

ADOPTED: The document was endorsed for publication and testing on 24 June 2020

doi: 10.2903/j.efsa.2020.6221

## Draft for internal testing

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

EFSA Scientific Committee,  
Simon More, Vasileos Bambidis, Diane Benford, Claude Bragard, Antonio Hernandez-Jerez, Susanne Hougaard Bennekou, Kostas Koutsoumanis, Kyriaki Machera, Hanspeter Naegeli, Soren Saxmose Nielsen, Josef R Schlatter, Dieter Schrenk, Vittorio Silano, Dominique Turck, Maged Younes, Tony Fletcher, Matthias Greiner, Evangelia Ntzani, Neil Pearce, Marco Vinceti, Laura Ciccolallo, Marios Georgiadis, Andrea Gervelmeyer and Thorhallur I Halldorsson

### 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 of the different epidemiological study designs. It describes key epidemiological concepts relevant for evidence appraisal. Regarding study reliability, measures of association, exposure assessment, statistical inferences, systematic error and effect modification are explained. Regarding study relevance, the guidance describes the concept of external validity. The principles of appraising epidemiological studies are illustrated, and an overview of Risk of Bias (RoB) tools is given. A decision tree is developed to assist in the selection of the appropriate Risk of Bias tool, depending on study question, population and design. The customisation of the study appraisal process is explained, detailing the use of RoB tools and assessing the risk of bias in the body of evidence. Several examples of appraising experimental and observational studies using a Risk of Bias tool are annexed to the document to illustrate the application of the approach. This document constitutes a draft that will be applied in EFSA's assessments during a 1-year pilot phase and be revised and complemented as necessary. Before finalisation of the document, a public consultation will be launched.

© 2020 European Food Safety Authority. *EFSA Journal* published by John Wiley and Sons Ltd on behalf of European Food Safety Authority.

**Keywords:** Epidemiological studies, evidence appraisal, risk assessment, exposure assessment, hazard characterisation

**Requestor:** EFSA

**Question number:** EFSA-Q-2019-00199

**Correspondence:** [sc.secretariat@efsa.europa.eu](mailto:sc.secretariat@efsa.europa.eu)

**Panel members:** Vasileos Bambidis, Diane Benford, Claude Bragard, Thorhallur I Halldorsson, Antonio Hernandez-Jerez, Susanne Hougaard Bennekou, Kostas Koutsoumanis, Kyriaki Machera, Simon More, Hanspeter Naegeli, Soren Saxmose Nielsen, Josef R Schlatter, Dieter Schrenk, Vittorio Silano, Dominique Turck and Maged Younes.

**Acknowledgments:** The Panel wishes to thank the members of the Working Group on Epidemiological studies: Thorhallur I. Halldorsson (Chair), Tony Fletcher, Matthias Greiner, Susanne Hougaard Bennekou, Evangelia Ntzani, Neil Pearce, Marco Vinceti, for the preparatory work on this scientific opinion, Henning Thole (Hearing Expert) and Marco Zeilmaker (Hearing Expert) for their input, and the EFSA staff members Laura Ciccolallo, Marios Georgiadis and Andrea Gervelmeyer for the support provided to this scientific opinion.

**Suggested citation:** EFSA Scientific Committee, More S, Bambidis V, Benford D, Bragard C, Hernandez-Jerez A, Bennekou SH, Koutsoumanis K, Machera K, Naegeli H, Nielsen SS, Schlatter JR, Schrenk D, Silano V, Turck D, Younes M, Fletcher T, Greiner M, Ntzani E, Pearce N, Vinceti M, Ciccolallo L, Georgiadis M, Gervelmeyer A and Halldorsson TI, 2020. Draft for internal testing Scientific Committee guidance on appraising and integrating evidence from epidemiological studies for use in EFSA's scientific assessments. *EFSA Journal* 2020;18(8):6221, 83 pp. <https://doi.org/10.2903/j.efsa.2020.6221>

**ISSN:** 1831-4732

© 2020 European Food Safety Authority. *EFSA Journal* published by John Wiley and Sons Ltd on behalf of European Food Safety Authority.

This is an open access article under the terms of the [Creative Commons Attribution-NoDerivs License](https://creativecommons.org/licenses/by-nd/4.0/), which permits use and distribution in any medium, provided the original work is properly cited and no modifications or adaptations are made.



The EFSA Journal is a publication of the European Food Safety Authority, an agency of the European Union.



## Table of contents

Abstract.....	1
1. Introduction.....	4
1.1. Background and Terms of Reference as provided by the requestor.....	4
1.2. Interpretation of the Terms of Reference.....	4
2. Audience and degree of obligation.....	5
3. Data and methodologies.....	5
4. Assessment.....	5
4.1. Introduction on epidemiological studies.....	6
4.1.1. Descriptive epidemiological studies.....	7
4.1.2. Analytical epidemiological studies.....	7
4.1.2.1. Experimental studies.....	7
4.1.2.2. Non-experimental epidemiological studies.....	8
4.1.3. Epidemiological studies in animals.....	11
4.1.4. Epidemiological studies in plants.....	11
4.1.5. Cause and effect.....	12
4.1.5.1. Existing frameworks on causality.....	12
4.1.5.2. Experimental studies and causality: strengths and limitations.....	13
4.1.5.3. Observational studies and causality: strengths and limitations.....	13
4.2. Key epidemiological concepts relevant for evidence appraisal.....	14
4.2.1. Reliability of studies.....	14
4.2.1.1. Use and interpretation of measures of frequency and measures of association.....	14
4.2.1.2. Exposure assessment.....	16
4.2.1.3. Statistical inference for effect measures in epidemiological studies.....	17
4.2.1.4. Systematic error (bias).....	17
4.2.1.4.1. Information bias.....	18
4.2.1.4.2. Confounding.....	18
4.2.1.4.3. Selection bias.....	19
4.2.1.5. Effect modification/Interaction.....	20
4.2.2. Relevance.....	20
4.2.2.1. External validity.....	20
4.2.2.2. External vs. internal validity.....	21
4.2.2.3. Summary and conclusions.....	22
4.3. Appraisal of epidemiological studies.....	23
4.3.1. Background.....	23
4.3.2. Appraisal principles.....	24
4.3.3. Appraisal and RoB Tools: Development and overview.....	25
4.3.4. Customisation of the study appraisal process.....	31
4.3.5. Use of RoB tools for assessing individual studies.....	31
4.3.6. Assessing risk of bias in the body of evidence.....	34
4.3.6.1. Overall assessment of the risk of bias.....	34
4.3.6.2. Using causal inference by triangulation.....	35
4.4. Use of epidemiological evidence for specific scientific assessment questions.....	35
5. Conclusions.....	35
6. Recommendations.....	36
References.....	37
Glossary.....	44
Abbreviations.....	47
Annexes.....	49
Annex 1 – Bradford Hill viewpoints.....	49
Annex 2 – Hypothesis Testing vs. estimation.....	50
Annex 3 – Random Error and Statistical Precision.....	52
Annex 4 – Recent and relevant inventories and reviews of critical appraisal tools.....	53
Annex 5 – Appraisal of different studies using a risk of bias tool.....	57
Annex 6 – Categorisation of continuous exposures: analysis and interpretation.....	79

## 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 randomized 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 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 characterized 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:

- 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:

- Infosessions
- 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, i.e. nutrition, toxicology, animal and plant health as well as biological hazards. It requires a good understanding of the strengths and weaknesses 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 Term of Reference (ToR)**.

To address the **second ToR**, the guidance will explain (1) how to make inference on the three main types of bias, 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. Furthermore 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 the setting of reference values or health-based guidance values, 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.

In terms of scope, this guidance will cover experimental and observational studies with humans, animals and plants as target populations, excluding studies of laboratory animals and in-vitro studies. Where relevant guidance documents already exist, e.g. OECD guidelines, the guidance will refer to these.

## 2. Audience and degree of obligation

This guidance is aimed at all those contributing to EFSA assessments and 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. The document provides guidance on the general principles of the appraising epidemiological evidence, but 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 (Rothman et al., 2012) and animal (Dohoo et al., 2009) and plant disease (Madden et al., 2007; Cooke et al., 2006) epidemiology, as well as textbooks covering more focused topics including nutritional epidemiology (Willett, 2013). Published papers, book chapters and reports from the biomedical literature are referred to where appropriate in support of arguments, statements and examples.

Concerning methodology, for this guidance an extensive literature search or systematic review of the epidemiological literature was not considered relevant, as 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 by the working group to focus the guidance on these needs. Further clarifications of panel needs were obtained where needed.

## 4. Assessment

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 four-step approach

(plan/carry out/verify/report) for their fulfilment (EFSA, 2015a), 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 predefined 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, 2018a). 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, 2017a), 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 the causality for exposure-health associations.

The Weight of Evidence (WoE) guidance document (EFSA Scientific Committee, 2017b) 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. In line with the concepts and approaches set out in these guidances, this document can be considered an extension that addresses specific needs of guidance on appraisal of evidence from epidemiological studies.

#### 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 analyzing and collecting 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 and 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 but also the study of any health-related state, including distributions of biomarkers (or surrogate endpoints) in the broadest sense. Indeed, epidemiology provides a set of tools and methodologies to describe outcomes of interest in well-defined populations. Outcomes can be of a variety of types (e.g. infection, disease, immunity, presence of a characteristic, death etc.). Studies on health-related states may also cover outcome distributions where a direct link with health is not always well characterised. The growing number of studies in humans and animals linking environmental exposures to the composition of gut microbiota (Clemente et al., 2012; Snedeker and Hay, 2012) is one example of such studies.

The Dictionary of Epidemiology definition is relevant for the study of health-related states 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 methodological challenges also exist for non-experimental studies, regardless of the study population.

Epidemiological studies can be broadly classified as either descriptive or analytical. In descriptive studies, patterns of health-related states are described across one or more factors, such as over time and place, while in analytical studies, relationships between identifiable factors and health-related states are quantified. Analytical studies can be classified as either experimental or non-experimental studies, with the latter often referred to as observational studies.

#### 4.1.1. Descriptive epidemiological studies

Descriptive epidemiological studies have the objective of describing the occurrence of exposure or outcome in a population (e.g. humans, animals, plants and the same methodology may even apply to non-living items like food) over factors such as time, gender or 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. An example of such studies of relevance for EFSA are reports on trends and sources of zoonoses, zoonotic agents and food-borne outbreaks (e.g. EFSA and ECDC, 2018); reports on antimicrobial resistance (EFSA and ECDC, 2020); surveys for plant harmful organisms ('pests') relevant to the EU's plant health policy for which EFSA provides survey data sheets (EFSA, 2020b) and reports on pesticide residues in food (EFSA, 2020c) and National Dietary Surveys (EFSA, 2018b).

#### 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.

##### 4.1.2.1. Experimental studies

Experimental studies are primarily confined to experiments where the exposure conditions are modified 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 the 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, neighbourhoods, farms and villages. Cluster randomisation may be chosen to serve convenience, to avoid departures from treatment, or to assess group effects. At least for human studies, complexity and overhead in terms of recruiting a sufficient number of 'units' is generally less when the 'unit' is the individual, compared to when the 'unit' is the cluster.<sup>1</sup> Examples of experimental studies of randomised design include pharmaceutical trials, nutritional (or supplemental) trials and toxicological studies in experimental animals. The sample size needs to be sufficiently large to cover all the confounders and effect modifiers identified. In humans, variability in lifestyle and genetic factors is generally high. As a result, relatively large sample sizes are needed, compared to some experimental studies in animals.

In RCTs, random treatment allocation alone is, however, not sufficient to achieve valid results. Blinding the investigator and caretakers (e.g. in the case of children and animals) to treatment assignment 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 human participants, being blinded to the treatment received is equally important to avoid bias from selective dropout, changes in behaviour or departures<sup>2</sup> from the assigned treatment (Dodd et al., 2011, 2012). When both the investigator and participants are blinded to treatment, these studies are referred to as **double blind RCTs**. If appropriately designed and conducted, they should 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 nutritional, lifestyle and other **preventive interventions** that test

<sup>1</sup> Observations within the same cluster tend to be correlated. Therefore, in principle, the same number of clusters as compared to individual may need to be randomised to achieve balance across intervention groups. Recruiting clusters into a study is generally more complex than recruiting individuals.

<sup>2</sup> Departures here include participants' non-adherence to treatment (not complying or simply doing something else) or changes in the prescribed treatment by the investigator.

the efficacy of treatments aimed at reducing risk to health. For example, for many nutrients including fish oils, salts, nutritive and non-nutritive sweeteners and most foods, participant blinding is difficult or impossible to achieve (but blinding of the investigator can often still 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 departures from the assigned treatment. Another design is the **randomised crossover trial** where each participant receives both (or all) treatments in a random 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 which it is assigned. This design is often used when comparing the effects of different treatments on clinical biomarkers such as blood pressure or 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.

**Preclinical trials:** By convention, Phase 0, I and II clinical trials in humans are designed to assess safety, pharmacokinetics and -dynamics of pharmaceuticals in humans prior to conducting 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 microdoses 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. Furthermore, 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 subgroups (patients) and they provide valuable information on both pharmacokinetics and tolerability (acute toxicity) of the test substance. 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 rarely achieved. As a result, bias (confounding) may occur. **Phase II trials** are designed to test therapeutic doses of the test substance often in sensitive individuals (patients). These trials are of smaller scale (sample size) than Phase III trials (RCTs) and vary by design in terms of use of controls (no controls, currently preferred medication or placebo). Although phase 0, I and II trials are mostly used in relation to testing of pharmaceuticals, these designs (or variants of these designs) often cover exposures falling under EFSA remit. An example of such studies include 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 substances at low doses, such as for vitamins and antioxidants, have shown unexpected harmful effects (The Alpha-Tocopherol Beta Carotene Cancer Prevention Study Group, 1994; Lippman et al., 2009; Kristal et al., 2014). 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 has no control over the circumstances or 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) 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 derived 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 nmol/L perform better on physical function tests than those with vitamin D status below 50 nmol/L. Such an association may be confounded by the simple fact that those participants who are healthier (less frail) may spend more time outdoors. In practice, it is usually impossible to record and fully account for all factors that may influence the outcome. However, by replicating findings across different study populations and with support from other experimental findings *in vivo* and/or *in vitro*, a stronger case for or against causality can be made.

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

In **cohort study designs**,<sup>3</sup> a source population is defined, and participants (the study population) are recruited and classified according to their exposure. 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 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. After sufficient follow-up time, depending on the induction period of the outcome (or disease), it can then be examined if the exposure may have contributed to the development of the outcome. In general, cohort studies are more expensive 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 in order to test many different hypotheses, covering many exposures and outcomes, including rare diseases, that are studied over time (e.g. the Avon Longitudinal Study of Parents and Children,<sup>4</sup> Biobank UK<sup>5</sup>).

**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 (controls) are recruited from the same source population. Thus, a case-control study involves studying cases (from the source population over the risk period) and a sample of non-cases. The distribution of past or current exposures among cases and non-cases (controls) is then compared, taking confounding factors into consideration. 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 – over a defined period of time – are included in the study, whereas 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.

Cohort studies can be **prospective, historical or combination of both** (ambispective/ambidirectional). 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

<sup>3</sup> 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.

<sup>4</sup> <http://www.bristol.ac.uk/alspac/>

<sup>5</sup> <https://www.ukbiobank.ac.uk/>

relevant exposures occurred. In ambispective or ambidirectional 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).

Similarly, case-control studies may be based on historical records or may involve interviews about historical 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, 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, i.e. reverse causation. However, this is a problem in some but not all case-control studies.

Assessing the exposure prospectively prior to the occurrence of the outcome should guard against reverse causation. It is, however, a common misunderstanding that case-control studies are always of lower quality. 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. If past exposures can be assessed in such manner, with appropriate temporal separation in relation to the outcome assessment, the risk of bias due to the exposure assessment is most often at least comparable to that of a prospective design. Thus, for case-control studies the risk of bias is largely determined by how and when the exposure was assessed (retrospectively based on participant recall, cross-sectional or assessment of past exposures from high quality records). The case-control sampling is irrelevant in this context and it only relates to how participants were recruited.

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 are recruited at one specific point in time, and information on both outcome and exposure is ascertained simultaneously. By design, it is usually 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 may influence the exposure, resulting in reverse causation. The risk of such bias strongly depends on the health state under consideration, and this has to be evaluated on a study by study basis. Despite this methodological flaw, a cross-sectional design may often be appropriate, such as in cases when exposure has rapid short-term effects. For example, for relatively rare exposures such as consumption of glycyrrhetic 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 for risk factors that do not change (e.g. blood type, genetic factors, etc.).

Finally, in **ecological studies**, the units of observation are groups of participants, defined for example by region or community (i.e. populations of countries). Health-related states and exposures are measured across units, 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 (both in terms of exposure and in terms of outcome) cannot be ascertained, the 'ecologic fallacy' may be produced by the fact that the relationship between averages of population exposures and outcomes may not represent the relationship between exposure and outcomes at the individual level. Very specific sources of bias should be assessed for ecological studies. However, despite these potential problems, in rare cases ecological studies have the advantage of achieving large exposure gradients, as exposure to certain nutrients or contaminants are generally greater across different units of observation than within individual units (Weisskopf and Webster, 2017).

In summary, each of the observational study designs reviewed above have their strengths and weaknesses. For certain exposures and outcomes, cross-sectional and ecological study designs can provide relevant information for risk assessments to complement to other lines of evidence.

### 4.1.3. Epidemiological studies in animals

The 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 animal 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 effects, etc.) requires consideration in the design of the study and use of appropriate statistical methodology when analyzing its results. Studies on wildlife are typically 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 denominator in measures of disease frequency. Population size, on the other hand, is usually not a challenge in observational studies done within 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 animal 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 and Heesterbeek, 2000), and invasive species and entomology (e.g. Cock and 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 disease epidemiology, and to a lesser extent animal disease, 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 550 plant species (EFSA, 2020d). Despite this

difference, the One Health concept, which has been used to unify human and animal disease 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. Uwamahoro et al., 2018).

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 confounded by the limited number of epidemiological studies in plant health from which to synthesise information. Although 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 4.1.5.2–4.1.5.3. 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 (Rothman et al., 2008; Vandenbroucke et al., 2016). Theoretical frameworks include the simple but much cited viewpoints formulated by Austin Bradford Hill (1965); and more elaborate theoretical framework such as the Sufficient-Component Cause Model (Rothman et al., 2008). The subject of causality has also been elaborated by Pearl (2009) and Pearl and Mackenzie (2018).

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 study. 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 observation, 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?

The nine viewpoints and the questions to answer when assessing if they are met are listed in Annex 1 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 WoE (Meek et al., 2014). They can also be used to assess the WoE that supports an investigated Adverse Outcome Pathway, and for making a judgement on how strong the evidence is to support a particular investigated mode of action (Gross et al., 2017). The Bradford Hill viewpoints have also been used to

develop approaches to evaluate the confidence in a whole body of evidence when making inferences of causality (GRADE 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.

