative to Group 1.

Fibre intake (g/day)		Unadjusted	Adjusted[1]
Group	**Median (range)**	**β (95% CI)**	**β (95% CI)**
1 (*n* = 32)	7 (4–< 10)	Referent (3.71[2])	Referent
2 (*n* = 56)	13 (10–< 15)	−0.11 (−0.50, 0.28)	−0.17 (−0.54, 0,21)
3 (*n* = 51)	17 (15–< 20)	−0.32 (−0.71, 0.08)	−0.36 (−0,74, 0.03))
4 (*n* = 39)	22 (20–35)	−0.49 (−0.91, −0.07)	−0.48 (−0.88, −0.07)
P for trend[3]		0.01	0.01

[1]Adjusted for age, body mass index, total energy intake and recruitment centre.

[2]The mean total-cholesterol:HDl ration is 3.71 in group 1.

[3]*T*-test with the categorical fibre variable modelled as a continuous variable using the median intake in each category (e.g. 7, 13, 17 and 22 g/day).

The adjusted results in Table G.3 show that, compared to group 1, the total-cholesterol:HDL ratio is on average 0.17, 0.36 and 0.48 lower in groups 2, 3, and 4, respectively. That is when fibre intake is increased from a median of 7 to 22 g/day, the decrease in total-cholesterol:HDL is on average 0.48, which is about 13% of the mean ratio (see Table G.1).

In this example it appears that the assumption of linearity may be justified. Going back to Table G.2, the adjusted slope was −0.35 per 10 g-increase in fibres. The difference in intake between group 1 (median of 7 g/day) and group 4 (median of 22 g/day) is 15 g of fibre, which would be a decrease of 0.51 based on the adjusted linear slope. The estimate based on the categorical exposure is a 0.48 decrease, which is in practical terms almost the same number. If the association would not have been linear, this consistency would not have been present.

In terms of dose–response, the p-value for trend in Table G.3 is obtained by entering the categorical variables (4 values) as a continuous variable in the regression model where the median values in each group have been assigned to each observation. For both the unadjusted and adjusted value, the *p*-value for trend is 0.01 compared to 0.02 and 0.005 for the unadjusted and adjusted values in Table G.2, where the exposure variable was entered as a continuous variable in the regression model. Generally, the continuous exposure estimate is a stronger test, which is reflected by a lower *p*-value for the continuous exposure variable in this case.

To complete this example, we end with some general conclusions:

Why are continuous exposures sometimes modelled as categorical in observational studies?

- **Because the results are easy to explain:** From a public health point of view, it may be easier to explain and communicate that the total-cholesterol:HDL ratio decreases by 0.48 among those with high (> 20 g/day) compared to low (< 10 g/day) fibre intake compared to saying that the ratio decreases by 0.35 per 10 g increase in fibre intake under the assumption that the association is linear across the full exposure range.
- **Simple way to assess deviation from linearity:** In cases where the dose–response relationship is not linear, the categorical presentation of results as in Table G.3 gives a more readily interpretable estimate of the dose–response compared to presenting parameters of a non-linear function along with their confidence intervals. The high between-person variability makes graphical presentation of results a less feasible option compared to other study populations where variability is lower (e.g. controlled studies in inbred experimental animals).
- Suitability for incorporation of the results in dose–response meta-analysis (Crippa et al., 2018; Crippa & Orsini, 2016; Orsini et al., 2012).

What are the limitations?

- **Loss of information:** Collapsing continuous exposures into categorical exposures means loss of information (precision). Fewer categories mean greater loss of precision as the exposure across broader range is assigned the same value within

each category. This generally means that it is more difficult to detect an association and effect estimates tend to be biased towards the NULL.

- **False positives:** One criticism of using categorical exposures is that such analyses are prone to false positives (Bennette & Vickers, 2012). The pairwise comparison relative to a referent exposure category results in $n-1$ number of comparisons when the exposure has been divided into n categories. If correctly done, this problem can easily be avoided by first testing formally if there is an overall dose–response based on a single predefined test. Such a test could be a simple *t*-test as used in Tables G.1, G.2 where exposure was entered as the original continuous variable or grouped continuous variable in the regression model. Alternatively, in Table G.3 an *F*-test testing if all four groups are equal could have been used for the categorical exposure. A nonlinear function could also have been used for the continuous exposure variable. The key issue is to predefine a priori how a single test for dose–response will be done. The level of significance in each category relative to some referent should not be interpreted as a measure of dose–response (in the same way as estimating the NOAEL relative to a control dose group in a toxicological experiment is considered more uncertain compared to identifying a RP by performing a benchmark dose analyses).

Other general considerations

- In the example above the continuous and categorical exposure estimates gave similar results as the association is close to linear. When associations deviate from linearity the use of categorical exposure provide a simple and immediate way of presenting such relationships without the need for formal mathematical functions.
- Often exposure is broken down into equally sized quantiles (tertiles, quartiles of quintiles). The advantage of that approach is that the same number of observations in each quantile gives the same precision across groups. This comes at the expense that the exposure range within each quantile depends on the underlying exposure distribution that differs between different studies, thus reducing their comparability, and generates categories largely differing in their width and likely to be less biologically meaningful (Rothman, 2014). In the example above pre-defined cutoffs were used (not quantiles) and the precision (number of observations) across categories was therefore not equal.
- Use of categorical or quantile exposures in observational studies should not be confused with 'dose-groups' as used in controlled animal experiments. In simple terms, there are no similarities as no dose (exposure) is assigned.