TECHNICAL REPORT

APPROVED:

doi:10.2903/sp.efsa.20YY.EN-NNNN

Draft Bisphenol A (BPA) hazard assessment protocol

European Food Safety Authority (EFSA),
Ursula Gundert-Remy, Johanna Bodin, Cristina Bosetti, Rex FitzGerald, Annika Hanberg, Ulla Hass, Carlijn Hooijmans, Andrew A. Rooney, Christophe Rousselle, Henk van Loveren, Detlef Wölfle, Fulvio Barizzone, Cristina Croera, Claudio Putzu and Anna F. Castoldi

Endorsed for public consultation on 14 June 2017[1]

Abstract

To ensure an efficient, transparent and methodologically rigorous re-assessment of the safety for consumers of bisphenol A (BPA), the European Food Safety Authority (EFSA) has undertaken the task to develop a protocol detailing *a priori* the approach and methodology for performing BPA hazard identification and characterisation. The general aim of this hazard assessment will be to assess whether the new scientific evidence (published from 2013 onwards and not previously appraised by EFSA) still supports the current temporary Tolerable Daily Intake (t-TDI) for BPA of 4 µg/kg bw per day. In line with the principles highlighted in the EFSA project on promoting methods for evidence use in scientific assessments (Prometheus, https://www.efsa.europa.eu/en/efsajournal/pub/4121), the protocol states upfront and in detail the methods and/or the criteria that will be used in the planned BPA re-evaluation for data collection, study inclusion, evidence appraisal and integration. To pursue the goal of openness, this draft protocol will be subjected to a web-based public consultation and will be presented publicly in a stakeholder event. All the relevant comments and feedback received through these procedures will be considered and included in the revision of the protocol which will be implemented in its final form in the next BPA re-evaluation.

Key words:

Requestor: EFSA

Question number: EFSA-Q-2017-00491

Correspondence: FIP@efsa.europa.eu

[1] by the EFSA Panel on Food Contact Materials, Enzymes, Flavourings and Processing Aids (CEF)

Acknowledgements: EFSA wishes to thank Trine Husøy, Irene Munoz Guajardo and Camilla Smeraldi for the support provided to this scientific output.

Suggested citation: EFSA (European Food Safety Authority), Gundert-Remy U, Bodin J, Bosetti C, FitzGerald R, Hanberg A, Hass U, Hooijmans C, Rooney AA, Rousselle C, van Loveren H, Wölfle D, Barizzone F, Croera C, Putzu C and Castoldi AF, 2017. Draft Bisphenol A (BPA) hazard assessment protocol. EFSA supporting publication 20YY:EN-NNNN. 58 pp. doi:10.2903/sp.efsa.20YY.EN-NNNN

ISSN: 2397-8325

Table of contents

1. Introduction

The development of this protocol detailing the strategy for the hazard assessment of BPA (hazard identification and characterisation) was initiated as an EFSA self-task, as described in mandate M-2016-0207 (EFSA-Q-2016-00673). This was triggered by the need to ensure that the EFSA Panel on Food Contact Materials, Enzymes, Flavourings and Processing Aids (CEF Panel) is prepared for the upcoming re-evaluation of the safety for consumers of BPA, once the results of the two-year US National Toxicology Programme (NTP)/ Food and Drugs Administration (FDA) toxicity study become available in 2017/2018.

After the initiation of this work, EFSA received an additional mandate from the European Commission (EC, EFSA-Q-2016-00635) to re-evaluate the safety for consumers of BPA, which requires setting-up a BPA hazard assessment protocol as a first step.

These two independent mandates from EFSA and the European Commission are reported in Sections 1.1 and 1.2, respectively.

1.1. Background and Terms of Reference as provided by EFSA

This work aims to ensure that the CEF Panel will be fully prepared to engage in a re-evaluation of the safety for consumers of BPA (to set a full TDI) when the two-year ongoing NTP study report becomes available in 2017. In its latest risk assessment published in 2015, the CEF Panel reduced and set the TDI for BPA on a temporary basis to account for uncertainties related to possible BPA effects at low doses on mammary gland, reproductive, neurological, immune and/or metabolic systems, thus committing to a re-evaluation of the TDI in light of the new data available. Although the NTP study design covers all the most controversial issues, at the same time the extensive body of new literature that is being published on BPA cannot be ignored, and this is deemed appropriate for the applicability of a defined protocol in the context of the EFSA PROmoting METHods for Evidence Use in Science (Prometheus) project (EFSA, 2015) .

The Assessment and Methodological support (AMU) Unit will assist in the methodology and design of the protocol to be followed for the risk assessment. The sensitivity of the topic at EU level would also benefit from an early involvement of some Member States and/ or sister agencies.

The Food Ingredients & Packaging (FIP) Unit should ensure that the CEF Panel is fully prepared to engage in a re-evaluation of the safety for consumers of BPA (setting a full TDI) in compliance with the principles of Prometheus, when the two-year ongoing NTP study report becomes available in 2017.

Terms of reference

To ensure preparedness in view of an upcoming evaluation in 2017, the FIP Unit is invited to develop a protocol detailing the strategy for the hazard assessment of BPA (hazard identification and characterisation) to be endorsed by the CEF Panel. The protocol should also define *a priori* how the new evidence will be appraised for relevance and reliability.

1.2. Background and Terms of Reference as provided by the European Commission

EFSA has accepted a mandate upon request from the European Commission to perform a re-evaluation of the risks to public health related to the presence of bisphenol A (BPA) in foodstuffs and protocol for the risk assessment strategy. The background to this mandate, as provided by the EC is the following:

"In 2015 you published an opinion setting out a new temporary Tolerable Daily Intake (t-TDI) for BPA. Recently you received a request for the re-evaluation of this TDI from the Dutch authorities. Following the new temporary TDI and pending the outcome of the discussions following the Dutch request, the

Commission plans to hold a vote on a draft Commission Regulation in the Standing Committee on Plants, Animals, Food and Feed. This draft Regulation would lower the specific migration limit for BPA from plastic food contact materials and would apply the same limit to food contact varnishes and coatings.

The on-going discussions however highlight the need of additional information on various toxicological aspects. The study which was notified to EFSA by the Dutch authorities concerns the potential effects of BPA on the immune system. However more data are needed about other toxicological endpoints of BPA, including those relating to the mammary gland, reproductive, metabolic, and neurological systems. In your 2015 opinion which set the t-TDI, these endpoints were taken into account by means of an uncertainty evaluation.

In the mentioned opinion, EFSA assigned the temporary status to the TDI in recognition of the partially uncertain toxicology and because of its awareness of ongoing studies addressing the uncertainties. Therefore it is appropriate that the risk assessment you published in 2015 is refined.

It is essential that well-defined and transparent scientific criteria concerning the selection of the new scientific studies are laid down in advance of the re-evaluation. This would enable a comprehensive assessment of all relevant and adequate studies, and avoid the need to react to ad-hoc requests concerning individual scientific studies. The efficiency of work would thus be maximised.

My services have taken due note of the work that you have already undertaken in this respect and welcome the establishment of an ad hoc Working Group of experts including those from EFSA, external experts and those from Member States to set clear review criteria for the scientific evidence on BPA. Therefore, taking into account the timing for the activities involved in this work as foreseen by EFSA, including a public consultation, as the first part of this mandate the Commission therefore kindly requests EFSA:

- *To establish a protocol detailing the criteria for new study inclusion and for toxicological evidence appraisal for the re-evaluation of BPA as soon as possible, to ensure an efficient and transparent re-assessment of BPA.*

Once this work is complete, the Commission will kindly request EFSA the second part of this mandate:

- *To re-evaluate the risks to public health related to the presence of BPA in foodstuffs, taking into account the results of all relevant scientific data insofar as it meets the criteria laid down in the protocol mentioned above and in line with the terms of reference set out in the annex to this letter.*

Whilst we consider it important to send this mandate now, the Commission views it as premature at this stage to establish a deadline for the completion of the re-evaluation. Therefore, the Commission is asking you to inform us on a feasible timeline for the second part of this mandate.

The present mandate does not include the re-evaluation of the exposure to BPA. At present the Commission considers that there is no justification for such a re-evaluation. If this changes in the future, the Commission will provide you with a specific mandate.

Terms of Reference

In accordance with Article 29(1)(a) of Regulation (EC) No 178/2002, the European Commission asks EFSA to:

- *establish a protocol detailing the criteria for new study inclusion and for toxicological evidence appraisal for the re-evaluation of BPA, to ensure an efficient and transparent re-assessment of BPA;*
- *re-evaluate the risks to public health related to the presence of bisphenol A (BPA) in foodstuffs. In particular, the re-evaluation should take into consideration new data available from the results of the US NTP/ FDA study due in 2017 as well as all other new available information not previously evaluated by EFSA and which fulfil the criteria laid down in an established protocol. This re-evaluation should seek to clarify the remaining uncertainties concerning the toxicological endpoints of BPA, especially those concerning the mammary gland, reproductive, metabolic,*

neurobehavioural and immune systems and to establish a full tolerable daily intake (TDI) on the basis of the new information available."

1.3. Interpretation of the Terms of Reference

To address both mandates, the protocol should define *a priori* the following processes inherent to BPA hazard identification and characterisation:

- problem formulation (Section 2)
- gathering the evidence (Section 3)
- selecting the evidence (Section 4)
- collecting the data from the included studies (Section 5)
- appraising and weighing the evidence (Sections 6-8 and Appendices B-C)
- synthesis of the results (Appendix D)
- hazard characterisation (Section 10)
- uncertainty analysis (Section 11)

Protocol development is part of the on-going EFSA Prometheus project (EFSA, 2015) aimed at further enhancing the methodological rigour, transparency and openness of EFSA scientific assessments. In this context, the hazard assessment of BPA was chosen as a case-study to test the importance of performing the assessment in two separate steps: 1) planning (protocol development) and 2) implementation of the protocol.

2. Problem formulation

2.1. Objectives of the hazard assessment

The general aim of this hazard assessment is to assess whether the new scientific evidence (published after 31/12/2012, and not previously appraised by the EFSA CEF Panel in 2015 and 2016) still supports the current temporary TDI (t-TDI) for BPA of 4 µg/kg bw per day.

More specifically, the evaluation will cover:

(i) the adverse effects in humans associated with the exposure to BPA *via* any route;

(ii) the adverse effects in animals after oral exposure to BPA at doses below the cut-off of 10 mg/kg bw per day (based on the benchmark dose lower confidence interval ($BMDL_{10}$) in mice used by the EFSA CEF Panel to set the t-TDI in 2015) or after subcutaneous exposure to BPA doses below the cut-off of 0.5 mg/kg bw per day (based on the ratio of oral bioavailability and of subcutaneous systemic availability in mice) or after dermal BPA exposure at any dose;

(iii) the human and animal toxicokinetics of BPA.

The scientific evidence needed to directly address these objectives will be dealt with by applying a narrative or a systematic approach as explained in details in the following sections.

The evaluation will deal with evidence available after the closing date of the literature search of the previous EFSA BPA risk assessment (EFSA CEF Panel, 2015). The few studies published after 31/12/2012 and already appraised by the CEF Panel in its 2015 opinion or in its 2016 statement on immunotoxicity (EFSA CEF Panel, 2016) will not be re-appraised. The conclusions from these two previous EFSA assessments of BPA will be the starting point of the new assessment.

2.2. Target population

The target population of the hazard assessment is the EU general population, including specific vulnerable groups (unborn children and breast-fed infants).

2.3. Chemical of concern

The target chemical substance is bisphenol A (BPA; chemical formula $C_{15}H_{16}O_2$, CAS No 80-05-7 and EC No 201-245-8). BPA derivatives will not be object of the assessment.

2.4. Endpoints relevant to the hazard assessment

The potential adverse effects of BPA have been extensively characterised in previous risk assessments by EFSA (EFSA, 2006, 2008; EFSA CEF Panel, 2010, 2011, 2015) and international bodies such as WHO/FAO (2011) and US FDA (2013). Table 1 summarises the outcome of the Weight of Evidence (WoE) approach carried out in the 2015 EFSA comprehensive review on BPA (EFSA CEF Panel, 2015) and the degree of likelihood for the effects under consideration on the basis of the then available human and animal evidence.

Table 1: Likelihood[2] for BPA effects (as taken from outcome of WoE approach in EFSA CEF Panel, 2015)

Effects classified as "Likely"
• General toxicity (liver and kidney weight increase) • Mammary gland proliferative changes
Effects classified "As likely as not"
• Reproductive and developmental effects • Neurological, neurobehavioural and neuroendocrine effects • Immune effects • Cardiovascular effects • Metabolic effects • Carcinogenicity
Effects classified as "Unlikely"
• Genotoxicity

In the 2015 BPA risk assessment (EFSA CEF Panel, 2015), only the "Likely" effects of BPA (increase of liver and kidney weight and mammary gland proliferation) were brought forward for dose-response analysis and for defining the reference point for the health-based guidance value. The effects classified as "As likely as not" were considered in the uncertainty analysis and were taken into account in the definition of an extra factor for the derivation of the t-TDI.

The mean relative kidney weight increase in the two generation study in mice by Tyl *et al*. (2006, 2008), for which a BMDL10 of 8.96 mg/kg bw per day was calculated, was used as the basis of a revised TDI (EFSA CEF Panel, 2015).

This dose in mice was extrapolated to an oral Human Equivalent Dose (HED) using the so called HED approach. This approach could be used in the 2015 EFSA CEF Panel opinion on BPA because of the availability for this chemical of (i) a solid base of toxicokinetic data in various laboratory animal species providing internal dose metrics for neonatal-to-adult stages and for different routes of exposure; (ii) physiologically-based pharmacokinetic (PBPK) models predicting internal exposures in laboratory animals and humans in a route-specific manner. In 2015, the HED value of 609 µg/kg bw per day was obtained by multiplying the mice BMDL10 by the Human Equivalent Dose Factor (HEDF) of 0.068 for oral exposure of adult mice. This HED was taken as the reference point for setting the new health-based guidance value for BPA. A t-TDI of 4 µg BPA/kg bw per day was obtained by dividing the HED by an overall uncertainty factor of 150 to account for intra-species differences (factor

[2] It is important to emphasise that the WoE approach referred specifically to hazard identification, i.e. it referred to the likelihood of an association between BPA exposure at any dose and the effect under consideration and not to the likelihood or frequency of the effect actually occurring in humans.

of 10), inter-species toxicodynamic differences (factor of 2.5) and uncertainties in the database regarding mammary gland, reproductive, neurobehavioural, immune, and metabolic systems (extra factor of 6). Notably, the default uncertainty factor of 4 for interspecies kinetic differences was already accounted for by the use of the chemical-specific approach, in which the ratio of the Area Under the Curve (AUC) in animals to the AUC in humans was used to adjust the external doses in animals to the external doses in humans.