#### **4.1.5.2. Experimental studies and causality: strengths and limitations**

Experimental studies (also named 'intervention studies'), when they are feasible, 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 optimal 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 risk of bias is low meaning that double blindness can be achieved, compliance is high, and dropout is low. These conditions can more easily be met in nutritional interventions with vitamins, minerals and 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 very modest and far from the goals set out for dietary changes in the beginning of the intervention. In such studies, compliance may decrease considerably over time, thus hampering the reliability of long-term intervention studies. Before concluding that results from such trials provide 'no' or 'limited' evidence (Zeraatkar et al., 2019), relevant questions often not asked are if one would expect such modest changes to have a measurable effect on the outcome under study, as compliance was good, if the intervention lasted enough to induce some effects, and if participants (and investigators) were blinded to the treatment allocation. That is, despite superiority in design, 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, in the end, weaken evidence for or against causality.

#### **4.1.5.3. Observational studies and causality: strengths and limitations**

Compared to experimental studies, observational studies are more prone to bias, particularly confounding. It is therefore more difficult to make strong statements on causality based on their results, particularly if based on a single or very few studies. Replication of findings from different study populations, where confounding factors due to lifestyle habits may differ, and taking other lines of evidence into consideration is usually needed to build a stronger case for causality (EFSA, 2017b). It is also a common belief 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 such as smoking, whereas case-control studies often do have this information. Thus, if the case-control studies indicate that there is little or no confounding by smoking, then this strengthens the evidence from the cohort studies, and indicates that any observed increased risks are unlikely to be due to confounding.

There are also examples where observational studies can be considered better suited than experimental studies to identify causal relationships, such as when assessing the safety of food supplements, food additives, pesticides or pharmaceuticals post-marketing. One famous example is the marketing of oral contraceptives in the 1960s. 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 lack power to detect. Based on these findings, the ethinylestradiol dosage in these pills was reduced substantially, which was associated with less side effect 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 that is conducted and analyzed in a way that minimises risk of bias and chance findings is said to have 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 on main sources of biases. A more practical application of these concepts is then introduced in Section 4.3.

### 4.2.1. Reliability of studies

#### 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 **binary outcome measures** in epidemiology are the prevalence and incidence:

- **Prevalence** refers to the proportion of cases in the 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.

Prevalence and incidence rate are useful measures for describing how frequent a given outcome occurs (at a certain point in time) and the rate at which it is occurring (over time). For frequency measures, the more common approach in epidemiology is to use rate ratios or risk ratios as measures of effect. The most common measures are explained in **Box 1**. More complete descriptions can be found elsewhere (Dohoo et al., 2009; Rothman et al., 2012).

#### **Box 1:** Measures of effect for frequency outcomes

**Measures of effect** are indexes that summarize the strength of the link between exposures and outcome. Effect measures can be expressed in both relative and absolute terms

**Measures of effect 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 ( $p$ ) of the event occurring in group 1 and 2 is then:

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

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

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

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

When the probability of an event is time dependent (changes over time) the effect measure becomes the **rate ratio (also sometimes called the hazard ratio)**: That is, the number of new cases (events) occurring divided by the number of person-years at risk (for example, if 10 people are each followed for 10 years, this involves 100 person-years of follow-up)

Then the **rate ratio** is defined as

$$\text{Rate Ratio} = \frac{\lambda_1}{\lambda_2}, \text{ where } \lambda_1 \text{ and } \lambda_2 \text{ is the rate in groups 1 and 2, respectively}$$

The risk ratio refers to the proportion of participants who experience the event over the follow-up period, whereas the rate ratio takes into account the person-time of follow-up. 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 risk measures** such as 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 usually involve studying all 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 be estimated based on how subjects are recruited. The outcome measure in a case-control study is therefore the 'odds ratio'. That is, the odds of an event occurring is the probability of the event ( $p$ ) divided by the probability of no event ( $1 - p$ ). The odds ratio is then the odds of an event in group 1 divided by the odds of the event in group 2:

$$\text{Odds Ratio, OR} = \frac{\text{odds}_1}{\text{odds}_2} = \frac{\frac{p_1}{1-p_1}}{\frac{p_2}{1-p_2}}$$

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 exaggerates the estimate of the relationship between exposure and outcome. However, for rare outcomes (< 10%), the value of the OR is not too different from the RR.

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 waste water 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 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 4.00 (OR is 4.01). This effect size reflects a 'dramatic' shift in an outcome that substantially impact on the families involved.

To take another example, let us say that in a randomised controlled trial 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 is 34.5%, but 28.8% among controls (placebo). The risk difference here is 5.7% which on an individual or population basis could be considered as biologically relevant. However, in terms of interpretation, this outcome should be reversible and of modest clinical relevance in the short (but not long-) term. The RR here is only 1.20 (OR is 1.32).

To conclude, absolute risk measures are the most relevant measure 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 the absolute relevance 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.

The approach of modelling absolute risk is also used in one particular tool in risk assessment: benchmark dose modelling (BMD) (EFSA Scientific Committee, 2017c). This approach was developed for toxicological studies with different groups of 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%, can be read across. When applied to multiple chemicals, a standardised result can be reached: what dose causes 5% of the animals to show this adverse health effect. 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, 2010). BMD modelling is useful as a tool for setting health-based guidance values but is not of itself a tool for assessing causality, which needs to be done based on integrating the strands of evidence from multiple studies.

#### 4.2.1.2. Exposure assessment

In controlled experimental animal studies, the investigator usually has full 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 rare mistakes. In humans, similar control over exposure conditions may be achieved in highly controlled metabolic trials that can, for ethical and practical reasons, at most stretch over a few days. For other experimental studies, including many RCTs, the investigator has much less control, as exposure is only assigned and partly 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). In most RCTs, some departures are more the rule than the exception and depend on the complexity or intensity of treatment. Therefore, compliance to the treatment allocation should be carefully measured, whenever possible, throughout the experiment. Exposure misclassification due to departures from randomisation can be differential or non-differential, the former potentially leading to biased results without knowledge of the direction of the bias, while the second generally tends to distort the measured effects towards the Null. In contrast, non-differential misclassification cannot create an association and cannot be adjusted/accounted for.

In observational epidemiologic studies (of humans, animals or plants), the investigator has no control over 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 assume that its duration and amount are consistent with a causal effect, that is biologically plausible. Most often, the longer the time period that the exposure assessment covers (or can be assumed to cover), the better it is, at least for diseases with a long induction and latent period. 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 cardiovascular disease. 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, in relation with short-term endpoints. 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 reliable 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, risk of bias 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. However, this objective biomarker has not always led to new major achievements in our understanding of the causal relationship between smoking and health.

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 optimal 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 an acceptable estimate, with only a limited amount of systematic bias. By considering the strength and the limitations of the methods applied for exposure assessment, for example through careful validation of the assessment method, a more appropriate use of the available evidence can be made in the risk assessment process.

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 the pros and cons of this approach are discussed in some detail in Annex 6.

#### 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. In order to make inferences about such parameters, uncertainties around the statistical (or central) estimate need to be considered. This is done by estimating confidence interval which accounts for random measurement errors<sup>6</sup> 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 a growing consensus that significance testing involving arbitrary cut-points (e.g.  $p < 0.05$ ) is not appropriate. For further discussion on this issue, the reader is directed to Annex 2. 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 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. This phenomenon has been discussed in the context of the reproducibility crisis in science (Ioannidis, 2005). A similar bias would result from a selection process of published studies for evidence integration (e.g. systematic reviews). 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. A worldwide voluntary pre-registration of studies involving animals has been launched recently (Bert et al., 2019). A preregistration of observational epidemiological studies can be assumed to have similar positive effects but is not yet established.

#### 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

<sup>6</sup> Random errors in experimental 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 Annex 3).

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 misclassification 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. In most cases, non-differential misclassification of exposure biases the relative risk estimate towards the NULL. In most cases, non-differential misclassification tends to reduce the size of the effect which is of particular concern in studies which find a negligible association (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 nested case-control study of lung cancer, with a control group selected from among members of the cohort without the health outcome of concern, the recall of past exposures in controls might differ from that of the cases, leading to differential misclassification. This could bias the odds ratio towards or away from the NULL (value of 1.0), depending on whether members of the cohort who did not develop lung cancer were more or less likely to recall such exposure than the cases (Pearce, 2005).

#### 4.2.1.4.2. Confounding

The term 'confounding' describes a mechanism that may result in invalid inferences about an investigated risk factor. 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 typically away from the NULL (towards significant effect). 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<sup>7</sup> in the interpretation of the study results.

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. RCTs 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 due to small study size, uneven dropout or compliance across treatment groups. In observational studies, where exposure to the main study factor is not controlled by design, 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 (more precisely generalised linear models with link function suitable for the given response variable) 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

<sup>7</sup> Residual confounding is the distortion that remains after incomplete confounder control in the design and/or analyses of the study.

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. For example, in CVD epidemiology, a different set of confounding variables will be required compared to infectious disease epidemiology.

While the technique of matching in cohort studies can be used to prevent confounding from occurring, in case-control studies, it may also lead to the opposite result, i.e. introduce confounding, i.e. 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) 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).

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 confounder 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 (Rothman et al., 2012). 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 et al. (1992a,b,c).

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. over-representation 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 of 3.0 in men and 1.5 in women, gender is an effect modifier because the effect of the exposure is 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, according 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, while the second deals with the possibility that the association we find between a risk factor and endpoint is not actually causally related to that factors, but derives from factors co-varying with the primary exposure under investigation. Therefore, understanding of effect modification is necessary to characterise causal association, interactions and susceptibilities, thus being important in risk assessment. Confounding on the other hand is a bias that must be minimised when planning a study or removed at the analysis stage.

Effect modification may be assessed either as statistical interaction or biological interaction (Rothman et al., 2012). Statistical interaction is just departure from the basic form of a statistical model and is therefore dependent on the metrics used in the statistical model. By changing the scales of measurement, one can introduce or remove statistical interaction. 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 endpoint. 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.

#### 4.2.2. Relevance

Generally, epidemiological studies conducted in the target species have clear advantages in terms of relevance over studies conducted in non-target species, as uncertainties due to between-species extrapolation are eliminated. Second, 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 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 risk 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 (perhaps smaller) 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, then 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 4.2.2.2). In most studies, the 'target population' is left undefined, with the implication that the findings are intended to apply to 'everyone on the planet'. In fact, there is no need to invoke some hypothetical target population to validly design and analyze a study. In more simple terms, generalisability is a matter of scientific inference, not statistical generalisation (Rothman et al., 2012). When studying exposure-health relationships in the target species, 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 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 at the slaughterhouse due to their infection or due to old age (in which case they may have more advanced infections, since bovine paratuberculosis is a chronic disease with slow progression). 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). For all these reasons, evaluations of Sensitivity and Specificity of diagnostic tests for bovine paratuberculosis (or other chronic progressive infectious diseases in animals) using populations in slaughterhouses might lead to spectrum bias, since they have not been conducted in a representative sample of the target population (Nielsen et al., 2011).

Spectrum bias occurs if the distribution of stages or severity among cases ('case mix') included in a study is not representative for the case mix in the source population. This will not create a bias, but it may mean that the study findings apply only to that particular case mix (e.g. severe cases), and not to all cases. For example, most studies of COVID-19 currently involve symptomatic cases (since most people are only tested if they have symptoms), and the findings may not apply to milder asymptomatic cases. Spectrum bias is an important aspect in quality assessments of studies of diagnostic accuracy (Whiting et al., 2003). For example, if a diagnostic test is trialled comparing severe COVID-19 cases with healthy controls, it may perform very well, but the performance may be less optimal in a general population survey involving many mild asymptomatic cases. Although the terminology seems restricted to the area of diagnostic validation studies, spectrum bias should always be considered when subjects are sampled stratified by disease status. In principle, this applies to case-control studies and thus constitutes a relevant mechanism of a selection bias.

#### **4.2.2.2. External vs. internal validity**

In praxis, it is difficult to separate external validity from internal validity (Steckler and McLeroy, 2007). For example, controlled experimental studies have, for reasons explained above, generally lower risk of bias 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. Eliminating uncertainty in terms of external validity means relying on human observational studies, which generally have lower internal validity. Even within the same target population similar opposing forces are often present. For example, RCTs testing the effect of pharmaceuticals or individual nutrients on health are usually conducted in selective

populations that are most likely to benefit from treatment (pharmaceuticals) or demonstrate some treatment effects (nutritional RCTs). Such a selection may, however, hamper extrapolations to the more general population. For example, results from an RCT showing modest increase 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 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, which is associated with more diverse exposure to various other nutrients than beta carotene supplementation alone. Similarly, findings on increased mortality in postmenopausal women with underlying cardiovascular disease following supplementation with vitamin C and E (Waters et al., 2002) could be considered to have modest to low external validity for the general population. In summary, external validity and internal validity are often inversely related (Steckler and McLeroy, 2007) and in terms of making conclusions on causality both factors need to be considered.

#### 4.2.2.3. Summary and conclusions

In Section 4.1, a brief description of different experimental and non-experimental studies was given, highlighting their main strengths and weakness. In this section different types of biases that may occur in each of these designs have been explained. When interpreting and appraising epidemiological studies 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 not always justified as the examples and discussions above have tried to highlight. In fact, all study designs are more (or less) prone to typical biases.<sup>8</sup> However, the basis of such ascertainment is mainly conceptual or anecdotal. 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, or that the results of a study are entirely unreliable based on the existence of some extent of bias. Some typical biases that may occur in different studies designs are briefly summarised in Table 1.

**Table 1:** Study designs and typical biases

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

RCT: randomised controlled trial.

**General note on Table 1:** 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.

- a) Selection bias at baseline is not usually a concern in RCTs if exposure is adequately randomised and allocation is 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.
- b) Selection bias may occur from the way in which the study participants are selected (or select themselves) from the source population (i.e. if selection or response differs by both exposure and disease status). Selection bias can also occur due to loss to follow-up.
- c) Selection bias may occur from the way in which the study participants are selected (or select themselves) from the source population (i.e. if selection or response differs by both exposure and disease status).
- d) Selection bias is not a problem in ecological studies provided that they cover an entire defined population; selection bias can occur if not all of the relevant cases are identified in the population, or only a subgroup has exposure information.
- e) Confounding may occur if allocation concealment and randomisation fail or the study size is too low, thus making treatment groups not comparable at baseline.

<sup>8</sup> <https://www.hkmj.org/system/files/hkm1206p226.pdf>

- f) Confounding can occur in all observational designs.
- g) Information bias on exposure (i.e. exposure misclassification) occurs if the treatment groups 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).
- h) Information bias may occur 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). 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.
- i) Information bias may occur due to misclassification of exposure or the outcome. Information bias may particularly occur 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) and cannot explain 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).
- j) 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 strength of these biases depends very much on the question under study (area of the study), e.g. in some nutritional studies assessing nutrient intake from all foods recall may be different in cases and controls leading to possible reverse causation. For other dietary habits that are easier to recall compared to the whole diet, like consumption of coffee, tea or alcoholic beverages, the risk of such bias may be of much less concern.

### 4.3. Appraisal of epidemiological studies

In bringing together evidence, issues of quality, consistency and complementarity all play a role. This chapter focusses on tools for assessing characteristics of individual studies to enable their quality to be assessed in a thorough and consistent way. Assembling the evidence from multiple epidemiological studies serves three needs; firstly, the hazard assessment, i.e. contributing to the strength of evidence towards the assessment of causality in an association; second, the quantitative data to characterise the exposure–response relationship; third, possible biases can be better assessed by considering multiple studies. Evidence is more convincing if there is consistency in effect estimates across different studies, populations and study designs, particularly if different study designs or settings might create bias in different directions (triangulation, see Section 4.3.6.2). Consistency across studies may be highlighted in the summary of all relevant studies or be formally assessed in a meta-analysis.

#### 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). An established common feature in any type of decision-making process is the efficient retrieval, organisation and integration of the available evidence on a specific question or set of terms of reference (Langlois et al., 2018). In EFSA's risk assessment, the information sources related to the evidence retrieval are numerous and diverse opting for the inclusion of the totality of the evidence. In many assessments, available evidence stems from different epidemiological studies with **unique design attributes and population characteristics**. Case reports and case series, ecological studies, cross-sectional studies, case–control and cohort studies as well as interventional studies (randomised or not) all try to address pertinent questions. All these designs have a place in the risk assessment process (EFSA, 2017b).

The information derived from each piece of evidence, however, is not necessarily equally relevant and each design is possibly prone to different sources of bias as described in the previous section.

Moreover, the number of studies is usually on continuous rise, resulting in an accumulated evidence load of studies of varying design and sample size covering different health outcomes. For example, for data rich substances such as dioxins, perfluoroalkyl substances (PFAS) or bisphenol-A, there are hundreds of published studies addressing putative associations with different endpoint categories ranging from measures of cognitive function, cardiometabolic risk factors to cancer (EFSA CONTAM Panel, 2018). As a result, advanced information retrieval methods may need to be deployed in order to identify, extract and form preliminary clusters of relevant studies. Furthermore, such wealth of evidence often suffers from varying **replication validity**, i.e. the extent to which a postulated association can be replicated by an independent study. Replication validity can often be attributed to different parameters that are either inherent to the assessment question (e.g. exposure assessment limitations, outcome definitions, adopted analytical approach) or field based (e.g. selective reporting, multiple varied stakeholders, flawed use of methods) (Dickerman et al., 2019). Taking into consideration the above and given that the scenario of an evidence base consisting of only a single study is highly unlikely, for a successful integration to be achieved, the constructed evidence base must be organised in a way that assigns an appropriate role to each information piece. Examples of that could be the:

- supportive use of cross-sectional studies in hazard identification,
- use of case-control studies for the identification of potential critical endpoints in the hazard identification process,
- the use of cohort studies in the risk characterisation stage,
- use of randomised nutritional studies in setting daily reference values, establishing a health claim, etc.

The process of assigning such role to each piece or body of evidence is complex and specific to each research 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 EFSA guidance documents (EFSA, 2015a,b, 2017b, 2020c).

#### 4.3.2. Appraisal principles

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

- a) identification of the key elements of the research question under study
- b) assessment of internal validity (risk of bias)
- c) summarisation of the study results.

Clarifying the **key elements of a research 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 that works well is the PICO(T) (**P**opulation, **I**ntervention, **C**omparator, **O**utcomes, **T**ime, if applicable) approach originally implemented for intervention studies. For observational evidence, PICO(T) has been adapted to include the **P**opulation description, the **E**xposure, the **C**omparator, the **O**utcome (endpoint) and follow up **T**ime (PECO(T); Morgan et al., 2019)). For other relevant questions, the **PO** (Population and Outcome: descriptive studies) and the **PIT** (Population, Index test, Target condition: test accuracy studies) can be used for the same purpose (EFSA, 2010). The PICO(T) and its variations provide a structured formulation that can be used to

- designing the literature search strategy,
- identifying the study designs that 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 best comparison among the usually large number of performed comparisons, and
- organise and prioritise the relevant endpoints and follow up time points thereof.

**Internal validity** is the extent to which a piece of evidence provides an unbiased estimate of the causal 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, particularly in terms of the likelihood, magnitude 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 results** of a study basically pertains to the magnitude of the effect and the precision of the point estimate. However, while clarifying what are the main results of the study, various quantitative 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, which can help assessing the influence of the confounders adjusted for.

The next three subsections focus on use of risk of bias (RoB) tools for assessing internal validity of individual studies and on summarising their results in a systematic manner (points b) and c) above). Before giving specific guidance on this issue, a brief description on the development of appraisal and RoB tools and an overview of existing tools is given.