2.5. Identification of the hazard assessment sub-questions

This section illustrates the hazard assessment sub-questions to be answered and the review approach, *i.e.* narrative *vs.* systematic[3], to follow for the new BPA re-evaluation (Table 2).

In the hazard assessment a full systematic process will be followed to address human and animal evidence of outcomes related to the exposure of BPA focussing on those effects considered "Likely" and "As likely as not" in 2015 (Tables 1 and 2), with the intention to decrease the uncertainty surrounding the conclusions.

Since the conclusions of the 2015 BPA opinion (EFSA CEF Panel, 2015) were underpinned by a thorough review of toxicokinetic data in different animal species and PBPK models to derive an oral HED, new evidence addressing BPA toxicokinetics in humans and animals will be reviewed - using a narrative approach - to evaluate whether the previously used HEDF should be changed. The definition of the various HEDF to be used for dose extrapolation from animal to human according to species, exposure time and route will be determined at the WoE step to ensure comparability of the effects across different studies and species.

Since the evidence from the EFSA CEF Panel 2015 opinion was not conclusive with regards to the mode of action of BPA, new studies that could potentially help elucidating this aspect will be part of the review and will be dealt with narratively.

Additional sub-questions will refer to the assessment of the dose-response relationship and an evaluation of possible uncertainties, for example those derived from consideration of the toxicokinetic and toxicodynamic properties of BPA and from considerations of inter-species variability in case animal data are being used for deriving a health-based guidance value.

[3]For a comparison between a systematic and a narrative review, the reader should refer to Table 2 of the Guidance of EFSA (2010): Application of systematic review methodology to food and feed safety assessments to support decision making. EFSA Journal 2010; 8(6):1637. [90 pp.]. doi:10.2903/j.efsa.2010.1637. Available online: http://onlinelibrary.wiley.com/doi/10.2903/j.efsa.2010.1637/epdf

Table 2: Hazard assessment sub-questions

Q#	Hazard assessment step	Hazard assessment sub-questions	Approach
1	Hazard Identification	Is there new evidence with regards to any association between exposure to BPA at any pre- and/or postnatal life stage and general toxicity *(e.g.* liver and kidney), or reproductive and developmental, neurological, immune, cardiovascular, metabolic, mammary gland or carcinogenic outcomes in humans?	Systematic
2	Hazard Identification	Is there new evidence with regards to any association between oral (below 10 mg BPA/kg bw per day), subcutaneous (below 0.5 mg/kg bw per day) or dermal exposure to BPA at any pre- and/or postnatal life stage and general toxicity *(e.g.* liver and kidney) or reproductive/developmental, neurological, immune, cardiovascular, metabolic mammary gland or carcinogenic outcomes in mammalian animals?	Systematic
3	Hazard Identification	Is there new evidence with regards to BPA genotoxicity *in vitro* or *in vivo*?	Narrative
4	Hazard Identification	Is there new evidence with regards to an association between exposure to BPA at any pre- and/or postnatal life stage and any outcome not mentioned in Q1 in humans?	Systematic
5	Hazard Identification	Is there new evidence with regards to an association between exposure to BPA at any pre- and/or postnatal life stage and any outcome not mentioned in Q2 in mammalian animals?	Systematic
6	Hazard Identification	What is the new evidence with regards to the mode of action (MoA) of BPA arising from *in vitro* studies at concentrations lower than 100 nM[4] ?	Narrative
7	Hazard Identification	Is there new evidence with regards to the MoA of BPA arising from other MoA studies?	Narrative
8	Hazard characterisation	Is there new evidence with regards to BPA toxicokinetics in humans?	Narrative
9	Hazard characterisation	Is there new evidence with regards to BPA toxicokinetics in experimental mammalian animal species/strains?	Narrative
10	Hazard characterisation	Does the new evidence on the toxicokinetics of BPA in humans and experimental mammalian animals still support the same HED factors used in the 2015 EFSA opinion on BPA?	Informed by sub-questions 8 & 9
11	Hazard characterisation	What is the dose-response relationship for relevant outcomes in humans?	Informed by sub-questions 1 & 4
12	Hazard characterisation	What is the dose-response relationship for relevant outcomes in experimental animals according to the new evidence?	Informed by sub-questions 2, 3 & 5

3. Methods for gathering the evidence

3.1. Time span of evidence search

The evaluation will deal with new evidence available since 1 January 2013. The studies published in 2013 and already appraised by EFSA in its 2015 opinion on BPA or in its 2016 statement on immunotoxicity of BPA (EFSA CEF Panel, 2015 and 2016) will not be re-assessed in the re-evaluation.

[4] See rationale for choice of this cut-off concentration in Section 4.2.4.3

The proposed ending date is 31/12/2017 unless the publication of the NTP/FDA study is delayed.

3.2. Information sources

Literature searches will be conducted in the following bibliographic databases (see Appendix A):

- PubMed
- Web of Science™ Core Collection
- Scopus
- Toxline + DART (TOXNET platform)

Furthermore, EFSA will carry out a call for data in order to gather study reports and other information.

An open search strategy will be used, including only the terms "BPA" or "Bisphenol A" and synonyms with a view to capture as many records as possible.

The search strings proposed to be used for each database search are annexed in Appendix A and will be published in their final form in the BPA scientific opinion, to ensure transparency and reproducibility of the results.

3.3. Type of evidence

Only primary research studies will be considered for the assessment.

Reviews will only be used to check whether they contain additional references of primary studies that have not been captured by the literature search/call for data.

Comments, letters to the editors, book chapters, poster and/or conference abstracts will be excluded.

3.4. Management of the information

The evidence retrieved from each bibliographic database or obtained through the call for data will be imported in the bibliographic reference management software EndNote and combined together. A first removal of duplicates will be done at this step using the functionality available in the EndNote reference manager software.

The EndNote file obtained from the merge of the records retrieved from the different sources of information will be uploaded into an online systematic review tool, DistillerSR[5], for the subsequent steps of the review.

Following uploading of the records into DistillerSR, removal of duplicates will again be undertaken, using the Duplicate Detection feature of the tool.

4. Methods for selecting the studies

4.1. Screening of titles and abstracts

The titles, and where available, the abstracts identified in the searches described in Section 3 and Appendix A will be screened for relevance to the general scope of the assessment: is the paper relevant to (i) exposure to humans OR (ii) exposure to animals OR (iii) mode of action.

The screening of titles and abstract will be performed by two independent reviewers.

The Distiller SR tool will allow for the identification of potential disagreements between the two reviewers on study eligibility.

In case of disagreement between the two independent reviewers, the paper will be automatically brought to the next screening phase, *i.e.* at the level of full text.

[5] DistillerSR - https://www.evidencepartners.com/products/distillersr-systematic-review-software/

4.2. Examining full-text reports for eligibility of studies

For records that pass the first screening based on titles and abstracts, the full text will undergo a second screening against the inclusion criteria by means of two independent reviewers.

This step will also serve for the first categorisation of the studies into the different health outcome categories identified in the sub-questions in Table 2.

The possibility of an "unclear" reply is no longer foreseen at this stage because all the information needed for taking a decision will be available in the full text.

In case of disagreement, the two independent reviewers will discuss the paper in order to find a solution to solve it. In case an agreement between the two reviewers cannot be found the paper will be brought to the attention of the Working Group (WG) on BPA assessment for the final decision.

Study reports and other information made available through the call for data to EFSA will also follow this procedure.

4.2.1. Availability of full-text and language

Availability of the full text in English will be a pre-requisite for an article to be included in the assessment. Furthermore, authors of publications other than in English will have the opportunity to submit their full text articles translated in English for consideration by EFSA through the call for data. Such translated studies will undergo the same appraisal process as the other literature as far as they me*et all* the other applied inclusion criteria.

The reviewers will thus be asked to reply to these questions:

- Is the full text available?
- Is the full text in English?

If the answer to one of these two questions is "No", the record will be excluded from the assessment. If the answer to both questions is "Yes", the reviewers will be prompted to reply to the next question.

The decision of excluding publications in languages other than English may be a source of uncertainty, but it is based on the available EFSA resources

4.2.2. Type of studies

The reviewers will be asked to reply to the following question:

- Is the paper a primary or a secondary study?

If the answer to the question is "primary", the reviewers will be prompted to reply to the following question.

If the answer to the question is "secondary", the record will be excluded from the assessment but it will be used to check whether it contains additional references of primary studies that have not been captured by the literature search/call for data.

If the answer to the question is "other", the record will be excluded from the assessment.

4.2.3. Outcomes of interest

In the first instance the reviewers will be asked to confirm that the record relates to a study reporting information considered relevant to the review question *i.e.* on BPA exposure in humans or in animals or on the mode of action of BPA *(e.g. in vitro*, cell cultures, specific molecular pathways). Primary studies that are not aimed at studying effects associated with exposure to BPA *(e.g.* human biomonitoring studies) will be excluded at this step.

If the answer to the question is "Yes", the reviewers will be prompted to reply to the following question.

If the answer to the question is "No", the record will be excluded from the assessment.

The reviewers will then classify the studies considered relevant for the assessment as providing information on:

- Effects associated with human exposure to BPA
- Effects associated with animal exposure to BPA
- Mode of action (*in vitro,* non-mammalian animals, microbiota, *etc*.)

The effects considered relevant for the assessment will be classified in the following health outcome categories: general toxicity *(e.g.* liver and kidney), reproductive, developmental, neurological, immune, cardiovascular, metabolic, mammary gland or carcinogenic, genotoxic or any other effects with the addition of toxicokinetic studies.

One source may report on more than one outcome of interest and each outcome will be assessed separately.

4.2.4. Exposure of interest

4.2.4.1. Human studies

For human studies, all types of exposure to BPA (alone or in mixtures) will be considered, including occupational exposure scenario.

4.2.4.2. Animal studies

For animal studies to be considered for the assessment, exposure to BPA (not given as a part of a mixture) via the oral, subcutaneous or dermal routes will be investigated.

For oral studies, at least one of the doses tested must be below the oral cut-off value of 10 mg BPA/kg bw per day (based on the BMDL10 of 8.96 mg/kg bw per day used for the EFSA t-TDI in 2015) given that the main focus of the new BPA hazard assessment will be on low dose effects.

For subcutaneous studies a cut-off of 0.5 mg/kg bw per day (based on the ratio of oral bioavailability and of subcutaneous systemic availability in mice) for at least one of the tested subcutaneous doses will be applied.

For dermal studies, no cut-off values will be set for study inclusion because of the unavailability of the necessary toxicokinetic data for the calculation of the route-specific conversion factor.

Studies via the inhalation route will not be included as this is not considered as a relevant route of BPA exposure for consumers.

4.2.4.3. Mode of action studies

Studies that investigate possible mode of action of BPA must be conducted using BPA alone at concentrations that are considered to be in a toxicologically relevant range; hence *in vitro* studies will be considered only if at least one of the concentrations tested is below 100 nM. In defining this cut-off concentration, we have considered the concentration of unconjugated BPA in humans, as published by Thayer *et al.* (2015), at the exposure levels identified in the 2015 EFSA CEF Panel opinion. In addition a factor of 10 has been applied to account for the amount possibly being absorbed by the experimental devices.

For *in vitro* genotoxicity studies, this cut-off value will not be applied to ensure that this MoA can become manifest.

Concerning non mammal animal models *(e.g.* zebrafish) or other *in vivo* studies, no cut-off doses will be applied, hence studies will be included for mode of action analysis, if applicable.

4.2.5. Inclusion/exclusion criteria for human, animal and MoA studies

Tables 3, 4 and 5 schematically list the criteria for including or excluding from the review human, animal and MoA studies, respectively.

Only epidemiological studies with cohort and case-control designs will be systematically appraised for humans. Studies with a cross-sectional design bear some limitations in relation to the scope of the BPA review and therefore will only be considered in case of need for supporting information in a narrative manner.

Studies reporting either levels of unconjugated or conjugated BPA will be considered relevant, taking into consideration the limit of detection for the unconjugated BPA and the existing exposures.

Studies in which levels of BPA have been measured in human biological samples only once will not be included as exposure assessment is uncertain, with the exception of studies in pregnant women, which could be relevant for time-windows of exposure.

Table 3: Inclusion/exclusion criteria related to human studies

Sub-question 1: Is there an association between exposure to BPA at any pre- and/or postnatal life stage and general toxicity *(e.g.* liver and kidney), or reproductive and developmental, neurological, immune, cardiovascular, metabolic, mammary gland or carcinogenic outcomes in humans? Sub-question 3: Is there new evidence with regards to BPA genotoxicity *in vitro* or *in vivo*? (narrative approach)? Sub-question 4: Is there an association between exposure to BPA at any pre- and/or postnatal life stage and any outcome not mentioned in Q1 in humans? Sub-question 8: Is there new evidence with regards to BPA toxicokinetics in humans (narrative approach)? Sub-question 11: What is the BPA dose-response relationship for relevant outcomes in humans?		
Study design	**In**	Cohort studies Case-control studies (retrospective and nested) Toxicokinetic studies on any route of exposure (narrative approach)
	Out	Cross sectional studies Animal studies *In vitro* studies
Population	**In**	All populations groups, all ages, males and females
	Out	/
Exposure/ intervention	**In**	All routes of exposure All studies during pregnancy including those with single spot urine samples Studies in which levels of BPA have been measured in human biological samples more than once
	Out	Biomonitoring studies Studies with single spot urine samples in non-pregnant individuals
Language	**In**	English
	Out	Other languages
Time	**In**	From 01/01/2013 (except those which were already included in the 2015 opinion)
	Out	Before 2013
Publication type	**In**	Primary research studies *(i.e.* studies generating new data)
	Out	Secondary studies* Expert opinions, editorials, and letters to the editor PhD Theses Extended abstracts, conference proceedings

* they will be used to obtain additional references of primary research studies

Table 4: Inclusion/exclusion criteria related to experimental mammalian animal studies

Sub-question 2: Is there new evidence with regards to any association between oral (below 10 mg BPA/kg bw per day), subcutaneous (below 0.5 mg/kg bw per day) or dermal exposure to BPA at any pre- and/or postnatal life stage and general toxicity (*e.g.* liver and kidney) or reproductive/developmental, neurological, immune, cardiovascular, metabolic mammary gland or carcinogenic outcomes in mammalian animals? Sub-question 3: Is there new evidence with regards to BPA genotoxicity *in vivo* at any dose (narrative approach)? Sub-question 5: Is there an association between exposure to BPA at any pre- and/or postnatal life stage and any outcome not mentioned in Q2 in mammalian animals? Sub-question 9: Is there new evidence with regards to BPA toxicokinetics in experimental mammalian animal species/strains (narrative approach)? Sub-question 12: What is the dose-response relationship for relevant outcomes in experimental animals according to the new evidence??		
Study design	In	*In vivo* studies on animals not examining MoA Toxicokinetic studies (narrative approach)
	Out	Human studies *In vitro* studies
Population	In	All mammalian animals
	Out	Non-mammalian animals
Exposure/ intervention	In	Oral, sub-cutaneous and dermal. Studies in which levels of BPA have been measured in biological samples At least one tested dose below the cut-off of 10 mg/kg bw per day for orall studies or 0.5 mg/kg bw per day for subcutaneous studies (no cut-off is applied to dermal studies) All *in vivo* genotoxicity studies with no cut-off dose
	Out	Exposure routes other than oral, dermal, or subcutaneous Mixtures
Language	In	English
	Out	Other languages
Time	In	From 01/01/2013 (except those already included in the 2015 opinion and the 2016 statement on BPA immunotoxicity)
	Out	Before 2013
Publication type	In	Primary research studies *(i.e.* studies generating new data)
	Out	Secondary studies* Expert opinions, editorials, and letters to the editor PhD Theses Extended abstracts, conference proceedings

* they will be used to obtain additional references of primary research studies

401 **Table 5:** Inclusion/exclusion criteria related to MoA studies

Sub-question 3: Is there new evidence with regards to BPA genotoxicity *in vitro* at any concentration (narrative approach)? Sub-question 6: What is the evidence of the MoA of BPA arising from *in vitro* studies at concentrations lower than 100 nM (narrative approach)? Sub-question 7: What is the evidence of the MoA of BPA arising from other studies (narrative approach)?		
Study design	In	*In vitro* studies *In vivo* studies on MoA in humans, mammalian and non-mammalian animals
	Out	Human studies or *in vivo* studies not examining MoA
Exposure/ intervention	In	At least one concentration below the cut-off of 100 nM for *in vitro* studies (except for *in vitro* genotoxicity studies) All *in vitro* genotoxicity studies All routes of exposure for *in vivo* studies
	Out	Mixtures *In vitro* studies (except for *in vitro* genotoxicity studies) testing BPA only above 100 nM
Language	In	English
	Out	Other languages
Time	In	From 01/01/2013 (except those already included in the 2015 opinion)
	Out	Before 2013
Publication type	In	Primary research studies *(i.e.* studies generating new data)
	Out	Expert opinions, editorials, and letters to the editor PhD Theses Extended abstracts, conference proceedings Secondary studies

402 # 5. Methods for collecting the data from the included studies

403 ## 5.1. Data extraction

404 Pre-defined data extraction forms (see Tables 6-7) will be used for collecting the data from the
405 individual studies undergoing a systematic review approach and an internal validity appraisal. These
406 extraction forms will be implemented using Distillers SR, the same software used in the previously
407 described steps.

408

409 **Table 6:** Data extraction from human studies

Study ID	Reference:
	Study name and acronym (if applicable):
	Total number of subjects:
	Health outcome category[(a)]:
Funding	Funding source(s):
Study design	Cohort study Case control study
	Type of blinding:
	Year the study was conducted (start):
	Duration/length of follow-up:
	Dates of sampling (when relevant):
	Dates of analysis of BPA/BPA-conjugates in the samples:
Subjects	Number of participants in the study: Participation rates (%):
	Number of exposed/non exposed subjects : Follow-up rates by group (%):
	Sex (male/female):
	Geography (country, region, state, *etc.*):
	Age at exposure
	Race and ethnicity
	Socioeconomic background
	Confounders and other variables[(b)] as reported
	Outcome assessment *(e.g.* mean, median, measures of variance as presented in paper such as SD, SEM, 75th/90th/95th percentile, minimum/maximum):
	Inclusion and exclusion criteria:
Intervention/exposure	Exposure: - Measured levels in human biological samples *(e.g.* breast milk, blood, urine) and method used (Validation of the method, measures to avoid contamination of samples, *etc.*) - Estimated dietary exposure and method used (Validation of the method, measures to avoid contamination of samples, *etc.*)
Methods for endpoint assessment	Parameters measured (units of measure, measures of central tendency and dispersion, CI level)
	Diagnostic or method to measure health outcome (including self-reporting):
Statistical analysis	Statistical method used
Results	Measures of effect and corresponding confidence interval at each exposure level as reported in the paper, and for each sub-group when applicable: Were sub-groups analyses predefined (yes/no, if not, how was it justified?)? How were the variables treated (continuous or transformed or categorical)?
	Statistical test used: Modifying factors: other potential sources of bias considered in the analysis, and how they were considered:
	Shape of dose-response if reported by the authors *(e.g.* description of whether shape appears to be monotonic, non-monotonic, according to the study authors):

410 (a): General toxicity (*e.g.* liver and kidney), reproductive, developmental, neurological, immune, cardiovascular, metabolic *(e.g.*
411 diabetes, thyroid function, obesity), mammary gland or carcinogenic, genotoxic, other (more than one option should be
412 possible).
413 (b): Age, sex, race/ethnicity, education/sociodemographic characteristics, smoking status, BMI, dietary factors, alcohol
414 consumption, concurrent exposures (other chemicals/drugs)
415

Table 7: Data extraction from animal studies

Study ID	Reference:
	Year the study was conducted (start, if available):
	Health outcome category[(a)]
Funding	Funding source(s):
Type of study and guideline	Type of study [(b)]:
	GLP (yes/no): Guidelines studies (if yes specify):
Animal model	Species/(sub-)strain/line:
	Disease models *(e.g.* infection, diabetes, allergy, obesity, autoimmune disease):
Housing conditions and diet	Housing conditions (including cages, bottles, bedding)
	Diet name and source:
	Background levels of phytoestrogens in the diet (type and levels):
Exposure	BPA provider
	Compound purity (if available, specify impurities identified):
	Vehicle used: Dose regimen (dose level or concentration of BPA per group, and frequency):
	Route of administration (diet, gavage, subcutaneous):
	Period of exposure (pre-mating, mating, gestation, lactation, adult)
	Duration of the exposure:
Study design	Sex and age of the initially exposed animals:
	Number of groups/ number of animals per group:
	Randomisation procedures at start of the study:
	Reducing (culling) of litters and method:
	Number of pups per litter for next generation and methodology:
	Number of pups per litter/animals for certain measurements and methodology:
	Time of measurement/Observation period (premating, mating, gestation, lactation, adult):
	Endpoints measured:
	Methods to measure endpoint:
Statistical analysis	Statistical method used:
Results:	Concentration of the test compound in vehicle (analysed, stated, unclear):
	Documentation of details for dose conversion when conducted:
	Level of test compound(s) in tissue or blood:
	Results per dose or concentration *(e.g.* mean, median, frequency, measures of precision or variance):
	NOEL, NOAEL, LOEL, LOAEL, BMD/BMDL, and statistical significance of other dose levels (author's interpretation):
	Shape of dose response if reported by the authors *(e.g.* description of whether shape appears to be monotonic, non-monotonic, NA for single exposure or treatment group studies)

(a): General toxicity (*e.g.* liver and kidney), reproductive, developmental, neurological, immune, cardiovascular, metabolic *(e.g.* diabetes, thyroid function, obesity), mammary gland or carcinogenic, genotoxic, other (more than one option possible).

(b): *e.g.* acute, sub-acute (*i.e.* 4 weeks), subchronic (*i.e.* 13 weeks), chronic (*i.e.* 104 weeks), multigenerational, developmental, carcinogenicity.

6. Relevance of studies to the hazard sub-questions

After data extraction and study classification according to the health outcome category, each study will undergo the assessment of its relevance. For the sake of clarity, the reader should note that relevance will be assessed in two stages:

(i) Relevance of the studies to the sub-question (for both human and animal studies): as a first stage, the relevance will be evaluated for every endpoint in each individual study in relation to the specific hazard sub-question asked (see Table 2).

(ii) Relevance of effects for human health (for animal studies only): this assessment will take place after the WoE evaluation and will not be done for individual studies. It will be done for the effects identified in animals through the WoE analysis, for which an evaluation of the relevance to human health will be made by expert judgement (see Section 8).

The outcome of the evaluation of the relevance of the endpoint to the sub-question could be "yes (Y)", "unclear (U)" or "no (N)". Endpoints rated as having no relevance for any health outcome category will not be carried forward for internal validity appraisal. The judgement for individual study relevance will be included in the WoE table.

Each evaluation will be performed by two independent reviewers. In case of disagreement the reviewers will first discuss to try to find an agreement, and if this is not possible, the paper will be brought to the attention of the working group and EFSA staff who will take the final decision. The process will have to be fully documented, making sure a justification for the judgement is provided.

7. Internal validity of the studies

An overview of the whole process is provided schematically in Fig. 1.

For every endpoint in each study, two components of internal validity will be rated: (i) quality and (ii) risk of bias (RoB). The definition of what is meant by each term in the context of the present assessment is reported below.

Quality: Intrinsic ability of the methodology and experimental design of a study to provide accurate evidence with regards to the endpoint/effect under investigation.

Risk of bias (RoB): The systematic error caused by systematic differences (including environmental conditions or data handling, exposure, study design (blinding) between the control and experimental subjects (other than the (intervention or) exposure of interest).

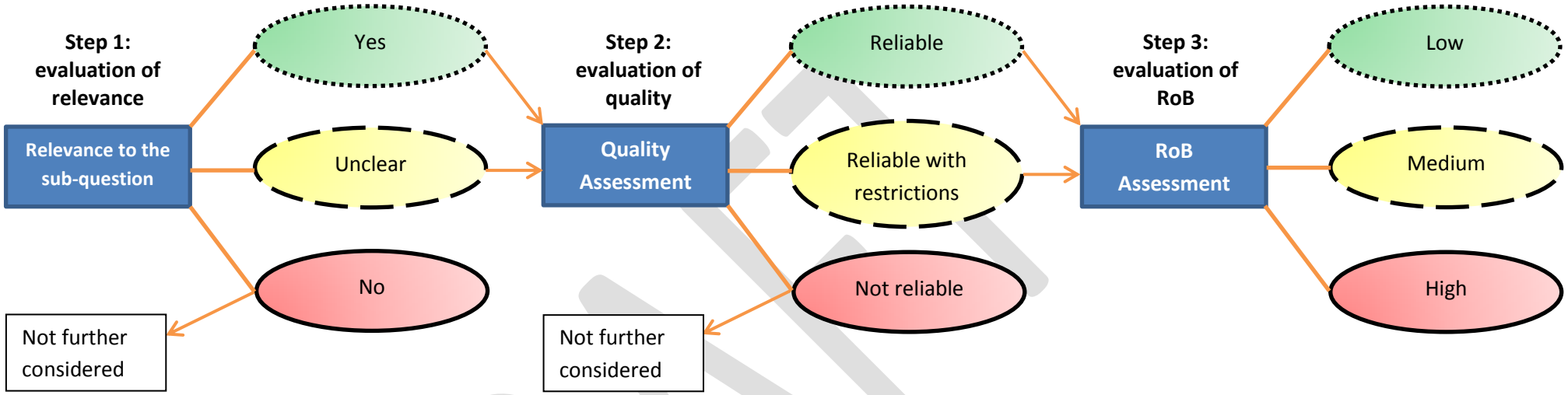
Each evaluation will be performed by two independent reviewers. In case of disagreements between the reviewers, they will discuss in order to find an agreement. If an agreement cannot be found, the paper will be brought to the attention of the working group and EFSA staff who will take the final decision. The appraisal will be conducted separately for each study by the outcome of interest reported. The evaluation of all studies will only be based on the reported information/data. Due to limited resources the study authors will not be contacted for clarifications or missing information.

The process (see flowchart below) will start with the evaluation of the study quality, at the end of which the study will be rated as either "Reliable without restrictions", "Reliable with restrictions" or "Not reliable". If the rating is "Not reliable", the study will not undergo any further assessment and this will be recorded.

For animal studies the appraisal of quality will be performed using criteria that were adapted from SciRAP (Beronius *et al.*, 2014, revised version at www.scirap.org, see Appendix B3). For human studies, since the criteria from SciRAP are not applicable, the assessment of quality will be performed using criteria that were adapted from the NTP Risk of Bias tool (NTP OHAT, 2015, see Appendices B1 and B2).

The quality assessment will be followed by the RoB evaluation using an adapted protocol from NTP developed separately for human and animal studies (NTP OHAT, 2015, see Appendix C). The appraisal of RoB will be performed only for those studies which have been rated for quality as "Reliable with or without restrictions". At the end of the RoB evaluation, the studies will be rated as "High RoB", "Medium RoB", or "Low RoB".

After an overall evaluation of the quality and RoB aspects has been made, the two distinct ratings will be combined to obtain three tiers of internal validity, from 1 to 3 corresponding to decreasing levels of internal validity. All the studies belonging to tiers 1-3 will be considered in the WoE analysis with different impacts on the conclusions drawn by expert judgement on the likelihood of an effect.