### 4.3.3. Appraisal and RoB Tools: Development and overview

Despite 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. Others have been the result of the efforts by risk assessment organisations, Health Technology Assessment (HTA) programs or governmental bodies which were interested in implementing structured and harmonised approaches in their own assessment processes (BfR, IARC, ECETOC, NIHS R&D HTA Programme, AHRQ, NTP-OHAT, EPA-IRIS, Navigation Guide, USDA-NESR). Currently, there are no agreed gold standards and no rigorous processes for developing such tools (see the table in Annex 4).

Critical appraisal tools have been developed for different purposes and contexts such as to 1) to appraise single studies; 2) to assess risk of bias in systematic reviews or health technology assessments; and 3) to inform the WoE in risk assessments (part of a grading systems). 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](#)),
- tools for different bias domains (e.g. [Cochrane RoB 2.0](#); [NTP-OHAT](#)).

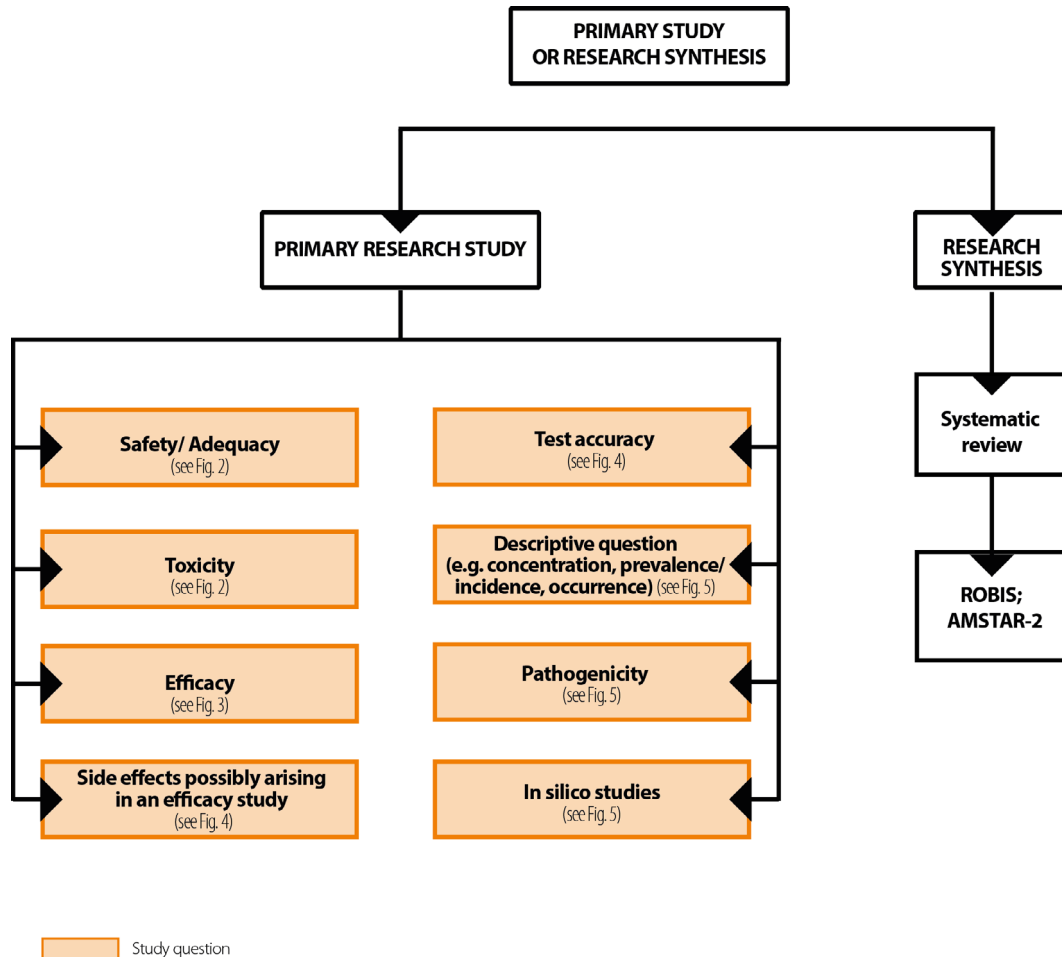
EFSA has piloted the NTP-OHAT tool in several of its scientific assessments since 2015. The Office of Health Assessment and Translation (OHAT) from the National Toxicological Program in the US (NTP) has outlined operating procedures for systematic review and evidence integration for conducting literature-based evaluations in environmental health and toxicology (Rooney et al., 2014). They have recently developed a RoB Tool that applies a parallel approach to the evaluation of risk of bias for human and animal studies, facilitating consideration of potential bias across evidence streams with common terminology and domains (NTP, 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 and Green, 2011), SYRCLE's RoB tool for animal studies (Hooijmans et al., 2014) and the Navigation Guide (Woodruff and 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. These questions are complemented by detailed criteria 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.

As far as ongoing efforts are concerned, the Committee for Evidence-Based Methods in Risk Assessment, which advises the German Federal Institute for Risk Assessment (BfR) on the establishment of scientific standards in the field of evidence-based methods, is currently compiling a transparent tool for the (rapid) quality assessment of observational epidemiological studies (EvaRisk-toolbox).

In Annex 4, a table showing a selection of recent inventories and reviews on critical appraisal tools is provided. This includes a description of their context, objectives and study designs

covered. The purpose of this table is to give an overview of other available tools than those mentioned above, including relevant frameworks and guidance on evaluating study quality.

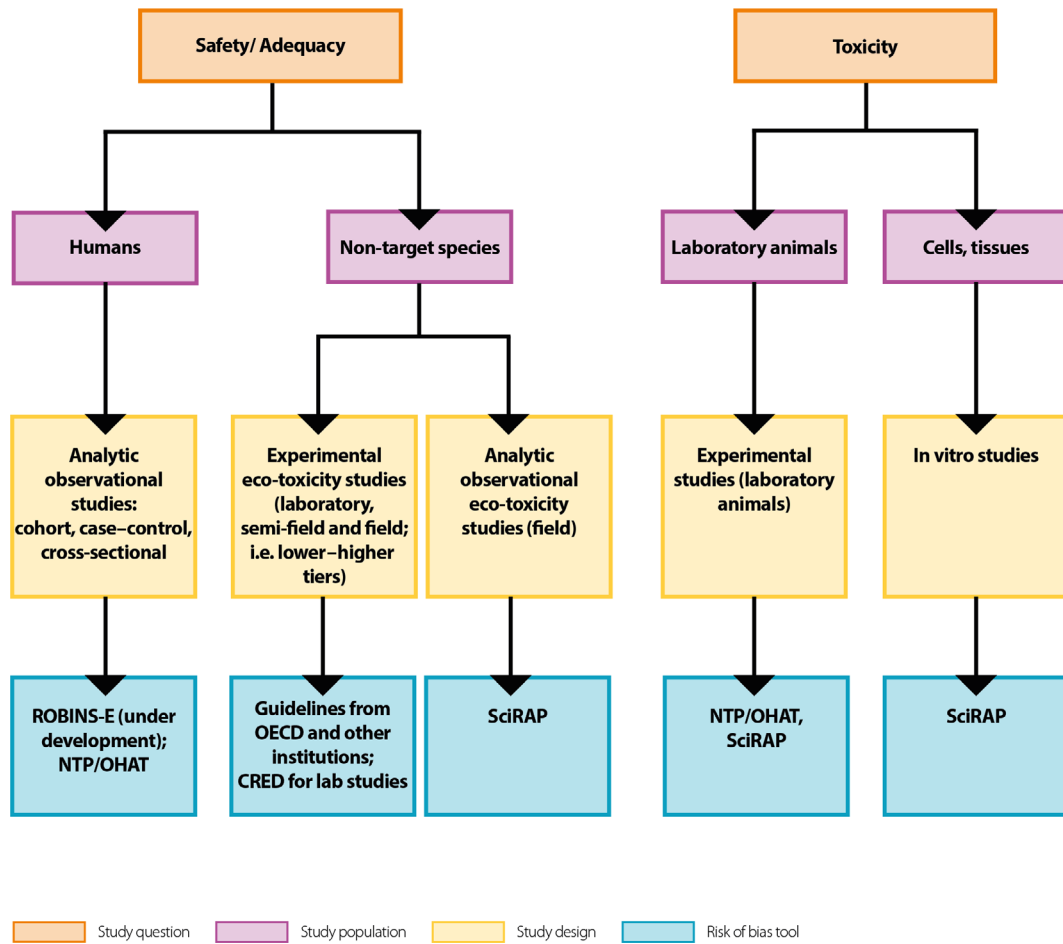
An overview of the RoB tools for research synthesis (i.e. systematic reviews of primary research studies) and for appraising individual primary research studies is provided in Figures 1–5. The tools for the appraisal of primary studies are presented stratified by study question, study population, and study design, allowing the readers to identify the tools available for their specific appraisal task. Where no specific tool is available,<sup>9</sup> 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).



**Figure 1:** Risk of bias tools for appraisal of research syntheses or primary research studies (the RoB tools for these are shown in Figures 2–5)<sup>10</sup>

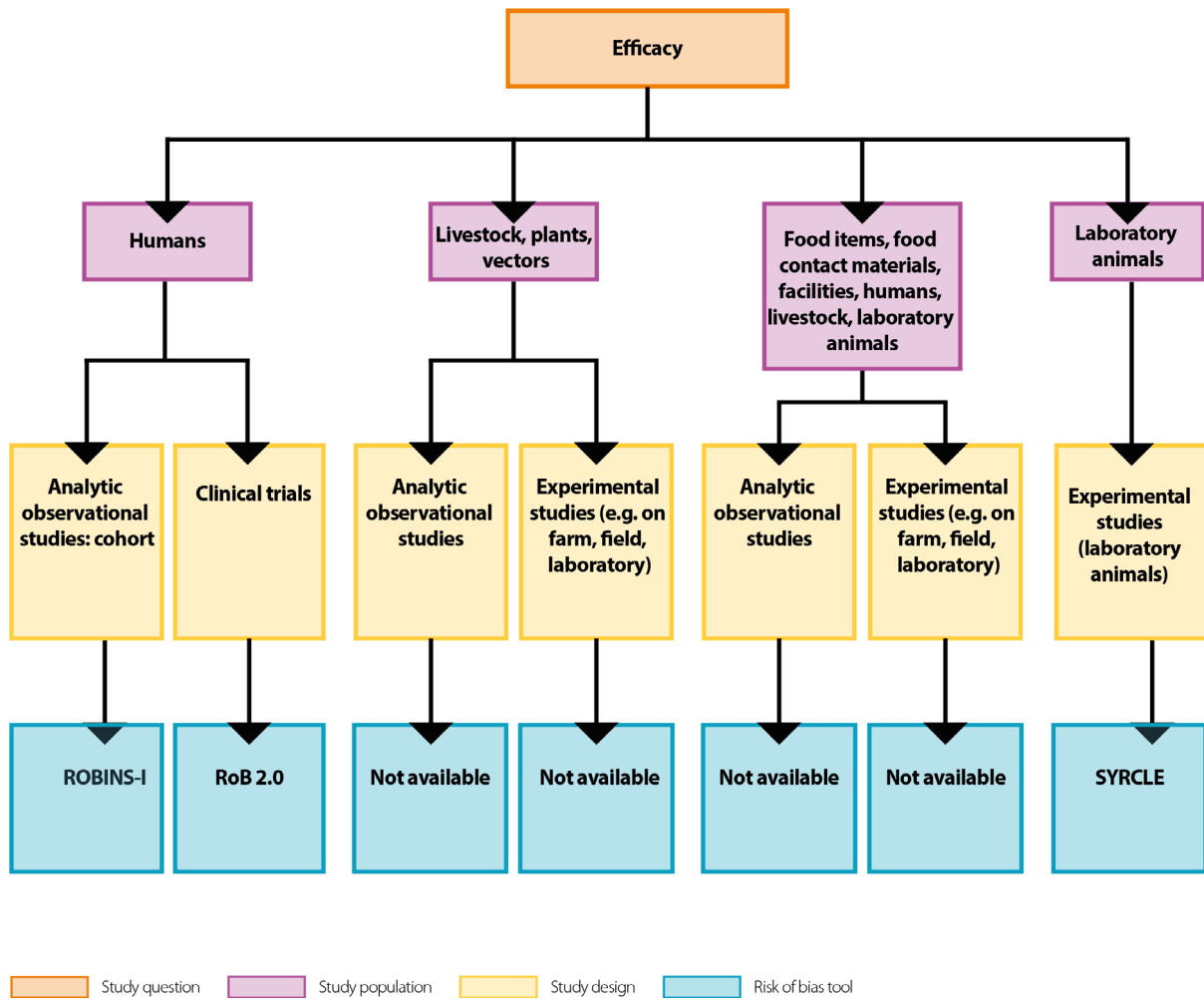
<sup>9</sup> 'Not available' in Figure 1 intends 'not yet available'; note that no comprehensive review has been carried out.

<sup>10</sup> AMSTAR-2, ROBIS



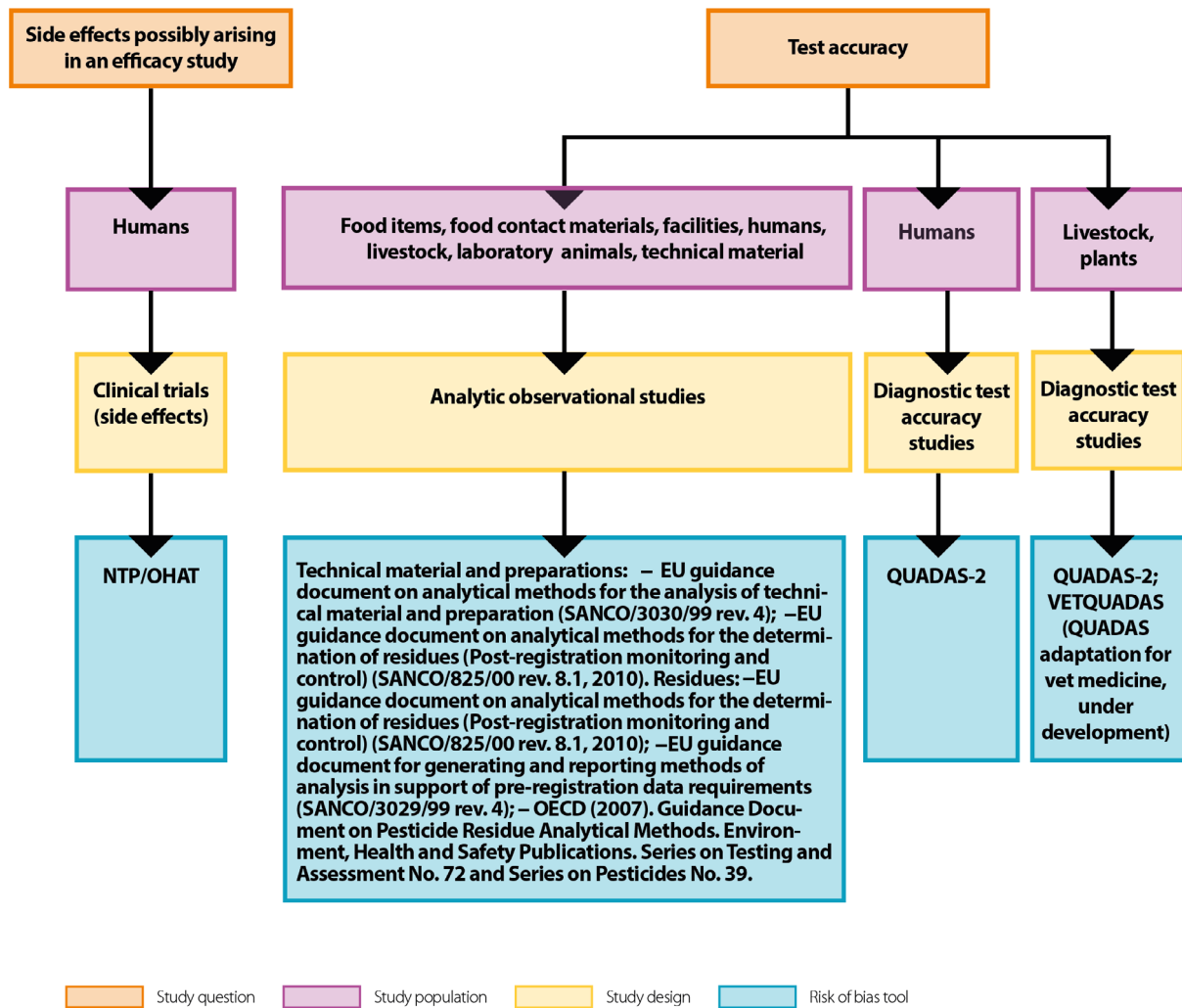
**Figure 2:** Risk of bias tools for appraisal of primary research studies assessing safety/adequacy or toxicity<sup>11</sup>

<sup>11</sup> ROBINS-E, CRED for lab studies, SciRAP, NTP-OHAT



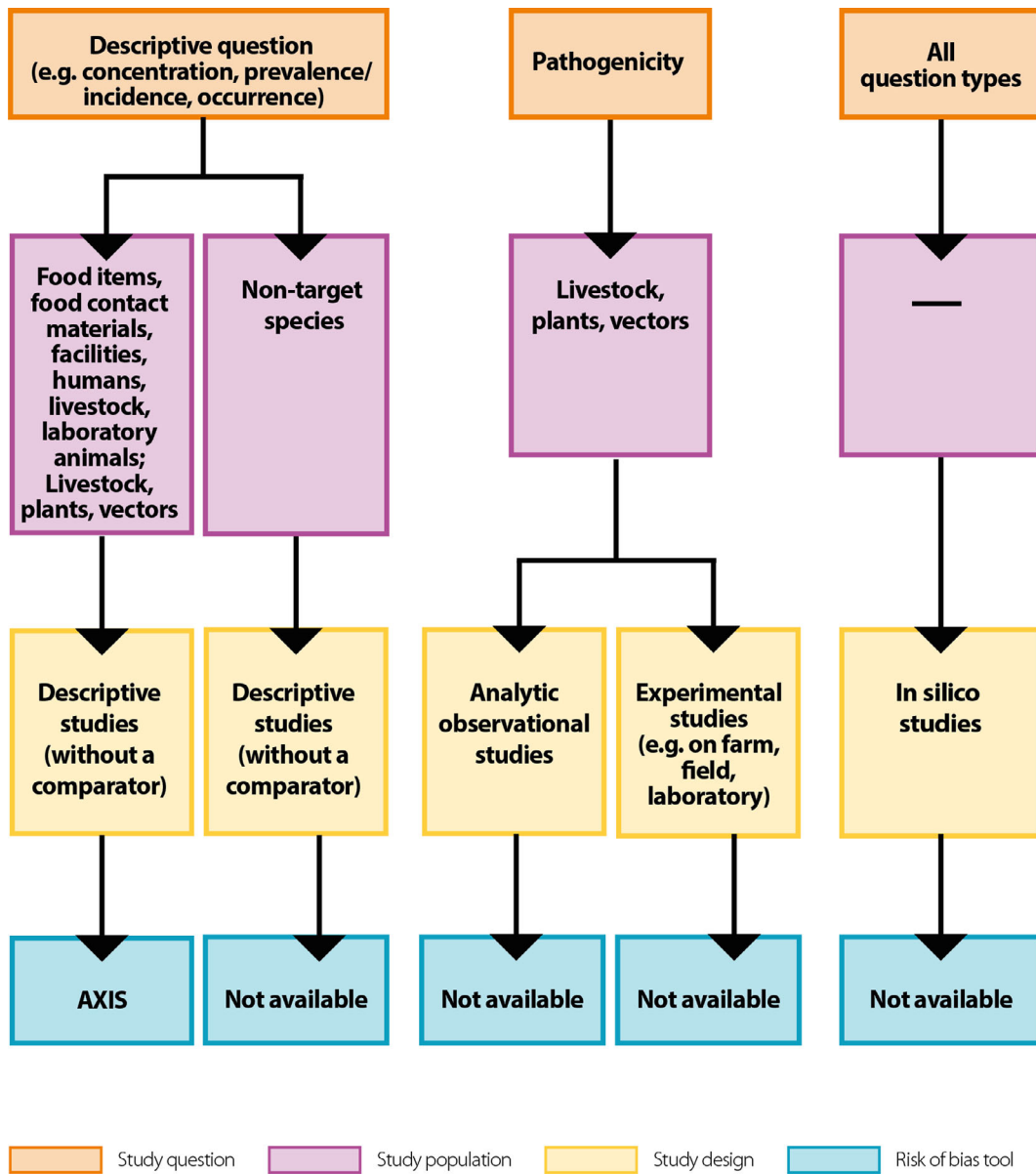
**Figure 3:** Risk of bias tools for appraisal of primary research studies assessing efficacy<sup>12</sup>

<sup>12</sup> ROBINS-I, RoB 2.0, Syrcle



**Figure 4:** Risk of bias tools for appraisal of primary research studies assessing side effects possibly arising in efficacy studies or test accuracy<sup>13</sup>

<sup>13</sup> NTP-OHAT, QUADAS-2, VETQUADAS



**Figure 5:** Risk of bias tools for appraisal of primary research studies assessing descriptive questions, pathogenicity or in silico studies<sup>14</sup>

<sup>14</sup> AXIS

#### 4.3.4. Customisation of the study appraisal process

The study appraisal process can and should be tailored accordingly to serve the specific purpose of the risk assessment it is used for. Following the structure proposed before, the first level of customisation is implemented at the **PICOT/PECOT/PO/PIT** level as elaborately described in the relevant draft EFSA guidance (EFSA, 2010). Predefining the population groups that are to be included in the risk assessment will greatly facilitate not only the appraisal process, through successful clustering of the studies, but also the integration process that follows. For example, with regard to exposure assessment, contemplating upon a compound group and its constituents at the beginning of the appraisal process can readily identify the need to break down the risk assessment effort into a number of sub-questions. In other cases, a mixtures' approach can be decided upon upfront which requires a different assessment of the evidence in terms of exposure. Endpoint definition and categorisation is another important task; adequately capturing endpoint categories will prove valuable during the integration process of different lines of evidence and endpoints (main and surrogate) (EFSA, 2020c).

Having framed the research question and related sub-questions that are relevant to the risk assessment, a decision needs to be made as to which extent the study appraisal is going to be generic or specific to the **study design** and as to whether the same weight and depth will be applied across all study designs. 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. For example, if there is a large number of cohort and/or case-control studies for a given risk assessment, it could be justified to evaluate cross sectional studies as supportive information. That is, to save effort, knowing that the validity of the risk assessment is not jeopardised, the study appraisal of cross-sectional studies could be performed using a smaller number of appraisal items and focusing on the bias domains that emerge as important.

Moving to the customisation of the **risk of bias 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 randomisation and allocation concealment and what should be considered as acceptable compliance on case by case basis.

Exposure assessment is another domain where an 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 assigning weight 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 example for cohort studies in Annex 5). Similar to the exposure assessment, the endpoint (outcome) ascertainment is a feature that should receive the same attention.

Reporting of the **study results** can be overwhelming if left unsupervised during the study appraisal of observational studies. Multiple analyses on numerous compounds and various outcomes can skew judgement and should be thought through in terms of interpretation and properly addressed. Finally, **applicability** items, albeit more subjective compared to the previous stages of the appraisal, are valuable tools for the selection of the critical endpoint.

#### 4.3.5. 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 be aware of the fact that many existing 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 development and use, some customisation (or tailoring) of these tools prior to use is usually needed. Moreover, it should be recognised that there will usually be no 'definitive' study – different studies may have different strengths and weaknesses, or

potential biases in different directions. Thus, one should always consider all of the evidence when assessing causality. Used on the basis of those principles, RoB tools provide a transparent structure for considering different types of bias, which is an important background information for evidence synthesis and assessment of uncertainty. Another point to consider is that, for a particular hypothesis, the issues 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. Cohort studies may not include smoking data, whereas case-control studies may include this – if smoking is not a confounder in the case-control studies, it is also not likely to be a confounder in cohort studies conducted in the same population. On the other hand, if chemical exposure occurs occupationally, then confounding is unlikely to be important (there is little or no confounding in comparisons of different groups of manual workers), but other potential biases (e.g. the healthy worker effect) may be of more concern. Thus, any use of RoB tools should focus on the issues of most concern for the hypothesis and context under study.

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).<sup>15</sup> When applying RoB tools to individual studies understanding the different process that may lead to different biases, their complexity and knowing what to look for is fundamental for facilitating proper use. Below a short, non-comprehensive 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 straight forward 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 risk of bias.<sup>16</sup> 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 recruited) is usually missing. Selection bias can occur **in both study designs** 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 the two groups.
- **Confounding:** Assuming that the randomisation process and study size are appropriate (no confounding from these sources), bias due to confounding in **RCTs** may still occur if there are **deviations from intended treatment**.<sup>17</sup> Such bias may occur if participants or investigators are not blinded to treatment allocation. However, even when blinding is successfully implemented, differential treatment cannot be excluded (e.g. despite attempts to conceal treatment, participant assigned to treatment 'A' may for different reasons stop treatment; find out what the treatment was and do something else; and/or the investigator may find out what the treatment is and that may influence how the participant is treated and/or assessed). The intention to blind participants and investigators can easily be evaluated by study reporting, but how successfully 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. If known confounders are accounted for, confounding due to unidentified factors or improper confounder control (modelling) 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).

<sup>15</sup> 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.

<sup>16</sup> Selection bias as defined here for RCTs would be source of confounding that is deviation from randomization at baseline may result in imbalance in factors that can confound the results for the experiment.

<sup>17</sup> This is often referred to as performance bias.



- **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, if the assessor is not experienced. 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 and USDA Nutrition Evidence Systematic reviews (**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–control studies (unlike NTP-OHAT and USDA-NESR). The purpose of this example is to highlight that formulations used to capture possible selections bias can be quite different. Without further considerations, 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 RoB 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 randomization process?'
- Similarly, the **NTP-OHAT** asks: '1) was administered dose or exposure level adequately randomized, 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 risk of bias in observational studies (Viswanathan et al., 2013; Page et al., 2018)

**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, rather than sticking too tightly to the exact wording of individual 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 context of this section on use of RoB tools in perspective, Annex 5 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.

### 4.3.6. Assessing risk of bias in the body of evidence

#### 4.3.6.1. Overall assessment of the risk of bias

After assessing each 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 risk of bias) or
- numerical scoring.

Scoring here refers to the approach of assigning a numerical value to each risk of bias question that is then summarised in some way (perhaps using different weights) into an overall score. The use of scores (or scales) for assessing quality or risk of bias is currently explicitly discouraged by the Cochrane handbook (Higgins and 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; Schulz et al., 1995; Jüni et al., 1999).

Characterisation of study quality by ranking into tiers is generally more favoured. It is considered more transparent as it usually relies on an a priori selection of few well-defined attributes (bias questions) that determine the overall summary assessment (less data driven). The advantages of using summary assessments by ranking (into tiers) or possibly scoring is that it provides a simple summary of individual studies that is easy to compare across the whole body of evidence. Such summary can be done in the form of heat maps, tables or other formats. This allows for rapid evaluation by the user (and other readers) to assess the overall quality of several studies in terms of risk of bias. Presenting the body of evidence in such a format can be both transparent and helpful in terms of making decision on how to use that evidence later in the risk assessment.

A potential problem of relying on summary assessments by ranking into tiers (and, to a greater extent, scoring) is that judgements on the overall body of evidence should always be made by considering the **type**, and possible **direction and magnitude of** potential **biases** identified across different studies. Summarising risk of bias by tiers (or scoring) 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 risk of bias, 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 risk of bias. 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 (low risk), 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 but quite plausible, and the approach of taking both type and direction of bias into account compared to just looking at the risk of bias scoring can lead to different conclusions. 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.

The considerations reflected on here highlight once again the importance of integrating the assessment of magnitude (where possible) and direction of biases (and their composition) into the overall risk of bias characterisation of the body of evidence.

#### 4.3.6.2. Using causal inference by triangulation

One approach that has been suggested as a possible alternative to the use of RoB tools is causal inference by triangulation.. Some of the Bradford Hill viewpoints 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. For 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 checklist for risk of bias assessment likely would have led to the rejection of both components of the study as being at high risk of bias (albeit in opposite directions).

Triangulation is also consistent with the approach advocated by Savitz et al. (2019) who argue that risk of bias 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).

In summary, there are no universal rules for summarising the risk of bias in the body of evidence, and one cannot rank studies as 'good' or 'bad' generically just based on their study design. Rather, one should consider all the evidence, and take into account the Bradford Hill considerations. If there is reasonable consistent evidence of an increased risk, one needs to consider possible sources of bias, using information from all the relevant studies (for each type of bias), and also using a triangulation approach where possible. One can then make an overall informed judgement, taking all the evidence into account. As with any other scientific field, this involves expert judgement and needs to be made by appropriately qualified experts and committees.

#### 4.4. Use of epidemiological evidence for specific scientific assessment questions

This section will be written after the end of the testing period.

### 5. Conclusions

These are preliminary conclusions pending the finalisation of the guidance document.

- 1) Epidemiology is the study of occurrence and distribution of health-related events, states and processes in specified populations.
- 2) This guidance document is intended to guide EFSA panels and staff in the appraisal and integration of evidence from epidemiological studies for use in EFSA's scientific assessments.
- 3) Appropriately designed and conducted experimental studies in humans can provide an unbiased measure of effect. Yet, such studies are confined to interventions expected to

- provide benefits. In particular, dietary and lifestyle interventions that require substantial participant commitment may suffer from certain types of biases relating to difficulties in achieving blinding or compliance over time.
- 4) In observational epidemiological studies (of humans, animals or plants), the assessment of exposure must rely on validated biomarkers or other pertinent measures. What can be considered as 'acceptable' or 'valid' depends on the exposure and outcome under study. All methods have their limitations but by considering both their strength and weaknesses a more appropriate use of the available evidence can be made in the risk assessment process.
  - 5) Regardless of design, all studies are prone to different sources of bias. Evidence appraisal, including risk of bias assessment, is a necessary step to reach conclusions about the potential causality underlying exposure-outcome associations.
  - 6) Risk of bias tools provide a transparent and structured approach for assessing biases, which is an important background information for evidence synthesis and assessment of uncertainty. Use of such tools generally requires some form of training and tailoring before use.
  - 7) No universal rules for summarising the risk of bias in the body of evidence exist, and one cannot rank studies as 'good' or 'bad' generically just based on their study design.
  - 8) Characterisation of study quality by ranking into tiers can be helpful in preparing a summary of individual studies for comparison across the whole body of evidence. It is, however, important to keep in mind that both within and across tiers potential sources of bias often differ between studies. This has to be considered carefully in any evidence synthesis.
  - 9) Risk of bias assessment is only one of several steps in evidence synthesis and integration. A thorough evaluation of any association or effect reported in a study also requires careful weighing of the actual effect size and statistical precision, its biological relevance, the importance of the assessed endpoint and the absolute relevance of the association for the individual and the community/population.
  - 10) The method of triangulation, which builds on the work of Bradford Hill on causation, provides a structured approach to assess causality in evidence synthesis.

## 6. Recommendations

These are preliminary recommendations pending the finalisation of the guidance document.

- 1) Risk of bias tools have a long history of use for randomised controlled trials 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. animals and plants). It is recommended that EFSA collaborates at the European and international level with relevant organisations and initiatives to harmonise developments in this area.
- 2) Risk of bias tools provide a structured way to identify different biases that may occur to varying degree in different studies. The key elements to capture within each study are the source, magnitude and direction of possible biases, which often differ across studies. That complexity cannot be accurately captured by assigning weights to different bias domains to provide a numerical score of study quality.
- 3) When making the overall assessment of the body of evidence, all evidence should be considered. Judgements on the overall body of evidence should always be made by considering the type, and possible direction and magnitude of potential biases identified across different studies, for example by using a triangulation approach.
- 4) Concerning risk characterisation, benchmark dose modelling is increasingly being used to derive health-based guidance values from epidemiological studies in humans. For this type of evidence, specific guidance is needed to address differences in variability, design and conduct compared to controlled toxicological experiments in animals to which this methodology has traditionally been applied.
- 5) Further work and consensus are needed on how to integrate evidence from human epidemiological studies in chemical risk assessment that has been designed around use of experimental studies in animals.
- 6) 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 closer collaboration and training among experts and staff across panels.

## References

- Adami HO, Berry CL, Breckenridge CB, Smith LL, Swenberg JA, Trichopoulos D, Weiss NS and Pastoor TP, 2011. Toxicology and epidemiology: improving the science with a framework for combining toxicological and epidemiological evidence to establish causal inference. *Toxicological Sciences*, 122, 223–234. <https://doi.org/10.1093/toxsci/kfr113>
- 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 DG, Gore SM, Gardner MJ and Pocock SJ, 1983. Statistical guidelines for contributors to medical journals. *British Medical Journal*, 286, 1489–1493. <https://doi.org/10.1136/bmj.286.6376.1489>
- Altman RD, Lang AE and Postuma RB, 2011. Caffeine in Parkinson's Disease: a pilot open-label, Dose-Escalation Study. *Movement Disorders*, 26, 2427–2431. <https://doi.org/10.1002/mds.23873>
- Amrhein V, Greenland S and McShane B, 2019. Scientists rise up against statistical significance. *Nature*, 567, 305–307. <https://doi.org/10.1038/d41586-019-00857-9>
- Apelberg BJ, Witter FR, Herbstman JB, Calafat AM, Halden RU, Needham LL and Goldman LR, 2007. Cord Serum Concentrations of Perfluorooctane Sulfonate (PFOS) and Perfluorooctanoate (PFOA) in Relation to Weight and Size at Birth. *Environmental Health Perspectives*, 115, 1670–1676. <https://doi.org/10.1289/ehp.10334>
- Armitage JM, Macleod M and Cousins IT, 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, 5830–5836. <https://doi.org/10.1021/es900753y>
- Balzer J, Rassaf T, Heiss C, Kleinbongard P, Lauer T, Merx M, Heussen N, Gross HB, Keen CL, Schroeter H and 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, 2141–2149. <https://doi.org/10.1016/j.jacc.2008.01.059>
- Bennette C and 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 JL and Woodruff TJ, 2018. The risk of bias in observational studies of exposures (ROBINS-E) tool: concerns arising from application to observational studies of exposures. *Systematic Review*, 7, 242. <https://doi.org/10.1186/s13643-018-0915-2>
- Bert B, Heintz C, Chmielewska J, Schwarz F, Grune B, Hensel A, Greiner M and Schönfelder G, 2019. Refining animal research: the animal study registry. *PLoS Biology*, 17, e3000463. <https://doi.org/10.1371/journal.pbio.3000463>
- Boa E, Danielsen S and Haesen S, 2015. Better together – identifying the benefits of a closer integration between plant health, agriculture and one health. In Zinsstag J, Schelling E, Waltner-Townsend D, Whittaker M and Tanner M, eds. Published in *One Health: the added value of integrated health approaches*. CABI Publishing, Wallingford.
- Clemente JC, Ursell LK, Wegener Parfrey L and Knight R, 2012. The impact of the gut microbiota on human health: an integrative view. *Cell*, 148, 1258–1270. <https://doi.org/10.1016/j.cell.2012.01.035>
- Cock MJW and Wittenberg R, 2001. *Invasive alien species: a toolkit of best prevention and management practices*. CAB International.
- Cooke BM, Jones DG and Kayle B, 2006. *The Epidemiology of Plant Diseases*, 2nd Edition. Springer.
- Crippa A and 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>
- Crippa A, Discacciati A, Bottai M, Spiegelman D and Orsini N, 2018. One-stage dose-response meta-analysis for aggregated data. *Statistical Methods in Medical Research*, 28, 1579–1596. <https://doi.org/10.1177/0962280218773122>
- Deeks JJ, Dinnes J, D'Amico R, Sowden AJ, Sakarovich C and Song F, 2003. Evaluating non-randomised intervention studies. *Health Technology Assessment*, 7. <https://doi.org/10.3310/hta7270>
- Dhont M, 2010. History of oral contraception. *The European Journal of Contraception & Reproductive Health Care*, 15, S12–S18. <https://doi.org/10.3109/13625187.2010.513071>
- Dickerman BA, García-Albéniz X, Logan RW, Denaxas S and Hernán MA, 2019. Avoidable flaws in observational analyses: an application to statins and cancer. *Nature Medicine*, 25, 1601–1606. <https://doi.org/10.1038/s41591-019-0597-x>
- Diefenbach MA, Miller-Halegoua S and Bowen DJ, eds, 2016. *Handbook of Health Decision Science*, 1st Edition. Springer Science+Business Media LLC, New York.
- Diekmann O and Heesterbeek JAP, 2000. *Mathematical Epidemiology of Infectious Diseases: Model Building, Analysis and Interpretation*. Wiley.
- Dodd S, White I and Williamson P, 2011. Departure from treatment protocol in published randomised controlled trials: a review. *Trials*, 12, A129. <https://doi.org/10.1186/1745-6215-12-S1-A129>