Step 4:setting tiers of internal validity

		Quality rating	
		Reliable without restrictions	**Reliable with restrictions**
Risk of Bias rating	**Low RoB**	**Tier 1**	**Tier 2**
	Medium RoB	**Tier 2**	**Tier 3**
	High RoB	**Tier 3**	**Not further considered**

Figure 1: Individual study appraisal summary scheme done separately for each endpoint

7.1. Quality appraisal of human studies

The appraisal tool developed for the assessment of the quality of human studies is presented in Table 8. The evaluation of quality will be performed focussing on four questions that EFSA experts have extrapolated and adapted from the NTP OHAT Risk of Bias rating tool for human studies (NTP OHAT, 2015). Three out of four questions are considered as key (A-C) with regards to the overall quality assessment. The question on statistics was deemed as non-key due to the difficulties in explicitly formalising the criteria for an appropriate statistical method. In line with the NTP methodology, each question can have four possible answers ranging from "Definitely appropriate (++)" or "Probably appropriate (+)" to "Probably not appropriate (-)" or "Definitely not appropriate (--)".

The instructions on how to rate each quality aspect can be found in Appendices B1 and B2.

Table 8: Quality appraisal tool for human studies (case-control and cohort study design)

#	Key Q	Question	Domain	Rating (++, +, -, --)
1	**A**	Can we be confident in the exposure characterisation (methods)?	Detection	
2	**B**	Can we be confident in the outcome assessment (methods)?	Detection	
3	**C**	Was the time-window between exposure and outcome assessment appropriate?	Exposure	
4		Do the statistical methods seem appropriate?	Other	
Overall assessment of quality				• **Reliable without restrictions (R)** • **Reliable with restrictions (RR)**

The ratings of the quality key and non-key questions (++, +, -, --) will be integrated to obtain an overall study quality rating *(i.e.* "Reliable without restrictions", "Reliable with restrictions", "Not reliable), as indicated below. This scheme will be tested in a pilot phase during the public consultation and might be modified according to the results and public consultation outcome.

Reliable without restrictions:

- At least two key questions (1-3) are scored with ++ and no key question is scored with - /- -
  AND
- the non-key question 4 is scored with + /++

Reliable with restrictions:

- All the other combinations not falling under either "Reliable without restrictions" or "Not reliable"

Not reliable:

- At least one key question (1-3) is scored with - / --

OR

- the non-key question 4 is scored with –

If the study quality rating is "Not reliable", the study will not undergo the RoB evaluation and will be excluded from the assessment. This will be recorded.

Only the studies that will be rated as "Reliable (with or without restrictions)" will be evaluated for RoB.

7.2. Risk of bias appraisal for human studies

The questions that will address the RoB of human studies are presented in Table 9. Two out six questions are considered key. Each individual element will be rated ("Definitely low RoB (++)", "Probably low RoB (+)", "Probably high RoB (-)", "Definitely high RoB (--)"). Whenever one of the elements to be appraised for RoB is not reported, this will be by default judged as "Probably high RoB". However, when there is indirect evidence that the element to be appraised was implemented in the correct way or would have not appreciably affected the results, a categorisation of "Probably low RoB" should be given.

The instructions on how to rate each RoB aspect can be found in Appendices C1 and C2.

Table 9: Risk of bias appraisal tool for human studies (case-control and cohort study design)

#	Key Q	Question	Domain	Rating (++, +, -, --)
1	**A**	Did selection of study participants result in appropriate comparison groups?	Selection	
2	**B**	Did the study design or analysis account for important confounding and modifying variables?	Confounding	
3		Were outcome data completely reported without attrition or exclusion of experimental units from analysis?	Attrition	
4		Was the exposure characterised consistently across study groups?	Detection	
5		Was the outcome assessment adequately blinded and consistent across study groups?	Detection	
6		Were all measured outcomes reported?	Selective reporting	
Overall rating				• **Low RoB** • **Medium RoB** • **High RoB**

The ratings of the RoB key and non-key questions (++, +, -, --) will be integrated to obtain an overall study RoB rating ("Low RoB", "Medium RoB", or "High RoB") as follows.

Low RoB:

- All the key questions (1, 2) are scored with +/++
  AND
- No more than two non-key questions (≤2) (3, 4, 5, 6) are scored with -
  AND
- No non-key question is scored with --

Medium RoB:

All the other combinations not included either in "Low RoB" or in "High RoB"

High RoB:

- One key question is scored with -/--
  OR
- Any non-key question is scored with --

7.3. Quality appraisal of animal studies

The appraisal tool developed for the assessment of the quality of animal studies is presented in Table 10. The quality criteria which have been taken from the SciRAP tool have been adapted by the experts for the purpose of BPA safety assessment.

Six aspects were considered as having a higher relative weight and were identified as key (A-F) when determining the overall quality rating.

Each quality aspect will be evaluated as "Fulfilled" (F), "Partially Fulfilled" (PF) or "Not Fulfilled" (NF).

The instruction on how to obtain these ratings can be found in Appendix B3.

Table 10: Quality appraisal tool (adapted from SciRAP)

#	Key Q	Quality aspect	Rating [Fulfilled (F), Partially Fulfilled (PF) Not fulfilled (NF)]
1		The test compound or mixture was unlikely to contain any impurities that may significantly have affected its toxicity.	
2	**A**	A concurrent negative control group was included.	
3	**B**	A reliable and sensitive animal model was used for investigating the test compound and selected endpoints.	
4		Animals were individually identified.	
5		Housing conditions (temperature, relative humidity, light-dark cycle) were appropriate for the study type and animal model.	
6		The number of animals per sex in each cage was appropriate for the study type and animal model.	
7		The test system is unlikely to contain contaminants that could affect study results, such as phytoestrogens and estrogenic contamination.	
8		An adequate number of doses was selected	
9	**C**	The timing and duration of administration do not seem to be inappropriate for investigating the included endpoints.	
10	**D**	Reliable and sensitive test methods were used for investigating the selected endpoints.	
11	**E**	Measurements do not seem to have been collected at unsuitable time point in order to generate sensitive, valid and reliable data.	
12	**F**	The statistical methods have been clearly described and do not seem inappropriate, unusual or unfamiliar and a sufficient number of animals per dose group was used	
Overall assessment of quality			• **Reliable without restrictions (R)** • **Reliable with restrictions (RR)** • **Not reliable (NR)**

The following criteria were set in order to rate the study quality as either "Reliable without restrictions", "Reliable with restrictions" or "Not reliable".

Reliable without restrictions:

- All the key questions (2, 3 , 9, 10, 11, 12) are FULFILLED
  AND
- All the non-key questions (1,4,5,6, 7, 8) are at least PARTIALLY fulfilled

Reliable with restrictions:

All the other combinations not falling under either "Reliable without restrictions" or "Not Reliable"

Not reliable:

- One of the key questions (2, 3, 9, 10, 11, 12) is NOT fulfilled
  OR
- All the key questions (2, 3, 9, 10, 11, 12) are at least PARTIALLY fulfilled
  AND
- More than 3 non-key questions (1, 4, 5, 6, 7, 8) are NOT fulfilled.

As was the case for the appraisal of human studies, if the study quality is rated as "Not reliable" the study will not be evaluated for RoB and will be excluded from the assessment.

7.4. Risk of bias appraisal for animal studies

The questions that will address the RoB of animal studies are presented in Table 11. Each individual element will be rated ("Definitely low RoB (++)", "Probably low RoB (+)", "Probably high RoB (-)", "Definitely high RoB (--)").

The instructions on how to rate each RoB aspect can be found in Appendix C3.

Table 11: Risk of bias tool for animal studies

#	Key Q	Question	Domain	Rating (++, +, -, --)
1	**A**	Was administered dose or exposure level adequately randomised?	Selection	
2		Was allocation to study group adequately concealed	Selection	
3		Were experimental conditions identical across study groups?	Performance	
4	**B**	Were outcome data completely reported without attrition or exclusion from analysis?	Attrition	
5		Can we be confident in the exposure characterisation?	Detection	
6	**C**	Can we be confident in the outcome assessment?	Detection	
7		Were all measured outcomes reported?	Selective Reporting	
Overall rating				• **Low RoB** • **Medium RoB** • **High RoB**

As a last step, the scores for the different questions will be integrated to obtain the study overall RoB estimate ("High RoB", "Medium RoB" or "Low RoB").

Three key questions, with higher relative weight when determining the overall RoB rating, were selected. The rules to follow to obtain the integrated RoB rating are shown below.

Low RoB:

- All the key questions are scored with + /++

AND

- None of the non-key questions are scored with - - and no more than two with -.

Medium RoB:

- All the other combinations not falling under either "High RoB" or "Low RoB".

High RoB:

- Any key question is scored with a – /--

OR

- More than two non-key questions are scored with a – / --.

7.5. Conclusion on internal validity of human and animal studies

After an overall evaluation of the quality and RoB aspects has been made, the two distinct ratings will be combined to obtain three tiers of internal validity, from 1 to 3 corresponding to decreasing levels of internal validity, as shown in Table 12.

Please note when the quality appraisal of a study is rated as "Not reliable", the study will not undergo the RoB evaluation and will then be excluded from the assessment ("Not further considered" in the matrix below). Similarly, studies rated as "Reliable with restrictions" with "High RoB" will be excluded from any further assessment. Only studies considered as "Reliable without restrictions" and with "Low RoB" will be allocated to Tier 1. Those "Reliable without restrictions" will be allocated to either Tiers 2 or 3 depending on the rating of the RoB as "Medium" or "Low", respectively. The studies "Reliable with restrictions" will be classified in Tier 2 if the RoB is rated as "Low" or Tier 3 if RoB is rated as "Medium".

Table 12: Internal validity

		Quality rating		
		Reliable without restrictions	**Reliable with restrictions**	**Not reliable**
Risk of Bias rating	**Low RoB**	Tier 1	Tier 2	Not further considered
	Medium RoB	Tier 2	Tier 3	Not further considered
	High RoB	Tier 3	Not further considered	Not further considered

8. Weight of evidence approach

Following the appraisal of the individual human and animal studies for internal validity and relevance to the questions, the experts will evaluate the confidence in the overall body of evidence by applying a WoE approach supported by a graphical representation of the study features, results and appraisal. This analysis will be performed in accordance with the draft guidance on the use of the WoE approach in scientific assessments by the EFSA Scientific Committee (2017). The WoE evaluation will allow an estimation of the overall likelihood that BPA is hazardous with respect to certain toxicological endpoints in human/animal studies, considering the findings for all of the exposure levels examined.

In order to be able to combine the whole body of evidence from animal studies into a single figure, potential differences in internal exposure have to be considered due to interspecies toxicokinetic differences. To deal with such differences, the HED concept has been used in the previous EFSA CEF Panel opinion on BPA (2015), as explained in Section 2.4 of this protocol. To apply this concept in the re-evaluation, data on the area under the plasma concentration-time profile (AUC) for animal species and humans will be obtained from all the then available publications or other sources: after adjusting for the same dose and assuming linear kinetics, the ratio of the AUCs in animal species and humans will be calculated. This factor, the HEDF, will be used to convert the doses in the animal studies to the corresponding human doses, thus enabling to compare the doses at which effects are observed in various species.

For the calculation of the HEDFs in the EFSA CEF Panel opinion in 2015, the AUC in humans was derived from PBPK-simulation. In view of the BPA re-evaluation new toxicokinetic studies with BPA in human volunteers have become available and these will be used for determining the AUC in humans. Similarly, new toxicokinetic data in animals, if any, will be considered for AUC determinations. The HEDFs will be established in accordance with the new knowledge on the kinetics of BPA in humans and/or animals.

As defined in the EFSA Scientific Committee draft guidance (2017a) "A WoE assessment is a process in which evidence is integrated to determine the relative support for possible answers to a question". That guidance considers the WoE assessment as comprising three basic steps: (i) assembling the evidence, (ii) weighing the evidence, and (iii) integrating the evidence, as described below.

(i) Assembling the evidence: the experts will sort the studies according to the human or animal health outcome category, *e.g.* reproductive/developmental effects.

All the endpoints tested in the studies referring to a certain health outcome category will be either classified as "apical" endpoints (*e.g.* infertility) or "intermediate" endpoints (*e.g.* anogenital distance (AGD)) by two independent reviewers specialised in the area. By apical endpoint it is intended an observable outcome in a whole organism, such as a clinical sign or pathologic state, that is indicative of a disease state that can result from exposure to a toxicant (Krewski *et al.*, 2011). Intermediate endpoints are events occurring at a step between the molecular initiating event and the apical outcome: they are toxicologically relevant to the apical outcome (a necessary element of the mode of action or a biomarker of effect (see *e.g.* OECD, 2008)) and are experimentally quantifiable.

A further sorting of the studies will also take into consideration the internal validity score of the individual studies, followed by the relevance to the question judgement. To summarise the evidence in animals, the animal studies addressing a given health outcome/endpoint will be plotted in a single graph (see Fig. 2 as an example) containing, for each study, information on the life stage of the animals at treatment onset, duration of the treatment and sampling time for measurements, the doses tested, the magnitude and statistical significance of the effects at any dose (filled *vs.* empty symbol), the study validity tier, the rating (*i.e.* medium or high) of the study relevance to the question, and the reference. As described above, all doses on the x-axis of the graph will be converted into HED. The magnitude of the effects caused by BPA at each dose will be standardised to the effect size in the control group to enable a comparison of the magnitude of the effects across different doses and studies. The direction of the effect (increase *vs.* decrease) will also be graphically represented (round *vs.* diamond shape) to compare inter-study consistency.

Similar graphs will also be created for human studies. BPA exposure will be expressed as quartiles and changes in the measured parameters will be expressed relatively to the control levels.