- Dodd S, White I and 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 and Stryhn H, 2009. *Veterinary Epidemiological Research*, 2nd Edition. VER Inc.
- Duffield-Lilloco AJ, Dalkin BL, Reid ME, Turnbull BW, Slate EH, Jacobs ET, Marshall JR and Clark LC for the 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, 608–612. <https://doi.org/10.1046/j.1464-410X.2003.04167.x>
- EFSA (European Food Safety Authority), 2010. Application of systematic review methodology to food and feed safety assessments to support decision making. *EFSA Journal* 2010;8(6):1637. <https://doi.org/10.2903/j.efsa.2010.1637>
- EFSA (European Food Safety Authority), 2015a. Scientific report on principles and process for dealing with data and evidence in scientific assessments. *EFSA Journal* 2015;13(5):4121, 35 pp. <https://doi.org/10.2903/j.efsa.2015.4121>
- EFSA (European Food Safety Authority), 2015b. Editorial: increasing robustness, transparency and openness of scientific assessments. *EFSA Journal* 2015;13(3):e13031, 3 pp. <https://doi.org/10.2903/j.efsa.2015.e13031>
- 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 HD, Hoogenboom RL, Messean A, Naska A, Navajas Navarro M, Noerrung B, Ockleford C, Wallace RJ, Younes M, Abuntori B, Alvarez F, Aryeetey M, Baldinelli F, Barucci F, Bau A, Binaglia M, Broglia A, Castoldi AF, Christoph E, De Sesmaisons-Lecarré A, Georgiadis N, Gervelmeyer A, Istace F, López-Gálvez G, Manini P, Maurici D, Merten C, Messens W, Mosbach-Schulz O, Putzu C, Ramos Bordajandi L, Smeraldi C, Tiramani M, Valtueña Martínez S, Vos S, Hardy AR, Hugas M, Kleiner J, De Seze G, Verhagen H and Verloo D, 2018a. Implementation of PROMETHEUS 4-step approach for evidence use in EFSA scientific assessments: benefits, issues, needs and solutions. *EFSA supporting publication* 2018;EN-1395, 31 pp. <https://doi.org/10.2903/sp.efsa.2018.EN-1395>
- EFSA (European Food Safety Authority) and State General Laboratory of Cyprus, Research and Education Institute of Child Health, Yiannopoulos S, Ioannou-Kakouri E, Kanari P, Anastasi A, Agathocleous M, Kakoulli A, Christodoulidou M, Papoutsou S, Savva S, Hadjigeorgiou C, Solea T and Tornaritis M, 2018b. National dietary survey on the adult population of Cyprus. *EFSA supporting publication* 2018;EN-1458, 25 pp. <https://doi.org/10.2903/sp.efsa.2018.EN-1458>
- EFSA (European Food Safety Authority), Schrader G, Kinkar M and Vos S, 2020a. Pest survey card on *Agrilus anxius*. *EFSA supporting publication* 2020;EN-1777, 23 pp. <https://doi.org/10.2903/sp.efsa.2020.en-1777>
- EFSA (European Food Safety Authority), Medina-Pastor P and Triacchini G, 2020b. The 2018 European Union report on pesticide residues in food. *EFSA Journal* 2020;18(4):6057, 103 pp. <https://doi.org/10.2903/j.efsa.2020.6057>
- EFSA (European Food Safety Authority), Martino L, Aiassa E, Halldórsson TI, Koutsoumanis PK, Naegeli H, Baert K, Baldinelli F, Devos Y, Lodi F, Lostia A, Manini P, Merten C, Messens W, Rizzi V, Tarazona J, Titz A and Vos S, 2020c. Draft framework for protocol development for EFSA's scientific assessments. *EFSA supporting publication* 2020;EN-1843, 46 pp. <https://doi.org/10.2903/sp.efsa.2020.EN-1843>
- EFSA (European Food Safety Authority), 2020d. Scientific report on the update of the *Xylella* spp. host plant database – systematic literature search up to 30 June 2019. *EFSA Journal* 2020;18(4):6114. <https://doi.org/10.2903/j.efsa.2020.6114>
- EFSA CONTAM Panel (EFSA Panel on Contaminants in the Food Chain) Knutsen HK, Alexander J, Barregård L, Bignami M, Brüschweiler B, Ceccatelli S, Cottrill B, Dinovi M, Edler L, Grasl-Kraupp B, Hogstrand C, Nebbia CS, Oswald IP, Petersen A, Rose M, Roudot AC, Schwerdtle T, Vleminckx C, Vollmer G, Wallace H, Fürst P, Håkansson H, Halldorsson T, Lundebye AK, Pohjanvirta R, Rylander L, Smith A, van Loveren H, Waalkens-Berendsen I, Zeilmaker M, Binaglia M, Gómez Ruiz JA, Horváth Z, Christoph E, Ciccolallo L, Ramos Bordajandi L, Steinkellner H and Hoogenboom L, 2018. Scientific opinion on the risk for animal and human health related to the presence of dioxins and dioxin-like PCBs in feed and food. *EFSA Journal* 2018;16(11):5333, 331 pp. <https://doi.org/10.2903/j.efsa.2018.5333>
- EFSA and ECDC (European Food Safety Authority and European Centre for Disease Prevention and Control), 2018. The European Union summary report on trends and sources of zoonoses, zoonotic agents and food-borne outbreaks in 2017. *EFSA Journal* 2018;16(12):5500, 262 pp. <https://doi.org/10.2903/j.efsa.2018.5500>
- EFSA and ECDC (European Food Safety Authority and European Centre for Disease Prevention and Control), 2020. The European Union Summary Report on Antimicrobial Resistance in zoonotic and indicator bacteria from humans, animals and food in 2017/2018. *EFSA Journal* 2020;18(3):6007, 166 pp. <https://doi.org/10.2903/j.efsa.2020.6007>
- 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 JA, MacLeod A, Navajas Navarro M, Niere B, Parnell S, Potting R, Rafoss T, Rossi V, Urek G, Van Bruggen A, Van Der Werf W, West J, Winter S, Hart A, Schans J, Schrader G, Suffert M, Kertesz V, Kozelska S, Mannino MR, Mosbach-Schulz O, Pautasso M, Stancanelli G, Tramontini S, Vos S and Gilioli G, 2018. Guidance on quantitative pest risk assessment. *EFSA Journal* 2018;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 JA, Fejer Justesen A, MacLeod A, Magnusson CS, Milonas P, Navas-Cortes JA, Parnell S, Reignault PL, Thulke H-H, Van der Werf W, Vicent Civera A, Yuen J, Zappala L, Jeger MJ, Gardi C, Mosbach-Schulz O, Preti S, Rosace MC, Stancanelli G and Potting R, 2019. Guidance on commodity risk assessment for the evaluation of high risk plants dossiers. *EFSA Journal* 2019;17(4):5668, 16 pp. <https://doi.org/10.2903/j.efsa.2019.5668>
- EFSA Scientific Committee, 2011. Statistical Significance and Biological Relevance. *EFSA Journal* 2011;9(9):2372, 17 pp. <https://doi.org/10.2903/j.efsa.2011.2372>
- EFSA Scientific Committee, Hardy A, Benford D, Halldorsson T, Jeger MJ, Knutsen HK, More S, Younes M, Naegeli H, Noteborn H, Ockleford C, Ricci A, Rychen G, Schlatter JR, Silano V, Solecki R, Turck D, Bresson J-L, Griffin J, Hougaard Benekou S, Van Loveren H, Luttik R, Messean A, Penninks A, Ru G, Stegeman JA, van der Werf W, Westendorf J, Woutersen RA, Barizzone F, Bottex B, Lanzoni A, Georgiadis N and Alexander J, 2017a. Guidance on biological relevance. *EFSA Journal* 2017;15(8):4970, 73 pp. <https://doi.org/10.2903/j.efsa.2017.4970>
- EFSA Scientific Committee, Hardy A, Benford D, Halldorsson T, Jeger MJ, Knutsen HK, More S, Naegeli H, Noteborn H, Ockleford C, Ricci A, Rychen G, Schlatter JR, Silano V, Solecki R, Turck D, Benfenati E, Chaudhry QM, Craig P, Frampton G, Greiner M, Hart A, Hogstrand C, Lambre C, Luttik R, Makowski D, Siani A, Wahlstroem H, Aguilera J, Dorne J-L, Fernandez Dumont A, Hempen M, Valtuena Martinez S, Martino L, Smeraldi C, Terron A, Georgiadis N and Younes M, 2017b. Scientific Opinion on the guidance on the use of the weight of evidence approach in scientific assessments. *EFSA Journal* 2017;15(8):4971, 66 pp. <https://doi.org/10.2903/j.efsa.2017.4971>
- EFSA Scientific Committee, Hardy A, Benford D, Halldorsson T, Jeger MJ, Knutsen KH, More S, Mortense NA, Naegeli H, Noteborn H, Ockleford C, Ricci A, Rychen G, Silano V, Solecki R, Turck D, Aerts M, Bodin L, Davis A, Edler L, Gundert-Remy U, Sand S, Slob W, Bottex B, Abrahantes JC, Marques DC, Kass G and Schlatter JR, 2017c. Update: guidance on the use of the benchmark dose approach in risk assessment. *EFSA Journal* 2017;15(1):4658, 41 pp. <https://doi.org/10.2903/j.efsa.2017.4658>
- Emerson JD, Burdick E, Hoaglin DC, Mosteller F and Chalmers TC, 1990. An empirical study of the possible relation of treatment differences to quality scores in controlled randomized clinical trials. *Controlled Clinical Trials*, 11, 339–352. [https://doi.org/10.1016/0197-2456\(90\)90175-2](https://doi.org/10.1016/0197-2456(90)90175-2)
- Fei C, McLaughlin JK, Tarone RE and Olsen J, 2007. Perfluorinated chemicals and fetal growth: a study within the danish national birth cohort. *Environmental Health Perspectives*, 115, 1677–1682. <https://doi.org/10.1289/ehp.10506>
- Genaidy AM, Lemasters GK, Lockey J, Succop P, Deddens J, Sobehi T and Dunning K, 2007. An epidemiological appraisal instrument - a tool for evaluation of epidemiological studies. *Ergonomics*, 50, 920–960. <https://doi.org/10.1080/00140130701237667>
- Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN and Altman DG, 2016. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *European Journal of Epidemiology*, 31, 337–350. <https://doi.org/10.1007/s10654-016-0149-3>
- Gross M, Green RM, Weltje L and Wheeler JR, 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 TI, Fei C, Olsen J, Lipworth L, McLaughlin JK and Olsen SF, 2008. Dietary predictors of perfluorinated chemicals: a study from the Danish National Birth Cohort. *Environmental Science & Technology*, 42, 8971–8977. <https://doi.org/10.1021/es801907r>
- Hartley L, May MD, Loveman E, Colquitt JL and Rees K, 2016. Dietary fibre for the primary prevention of cardiovascular disease. *Cochrane Database of Systematic Reviews*, 7. <https://doi.org/10.1002/14651858.CD011472.pub2>
- Hernán MA, Hernández-Díaz S and Robins JM, 2004. A structural approach to selection bias. *Epidemiology*, 15, 615–625. <https://doi.org/10.1097/01.ede.0000135174.63482.43>
- 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. Available online: [www.handbook.cochrane.org](http://www.handbook.cochrane.org)
- Hill AB, 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 CR, Rovers MM, de Vries RB, Leenaars M, Ritskes-Hoitinga M and Langendam MW, 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 BV, Van Horn L, Hsia J, Manson JE, Stefanick ML, Wassertheil-Smoller S, Kuller LH, LaCroix AZ, Langer RD, Lasser NL, Lewis CE, Limacher MC, Margolis KL, Mysiw WJ, Ockene JK, Parker LM, Perri MG, Phillips L, Prentice RL, Robbins J, Rossouw JE, Sarto GE, Schatz IJ, Snetselaar LG, Stevens VJ, Tinker LF, Trevisan M, Vitamins MZ, Anderson GL, Assaf AR, Bassford T, Beresford SA, Black HR, Brunner RL, Brzyski RG, Caan B, Chlebowski RT, Gass M, Granek I, Greenland P, Hays J, Heber D, Heiss G, Hendrix SL, Hubbell FA, Johnson KC and Kotchen JM, 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, 655–666. <https://doi.org/10.1001/jama.295.6.655>

- Ingelsson E, Schaefer EJ, Contois JH, McNamara JR, Sullivan L, Keyes MJ, Pencina MJ, Schoonmaker C, Wilson PW, D'Agostino RB and Vasan RS, 2007. Clinical Utility of Different Lipid Measures for Prediction of Coronary Heart Disease in Men and Women. *Journal of the American Medical Association*, 298, 776–785. <https://doi.org/10.1001/jama.298.7.776>
- Ioannidis JP, 2005. Why most published research findings are false. *PLoS Medicine*, 2, e124. Epub 2005/08/03. <https://doi.org/10.1371/journal.pmed.0020124>
- Johnson PI, Sutton P, Atchley DS, Koustas E, Lam J, Sen S, Robinson KA, Axelrad DA and Woodruff TJ, 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, 1028–1039. <https://doi.org/10.1289/ehp.1307893>
- Jüni P, Witschi A, Bloch R and Egger M, 1999. The hazards of scoring the quality of clinical trials for meta-analysis. *Journal of the American Medical Association*, 282, 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>
- Kilkenny C, Browne WJ, Cuthill IC, Emerson M and Altman DG, 2010. Improving bioscience research reporting: the ARRIVE guidelines for reporting animal research. *PLoS Biology*, 8, e1000412.
- Klimisch HJ, Andreae M and Tillmann U, 1997. A Systematic Approach for Evaluating the Quality of Experimental Toxicological and Ecotoxicological Data. *Regulatory Toxicology and Pharmacology*, 25, 1–5. <https://doi.org/10.1006/rtp.1996.1076>
- Koustas E, Lam J, Sutton P, Johnson PI, Atchley DS, Sen S, Robinson KA, Axelrad DA and Woodruff TJ, 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). 58 pp.
- Koustas E, Lam J, Sutton P, Johnson PI, Atchley DS, Sen S, Robinson KA, Axelrad DA and Woodruff TJ, 2014. The navigation guide – evidence-based medicine meets environmental health: systematic review of nonhuman evidence for PFOA effects on fetal growth. *Environmental Health Perspectives*, 122, 1015–1027. <https://doi.org/10.1289/ehp.1307177>
- Kristal AR, Darke AK, Morris JS, Tangen CM, Goodman PJ, Thompson IM, Meyskens FL, Goodman GE, Minasian LM, Parnes HL, Lippman SM and Klein EA, 2014. Baseline selenium status and effects of selenium and vitamin E supplementation on prostate cancer risk. *Journal National Cancer Institute*, 106, djt456. <https://doi.org/10.1093/jnci/djt456>
- Lam J, Koustas E, Sutton P, Johnson PI, Atchley DS, Sen S, Robinson KA, Axelrad DA and Woodruff TJ, 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, 1040–1051. <https://doi.org/10.1289/ehp.1307923>
- Langlois EV, Daniels K and Akl EA, eds, 2018. *Evidence synthesis for health policy and systems: a methods guide*. World Health Organization, Geneva, 2018. Licence: CC BY-NC-SA 3.0 IGO. Available online: <https://apps.who.int/iris/bitstream/handle/10665/275367/9789241514552-eng.pdf?ua=1>
- Lazcano G, Papuzinski C, Madrid E and Arancibia M, 2019. General concepts in biostatistics and clinical epidemiology: observational studies with cohort design. *Medwave*, 19, e7748. <https://doi.org/10.5867/medwave.2019.11.7748>
- Lenters V, Vermeulen R, Dogger S, Stayner L, Portengen L, Burdorf A and 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>
- Li Y, Fletcher T, Mucs D, Scott K, Lindh CH, Tallving P and Jakobsson K, 2018. Half-lives of PFOS, PFHxS and PFOA after end of exposure to contaminated drinking water. *Occupational and Environmental Medicine*, 75, 46–51. <https://doi.org/10.1136/oemed-2017-104651>
- Lilienfeld AM and Lilienfeld DE, 1980. *Foundations of epidemiology*, 2nd Edition. Oxford University Press, New York.
- Lippman SM, Goodman PJ, Klein EA, Parnes HL, Thompson Jr IM, Kristal AR, Santella RM, Probstfield JL, Moinpour CM, Albanes D, Taylor PR, Minasian LM, Hoque A, Thomas SM, Crowley JJ, Gaziano JM, Stanford JL, Cook ED, Fleshner NE, Lieber MM, Walther PJ, Khuri FL, Karp DD, Schwartz GG, Ford LG and Coltman Jr CA, 2005. Designing the Selenium and Vitamin E Cancer Prevention Trial (SELECT). *Journal of the National Cancer Institute*, 97, 94–102. <https://doi.org/10.1093/jnci/dji009>
- Lippman SM, Klein EA, Goodman PJ, Lucia MS, Thompson IM, Ford LG, Parnes HL, Minasian LM, Gaziano JM, Hartline JA, Parsons JK, Bearden JD, Crawford ED, Goodman ED, Claudio J, Winquist E, Cook ED, Karp DD, Walther P, Lieber MM, Kristal AR, Darke AK, Arnold KB, Ganz PA, Santella RM, Albanes D, Taylor PR, Probstfield JL, Jagpal TJ, Crowley JJ, Meyskens FL, Baker LH and Coltman CA, 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, 39–51. <https://doi.org/10.1001/jama.2008.864>
- Longnecker MP, Smith CS, Kissling GE, Hoppin JA, Butenhoff JL, Decker E, Ehresman DJ, Ellefson ME, Flaherty J, Gardner MS, Langlois E, Leblanc A, Lindstrom AB, Reagen WR, Strynar MJ and Studabaker WB, 2008. An interlaboratory study of perfluorinated alkyl compound levels in human plasma. *Environmental Research*, 107, 152–159. <https://doi.org/10.1016/j.envres.2008.01.005>



- Lynch HN, Goodman JE, Tabony JA and Rhomberg LR, 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 LV, Hughes G and van den Bosch F, 2007. *The Study of Plant Disease Epidemics*. The American Phytopathological Society.
- Mansournia MA, Higgins JPT, Sterne JAC and Hernán MA, 2017. Biases in randomized trials: a conversation between trialists and epidemiologists. *Epidemiology*, 28, 54–59. <https://doi.org/10.1097/EDE.0000000000000564>
- Meek ME, Palermo CM, Bachmann AN, North CM and Lewis RJ, 2014. Mode of action human relevance (species concordance) framework: evolution of the Bradford Hill considerations and comparative analysis of weight of evidence. *Journal of Applied Toxicology*, 34, 595–606. <https://doi.org/10.1002/jat.2984>
- Money CD, Tomenson JA, Penman MG, Boogaard PJ and Jeffrey LR, 2013. A systematic approach for evaluating and scoring human data. *Regulatory Toxicology and Pharmacology*, 66, 241–247. <https://doi.org/10.1016/j.yrtph.2013.03.011>
- Morgan RL, Thayer KA, Bero L, Bruce N, Falck-Ytter Y, Ghersi D, Guyatt G, Hooijmans C, Langendam M, Mandrioli D, Mustafa RA, Rehfues EA, Rooney AA, Shea B, Silbergeld EK, Sutton P, Wolfe MS, Woodruff TJ, Verbeek JH, Holloway AC, Santesso N and Schünemann HJ, 2016. GRADE: assessing the quality of evidence in environmental and occupational health. *Environment International*, 92–93, 611–616. <https://doi.org/10.1016/j.envint.2016.01.004>
- Morgan RL, Whaley P, Thayer KA and Schünemann HJ, 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>
- Morgan RL, Thayer KA, Santesso N, Holloway AC, Blain R, Eftim SE, Goldstone AE, Ross P, Ansari M, Akl EA, Filippini T, Hansell A, Meerpohl J, Mustafa RA, Verbeek J, Vinceti M, Whaley P and 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>
- Nielsen SS, Toft N and Gardner IA, 2011. Structured approach to design of diagnostic test evaluation studies for chronic progressive infections in animals. *Vector Microbiology*, 150, 115–125. <https://doi.org/10.1016/j.vetmic.2011.01.019>
- Noordzij M, van Diepen M, Caskey FC and Jager KJ, 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), 2013. *Draft OHAT Approach for Systematic Review and Evidence Integration for Literature-based Health Assessments*. Office of Health Assessment and Translation (OHAT). 15 pp.
- NTP (National Toxicology Program), 2015. *Handbook for Conducting a Literature-based Health Assessment Using OHAT Approach for Systematic Review and Evidence Integration*. Office of Health Assessment and Translation (OHAT). 98 pp.
- 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. Available online: <https://ntp.niehs.nih.gov/whatwestudy/assessments/noncancer/handbook/index.html>
- OECD (Organisation for Economic Co-operation and Development), 2005. *Guidance document on the validation and international acceptance of new or updated test methods for hazard assessment*. OECD series on testing and assessment Number 34. ENV/JM/MONO(2005)14
- Orsini N, Li R, Wolk A, Khudyakov P and Spiegelman D, 2012. Meta-analysis for linear and nonlinear dose-response relations: examples, an evaluation of approximations, and software. *American Journal of Epidemiology*, 175, 66–73. <https://doi.org/10.1093/aje/kwr265>
- Page MJ, McKenzie JE and Higgins JPT, 2018. Tools for assessing risk of reporting biases in studies and syntheses of studies: a systematic review. *British Medical Journal Open*, 8, e019703. <https://doi.org/10.1136/bmjopen-2017-019703>
- Pearce N, 2005. *A Short Introduction to Epidemiology*, 2nd Edition. Occasional Report Series No 2, Centre for Public Health Research, Massey University, Wellington, New Zealand. Available online: [https://courses.cit.cornell.edu/bionb4280/Intro\\_Epidemiology.pdf](https://courses.cit.cornell.edu/bionb4280/Intro_Epidemiology.pdf)
- Pearce N, 2016. Analysis of matched case-control studies. *BMJ*, 2016, 352. <https://doi.org/10.1136/bmj.i969>
- Pearce N, Smith AH, Howard JK, Sheppard RA, Giles HJ and Teague CA, 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, 75–83. <https://doi.org/10.1136/oem.43.2.75>
- Pearce N, Checkoway H and Kriebel D, 2007. Bias in occupational epidemiology studies. *Occupational and Environmental Medicine*, 64, 562–568. <https://doi.org/10.1136/oem.2006.026690>
- Pearl J, 2009. Causal inference in statistics: an overview. *Statistics survey*, 3, 96–146. <https://doi.org/10.1214/09-SS057>
- Pearl J and Mackenzie D, 2018. *The book of why: the new science of cause and effect*. Basic Books.
- Porta, 2014. *Dictionary of Epidemiology*, 6th Edition. Oxford University Press, New York.

- Poston L, Bell R, Croker H, Flynn AC, Godfrey KM, Goff L, Hayes L, Khazaezadeh N, Nelson SM, Oteng-Ntim E, Pasupathy D, Patel N, Robson SC, Sandall J, Sanders TAB, Sattar N, Seed PT, Wardle J, Whitworth MK and Briley AL and 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, 767–777. [https://doi.org/10.1016/S2213-8587\(15\)00227-2](https://doi.org/10.1016/S2213-8587(15)00227-2)
- Quigley JM, Thompson JC, Halfpenny NJ and Scott DA, 2019. Critical appraisal of nonrandomized studies - a review of recommended and commonly used tools. *Journal of Evaluation in Clinical Practice*, 25, 44–52. <https://doi.org/10.1111/jep.12889>
- Rooney AA, Boyles AL, Wolfe MS, Bucher JR and Thayer KA, 2014. Systematic review and evidence integration for literature-based environmental health science assessments. *Environmental Health Perspectives*, 122, 711–718. <https://doi.org/10.1289/ehp.1307972>
- Rothman KJ, 1976. Causes. *American Journal of Epidemiology*, 104, 587–592. <https://doi.org/10.1093/oxfordjournals.aje.a112335>
- Rothman KJ, 2014. Six persistent research misconceptions. *Journal of General Internal Medicine*, 29, 1060–1064. <https://doi.org/10.1007/s11606-013-2755-z>
- Rothman KJ, 2016. Disengaging from statistical significance. *European Journal of Epidemiology*, 31, 443–444. <https://doi.org/10.1007/s10654-016-0158-2>
- Rothman KJ, Greenland S and Lash TL, 2008. *Modern Epidemiology*. Lippincott, Williams & Wilkins, Philadelphia, PA.
- Rothman KJ, Greenland S and Lash TL, 2012. *Modern epidemiology*, 3rd Edition. Wolters Kluwer Health/Lippincott Williams & Wilkins, Philadelphia.
- Samuel GO, Hoffmann S, Wright RA, Lalu MM, Patlewicz G, Becker RA, DeGeorge GL, Fergusson D, Hartung T, Lewis RJ and Stephens ML, 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 ID and Higgins JP, 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, 666–676. <https://doi.org/10.1093/ije/dym018>
- Savitz DA, 2007. Guest editorial: biomarkers of perfluorinated chemicals and birth weight. *Environmental Health Perspectives*, 115, A528–A529. <https://doi.org/10.1289/ehp.10923>
- Savitz DA, Stein CR, Elston B, Wellenius GA, Bartell SM, Shin HM, Vieira VM and Fletcher T, 2012. Relationship of perfluorooctanoic acid exposure to pregnancy outcome based on birth records in the mid-Ohio Valley. *Environmental Health Perspectives*, 120, 1201–1207. <https://doi.org/10.1289/ehp.1104752>
- Savitz DA, Wellenius GA and Trikalinos TA, 2019. The problem with mechanistic risk of bias assessments in evidence synthesis of observational studies and a practical alternative: Assessing the impact of specific sources of potential bias. *American Journal of Epidemiology*, 188, 1581–1585. <https://doi.org/10.1093/aje/kwz131>
- Schulz KF, Chalmers I, Hayes RJ and Altman DG, 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, 408–412. <https://doi.org/10.1001/jama.273.5.408>
- Shamliyan T, Kane RL and 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 for Clinical Epidemiology*, 63, 1061–1070. <https://doi.org/10.1016/j.jclinepi.2010.04.014>
- Shea BJ, Grimshaw JM, Wells GA, Boers M, Andersson N, Hamel C, Porter AC, Tugwell P, Moher D and Bouter LM, 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>
- Shin HM, Vieira VM, Ryan PB, Steenland K and Bartell SM, 2011. Retrospective exposure estimation and predicted versus observed serum perfluorooctanoic acid concentrations for participants in the C8 Health Project. *Environmental Health Perspectives*, 119, 1760–1765. <https://doi.org/10.1289/ehp.1103729>
- Sigurjónsdóttir HA, Franzson L, Manhem K, Ragnarsson J, Sigurdsson G and Wallerstedt S, 2001. Liquorice-induced rise in blood pressure: a linear dose-response relationship. *Journal of Human Hypertension*, 15, 549–552. <https://doi.org/10.1038/sj.jhh.1001215>
- Snedeker SM and Hay AG, 2012. Do interactions between gut ecology and environmental chemicals contribute to obesity and diabetes? *Environmental Health Perspectives*, 120, 332–339. <https://doi.org/10.1289/ehp.1104204>
- Sommar JN, Pettersson-Kymmer U, Lundh T, Svensson O, Hallmans G and Bergdahl IA, 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 and McLeroy KR, 2007. The importance of external validity. *American Journal of Public Health*, 98, 9–10. <https://doi.org/10.2105/ajph.2007.126847>
- Steenland K, Barry V and Savitz D, 2018. Serum perfluorooctanoic acid and birthweight: an updated meta-analysis with bias analysis. *Epidemiology*, 29, 765–776. <https://doi.org/10.1097/EDE.0000000000000903>

- Sterne JAC, Hernán MA, Reeves BC, Savović J, Berkman ND, Viswanathan M, Henry D, Altman DG, Ansari MT, Boutron I, Carpenter JR, Chan AW, Churchill R, Deeks JJ, Hróbjartsson A, Kirkham J, Jüni P, Loke YK, Pigott TD, Ramsay CR, Regidor D, Rothstein HR, Sandhu L, Santaguida PL, Schünemann HJ, Shea B, Shrier I, Tugwell P, Turner L, Valentine JC, Waddington H, Waters E, Wells GA, Whiting PF and Higgins JPT, 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 JL, Armitage JM, Lancaster T, Silagy CA, Fowler GH and Neil HA, 1998. Systematic review of dietary intervention trials to lower blood total cholesterol in free-living subjects. *British Medical Journal*, 316, 1213–1220. <https://doi.org/10.1136/bmj.316.7139.1213>
- 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, 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. 142 pp.
- Uwamahoro F, Berlin A, Bucagu C, Bylund H and Yuen J, 2018. Potato bacterial wilt in Rwanda: occurrence, risk factors, farmers' knowledge and attitudes. *Food Security*, 10, 1221–1235. <https://doi.org/10.1007/s12571-018-0834-z>
- Vandenbroucke JP, Broadbent A and Pearce N, 2016. Causality and causal inference in epidemiology: the need for a pluralistic approach. *International Journal of Epidemiology*, 45, 1776–1786. <https://doi.org/10.1093/ije/dyv341>
- Verner MA, Loccisano AE, Morken NH, Yoon M, Wu H, McDougall R, Maisonet M, Marcus M, Kishi R, Miyashita C, Chen MH, Hsieh WS, Andersen ME, Clewell HJ III and Longnecker MP, 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, 1317–1324. <https://doi.org/10.1289/ehp.1408837>
- Viswanathan M, Ansari M, Berkman ND, Chang S, Hartling L, McPheeters LM, Santaguida PL, Shamliyan T, Singh K, Tsertsivadze A and Treadwell JR, 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. Available online: [www.effectivehealthcare.ahrq.gov/](http://www.effectivehealthcare.ahrq.gov/)
- Viswanathan M, Berkman ND, Dryden DM and 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. Available online: [www.effectivehealthcare.ahrq.gov/](http://www.effectivehealthcare.ahrq.gov/)
- Vlaanderen J, Lan Q, Kromhout H, Rothman N and 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, 159–167. <https://doi.org/10.1289/ehp.1002318>
- Völkel W, Colnot T, Csanády GA, Filser JG and Dekant W, 2002. Metabolism and kinetics of bisphenol a in humans at low doses following oral administration. *Chemical Research in Toxicology*, 15, 1281–1287. <https://doi.org/10.1021/tx025548t>
- von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC and Vandenbroucke JC, 2007. Strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. *British Medical Journal*, 335, 806–808. <https://doi.org/10.1136/bmj.39335.541782.AD>
- Wacholder S, McLaughlin JK, Silverman DT and Mandel JS, 1992a. Selection of controls in case-control studies. I. Principles. *American Journal of Epidemiology*, 135, 1019–1028. <https://doi.org/10.1093/oxfordjournals.aje.a116396>
- Wacholder S, Silverman DT, McLaughlin JK and Mandel JS, 1992b. Selection of controls in case-control studies. II. Types of controls. *American Journal of Epidemiology*, 135, 1029–1041. <https://doi.org/10.1093/oxfordjournals.aje.a116397>
- Wacholder S, Silverman DT, McLaughlin JK and Mandel JS, 1992c. Selection of controls in case-control studies. III. Design options. *American Journal of Epidemiology*, 135, 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 and Rooney A, 2019. Woodruff T and Bero L. A systematic review, Tools for assessing methodological quality of human observational studies. NHMRC Available online: <https://nhmrc.gov.au/guidelinesforguidelines/develop/assessing-risk-bias>
- Warrington S, Lee C, Otabe A, Narita T, Polnjak O, Pirags V and 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>
- Waters DD, Alderman EL, Hsia J, Howard BV, Cobb FR, Rogers WJ, Ouyang P, Thompson P, Tardif JC, Higginson L, Bittner V, Steffes M, Gordon DJ, Proschan M, Younes N and Verter JI, 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, 2432–2440. <https://doi.org/10.1001/jama.288.19.2432>

- Weisskopf MG and Webster TF, 2017. Trade-offs of personal versus more proxy exposure measures in environmental epidemiology. *Epidemiology*, 28, 635–643. <https://doi.org/10.1097/EDE.0000000000000686>
- Whiting P, Rutjes AWS, Reitsma JB, Bossuyt PMM and Kleijnen J, 2003. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Medical Research Methodology*, 3, 25. <https://doi.org/10.1186/1471-2288-3-25>
- WHO (World Health Organization), 2010. Environmental Health Criteria 240. Principles and Methods for the Risk Assessment of Chemicals in Food. Chapter 5: Dose-Response Assessment and Derivation of Health-Based Guidance Values. WHO, Geneva, 34 pp. Available online: <https://www.who.int/publications/i/item/principles-and-methods-for-the-risk-assessment-of-chemicals-in-food>
- Willet W, 2013. *Nutritional Epidemiology*. Monographs in Epidemiology and Biostatistics, OUP USA.
- Woodruff TJ and 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, 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, 2191–2194. <https://doi.org/10.1001/jama.2013.281053>
- Zeraatkar D, Johnston BC, Bartoszko J, Cheung K, Bala MM, Valli C, Rabassa M, Sit D, Milio K, Sadeghirad B, Agarwal A, Zea AM, Lee Y, Han MA, Vernooij RWM, Alonso-Coello P, Guyatt GH and El Dib R, 2019. Effect of lower versus higher red meat intake on cardiometabolic and cancer outcomes: a systematic review of randomized trials. *Annals of Internal Medicine*, 71, 721–731. <https://doi.org/10.7326/M19-0622>

## 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, 2015a). 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
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

AMSTAR	assessment of multiple systematic reviews
ARRIVE	Animal Research: Reporting of In Vivo Experiments
BfR	German Federal Institute for Risk Assessment
BMD	benchmark dose modelling
BMI	body mass index

CI	confidence interval
CVD	cardiovascular disease
EKE	Expert Knowledge Elicitation
GDM	gestational diabetes
HDL	high-density lipoproteins
HTA	Health Technology Assessment
ISPM	International Standards for Phytosanitary Measures
NESR	Nutrition Evidence Systematic reviews
NRIS	Non-Randomised Intervention Studies
NOAEL	no-observed-adverse-effect-level
NTP	National Toxicological Program
OECD	Organisation for Economic Co-operation and Development
OHAT	Office of Health Assessment and Translation
PECO(T)	Population description, the Exposure, the Comparator, the Outcome (endpoint) and follow up Time
PFAS	perfluoroalkyl substances
PICO(T)	Population, Intervention, Comparator, Outcomes, Time
PIT	Population, Index test, Target condition
PLH	EFSA Plant Health
PO	Population and Outcome
PROMETHEUS	PRomoting METHods for Evidence Use in Scientific assessments
QPRA	quantitative pest risk assessments
RCT	randomised controlled trial
RD	risk difference
RoB	risk of bias
RR	risk ratio
STROBE	strengthening the reporting of observational studies in epidemiology
ToxRTool	Toxicological Data Reliability Assessment Tool
WoE	Weight of Evidence
WG	Working Group



## Annexes

### Annex 1 – Bradford Hill viewpoints

In order to assess the strength of evidence that exposure 'A' can cause a health outcome, Austin Bradford 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 synthesising evidence across different types of study. Some are simple, such as 'does the exposure precede the health outcome, but within most viewpoints there would be a range of 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, 2017b) 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 specific exposure A limited to one specific 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: Austin Bradford Hill, 1965. The environment and disease: Association or causation? Proceedings of the Royal Society of Medicine, 58, 295–300.

## Annex 2 – Hypothesis testing vs. estimation

In statistical hypothesis testing, a test statistic (e.g. risk difference) 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. risk difference being zero) 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 ( $\alpha$ ), 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 ( $\beta$ ), 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 ( $\alpha$ ) 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. This 'mixing' of effect size with precision, is a very undesirable characteristic. 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. 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 (CI) 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% CI for a risk difference (RD) of 0.01–0.03, 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. For this reason, statistical estimation should always be preferred over hypothesis testing, at least whenever it is feasible to evaluate both.

Indeed, emphasis on precision of effect measures with confidence intervals is gradually replacing the approach that focuses on hypothesis testing (Greenland et al., 2016; Rothman, 2016; Amrhein et al., 2019). As Altman and colleagues stated in 1983 (Altman et al., 1983), 'There is a close relation

between the result of a test of significance and the associated confidence interval: if the difference between treatments is significant at the 5% level, then the associated 95% confidence interval excludes the zero difference. 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. 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.

## Annex 3 – 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 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.

## Annex 4 – Recent and 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)	<ul style="list-style-type: none"> <li>• <b>Identify tools</b> commonly used to assess bias in NRIS</li> <li>• Determine those <b>recommended</b> by HTA bodies</li> </ul>	48 critical appraisal tools identified, among them those from Cochrane ( <b>RoB</b> , <b>ROBINS-I</b> ), the Centre for Reviews and Dissemination ( <b>CRD</b> ), and the Scottish Intercollegiate Guidelines Network ( <b>SIGN</b> )	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)	<ul style="list-style-type: none"> <li>• Identify, describe, categorise <b>key elements for appraisal</b> into domains</li> <li>• Develop <b>guidance on selecting risk of bias tools</b> for public health decision makers</li> </ul>	62 tools identified (with 17 categories of similar or overlapping items), full list available <a href="https://ntp.niehs.nih.gov/go/ohat_tools">here</a> ( <a href="https://ntp.niehs.nih.gov/go/ohat_tools">https://ntp.niehs.nih.gov/go/ohat_tools</a> )	<p>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:</p> <ol style="list-style-type: none"> <li>(1) the tool should have clear definitions for each item and be transparent regarding the empirical or theoretical basis for each domain,</li> <li>(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,</li> <li>(3) the ratings for each domain should be reported, rather than an overall score,</li> <li>(4) the tool should be rigorously and independently tested for usability and reliability</li> </ol>

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), <i>in vivo</i> studies, <i>in vitro</i> studies, Systematic reviews	<ul style="list-style-type: none"> <li>Critically evaluate several <b>available frameworks</b> for evaluating study quality</li> <li>Assess the criteria separately for human, animal, and <i>in vitro</i> studies as well as for systematic reviews and <b>evaluate commonalities</b> across disciplines</li> </ul>	10 systems for evaluating the quality of studies or systematic reviews were assessed: the <b>Klimisch system</b> (Klimisch et al., 1997); the <b>OECD Guidance Document (GD) 34</b> (OECD, 2005); the <b>Toxicological Data Reliability Assessment Tool</b> (ToxRTool) (European Commission, Undated); the approaches that have been used in recent <b>IRIS</b> systematic review documents (US EPA, 2013); the framework being developed by <b>NTP's OHAT</b> (NTP, 2013, 2015); Animal Research: Reporting of In Vivo Experiments ( <b>ARRIVE</b> ) guidelines for animal research (Kilkenny et al., 2010); the <b>Navigation Guide</b> for systematic reviews (Koustas et al., 2013, 2014; Woodruff and Sutton, 2014; Johnson et al., 2014; Lam et al., 2014); the 'assessment of multiple systematic reviews' ( <b>AMSTAR</b> ) system (Shea et al., 2007); the "strengthening the reporting of observational studies in epidemiology" ( <b>STROBE</b> ) system (von Elm et al., 2007); the <b>Systematic Approach for Scoring Human Data</b> , as developed by Money et al. (2013)	<p>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</p> <p>For <b>human studies</b> (especially observational epidemiologic ones): aspects of care in identifying and choosing study populations, investigating and adequately addressing potential confounding factors and avoiding selectively reporting results</p> <p>For <b>animal studies</b>: 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</p>

Publication	Context	Study designs covered	Purpose of the review	Further details	Authors' conclusions
Samuel et al. (2016)	Toxicology	Human observational designs (Cohort, Case-control, Cross-sectional), <i>in vivo</i> studies, <i>in vitro</i> studies, QSAR, Physico-chemical properties studies	<ul style="list-style-type: none"> <li>• <b>Scoping the available guidance to assess the methodological or reporting quality</b> of studies relevant to toxicology</li> <li>• Distil the <b>common elements</b> of these documents for each of the four study types</li> </ul>	<p>23 guidance documents for <i>in vitro</i> and <i>in vivo</i> studies included, 7 addressing methodological quality, two addressing reporting quality, and 14 addressing both methodological and reporting quality</p> <p>3 guidance documents for QSAR studies included</p> <p>3 guidance documents included for studies of physico-chemical properties</p> <p>12 publications in total identified as providing guidance on the assessment of human studies, including 10 on methodological quality, one on reporting quality, and one on mixed guidance</p>	<p>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</p>
Sanderson et al. (2007)	Systematic reviews	Observational designs (Cohort, Case-control, Cross-sectional – OBS), Systematic reviews	<ul style="list-style-type: none"> <li>• Provide an annotated <b>bibliography of tools</b> specifically designed to assess quality or susceptibility to bias in OBS epidemiological studies</li> <li>• Identify whether there is an <b>existing tool that could be recommended</b> for widespread use</li> </ul>	<p>86 tools reviewed, comprising 41 simple checklists, 12 checklists with additional summary judgements and 33 scales</p>	<p>Tools should be rigorously developed, evidence-based, valid, reliable and easy to use</p> <p>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</p> <p>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</p>

Publication	Context	Study designs covered	Purpose of the review	Further details	Authors' conclusions
Deeks et al. (2003)	Health Technology Assessment	Non-Randomised Intervention Studies (NRIS)	<ul style="list-style-type: none"> <li>• Review <b>empirical evidence of bias</b> associated with NRIS</li> <li>• Review the <b>content</b> of quality assessment tools for non-randomised studies</li> <li>• Review the <b>use</b> of quality assessment in systematic reviews of non-randomised studies</li> </ul>	182 tools identified; sixty 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



## Annex 5 – Appraisal of different studies using a risk of bias tool

### Background

The selected 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. For these (or other) examples, use of another RoB tool would in principle have highlighted the same considerations and challenges.