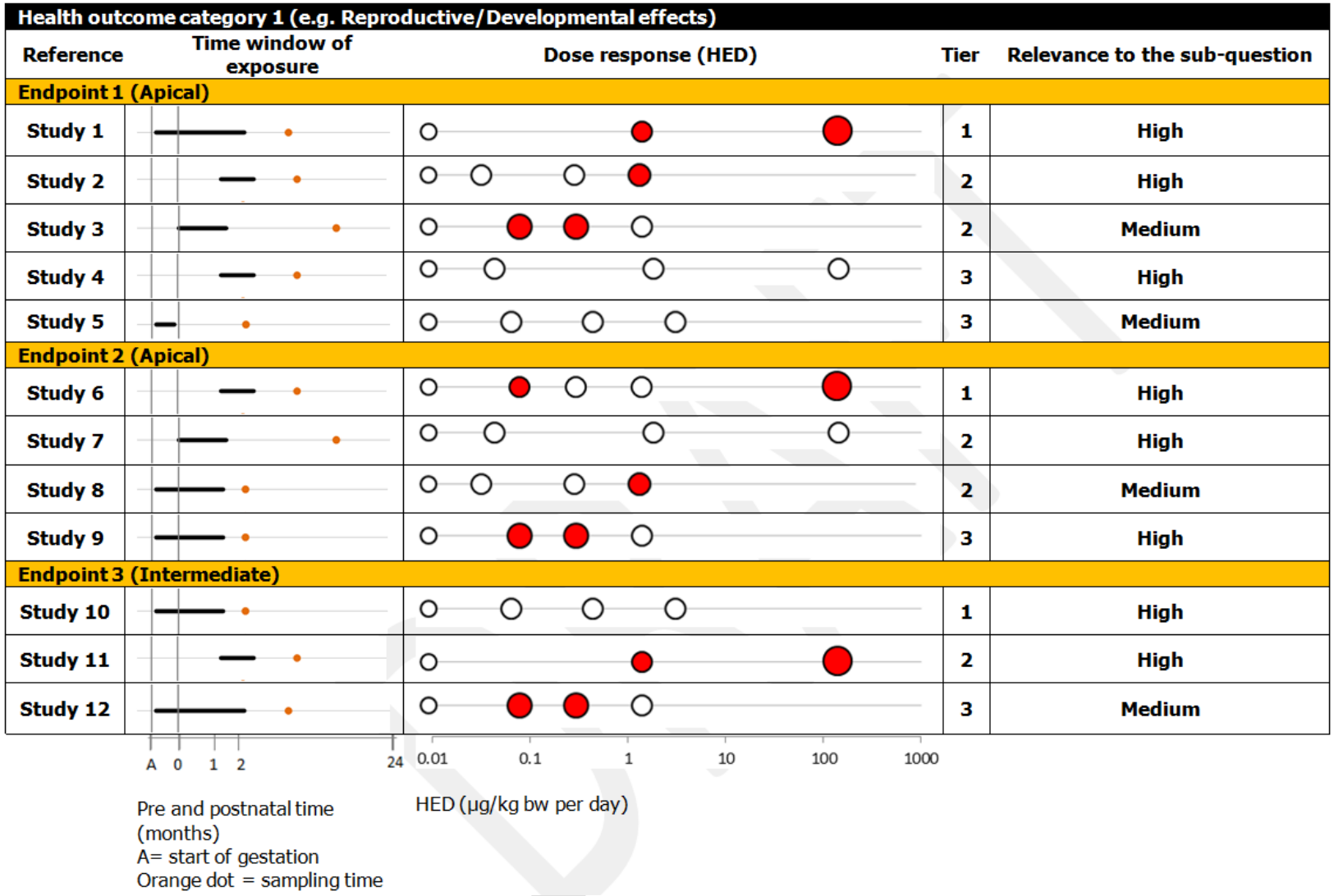


Figure 2: Graphical representation of the WoE analysis in the dose range of interest

ii) Weighing the evidence: based on expert judgement two independent reviewers will assess the confidence in the body of evidence using this graphical approach, which could allow, for each endpoint, to obtain an overall estimation of the likelihood of an effect being observed. The issues that would need to be considered, when making such estimation are:

a. the overall internal validity of the studies that show/don't show an effect

b. the consistency of the results between different studies within the same species/population or across species/populations)

c. the dose-response relationships

d. the magnitude of the effects

e. the biological plausibility of the effects on interrelated endpoints or MoA

f. the relevance of the results to the question of interest

Biological plausibility is a fundamental concept for data integration across different endpoints but within a certain health outcome category as it strengthens the causal inference for a certain effect. Indeed concordance of results between different endpoints on the same biological pathway known to lead to a certain toxicity or disease state increases the confidence in the body of evidence for a certain effect. If instead there are unexplained inconsistencies in the results concerning the same biological pathway, in principle priority should be given to the evidence arising from "apical" endpoints (*i.e.* overt effect or disease state). In all study types the apical endpoints are generally considered to be the most direct, or applicable, to the assessment of the health outcome *(e.g.* incidence of cancer of the mammary gland). Intermediate endpoints are relevant and can include key events, upstream indicators, risk factors, intermediate outcomes or measures related to the final endpoints, *e.g.* pre-neoplastic lesions. However, it is not always agreed what an intermediate and an apical endpoint are for every toxic effect. Also in some cases, intermediate endpoints may be as decisive as apical endpoints *e.g.* when they are key events in an adverse outcome pathway; in other cases, intermediate effects may be transient and therefore not as relevant as apical endpoints. Information on MoA of the target compounds and endpoints may also support this step. MoA studies in laboratory animals can establish the key events and their relationships required for the various adverse outcomes as a result of BPA exposure.

The outcome of this hazard identification step will be a conclusion on the likelihood that BPA is hazardous with respect to a certain health outcome in human and/or animal studies at any BPA exposure level. The scale of likelihood will comprise 3 categories, namely "Likely", "As likely as not" and "Unlikely" (see Table 13).

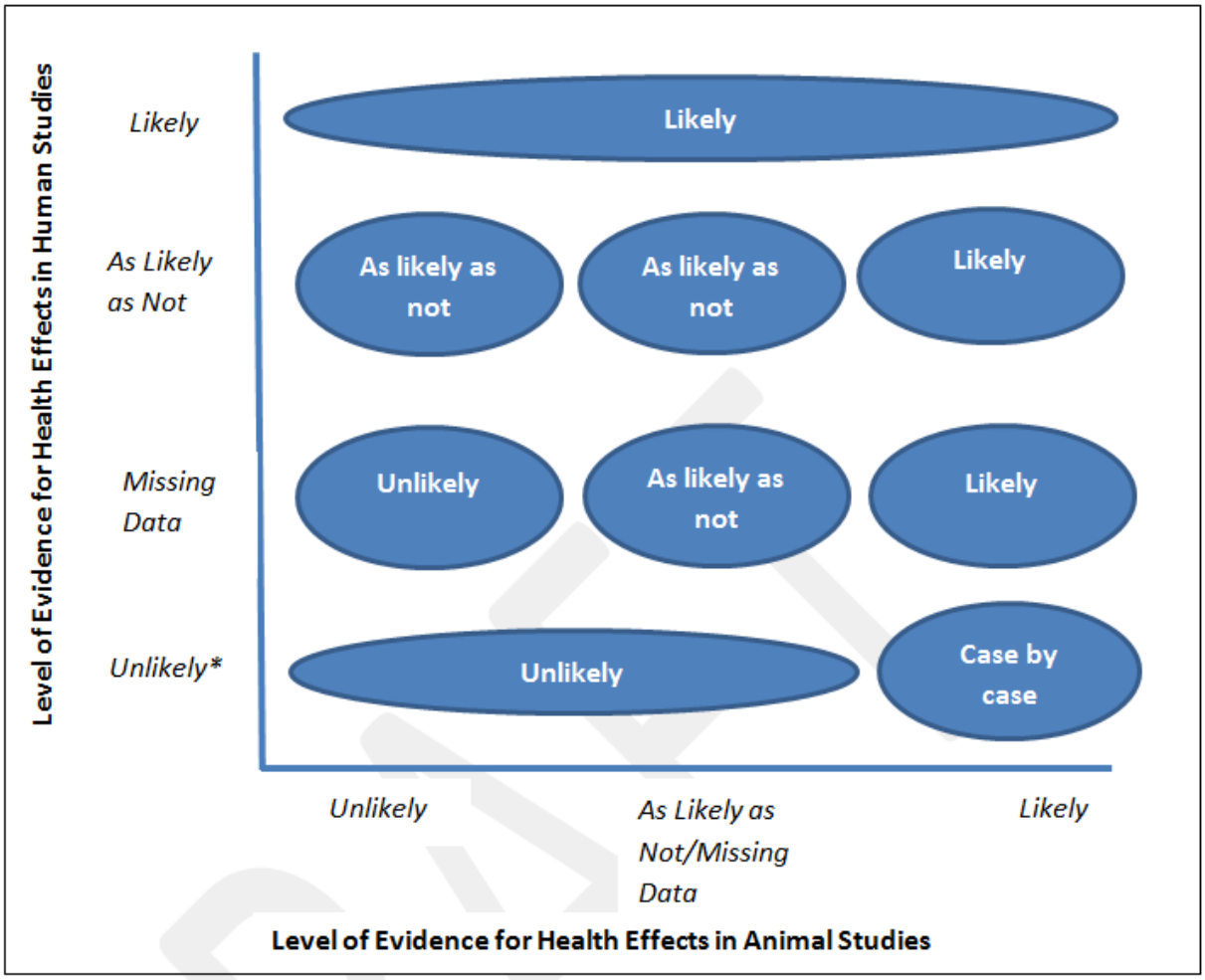
Table 13: Scale of likelihood for a given health outcome category

Category of likelihood	Percentage of likelihood	Comment
"Likely"	66-100%	
"As likely as not"	33—66%	It is about equally likely that BPA causes, or does not cause, the effect.
"Unlikely"	0-33%	

For the purpose of the current assessment, in the absence of studies addressing the apical endpoint and in the presence of studies showing "Likely" effects on intermediate endpoints, a BPA effect in that health outcome category would be concluded to be "Likely".

iii) Integrating the evidence: the overall conclusions of the WoE analysis applied to BPA hazard identification (in the dose range of interest) will be obtained by integrating the evidence from human and/or experimental animal studies according to the scheme shown in Fig. 3.

Adverse effects (apical and intermediate endpoints) for human and animal studies will be identified by the Working Group performing the risk assessment taking into account the previous scientific opinion issued by the CEF Panel (2015).



* It would be unrealistic to expect being able to derive a likelihood of "unlikely" from human studies on BPA-environmentally exposed subjects, however for consistency and to ease the understanding of the graph, this situation was included

Figure 3: Integrating the evidence from human and non-human animal studies

In the WoE approach, findings in animal studies have to be integrated with those in human studies. The integration of such findings follows the process described in Fig. 3.

The results of human studies are the most directly applicable to human health. Therefore findings in human studies with the level of evidence "Likely" will be rated as "Likely" in the final judgement irrespective of the outcome of the animal studies (Fig. 3). The same will apply for "Unlikely" effects in humans which will predominate over animal evidence except in the case of "Likely" effects in animals which will be treated on a case by case basis.

The "Likely" effects in animal studies will determine the final integrated human/animal likelihood judgement irrespective of whether human data is missing or judged "As likely as not" (Fig. 3).

If findings from animal studies indicate an "Unlikely" effect and (i) the evidence from human studies is "As likely as not" , the reviewers will conclude on an overall result of "As likely as not", while (ii) if there is missing data for humans the overall rating in the WoE will be "Unlikely".

9. Relevance and adversity of the effect for human health

An effect is considered "adverse" when leading to a change in the morphology, physiology, growth, development, reproduction or life span of an organism, system or (sub)population that results in an impairment of functional capacity to compensate for additional stress or an increase in susceptibility to other influences" (WHO, 2009).

Once the likelihood that BPA is hazardous has been assessed with respect to certain toxicological endpoints according to the overall body of evidence, the next step will be the assessment of the relevance of those effects seen in animal studies to human health and their adversity in humans, if they occur.

While relevance to humans is intrinsic in human studies, adversity of effects will be evaluated case-by-case by the experts performing the hazard assessment.

This evaluation will be performed by applying expert judgement, ensuring that a justification for the decision on the relevance and adversity is provided.

10. Method for performing hazard characterisation

By hazard characterisation it is intended the analysis of the dose-response relationship and the identification of a reference point (a benchmark dose (BMD), its lower confidence limit (BMDL) for a particular incidence/size of effect or a NOAEL) as a basis for a new TDI. Hazard characterisation will be performed for "Likely" effects (as assessed through the WoE analysis), using human or animal studies (depending on data availability and study appraisal outcome) showing adverse effects relevant to humans.

The animal studies supporting "Likely" effects that have been assigned relatively higher relevance and internal validity and include at least three test doses, will undergo such dose-response analysis. The lowest reference point will be considered for the possible derivation of a TDI.

In vivo studies considered not directly suitable for the derivation of the reference point and supporting "Likely" or "As likely as not" effects will be collectively considered in an uncertainty analysis, to possibly define the need of an extra factor to cover for uncertainties in the BPA database at low levels of exposure.

For human studies, due to methodological constraints, exposure can only be estimated by the sum of urinary conjugated and unconjugated BPA concentrations. These cannot be directly related to an internal/systemic concentration of the endocrine active fraction of BPA. The latter, which is the toxicologically relevant concentration for a reference point, highly depends on the route of exposure, *i.e.* oral *vs.* dermal. When possible, a dose – response relationship will be established and a reference point derived by appropriate statistical methods for human studies. Whereas there is no need for an inter-species assessment factor when using human data for deriving a TDI, an intra-species factor could be needed to adjust the observation for the whole population. An additional uncertainty factor might also be necessary to cover for uncertainty in the database.

For the human hazard characterisation, data on the toxicokinetics (ADME and PBPK modelling) will support the extrapolation of results from experimental animal studies to humans. This information is also important to determine which uncertainty factors have to be applied when establishing the health-based guidance value. It should be noted that the default factor of 4 for interspecies kinetic differences is already taken into consideration by the chemical-specific approach in which the ratio of AUCs in animals to the AUC in humans is used to adjust the external doses in animals to the external doses in humans. The remaining uncertainty factor should cover for inter-species difference in toxicodynamics (default factor is 2.5) and inter-individual variability in both toxicokinetics and toxicodynamics (the default factor being 10).

11. Methods for addressing the uncertainties

The evaluation of the inherent uncertainties will be performed in accordance with the upcoming EFSA Guidance on Uncertainty in EFSA Scientific Assessment of the EFSA Scientific Committee, which is currently published in its draft form (EFSA Scientific Committee, 2017b). According to the draft Guidance, uncertainty is used as a general term referring to all types of limitations in the knowledge available at the time an assessment is conducted and within the time and resources agreed for the assessment. Furthermore the draft Guidance recommends using quantitative expressions of uncertainty through verbal terms with quantitative definitions.

Several proposals are given in the draft Guidance on how the assessment of uncertainty could be performed. For the hazard characterisation step of BPA assessment, the informal expert knowledge elicitation could be the most appropriate for consistency and integration of the re-evaluation of BPA with the former BPA assessment (CEF Panel, 2015) where a quantitative uncertainty analysis was performed.