The NTP-OHAT tool is quite comprehensive and complex, and describing its full use is not straight forward when presenting short examples. Below we use the instructions as given in the NTP-OHAT manual. Following the manual in more detail, instructions for each risk of bias 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 risk of bias tool, we refer to the manual<sup>18</sup> (NTP, 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 risk of bias tool reflect a finer categorisation of biases that fall under either selection bias, information bias or 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 2.

For each of the 11 bias-questions, the risk of bias 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.
-  **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)

By weighing the ratings of the individual bias questions, an overall risk of bias summary for each study is derived. 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 risk of bias).

There are no fixed sets of rules on how to weigh different questions when assessing the risk of bias. Any decision of weighing is always partly arbitrary, but if done in a transparent way it provides a framework for condensing the judgement assigned to individual bias questions. Such a summary can be helpful if used correctly but it also has its pitfalls, as within each tier (low, medium and high risk of bias) 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 only.

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 risk of bias 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.

<sup>18</sup> [https://ntp.niehs.nih.gov/ntp/ohat/pubs/handbookmarch2019\\_508.pdf](https://ntp.niehs.nih.gov/ntp/ohat/pubs/handbookmarch2019_508.pdf)

**When reading the example below, it is important to keep in mind that this 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 discussions made.**

### Randomised Controlled Trials – risk of bias assessment

EXAMPLE 1: Selenium and cancer, full evaluation

To give an example on how to evaluate risk of bias for an RCT, we selected a randomised controlled trial on the 'Effect of Selenium and Vitamin E on Risk of Prostate Cancer and Other Cancers' by Lippman et al. (2009).<sup>19</sup>

This study was a 4-arm double-blind, randomised controlled trial of 35,633 men who were 50 years or older. These men were assigned to either placebo or supplementation with selenium, vitamin E, selenium + vitamin E. The study period was between 22 August 2001, and 23 October 2008. In this trial, no significant effect on prostate cancer was observed.

When reading the example, it is necessary to have the manuscript opened in parallel as we refer directly to individual tables and figures in the manuscript. A detailed argumentation and justification for the rating of each question is provided below; it makes reference to the generic instructions from the manual, which provide the criteria to choose a specific rating. For each bias question, only the most relevant rating options are shown (in this example, mostly ++ and + instructions). It is followed by a summary of the proposed ratings for individual bias questions of this study, along with a short explanation (Table 2 Summary of rating for Lippman et al., 2009).

#### Questions on selection bias

*Q1: Was administered dose or exposure level adequately randomised?*

Neither the manuscript, the pre-registered study protocol ([clinicaltrials.gov identifier: NCT00006392](https://clinicaltrials.gov/ct2/show/study/NCT00006392)) or previous methodological papers by the authors (Lippman et al., 2005) mention the specific method used to randomise participants. The only information given by the authors is that 'Participants were randomised in a randomised block scheme, in which the block was the study site'. However, it is clear from the manuscript that subjects were randomised, and Table 1 in the manuscript clearly shows that baseline characteristics of study participants are near identical across treatment groups. This would not have been the case if the method of allocation were non-random.

Based on Table 1 in the manuscript, the risk of bias therefore appears low and the relevant rating would be choosing between a '++' or a '+'.

++: 'There is direct evidence that subjects were allocated to any study group including controls using a method with a random component. Acceptable methods of randomization include: referring to a random number table, using a computer random number generator. . .'

+: 'There is indirect evidence that subjects were allocated to study groups using a method with a random component (i.e., authors state that allocation was random, without description of the method used), OR it is deemed that allocation without a clearly random component during the study would not appreciably bias results'

Following the instructions, the rating '+' would be most appropriate. Alternatively, similarities in reported characteristics of study participants (Table 1 of the manuscript) could be used as an argument for assigning the grade '++' because the result of the randomisation process can be evaluated despite the method not being reported. The selection between a '+' and '++' here depends on how strict on exact wording the assessor likes to be.

*Q2: Was allocation to study groups adequately concealed?*

In the manuscript, the authors report that subjects were 'randomly assigned to 4 groups (selenium, vitamin E, selenium + vitamin E, and placebo) in a double-blind fashion between August 22, 2001, and June 24, 2004'. Blinding of participants is also described in some detail:

To ensure that the quality of the blind was maintained, capsules received in each subsequent lot were compared with the previous lot and with matching capsules in the current shipment for their characteristics of weight, shape and size, colour and external marking, odour, and comparability of

<sup>19</sup> This is an open access publication: <https://jamanetwork.com/journals/jama/fullarticle/183163>

contents of opened capsules". This implies that capsules containing both active treatment and placebo were identical, which should ensure blinding of participants throughout the trial.

The authors also give some detail on when blinding of the investigator was first lifted after the study had finished in 24 June 2004: 'The trial was activated in July 2001 and follow-up blinded to the trial results ended on October 23, 2008'.

Given that the investigators report double-blinding and say that all capsules were identical, risk of bias due to allocation concealment appears low. The choice of rating is therefore again limited to either '+' or '++' which are defined in the instructions as:

'++': *There is direct evidence that at the time of recruitment the research personnel and subjects did not know what study group subjects were allocated to, and it is unlikely that they could have broken the blinding of allocation until after recruitment was complete and irrevocable. Acceptable methods used to ensure allocation concealment include central allocation (including telephone, web-based and pharmacy-controlled randomization); sequentially numbered drug containers of identical appearance; sequentially numbered, opaque, sealed envelopes; or equivalent methods.*

'+': *There is indirect evidence that the research personnel and subjects did not know what study group subjects were allocated to and it is unlikely that they could have broken the blinding of allocation until after recruitment was complete and irrevocable, OR it is deemed that lack of adequate allocation concealment would not appreciably bias results.*

The direct evidence for allocation concealment here could be the simple statement by the authors that identical capsules were administered, and that the randomisation code was not broken during the study period. This would qualify as '+'.

#### Questions on performance bias:

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

While the previous question covered concealment at randomisation, this question refers to blinding during the conduct of the study, i.e. if the research personnel and human subjects were blinded to the study group during the study.

In principle, the same information as provided by the authors in their manuscript is often used to assess both questions. Key information to look for here, in addition, are indications of departures from randomisation that could mean participants or investigators were aware of exposure status.

Referring to Q2 in selection bias above, the authors report that blinding of the investigator was first lifted after the study had finished in 24 June 2004: 'The trial was activated in July 2001 and follow-up blinded to the trial results ended on October 23, 2008'. Although not specifically mentioned, it is very unlikely that clinical or other research staff could have found out about participant treatment status (identical capsules should ensure that). The choice of rating here should therefore be limited to either '+' or '++', defined in the instructions as:

'++': *There is direct evidence that the subjects and research personnel were adequately blinded to study group, and it is unlikely that they could have broken the blinding during the study. Methods used to ensure blinding include central allocation; sequentially numbered drug containers of identical appearance; sequentially numbered, opaque, sealed envelopes; or equivalent methods.*

'+': *There is indirect evidence that the research personnel and subjects were adequately blinded to study group, and it is unlikely that they could have broken the blinding during the study, OR it is deemed that lack of adequate blinding during the study would not appreciably bias results.*

One could argue that a statement about all research personnel being blinded for the full duration of the trial, or other information in that direction, would be needed as 'direct evidence' to qualify for a ++. Based on such a strict evaluation, we give a rating of '+' (but '++' could have been justified).

#### Question on attrition/exclusion bias:

*Were outcome data complete without attrition or exclusion from analysis?*

In the manuscript, it is reported that 'All analyses were performed by using an intention-to-treat analysis in which men were classified according to the group to which they were randomized. All men were followed up until death or loss to follow-up'. Furthermore, in Figure 1 of the manuscript, all those receiving

treatment or placebo were included in the primary analyses and all reasons for exclusions prior to allocating treatment are given. Risk of bias due to missing outcome data therefore appears complete and the choice of rating is therefore limited to either '+' or '++', which are defined in the instructions as:

'++': *There is direct evidence that there was no loss of subjects during the study and outcome data were complete, OR loss of subjects (i.e., incomplete outcome data) was adequately addressed and reasons were documented when human subjects were removed from a study or analyses.... OR analyses (such as intention-to-treat analysis) in which missing data have been imputed using appropriate methods (ensuring that the 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).*

'+' : *There is indirect 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, OR it is deemed that the proportion lost to follow-up would not appreciably bias results (less than 20% in each group (Genaidy et al. 2007)).*

In this case, as all those receiving treatment were included in the primary intention-to-treat analyses, the appropriate rating appears to be "++".

### Questions on detection bias

Q1: *Can we be confident in the exposure characterisation?*

The exposure in this case is the allocation of capsules that contained either

- Placebo,
- Selenium (200 µg),
- Vitamin E (400 IU) or
- Selenium (200 µg) + Vitamin E (400 IU).

Participants were asked to take capsules daily for the duration of the experiment. Furthermore, self-reported compliance was recorded at regular intervals during the study period and was similar across treatment groups (Table 2 in the manuscript, upper panel). Compliance was also verified by measuring serum levels of selenium and Vitamin E (Cholesterol-adjusted  $\alpha$ -tocopherol) (Table 2 in the manuscript, lower panel).

Concerning quality of the capsules, the authors provided the following information:

As required by current good manufacturing practice, each lot of capsules was quarantined upon receipt until testing was performed to ensure that capsules labelled "active" by the manufacturer contained the appropriate active agent and that capsules labelled as "placebo" did not contain an active agent.

No further information was provided on the placebo. When considering appropriate rating, it is relevant to reflect on the four possible options as provided in the OHAT instructions:

'++': *There is direct evidence that the exposure (including purity and stability of the test substance and compliance with the treatment, if applicable) was independently characterized and purity confirmed generally as  $\geq 99\%$  for single substance or non-mixture evaluations (see NTP 2006 for example of study effects attributable to impurities of approximately 1%), AND that exposure was consistently administered (i.e., with the same method and timeframe) across treatment groups.*

'+' : *There is indirect evidence that the exposure (including purity and stability of the test substance and compliance with the treatment, if applicable) was independently characterized and purity confirmed generally as  $\geq 99\%$ \* (i.e., the supplier of the chemical provides documentation of the purity of the chemical), OR direct evidence that purity was independently confirmed as  $\geq 98\%$ 3 it is deemed that impurities of up to 2% would not appreciably bias results, AND there is indirect evidence that exposure was consistently administered (i.e., with the same method and timeframe) across treatment groups.*

'-': *There is indirect evidence that the exposure (including purity and stability of the test substance and compliance with the treatment, if applicable) was assessed using poorly validated methods, OR there is insufficient information provided about the validity of the exposure assessment method, but no evidence for concern (record "NR" as basis for answer).*

'--': *There is direct evidence that the exposure (including purity and stability of the test substance and compliance with the treatment, if applicable) was assessed using poorly validated methods.*

\* Note: purity thresholds should be developed for specific research questions and reflect empirical data for the substance and outcome under consideration when possible. Therefore, the appropriate cut-off purity value may be lower or higher than the values listed below for  $\geq 99\%$  defining the difference between 'definitely low' and 'probably low' or  $\geq 98\%$  defining the difference between 'probably low' and 'probably high' risk of bias.

Regarding the first condition referring to purity of the test substance, the exact text here would have had to be pre-tailored to reflect what is acceptable or needed in interventions with dietary supplements (where a pure active substance is usually not used). The key issue to consider is verification that the capsules contain the active ingredient in the correct amount. This was appropriately done. However, information on the placebo capsules (what substance) was not provided.

The second condition is that exposure should be consistently administered across treatment groups. What we know is that participants were given instructions to take the capsules daily. Furthermore, compliance was monitored on a regular basis, both by asking subjects and by measuring serum levels of selenium and vitamin E. Although self-reported compliance decreased from around 85% in the first year to roughly 70% in the 5th year, the level of compliance was similar in all groups over time. Furthermore, serum levels of selenium showed large exposure contrast between the groups receiving and not receiving selenium. Similar, but less pronounced differences were observed for vitamin E. Based on this information, there is direct evidence that exposure was consistently administered (by the participants themselves). Whether exposure contrast would be regarded as enough or compliance falling to 70% (self-reported) is acceptable is a matter of opinion.

Since the content of the placebo is poorly described, a rating of '+' seems appropriate.

Q2: *Can we be confident in the outcome assessment?*

The outcome is prostate cancer and other cancers; the assessment was described as follows by the authors:

Participants reported prostate cancers to the study site staff. Study staff obtained medical records supporting the diagnosis and abstracted the diagnostic method and clinical stage. Tissue and the corresponding pathology report were sent to the central pathology laboratory for confirmation. Gleason Score was based on central pathology review.

Based on that, we have the following options for rating:

'++': *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.*

+'': *There is indirect evidence that the outcome was assessed using acceptable methods (i.e., deemed valid and reliable but not the gold standard) (e.g., validity and reliability  $\geq 0.40$  Genaidy et al. 2007), AND subjects had been followed for the same length of time in all study groups, OR it is deemed that the outcome assessment methods used would not appreciably bias results, AND there is indirect 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, OR it is deemed that lack of adequate blinding of outcome assessors would not appreciably bias results, which is more likely to apply to objective outcome measures.*

'-': *There is indirect evidence that the outcome assessment method is an insensitive instrument (e.g., a questionnaire used to assess outcomes with no information on validation), OR the length of follow up differed by study group, OR there is indirect evidence that it was possible for outcome assessors (including study subjects if outcomes were self-reported) to infer the study group prior to reporting outcomes, OR there is insufficient information provided about blinding of outcome assessors (record "NR" as basis for answer).*

'---': *There is direct evidence that the outcome assessment method is an insensitive instrument, OR the length of follow up differed by study group, OR there is direct evidence for lack of adequate blinding of outcome assessors (including study subjects if outcomes were self-reported), including no blinding or incomplete blinding.*

Following the instructions, three issues need to be considered:

- 1) Appropriateness of the method: Information on cancers were based on clinical records, which is a 'gold standard'. However, it is the participants themselves who had to report having cancer and based on that their diagnosis were verified. It is possible that a participant reports (rightly or wrongly) that he has been diagnosed, but the medical records could not be retrieved for verification. This could result in some misclassification of the outcome assessment. This possibility is not addressed in the manuscript and there is no mentioning of any validation of their outcome assessment approach.
- 2) Comparable length of follow-up across treatment groups: As reported in the study the duration of treatment was the same across all treatment groups.
- 3) Blinding of the outcome assessor: As described above, blinding of the outcome assessor was indirectly reported by saying that the randomisation code was not broken until after the trial was completed. Another indirect evidence for blinding is the fact that all capsules were identical, so the assessors would not have been able to identify which treatment group cases and non-cases belonged to.

Overall the scoring here seems to justify a '+', but not '++', mostly because the completeness of the outcome assessments is not fully clear, but judged unlikely to be serious.

### Questions on Selective Reporting Bias

*Were all measured outcomes reported?*

For this trial, no significant differences in the primary outcomes (cancer) were observed. The issue of possible reporting bias therefore seems limited. Primary outcomes are clearly reported and information on baseline characteristics of study participants and measures of compliance are reported. However, if one examines the original study protocol (protocol ([clinicaltrials.gov](https://clinicaltrials.gov/ct2/show/study/NCT00006392) identifier: NCT00006392), then some of the outcomes reported as secondary in table 4, including diabetes and death from other sources than cancer or CVD, are not mentioned. In terms of rating, the following options are available:

'++': *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.*

'+': *There is indirect 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, OR analyses that had not been planned in advance (i.e., retrospective unplanned subgroup analyses) are clearly indicated as such and it is deemed that the unplanned analyses were appropriate and selective reporting would not appreciably bias results (e.g., appropriate analyses of an unexpected effect). This would include outcomes reported with insufficient detail such as only reporting that results were statistically significant (or not).*

'-': *There is indirect 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, OR and there is indirect evidence that unplanned analyses were included that may appreciably bias results, OR there is insufficient information provided about selective outcome reporting (record "NR" as basis for answer).*

'--': *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 not been reported. In addition to not reporting outcomes, this would include reporting outcomes based on composite score without individual outcome components or outcomes reported using measurements, analysis methods or subsets of the data (e.g., subscales) that were not pre-specified or reporting outcomes not pre-specified, or that unplanned analyses were included that would appreciably bias results.*

As all primary and secondary outcomes are clearly reported and the few additional outcomes whose results are reported have no relevance for the main findings, a '+' could be justified.

### Final Comments

The take home message from this example is first that the generic instruction text is not always fully in line with the elements that one is trying to capture. An example of that is the specification of purity of the test substance under exposure characterisation. This highlights the need to tailor the instructions before each assessment (as stated in the OHAT instructions).

Second, the choice between assigning a '+' or a '++' (or '-' or a '--') often depends on how authors write or their general reporting preferences. In many cases, for example, there is no data or evidence to evaluate if the lack of a sentence on the method of randomisation would lead to a higher risk of bias or not. The difference between a '+' or a '++' may often reflect issues in the quality of reporting, but not necessarily issues in the study conduct and therefore the real occurrence of biases.