The procedure should follow the minimal requirements with pre-defined questions and a pre-defined expert board; the process should be fully documented. The result of the uncertainty analysis will be a description of additional uncertainties not already covered in the form of subjective probabilities.

References

Beronius A, Molander L, Rudén C and Hanberg A, 2014. Facilitating the use of non-standard *in vivo* studies in health risk assessment of chemicals: a proposal to improve evaluation criteria and reporting. Journal of Applied Toxicology, 34(6), 607-617.

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

EFSA (European Food Safety Authority), 2006. Opinion of the Scientific Panel on Food Additives, Flavourings, Processing Aids and Materials in Contact with Food on a request from the Commission related to 2,2-bis(4-hydroxyphenyl)propane (Bisphenol A). The EFSA Journal 2006, 428, 1-75.

EFSA (European Food Safety Authority), 2008. Scientific Opinion of the Panel on Food Additives, Flavourings, Processing aids and Materials in Contact with Food (AFC) on a request from the Commission on the toxicokinetics of Bisphenol A. The EFSA Journal, 2008, 759, 1-10 pp.

EFSA CEF Panel (EFSA Panel on Food Contact Materials, Enzymes, Flavourings and Processing Aids), 2015. Scientific Opinion on the risks to public health related to the presence of bisphenol A (BPA) in foodstuffs: Executive summary. EFSA Journal 2015;13(1):3978, 23 pp. doi:10.2903/j.efsa.2015.3978.

EFSA CEF Panel (EFSA Panel on Food Contact Materials, Enzymes, Flavourings and Processing Aids) A statement on the developmental immunotoxicity of bisphenol A (BPA): answer to the question from the Dutch Ministry of Health, Welfare and Sport. EFSA Journal 2016;14(10):4580, 22 pp. doi:10.2903/j.efsa.2016.4580

EFSA Panel on Food Contact Materials Enzymes Flavourings and Processing Aids (CEF), 2010. Scientific Opinion on bisphenol A: evaluation of a study investigating its neurodevelopmentaltoxicity, review of recent scientific literature on its toxicity and advice on the Danish risk assessment of bisphenol A. EFSA Journal 2010;8(9):1829, 110 pp. doi:10.2903/j.efsa.2010.1829.

EFSA Panel on Food Contact Materials Enzymes Flavourings and Processing Aids (CEF), 2011. Statement on the ANSES reports on bisphenol A. EFSA Journal 2011; 9(12):2475, 10 pp. doi:10.2903/j.efsa.2011.2475.

EFSA Scientific Committee (SC), 2017a. DRAFT Guidance on the Use of the Weight of Evidence Approach in Scientific Assessments. Available online: https://www.efsa.europa.eu/sites/default/files/consultation/170306-0.pdf

EFSA Scientific Committee (SC), 2017b. DRAFT Guidance on Uncertainty in Scientific Assessments. Available online: https://www.efsa.europa.eu/sites/default/files/consultation/150618.pdf

FAO/WHO (Food and Agricultural Organisation of the United Nations and World Health Organisation), 2011. Joint FAO/WHO Expert Meeting to review toxicological and health aspects of Bisphenol A. 60 pp.

Krewski D, Westphal M, Al-Zoughool M, Croteau MC and Andersen ME, 2011. New directions in toxicity testing. Annu Rev Public Health, 32, 161–178. doi: 10.1146/annurev-publhealth-031210-101153.

OECD, 2013. Guidance document on developing and assessing adverse outcome pathwaysNo.184. Available online:http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=env/jm/mono%282013%296&doclanguage=en

Thayer KA., Doerge DR, Hunt D, Schurman SH, Twaddle NC, Churchwell MI, ... and Birnbaum LS, 2015. Pharmacokinetics of bisphenol A in humans following a single oral administration. Environment international, 83, 107-115.

Tyl RW, Myers CB and Marr MC, 2006. Draft Final Report: Two-generation reproductive toxicity evaluation of Bisphenol A (BPA; CAS No. 80-05-7) administered in the feed to CD-1® Swiss mice (modified OECD 416). RTI International Center for life Sciences and Toxicology, Research Triangle Park, NC, USA.

Tyl RW, Myers CB, Marr MC, Sloan CS, Castillo NP, Veselica MM, Seely JC, Dimond SS, Van Miller JP, Shiotsuka RN, Beyer D, Hentges SG and Waechter JM, 2008. Two-generation reproductive toxicity study of dietary bisphenol a in CD-1 (Swiss) mice. Toxicological Sciences, 104, 362-384.

Tyl RW, Myers CB, Marr MC, Thomas, BF, Keimowitz AR, Brine DR, Veselica MM, Fail PA, Chang TY, Seely JC, Joiner RL, Butala JH, Dimond SS, Cagen SZ, Shiotsuka RN, Stropp GD and Waechter JM, 2002. Three-generation reproductive toxicity study of dietary Bisphenol A in CD Sprague-Dawley rats. Toxicological Sciences, 68, 121-146.

US FDA (Food and Drug Administration), 2008. Draft assessment of Bisphenol A for use in food contact applications. DRAFT version 08/14/2008. 105 pp.

US FDA (Food and Drug Administration), 2013. Update on Bisphenol A (BPA) for use in food contact applications. Available from https://www.fda.gov/newsevents/publichealthfocus/ucm064437.htm

US National Toxicology Panel, 2015. Handbook for conducting a literature-based health assessment using OHAT approach for systematic review and evidence integration. Available online: http://ntp.niehs.nih.gov/ntp/ohat/pubs/handbookjan2015_508. pdf

WHO, 2009. IPCS Risk assessment terminology. Available from http://www.who.int/ipcs/methods/harmonisation/areas/ipcsterminologyparts1and2.pdf

Abbreviations

ADME	Absorption, Distribution, Metabolism And Excretion
AUC	Area Under The Curve
BMD	Benchmark Dose
BMDL	Benchmark Dose (Lower Confidence Limit)
BMDL10	Benchmark Dose (10% Lower Confidence Limit)
BMI	Body Mass Index
BPA	Bisphenol A
CEF Panel	Panel on Food Contact Materials, Enzymes, Flavourings And Processing Aids
CI	Confidence Interval
EC	European Commission
EFSA	European Food Safety Authority
FAO	Food And Agriculture Organisation of the United Nations
GLP	Good Laboratory Practices
HED	Human Equivalent Dose
HEDF	Human Equivalent Dose Factor
LOAEL	Lowest Observed Adverse Effect Level
LOEL	Lowest Observed Effect Level
MOA	Mode of Action
NOAEL	No Observed Adverse Effect Level
NOEL	No Observed Effect Level
NTP	National Toxicology Programme
OHAT	Office of Health Assessment and Translation
PBPK	Physiologically Based Pharmacokinetic Modelling
Prometheus	Promoting Methods for Evidence Use in Science
ROB	Risk of Bias
TDI	Tolerable Daily Intake
t-TDI	Temporary- Tolerable Daily Intake
FDA	Food and Drug Administration
WHO	World Health Organisation
WoE	Weight of Evidence

Appendix A – Search strings used for each database

Information sources

Information source	Platform	Dates
PubMed	National Library of Medicine	2013-present
Scopus	Scopus	2013-present
Web of Science Core Collection. Science Citation Expanded Index	Web of Science	2013-present
Web of Science Core Collection. Emerging Sources Citation Index (ESCI)	Web of Science	2015-present
Web of Science Core Collection. Current Chemical Reactions (CCR-EXPANDED)	Web of Science	2013-present
Web of Science Core Collection. Index Chemicus (IC)	Web of Science	2013-present
DART	Toxnet	2013-present
TOXLINE	Toxnet	2013-present

Search strategies

PubMed

Search	Query
#4	Search #1 NOT #2 Filters: Publication date from 2013/01/01
#3	Search #1 NOT #2
#2	Search "Comment" [Publication Type] OR "Editorial" [Publication Type] OR "Letter" [Publication Type]
#1	Search "bisphenol A" [Supplementary Concept] OR "bisphenol A"[tiab] OR BPA[tiab] OR "80 05 7"[tiab] OR "201 245 8"[tiab]

Scopus

Search	Query
#5	(CASREGNUMBER (80-05-7)) OR (TITLE-ABS-KEY ("bisphenol a" OR bpa OR "80 05 7" OR "201 245 8")) AND (PUBYEAR > 2012) AND (EXCLUDE (DOCTYPE , "cp") OR EXCLUDE (DOCTYPE , "ch") OR EXCLUDE (DOCTYPE , "le") OR EXCLUDE (DOCTYPE , "ed"))
#4	(CASREGNUMBER (80-05-7)) OR (TITLE-ABS-KEY ("bisphenol a" OR bpa OR "80 05 7" OR "201 245 8")) AND (PUBYEAR > 2012)
#3	(CASREGNUMBER (80-05-7)) OR (TITLE-ABS-KEY ("bisphenol a" OR bpa OR "80 05 7" OR "201 245 8"))
#2	TITLE-ABS-KEY ("bisphenol a" OR bpa OR "80 05 7" OR "201 245 8")
#1	CASREGNUMBER (80-05-7)

Web of Science Core Collection:

- Science Citation Expanded Index
- Emerging Sources Citation Index (ESCI)
- Current Chemical Reactions (CCR-EXPANDED)
- Index Chemicus (IC)

Search	Query
#2	TS=("bisphenol A" OR BPA OR "80 05 7" OR "201 245 8") Refined by: [excluding] DOCUMENT TYPES: (NEWS ITEM OR EDITORIAL MATERIAL OR MEETING ABSTRACT OR LETTER OR BOOK CHAPTER) Indexes=SCI-EXPANDED, ESCI, CCR-EXPANDED, IC Timespan=2013-20XX
#1	TS=("bisphenol A" OR BPA OR "80 05 7" OR "201 245 8") Indexes=SCI-EXPANDED, ESCI, CCR-EXPANDED, IC Timespan=2013-20XX

DART

Search	Query
# 1	(80-05-7 [rn] OR "bisphenol a" OR bpa OR "80 05 7" OR "201 245 8") AND 2013:20XX [yr]

TOXLINE

Search	Query
# 1	(80-05-7 [rn] OR "bisphenol a" OR bpa OR "80 05 7" OR "201 245 8") AND 2013:20XX [yr]

914

Appendix B – Guidelines for the assessment of quality

915 ## B.1. Human case-control studies

916

Question n°	Question	Rating	Explanation for expert judgement
1 Key question A	**Domain: Detection** Can we be confident in the exposure characterisation?	*++*	There is direct evidence that the exposure was assessed using well-established methods that directly measure exposure (*e.g.* measurement of the chemical in the environment 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 assessed using well-established methods that directly measure exposure (*e.g.* measurement of the chemical in the environment or measurement of the chemical in blood, plasma, urine, *etc.*), **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).
		-	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 method used for exposure assessment (record "NR" as basis for answer).
		--	There is direct evidence that the exposure was assessed using poorly validated methods, **OR** there is evidence of exposure misclassification (*e.g.* differential recall of self-reported exposure).
2 Key question B	**Domain: Detection** Can we be confident in the outcome Assessment?	++	There is direct evidence that the outcome was assessed in cases (*i.e.* case definition) using well-established methods (the gold standard).
		+	There is indirect evidence that the outcome was assessed in cases (*i.e.* case definition) using acceptable methods, **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.
		-	There is indirect evidence that the outcome was assessed in cases (*i.e.* case definition) using non acceptable methods, **OR** there is insufficient information provided about how cases were identified (record "NR" as basis for answer).
		--	There is direct evidence that the outcome was assessed in cases (*i.e.* case definition) using non-acceptable methods.

3 Key question C	**Domain: Exposure** Was the time-window between exposure and outcome assessment appropriate?	++	There is direct evidence that the time window was appropriate for the endpoint of interest.
		+	There is indirect evidence that the time window was appropriate for the endpoint of interest.
		-	There is indirect evidence that the time window was not appropriate for the endpoint of interest.
		--	There is direct evidence that the time window was not appropriate for the endpoint of interest.
4	**Domain: Others** Do the statistical methods seem appropriate?	++	The statistical methods have been described with enough detail and do seem appropriate, usual or familiar, (*i.e.* details on preliminary analyses to modify raw data before have been provided; variables used in the primary analyses are clearly identified and summarized with descriptive statistics; main methods for analyzing the primary objectives of the study are fully described; conformity of data to the assumptions of the test used to analyze them are verified; whether and how any allowance or adjustments were made for multiple comparisons have been indicated; if relevant, how any outlying data were treated in the analysis have been reported; whether tests were one- or two-tailed have been specified and use of one-tailed tests has been justified; alpha level (*e.g.* 0.05) that defines statistical significance has been reported; references for the statistical methods have been provided; the statistical software used has been specified).
		+	The statistical methods have not been described in detail, **AND** there is indirect evidence that statistical methods are appropriate, usual or familiar.
		-	The statistical methods have not been described in detail, **AND** there is indirect evidence that statistical methods are inappropriate, unusual or unfamiliar.
		--	The statistical methods have not been described in detail, **AND** there is direct evidence that statistical methods are inappropriate, unusual or unfamiliar.

917

918

B.2. Human cohort studies

Question n°	Question	Rating	Explanation for expert judgement
1 Key question A	**Domain: Detection** Can we be confident in the exposure characterisation?	++	There is direct evidence that the exposure was assessed using well-established methods that directly measure exposure (*e.g.* measurement of the chemical in the environment 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 assessed using well-established methods that directly measure exposure (*e.g.* measurement of the chemical in the environment or measurement of the chemical in blood, plasma, urine, *etc.*), **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).
		-	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 method used for exposure assessment (record "NR" as basis for answer).
		--	There is direct evidence that the exposure was assessed using poorly validated methods, **OR** evidence of exposure misclassification (*e.g.* differential recall of self-reported exposure).
2 Key question B	**Domain: Detection** Can we be confident in the outcome assessment?	++	There is direct evidence that the outcome was assessed using well-established methods (the gold standard).
		+	There is indirect evidence that the outcome was assessed using acceptable methods, **OR** it is deemed that the outcome assessment methods used would not appreciably bias results.
		-	There is indirect evidence that the outcome was assessed using a non-acceptable method (*e.g.* a questionnaire used to assess outcomes with no information on validation), **OR** there is insufficient information provided about the outcome assessment method (record "NR" as basis for answer).
		--	There is direct evidence that the outcome was assessed using a non-acceptable method.
3 Key question C	**Domain: Exposure** Was the time-window between exposure and	++	There is direct evidence that the time window was appropriate for the endpoint of interest.
		+	There is indirect evidence that the time window was appropriate for the endpoint of interest.
		-	There is indirect evidence that the time window was not appropriate for the endpoint of interest.

Question n°	Question	Rating	Explanation for expert judgement
	outcome assessment appropriate?	--	There is direct evidence that the time window was not appropriate for the endpoint of interest.
4	**Domain: Others** Do the statistical methods seem appropriate?	++	The statistical methods have been described with enough detail and do seem appropriate, usual or familiar (*i.e.* details on preliminary analyses to modify raw data before have been provided; variables used in the primary analyses are clearly identified and summarized with descriptive statistics; main methods for analysing the primary objectives of the study are fully described; conformity of data to the assumptions of the test used to analyse them are verified; whether and how any allowance or adjustments were made for multiple comparisons have been indicated; if relevant, how any outlying data were treated in the analysis have been reported; whether tests were one- or two-tailed have been specified and use of one-tailed tests has been justified; alpha level (*e.g.* 0.05) that defines statistical significance has been reported; references for the statistical methods have been provided; the statistical software used has been specified).
		+	The statistical methods have not been described in detail, **AND** there is indirect evidence that statistical methods are appropriate, usual or familiar.
		-	The statistical methods have not been described in detail, **AND** there is indirect evidence that statistical methods are inappropriate, unusual or unfamiliar.
		--	The statistical methods have not been described in detail, **AND** there is direct evidence that statistical methods are inappropriate, unusual or unfamiliar.

921

922

B.3. Animal experimental studies

Question	Explanation
1. The test compound was unlikely to contain any impurities that may significantly have affected its toxicity.	The purity of the test compound can potentially affect study results. Purity is also an important aspect to consider in terms of the relevance of the test compound to the compound being risk assessed. Ideally, the test chemical should be of the highest available purity. Significant impurities, or isomers of the test compound, are more likely to be present, and/or to impact toxicity for certain compounds. The measured toxicity of the test compound may then be due to the contaminant. In such cases information about the level of purity and composition is critical. How to judge this criterion: **Fulfilled** – The test compound has been clearly identified and characterized and is of sufficient purity. **Partially fulfilled** – The purity of the test compound has not been described and no information about its source is available, but it is assumed that it is unlikely that impurities are present that would significantly affect the results of the study. **Not fulfilled** – The test compound is likely to contain impurities that can affect study results
2. A concurrent negative control group was included. **Key aspect A**	A concurrent negative control group should always be included as it is critical for determining treatment-related effects. The negative control group can be either untreated or vehicle-treated. However, in studies where a vehicle is used to administer the test compound it is critical that a vehicle-treated control group is included. In certain cases, it may be useful to also include a completely untreated group for identification of any influence on results from the vehicle. Historical control data from the same laboratory using the same methods and relating to animals of the same strain, age and sex, and supplier, as those used in the study may be very useful. However, such data should not provide the only negative control data for statistical analyses as biological parameters in laboratory animals can vary significantly over time. Therefore, if a study includes only historical negative control data this criterion should be judged as "not fulfilled". How to judge this criterion: **Fulfilled** – a concurrent negative (vehicle-treated) control group was included. **Not fulfilled** – no negative control was included or only a historical negative control was referred to.
3. A reliable and sensitive animal model was used for investigating the test compound and selected endpoints. **Key aspect B**	The choice of animal model (test species, strain, sex, *etc.*) is based on a number of considerations, including knowledge regarding species differences in terms of pharmacology, repeat-dose toxicology, metabolism, toxicokinetics and route of administration. Rodents (rats or mice) are commonly recommended for *in vivo* testing in current OECD test guidelines and are well characterized in terms of the reliability and sensitivity, as well as relevance to humans of different biological parameters and endpoints. Thus, it is specifically important that the study authors have justified their choice of animal model if other species have been used. It should be noted that, for investigation of certain endpoints, other species may be more sensitive and preferable. For example, rabbits are commonly recommended for teratology studies (OECD 2008). Similarly, available information about species differences in the toxicokinetics of a compound may warrant testing in a specific species. The evaluator is referred to regulatory test guidelines *(e.g.* OECD or US EPA) for discussions of the most appropriate test species for different study types. Reliability, in this context, refers to whether the animal model has been shown to generate reproducible results for the type of endpoints investigated. The sensitivity of the animal model relates to the ability to detect changes in the endpoints investigated in the model. **Fulfilled** – The animal model used is not suspected to be insensitive or unreliable.

	Not fulfilled – there is available information that indicates that the animal model is either insensitive or clearly unreliable for studying the test compound or for investigating the endpoints considered. Or the expected outcome is lacking from concurrent positive controls, if included, indicating that the test methods or animal model is insensitive.
4. Animals were individually identified.	In order to ensure reliable administration of the test compound, allocation to treatment groups and different tests, as well as recording of observations and test results, it is important that animals are individually identified. **Fulfilled** – it is stated that animals were individually identified; the specific method for identification does not have to be described. **Partially fulfilled** – it is not clearly stated whether or not animals were individually identified, but it may be inferred from other information reported for the study design and conduct **Not fulfilled** – it is stated that animals were not individually identified, or this can be inferred from other information reported for the study design and conduct.
5. Housing conditions (temperature, relative humidity, light-dark cycle) were appropriate for the study type and animal model.	Housing conditions and handling may influence animal behaviour and physiological response to stress and, consequently, study results. Importantly, variability in housing conditions may lead to increased variability in results and decreased sensitivity of the tests conducted. Different housing conditions apply to different species and different types of studies. Descriptions of standard conditions may for example be found in OECD test guidelines relevant to different types of studies and in corresponding guidance documents (http://www.oecd.org/env/ehs/testing/oecdguidelinesforthetestingofchemicals.htm). Guidance is also provided in the US National Research Council's "Guide for the Care and Use of Laboratory Animals" (https://grants.nih.gov/grants/olaw/Guide-for-the-Care-and-use-of-laboratory-animals.pdf) Housing conditions are often incompletely reported in studies published in the peer-reviewed literature, therefore it might be useful to keep in mind that this criterion may often be judged as partially fulfilled for such studies, and the impact of lack of reporting on total study reliability should be carefully considered. **Fulfilled** – housing conditions have been fully described and were in line with standard recommendations relevant to the study type and animal model. **Partially fulfilled** – some of the housing conditions were in line with standard recommendations relevant to the study type and animal model. Others deviated from standard recommendations or were not reported. **Not fulfilled** – all housing conditions deviated from standard recommendations relevant to the study type and animal model.
6. The number of animals per sex in each cage were appropriate for the study type and animal model	The number of animals housed together may have an effect on behavior and other biological parameters. Generally, laboratory animals should be housed in pairs or groups, unless the species is naturally solitary. Crowding should also be avoided as it induces stress that affects *e.g.* hormone levels and development. Scientific and practical aspects connected to the type of study influence how animals are housed together. Recommendations and requirements for the number of animals per cage relevant for different study types can be found in OECD test guidelines and corresponding guidance documents. Single housing may be recommended in some cases, *e.g.* in acute toxicity tests and in inhalation studies using aerosol exposure. Individual housing may also be necessary *e.g.* for pregnant dams and for males after mating, as well as during certain procedures, such as the use of metabolism cages. When applied, single housing should be restricted to the shortest time possible (Morton and Hau 2011). Standardisation of litter size by culling is sometimes conducted. Descriptions and recommendations for this procedure are provided in OECD test guidelines for developmental toxicity studies. **Fulfilled** – the number of animals per sex and cage were in line with standard recommendations relevant to the study type and animal model. **Partially fulfilled** – the number of animals per cage deviated somewhat from standard

	recommendations relevant to the study type and animal model, however scientific and/or practical justifications for these deviations were provided. **Not fulfilled** – the number of animals per cage deviated significantly from standard recommendations relevant to the study type and animal model and no scientific or practical justification was provided.
7. The test system is unlikely to contain contaminants that could affect the study results, such as phytoestrogens and estrogenic contamination.	Materials used in cages, water bottles and any physical enrichment should be considered, *e.g.* in terms of releasing substances that may affect study results. It should be ensured as far as possible that feed and drinking water are free from phytoestrogens and estrogenic substances. Phytoestrogen content is specifically critical in studies where endocrine activity/disruption is being investigated. For guidance on appropriate phytoestrogen levels in feed see *e.g.* OECD TG 440 (OECD 2007b). Ideally, feed and water should be tested for the presence of contaminants and phytoestrogens. Similarly, the bedding material should be considered, especially if endocrine activity/disruption is being investigated, since it may contain naturally occurring estrogenic or antiestrogenic substances. *E.g.* corn cob appears to be antiestrogenic and affects cyclicity in rats (OECD 2007b). Specifically, phytoestrogen content should be minimized in the bedding material in these cases. A full report of possible contaminants is seldom provided in studies published in the peer-reviewed literature, therefore it might be useful to keep in mind that this criterion may often be judged as partially fulfilled for such studies, and the impact of lack of reporting on total study reliability should be carefully considered. **Fulfilled** – no contaminants that could have influenced study results are suspected and/or feed, water, bedding and other materials have been analyzed and controlled for relevant contaminants. **Partially fulfilled** – some contaminants have been controlled for or analyzed but there may potentially be other contaminants present. **Not fulfilled** – it is likely that the test system was contaminated in a way that could affect study results, *e.g.* a bedding material known to contain estrogenic or antiestrogenic substances was used in a study investigating endocrine endpoints.
8. An adequate number of dose levels was used.	**Fulfilled** – Three or more dose levels were used. **Partially fulfilled** – Two dose levels were used. **Not fulfilled** – One dose level was used.
9. The timing and duration of administration do not seem to be inappropriate for investigating the included endpoints. **Key aspect C**	OECD test guidelines and corresponding guidance provide recommendations for timing and duration of administration of the test compound for different types of studies. In general, the dosing regimen should "maximise the sensitivity of the test without significantly altering the accuracy and interpretability of the biological data obtained" (OECD 2002b). Timing and duration should be considered specifically in terms of covering sensitive periods of development *(e.g.* "period of male sexual differentiation in late gestation" (OECD 2008)). In certain cases, it is also relevant to consider timing of administration in relation to when measurements of toxicological outcomes are conducted. For example, when investigating effects on behavior the potential of the administration to produce acute effects on behavioral measures should be considered, especially where the test substance is administered directly to offspring daily (OECD 2008). **Fulfilled** – the timing and duration of administration of the test compound is in line with general recommendations for the study type, is not likely to interfere with the measurements conducted, and cover sensitive periods of development, where relevant. **Partially fulfilled** – the timing and duration of administration of the test compound deviates somewhat from standard recommendations, however a scientific or practical justification is provided and sensitive periods of development are covered. **Not fulfilled -** the timing and duration of administration of the test compound is significantly different from general recommendations for the study type without being

	justified, and/or is likely to directly interfere with toxicological outcomes/measurements, and/or do not cover sensitive periods of development, where relevant.
10. Reliable and sensitive test methods were used for investigating the selected endpoints. **Key aspect D**	The reliability of the methods refers to whether they are known to generate reproducible results for the type of endpoints investigated, *e.g.* if the methods have been validated across different laboratories. The sensitivity of the methods relates to the ability to detect changes in the endpoints investigated. Studies conducted according to standardized and validated test guidelines (such as OECD test guidelines) are often considered to be reliable and adequate for risk assessment. However, it is important to keep in mind that adherence to standardized test guidelines does not automatically ensure the sensitivity of the methods applied. Further, sensitivity of the methods may in some cases be influenced by how the protocols are utilized (OECD 2008). **Fulfilled** – there is no information that suggests that the test methods are insensitive or unreliable in this context. **Partially fulfilled** – it is suspected that one or more of the methods applied may be insensitive or unreliable. **Not fulfilled** – there is available information that indicates that one or more of the methods applied is either insensitive or clearly unreliable for studies of the test compound or for investigating the endpoints considered. Or the expected outcome is lacking from concurrent positive controls, if included, indicating that the methods or animal model is insensitive.
11. Measurements do not seem to have been collected at unsuitable time points in order to generate sensitive, valid and reliable data. **Key aspect E**	This criterion covers several aspects concerning the timing of measurements and collection of data. 1. Data should be collected at the relevant time point in relation to the time needed to detect treatment related effects. In regard to specific developmental effects, these may only become apparent at a certain age, relating *e.g.* to behavioral ontogeny or onset of puberty. In addition, the time point for measurements and data collection should be chosen to avoid influence from any acute effects of the test substance administration (OECD 2008). OECD test guidelines provide recommendations for the timing of measurements and data collection in different study types. 2. Data should be collected so that the time of day does not influence measurements. For example, responses in behavioral testing in nocturnal animals like mice and rats is likely to produce different behavior during the day than during the night. For such reasons reversed lighting conditions may be applied to test nocturnal animals during the day. **Fulfilled** – The timing of tests and measurements were appropriate to detect sensitive effects and there are no related aspects that are likely to influence the reliability of the results. **Partially fulfilled** – Some, but not all, aspects of timing were appropriate. Importantly, there are no critical issues that raise concern, **Not fulfilled** - The timing of tests and measurements were not appropriate. *E.g.* it is likely that sensitive treatment related effects have been missed, or there are other aspects that are likely to have influenced the reliability of the results,
12. The statistical methods have been clearly described and do not seem inappropriate, unusual or unfamiliar and a sufficient number	The choice of statistical analyses will depend on the type of study and the nature of the endpoints measured. OECD test guidelines and corresponding guidance documents provide some recommendations for statistical tests *(e.g.* Appendix IV of OECD's Guidance notes for analysis and evaluation of chronic toxicity and carcinogenicity studies, OECD 2002b) as well as for considerations to be made in statistical analyses of different types of tests. Evaluation of this criterion also includes considering if the correct statistical unit was used. For example, it is generally recommended that the litter (or dam) is the statistical unit in developmental toxicity studies to account for litter effects. Correlations across litter mates due to genetic and/or prenatal conditions can have considerable influence on the statistical

of animals per dose group was used **Key aspect F**	significance of results *(e.g.* Holson *et al*. 2008; Li *et al*. 2008). To control for litter effects, either only one pup per sex and litter is submitted to each test/measurement in the study, or all pups are examined and litter effects are accounted for in the statistical analyses. For certain endpoints, *e.g.* malformations, it might be warranted to examine all pups as it increases the statistical power and not all pups are identical. Similarly, examining many pups per litter greatly enhances the ability to detect low dose effects (OECD 2008). The size of litter effect varies depending on endpoint measured, dose (being larger at high dose levels), and chemical mode of action. In general, normality of the data should have been checked and the choice of parametric or non-parametric tests should have been based upon that result. Sample size should be large enough to ensure sufficient statistical power to detect any effects in the endpoints measured. This includes considerations of the background incidence and variability of the measured effects, as well as the method of analysis. Excessive losses of animals in treatment groups that could affect statistical power should be noted. OECD test guidelines provide recommendations for number of animals per treatment group for different study types and endpoint measurements. However, primary consideration should be given to justifications for sample size provided by study authors, if stated. **Fulfilled** – The statistical methods have been clearly described and do not seem inappropriate, unusual or unfamiliar. **AND** a sufficient number of animals was included in the different treatment groups and loss of animals during the study is not likely to have substantially affected statistical power. **Partially fulfilled** – Unusual or unfamiliar methods were applied in the statistical analyses but do not seem clearly inappropriate. **AND/OR** a lower than usual number of animals was used, which may have caused lower sensitivity/statistical power of the study. **Not fulfilled** – No statistical tests were used, or the tests used are clearly inappropriate for the study type and/or endpoints measured. **AND/OR** the number of animals in each treatment group was clearly insufficient or there was substantial loss of animals during the study that may have affected statistical power.