To summarise the overall assessment for this study, Table 2 below provides a summary of rating for each bias question along short notes that justify the rating. The two-first columns provide the grouping of each question into different types of bias as defined in Section 4.2 and the finer division into different bias categories as used in the NTP-OHAT risk of bias tool.

**Table 2:** Summary of rating for Lippman et al. (2009)

Type of bias as defined in Section 4.2	OHAT formulation of bias*	Question	Rating	Note**
Confounding	Selection bias	Was administered dose or exposure level adequately randomised?	+	Randomisation done, method of randomisation not reported (but could be verified in Table 1)
Information bias	Selection bias	Was allocation to study groups adequately concealed?	++	Clearly reported
Information bias	Performance Bias:	Were the research personnel and human subjects blinded to the study group during the study?	+	Clear from the manuscript, reason missing
Selection bias	Attrition/ Exclusion Bias	Were outcome data complete without attrition or exclusion from analysis?	++	All initially allocated to treatment were included in the primary analyses
<i>Information bias</i>	<i>Detection Bias</i>	Can we be confident in the exposure characterisation?	+	The placebo is not well characterised (description of what those capsules contained): Judged to be of limited concern as compliance was monitored and no change in selenium or Vitamin E observed among controls
<i>Information bias</i>	<i>Detection Bias</i>	Can we be confident in the outcome assessment?	+	Outcome assessment was based on initial self-report which was then confirmed by examining clinical records. Not clear if all cases would be detected in that way, but differences across groups should be non-differential
<i>Information bias</i>	<i>Selective Reporting Bias</i>	Were all measured outcomes reported?	+	Some outcomes such as diabetes are reported although not mentioned as outcome in the trial registration, since they were added later as safety issues. Minor issues and no substantial risk of bias associated with this reporting

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

\*\* : Note why best rating '+' was not given.

**EXAMPLE 2: Selenium and cancer, performance bias:**

The example provided above is a relatively simple as the selected study was well conducted and of high quality. It is therefore relevant for comparison to consider examples of studies that would be rated at a higher risk of bias on certain questions. One example of such an RCT whose findings may have been influenced by performance bias (in particular) and selection bias, we use the study by Duffield-Lillico et al. (2003). This study was a randomised, double-blind, placebo-controlled trial conducted in 1,312 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.<sup>20</sup> In this study, a protective effect of selenium supplementation on prostate cancer was observed.

The 1,312 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 3 (SS: selenium supplementation).

**Table 3:** Baseline characteristics of study participants (Duffield-Lillico et al., 2003)

Characteristic	SS	Placebo	TABLE 1
Participants randomized, n	457	470	<i>The baseline characteristics of the men participating in the NPC, by treatment group</i>
Mean (SD):			
age, years	64.9 (8.8)	63.7 (9.4)*	
BMI, kg/m <sup>2</sup>	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	<i>*Two-sample t-test, P &lt; 0.05; †PSA test within 6 months of randomization (694 men).</i>
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	

**Question on selection bias:**

*Was administered dose or exposure level adequately randomised?*

As can be seen in Table 3, 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 risk of bias, and selecting between the “–” or “—” rating seems justifiable:

<sup>20</sup> 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 and 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 fewer 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 personnel 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 behavioral 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 3: 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).<sup>21</sup>

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 1,555 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

<sup>21</sup> This is an open access publication: <https://pubmed.ncbi.nlm.nih.gov/26165396/>

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.

#### 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 4). 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 4:** Maternal nutritional and physical activity outcomes, by period of gestation (Duffield-Lillico et al., 2003)

	Standard care	Intervention	Mean difference (95% CI)	p
<b>Nutrition</b>				
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
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
<b>Physical activity</b>				
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
<p>Data are mean (SD) or median (IQR). Women with reported total energy <math>\leq 4.5</math> MJ/day or <math>\geq 20</math> 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. MET=metabolic equivalent of task. *Median difference (95% CI).</p>				
<b>Table 3: Maternal nutritional and physical activity outcomes, by period of gestation</b>				

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 risk of bias 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).<sup>22</sup>

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 1,400 subjects randomly selected from the Danish National Birth Cohort ( $n \sim 100.000$ ). 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).

<sup>22</sup> This is an open access publication: <https://pubmed.ncbi.nlm.nih.gov/18008003/>

**Table 5:** Adjusted regression coefficients ( $\beta$  (95% CI)) between PFOS and PFOA (ng/mL) in first maternal blood during pregnancy and birth weight (g) (Fei et al., 2007)

Strata	PFOS	PFOA
All <sup>a</sup>	-0.46 (-2.34 to 1.41)	-10.63 (-20.79 to -0.47)

<sup>a</sup>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, prepregnancy 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 6 from the manuscript (Fei et al., 2007) below):

**Table 6:** Plasma concentrations of PFOS, PFOA, and birth weight (mean ± SD) by characteristics of study subjects (n = 1,400)

Characteristic <sup>a</sup>	No. (%)	PFOS (ng/mL)	PFOA (ng/mL)	Birth weight (g)
<b>Maternal age at delivery (years)</b>				
< 25	118 (8.4)	38.6 ± 12.0	6.2 ± 2.1	3,573 ± 492
25–29	547 (39.1)	36.8 ± 12.8	6.0 ± 2.8	3,590 ± 521
30–34	504 (36.0)	33.9 ± 13.2	5.2 ± 2.2	3,676 ± 519
≥ 35	230 (16.4)	33.0 ± 12.7	5.1 ± 2.4	3,629 ± 581
<b>Parity</b>				
0	626 (44.7)	37.7 ± 13.0	6.6 ± 2.7	3,524 ± 514
1	508 (36.3)	33.2 ± 12.7	4.7 ± 1.9	3,689 ± 501
2	225 (16.1)	34.0 ± 12.6	4.8 ± 2.3	3,723 ± 580
≥ 3	41 (2.9)	30.5 ± 11.7	3.7 ± 1.6	3,862 ± 540
<b>Socio-occupational status</b>				
High	709 (50.8)	34.0 ± 12.7	5.6 ± 2.3	3,648 ± 527
Middle	566 (40.5)	36.6 ± 12.9	5.6 ± 2.8	3,603 ± 531
Low	121 (8.7)	36.5 ± 14.1	5.6 ± 2.3	3,606 ± 536
<b>Prepregnancy BMI (kg/m<sup>2</sup>)</b>				
< 18.5	58 (4.2)	33.1 ± 14.3	5.2 ± 2.2	3,396 ± 539
18.5–24.9	905 (66.2)	34.6 ± 12.9	5.5 ± 2.6	3,620 ± 506
25.0–29.9	299 (21.9)	36.3 ± 12.0	5.6 ± 2.3	3,638 ± 547
≥ 30.0	105 (7.7)	39.3 ± 14.4	6.1 ± 2.7	3,770 ± 542
<b>Smoking during the pregnancy</b>				
Nonsmoker	1,052 (75.1)	35.7 ± 13.3	5.6 ± 2.6	3,661 ± 514
Quit smoking	131 (9.4)	33.9 ± 11.6	5.8 ± 2.2	3,700 ± 592
1–9 cigarettes/day	109 (7.8)	35.5 ± 12.7	5.8 ± 2.6	3,434 ± 469
≥ 10 cigarettes/day	108 (7.7)	32.5 ± 11.9	4.9 ± 1.9	3,384 ± 560
<b>Sex</b>				
Female	690 (49.3)	35.3 ± 13.0	5.5 ± 2.4	3,582 ± 538
Male	710 (50.7)	35.2 ± 12.9	5.6 ± 2.7	3,668 ± 518

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 standardization, 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 1,400 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 characterization?*

*'++': 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).*

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.

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 intercalibration 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 7.

**Table 7:** Summary of rating for Fei et al. (2007)

Type of bias as defined in Section 4.2	OHAT formulation of bias <sup>(a)</sup>	Question	Rating	Remarks
Selection bias	Selection bias	Did selection of study participants result in appropriate comparison groups?	+	Random selection of 1,400 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 minimize 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 1,400 subjects



Type of bias as defined in Section 4.2	OHAT formulation of bias <sup>(a)</sup>	Question	Rating	Remarks
Information bias	Detection Bias	Can we be confident in the exposure characterization?	+	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

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

As Table 7 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)<sup>23</sup> 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 8 from the manuscript by Fei 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-significant 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. 1,400). 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).

**Table 8:** Estimated change in mean gestational age, birth weight, and birth size parameters with a change in PFOS or PFOA concentrations equal to one In-unit or from 25th to 75th percentile

Model	PFOS		PFOA	
	Change in end point (95% CI) per In-unit <sup>a</sup>	Per increase from the 25th to 75th percentile (IQR) <sup>b</sup>	Change in end point (95% CI) per In-unit <sup>a</sup>	Per increase from the 25th to 75th percentile (IQR) <sup>b</sup>
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 <sup>d</sup>	-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; In-unit = natural log unit.

<sup>a</sup>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. <sup>b</sup>Interquartile range is 3.4–7.9 ng/mL for PFOS and 1.2–2.1 ng/mL for PFOA. <sup>c</sup>Adjusted for maternal age, BMI, race, previous preterm birth, smoking, diabetes, and hypertension. <sup>d</sup>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).

<sup>23</sup> This is an open access publication: <https://pubmed.ncbi.nlm.nih.gov/18008002/>

## 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 foetuses 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).<sup>24</sup> The C8 cohort was established in 2005 after the 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 derived 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 -9 g (CI -20 to +2), equivalent to 0.1 g per ng/ml PFOA, essentially null.

<sup>24</sup> This is an open access publication: <https://ehp.niehs.nih.gov/doi/10.1289/ehp.1104752>

**Table 9:** Study II: PFOA and pregnancy outcome base on birth records linked to the C8 Health Project: association of PFOA with indicators of fetal growth, Mid-Ohio Valley, 1990–2004

Estimated PFOA	Term low birth weight				Term SGA				Change in term birth weight (g)		
	Term births ≥ 2,500 g (n)	Cases (n)	Crude OR	Adjusted <sup>a</sup> OR (95% CI)	Term, AGA (n)	Cases (n)	Crude OR	Adjusted <sup>a</sup> OR (95% CI)	Cases (n)	Crude difference	Adjusted <sup>a</sup> difference (95% CI)
<b>Uncalibrated</b>											
IQR(lnPFOA) <sup>b</sup> increase	4,043	99	0.87	1.04 (0.75, 1.44)	3,375	362	1.02	1.18 (0.97, 1.43)	4,142	0.97	-21.89 (-45.91, 2.13)
100-ng/mL increase	4,043	99	0.93	1.00 (0.82, 1.21)	3,375	362	1.01	1.07 (0.98, 1.17)	4,142	4.27	-9.14 (-20.30, 2.02)
< 40th percentile (3.9 to < 8.9 ng/mL)	1,629	40	1.0	1.0	1,356	144	1.0	1.0	1,669	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)
<b>Bayesian calibration</b>											
IQR(lnPFOA) <sup>b</sup> increase	4,043	99	0.97	1.16 (0.86, 1.58)	3,375	362	1.03	1.19 (1.00, 1.43)	4,142	7.78	-21.51 (-43.62, 0.61)
100-ng/mL increase	4,043	99	0.93	1.04 (0.85, 1.27)	3,375	362	0.98	1.06 (0.97, 1.16)	4,142	3.73	-18.55 (-31.31, -5.80)
< 40th percentile (3.9 to < 8.9 ng/mL)	1,624	42	1.0	1.0	1,358	148	1.0	1.0	1,666	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)
<b>Traditional calibration</b>											
IQR(lnPFOA) <sup>b</sup> increase	4,043	99	1.16	1.33 (1.04, 1.69)	3,375	362	1.05	1.17 (1.00, 1.36)	4,142	2.45	-16.90 (-34.89, 1.08)
100-ng/mL increase	4,043	99	1.00	1.07 (0.96, 1.18)	3,375	362	1.03	1.08 (1.01, 1.16)	4,142	4.54	-12.76 (-26.08, 0.57)
< 40th percentile (0.05 to < 11.4 ng/mL)	1,614	35	1.0	1.0	1,351	147	1.0	1.0	1,649	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)

AGA, appropriate for gestational age.

<sup>a</sup>Adjusted for maternal age, education, parity, smoking status, exposure year, state of residence, gestational age (term birth weight analysis only). <sup>b</sup>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]. <sup>c</sup>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]. <sup>d</sup>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 characterization?*

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 the bias from exposure could be downgraded '-' as did Johnson et al. (2014), or scored positively '+' as avoiding other biases.

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 of high quality. The study by Apelberg et al. (2007) would have been 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 meta-analysis 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 synthesizing 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)<sup>25</sup> carried out a nested case–control study to evaluate the association between exposure to cadmium (Ery-Cd) and low-trauma hip fracture risk.

#### 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.

<sup>25</sup> This is an open access publication: <https://link.springer.com/article/10.1007/s00223-013-9796-5>

### 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. Adjustment 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 (109/4,900). 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 characterization?*

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 study's 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 10.

**Table 10:** 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

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 (4,900) 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 analyzed 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 characterization?	++ 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 <sup>o</sup> 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.

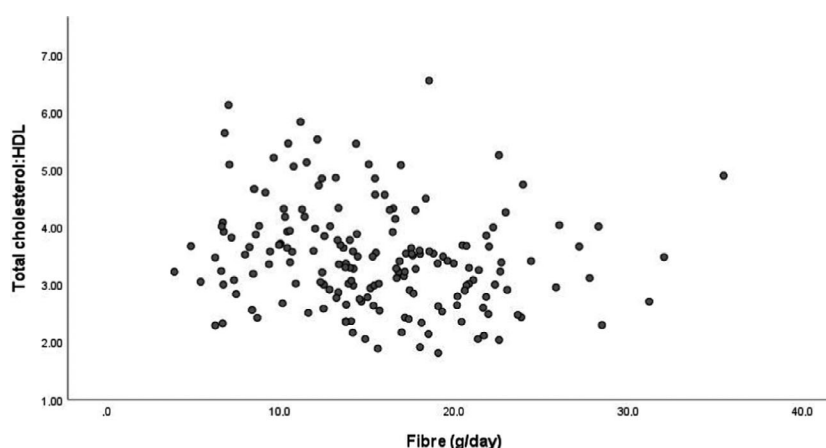
## Annex 6 – Categorization of continuous exposures: Analysis and interpretation

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 categorization 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 centres 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 analyzed 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. (Ingelsson et al., 2007; Hartley et al., 2016).

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 6. The distribution of exposure and outcome variables are shown in Table 11.



**Figure 6:** 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 years

**Table 11:** Distribution of the exposure and outcome variables

	Mean (SD)	Range
<b>Total-cholesterol: HDL</b>	3.5 (0.90)	1.8–6.6
<b>Total cholesterol (mg/dL)</b>	193 (33)	108–288
<b>HDL cholesterol (mg/dL)</b>	58 (14)	33–126
<b>Dietary fibre (g/day)</b>	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 12 below the regression model would be written as:

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

where  $x$  is fibre intake and  $\beta$  is the slope for the association between fibre intake and the total-cholesterol:HDL ratio; and  $\alpha$  is the value of that ratio at zero fibre intake. For this particular example the intercept  $\alpha$  gives the total-cholesterol: HDL ratio at 0 fibre intake, which does not exist in this population (the intercept  $\alpha$  is therefore of limited relevance). Table 12 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 12:** 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 $\beta$ (95% CI)	p for trend <sup>(a)</sup>	Adjusted <sup>(b)</sup> $\beta$ (95% CI)	p for trend <sup>(a)</sup>
-0.28 (-0.51, -0.06)	0.02	-0.35 (-0.60, -0.11)	0.005

(a): t-test with fibre intake entered as continuous variable in the crude regression model.

(b): Adjusted for age, body mass index, total energy intake and recruitment centre

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 6 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 12, some assumptions have to be made. One interpretation would be that the maximum possible increase in fibre intake in this population (see Table 11) 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 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. Secondly, 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.

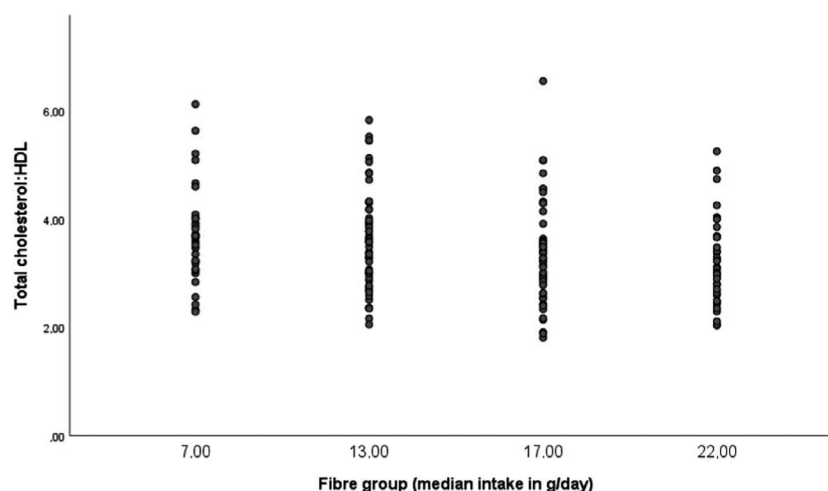
### Categorical exposure:

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 4 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 7 where the median intake in each fibre group is used. Figure 7 shows the same data as in Figure 6 but now participants in each group have been assigned the median value in their respective exposure category. As for Figure 6, 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 7:** Scatterplot of the cross-sectional association between dietary fibre intake categorized into 4 groups 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:

$$Y(x) = \alpha + \beta_2x_2 + \beta_3x_3 + \beta_4x_4$$

Here  $x_2$  is a binary 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  $\alpha$  represents the mean total-cholesterol:HDL ratio in group 1. The slopes  $\beta_2$ ,  $\beta_3$  and  $\beta_4$  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 13 (for the adjusted model more terms have been added to the model above).

**Table 13:** Association between dietary fibre intake and the total-cholesterol:HDL ratio. The dietary fibre variable has been divided into 4 categories and regression coefficients reflect the mean change relative to group 1

Fibre intake (g/day)		Unadjusted	Adjusted <sup>(a)</sup>
Group	Median (range)	$\beta$ (95% CI)	$\beta$ (95% CI)
1 (n = 32)	7 (4 – < 10)	Referent (3.71 <sup>(b)</sup> )	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 <sup>(c)</sup>		0.01	0.01

(a): Adjusted for age, body mass index, total energy intake and recruitment centre.

(b): The mean total-cholesterol: HDL ratio is 3.71 in group 1.

(c): t-test with the categorical fibre variable modelled as continuous variable using the median intake in each category (e.g. 7, 13, 17 and 22 g/day).

The adjusted results in Table 13 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 11).

In this example it appears that the assumption of linearity may be justified. Going back to Table 12, 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 13 is obtained by entering the categorical variables (4 values) as continuous variable in the regression model where the median values in each group has 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 12 above, 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 13 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 (Orsini et al., 2012; Crippa and Orsini, 2016; Crippa et al., 2018)

### 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 and Vickers, 2012). That is 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 11 and 12 above where exposure was entered as the original continuous variable or grouped continuous variable in the regression model. Alternatively, in Table 13 an F-test testing if all 4 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 identifying NOAEL relative to a control dose group in a toxicological experiment is considered more uncertain compared to identifying a reference point 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 or 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.