924

925

Appendix C – Guidelines for the assessment of risk of bias

926 ## C.1. Human case-control studies

927

Question n°	Question	Rating	Explanation for expert judgement
1 Key question A	**Domain: Selection** Did selection of study participants result in appropriate comparison groups?	++	There is direct evidence that cases and controls were similar *(e.g.* recruited from the same eligible population, with the same method of ascertainment using the same inclusion and exclusion criteria, and were of similar age, gender, ethnicity), recruited within the same time frame, and both cases and controls are described as having no history of the outcome. **Note**: A study will be considered at low risk of bias if baseline characteristics of the two comparison groups are not statistically different.
		+	There is indirect evidence that cases and controls were similar *(e.g.* recruited from the same eligible population, with the same method of ascertainment using the same inclusion and exclusion criteria, and were of similar age, gender and ethnicity), recruited within the same time frame, and both cases and controls are described as having no history of the outcome, **OR** differences between cases and controls are limited and would not appreciably bias the results.
		-	There is indirect evidence that controls were drawn from a very dissimilar population than cases or recruited within very different time frames, **OR** there is insufficient information provided about the appropriateness of controls (including rate of response reported for cases only).
		--	There is direct evidence that controls were drawn from a very dissimilar population than cases or recruited within very different time frames.
2 Key question B	**Domain: Confounding** Did the study design or analysis account for important confounding and modifying variables?	++	There is direct evidence that appropriate adjustments were made for primary covariates and confounders (including other exposures, if relevant) in the final analyses through the use of statistical models to reduce specific bias (including standardisation, matching of cases and controls, adjustment in multivariate model, stratification, propensity scoring, or other methods 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.
		+	There is indirect evidence that appropriate adjustments (including other exposures, if relevant) were made, **OR** it is deemed that not considering or only considering a partial list of covariates or confounders (including other exposures) 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).

		-	There is indirect evidence that the distribution of primary covariates and known confounders (including other exposures) differed between cases and controls and was not investigated further, **OR** there is insufficient information provided about the distribution of known confounders (including other exposures) in cases and controls (record "NR" as basis for answer), **OR** there is indirect evidence that primary covariates and confounders were assessed using measurements of unknown validity.
		--	There is direct evidence that the distribution of primary covariates and known confounders (including other exposures) differed between cases and controls, confounding was demonstrated, but 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**.**
3	**Domain: Attrition** Were outcome data completely reported without attrition or exclusion from analysis?	++	There is direct evidence that exclusion of subjects from analyses was adequately addressed, and reasons were documented when subjects were removed from the study or excluded from analyses.
		+	There is indirect evidence that exclusion of subjects from analyses was adequately addressed, and reasons were documented when subjects were removed from the study or excluded from analyses.
		-	There is indirect evidence that exclusion of subjects from analyses was not adequately addressed, **OR** there is insufficient information provided about why subjects were removed from the study or excluded from analyses (record "NR" as basis for answer).
		--	There is direct evidence that exclusion of subjects from analyses was not adequately addressed. Unacceptable handling of subject exclusion from analyses includes: reason for exclusion likely to be related to true outcome, with either imbalance in numbers or reasons for exclusion across study groups.
4	**Domain: Detection** Was the exposure characterised consistently across study groups?	++	There is direct evidence that exposure was consistently assessed (*i.e.* under the same method and time-frame) across groups.
		+	There is indirect evidence that exposure was consistently assessed (*i.e.* under the same method and time-frame) across groups. **OR** it is deemed that an inconsistent assessment of the exposure (*i.e.* under different methods and time-frames) across groups would not considerably bias the results.
		-	There is indirect evidence that exposure was not consistently assessed (*i.e.* under different methods and time-frames) across groups. **OR** there is insufficient information provided about the consistency of the exposure assessment (record "NR" as basis for answer).
		--	There is direct evidence that exposure was not consistently assessed (*i.e.* under different methods and time-frames) across groups.
5	**Domain: Detection** Was the blinding applied and measurement consistent across study groups?	*++*	There is direct evidence that the outcome assessors (including study subjects, if outcomes were self-reported) were adequately blinded to the exposure level.
		+	There is indirect evidence that the outcome assessors were adequately blinded to the exposure level when reporting outcomes, **OR** it is deemed that lack of adequate blinding of outcome assessors would not appreciably bias results (including that subjects self-reporting outcomes were likely not aware of reported links between the exposure and outcome or lack of blinding is unlikely to bias a particular outcome).

		-	There is indirect evidence that it was possible for outcome assessors to infer the exposure level prior to reporting outcomes (including that subjects self-reporting outcomes were likely aware of reported links between the exposure and outcome), **OR** there is insufficient information provided about blinding of outcome assessors (record "NR" as basis for answer).
		--	There is direct evidence that outcome assessors were aware of the exposure level prior to reporting outcomes (including that subjects self-reporting outcomes were aware of reported links between the exposure and outcome).
6	**Domain: Selective reporting** Were all measured outcomes reported?	++	There is direct evidence that all of the measured outcomes (apical and intermediate) 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 measured outcomes (apical and intermediate) 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.* 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 measured outcomes (apical and intermediate) outlined in the protocol, methods, abstract, and/or introduction (that are relevant for the evaluation) have not 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 measured outcomes (apical and intermediate) 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.

928

929

930 ## C.2. Human cohort studies

931

Question n°	Question	Rating	Explanation for expert judgement
1 Key question A	**Domain: Selection** Did selection of study participants result in appropriate comparison groups?	++	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. **Note**: A study will be considered at low risk of bias if baseline characteristics of exposure groups are not statistically different.
		+	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 exposure 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 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 very different participation/response rates.
2 Key question B	**Domain: Confounding** Did the study design or analysis account for important confounding and modifying variables?	++	There is direct evidence that appropriate adjustments or explicit considerations were made for primary covariates and confounders (including other exposures, if relevant) in the final analyses through the use of statistical models to reduce specific bias (including standardisation, matching, adjustment in multivariate model, stratification, propensity scoring, or other methods that were appropriately justified). Acceptable consideration of appropriate adjustment factors includes cases when the factor is not included in the final adjustment model because the author conducted analyses that indicated it did not need to be included, **AND** there is direct evidence that primary covariates and confounders were assessed using valid and reliable measurements.
		+	There is indirect evidence that appropriate adjustments (including other exposures, if relevant) were made, **OR** it is deemed that not considering or only considering a partial list of covariates or confounders (including other exposures) 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).

		-	There is indirect evidence that the distribution of primary covariates and known confounders (including other exposures) differed between the exposures groups and was not appropriately adjusted for in the final analyses, **OR** there is insufficient information provided about the distribution of known confounders (including other exposures) (record "NR" as basis for answer), **OR** there is indirect evidence that primary covariates and confounders were assessed using measurements of unknown validity.
		--	There is direct evidence that the distribution of primary covariates and known confounders (including other exposures) 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.
3	**Domain: Attrition** Were outcome data completely reported without attrition or exclusion from analysis?	++	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, **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.
		+	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. This would include reports of no statistical differences in characteristics of subjects lost to follow up or with unavailable records from those of the study participants. Generally, the higher the ratio of participants with missing data among participants with events, the greater potential there is for bias. For studies with a long duration of follow-up, some withdrawals for such reasons are inevitable.
		-	There is indirect evidence that loss of subjects (*i.e.* incomplete outcome data) was unacceptably large and not adequately addressed, **OR** there is insufficient information provided about numbers of subjects lost to follow-up.
		--	There is direct evidence that loss of subjects (*i.e.* incomplete outcome data) was unacceptably large and not adequately addressed. Unacceptable handling of subject attrition includes: reason for missing outcome data likely to be related to the outcome, with either imbalance in numbers or reasons for missing data across study groups; or potentially inappropriate application of imputation.
4 Key question C	**Domain: Detection** Was the outcome characterised consistently across exposure groups	++	There is direct evidence that outcome was consistently assessed (*i.e.* under the same method and time-frame) across exposure groups.
		+	There is indirect evidence that outcome was consistently assessed (*i.e.* under the same method and time-frame) across exposure groups, **OR** it is deemed that an inconsistent assessment of the outcome (*i.e.* under different methods and time-frames) across exposure groups would not considerably bias the results.

		-	There is indirect evidence that outcome was not consistently assessed (*i.e.* under different methods and time-frames) across exposure groups, **OR** there is insufficient information provided about the consistency of the outcome assessment (record "NR" as basis for answer).
		--	There is direct evidence that outcome was not consistently assessed (*i.e.* under different methods and time-frames) exposure across groups.
5 **Key question D**	**Domain: Detection** Was the blinding applied and measurement consistent across exposure groups?	++	There is direct evidence that the outcome assessors (including study subjects, if outcomes were self-reported) were adequately blinded to the exposure group, and it is unlikely that they could have broken the blinding prior to reporting outcomes, **AND** subjects had been followed for the same length of time in all exposure groups.
		+	There is indirect evidence that the outcome assessors (including study subjects, if outcomes were self-reported) were adequately blinded to the exposure group, and it is unlikely that they could have broken the blinding prior to reporting outcomes, **AND** subjects had been followed for the same length of time in all exposure groups. **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 it was possible for outcome assessors (including study subjects if outcomes were self-reported) to infer the exposure group prior to reporting outcomes, **OR** the length of follow-up differed by exposure group, **OR** there is insufficient information provided about blinding of outcome assessors (record "NR" as basis for answer).
		--	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, **OR** the length of follow-up differed by exposure group.
6	**Domain: Selective reporting** Were all measured outcomes reported?	++	There is direct evidence that all of the measured outcomes (apical and intermediate) 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 (apical and intermediate) 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.* 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 measured outcomes (apical and intermediate) outlined in the protocol, methods, abstract, and/or introduction (that are relevant for the evaluation) have *not* been reported, **OR** 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 measured outcomes (apical and intermediate) 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.

932

933

C.3. Animal experimental studies

Question n°	Question	Rating	Explanation for expert judgement
1 **Key question A**	**Domain: Selection** Was administered dose or exposure level adequately randomised?	++	There is direct evidence that animals were allocated to any study group including controls using a method with a random component *(if the study states that it was performed according to OECD test guidelines, randomisation is considered as done; **please note** in case the study reports that it was performed according to GLP, randomisation cannot be considered as done unless specified in other parts of the manuscript)*. **Note**: Acceptable methods of randomisation include: referring to a random number table, using a computer random number generator, coin tossing, shuffling cards or envelopes, throwing dice, or drawing of lots (Higgins and Green 2011). Restricted randomisation *(e.g.* blocked randomisation) to ensure particular allocation ratios will be considered low risk of bias. Similarly, stratified randomisation and minimisation approaches that attempt to minimize imbalance between groups on important prognostic factors *(e.g.* body weight) will be considered acceptable. This type of approach is used by NTP, *i.e.* random number generator with body weight as a covariate. **Note**: Investigator-selection of animals from a cage is not considered random allocation because animals may not have an equal chance of being selected, *e.g.* investigator selecting animals with this method may inadvertently choose healthier, easier to catch, or less aggressive animals.
		+	There is indirect evidence that animals were allocated to any study group including controls using a method with a random component *(e.g.* authors state that allocation was random, without description of the method used and/or a check for baseline characteristics support this assumption), **OR** it is deemed that allocation without a clearly random component during the study would not appreciably bias results. For example, approaches such as biased coin or urn randomisation, replacement randomisation, mixed randomisation, and maximal randomisation may require consultation with a statistician to determine risk-of-bias rating (Higgins and Green 2011).
		-	There is indirect evidence that animals were allocated to study groups using a method with a non-random component (*e.g.* a check for baseline characteristics support this assumption), **Note**: Non-random allocation methods may be systematic, but have the potential to allow researchers to anticipate the allocation of animals to study groups (Higgins and Green 2011). Such "quasi-random" methods include investigator-selection of animals from a cage, alternation, assignment based on shipment receipt date, date of birth, or animal number.
		--	There is direct evidence that animals were allocated to study groups using a non-random method including judgment of the investigator, the results of a laboratory test or a series of tests (Higgins and Green 2011), **OR** there is direct evidence that baseline characteristics differ significantly between groups.
2	**Domain: Selection** Was allocation to study group adequately	++	There is direct evidence that at the time of assigning study groups the research personnel did not know what group animals were allocated to, and it is unlikely that they could have broken the blinding of allocation until after assignment was complete and irrevocable. Acceptable methods used to ensure allocation concealment include sequentially numbered treatment containers of identical appearance or equivalent methods.

	concealed	+	There is indirect evidence that at the time of assigning study groups the research personnel did not know what group animals were allocated to and it is unlikely that they could have broken the blinding of allocation until after assignment was complete and irrevocable, **OR** it is deemed that lack of adequate allocation concealment would not appreciably affect the allocation of animals to different study groups.
		-	There is indirect evidence that at the time of assigning study groups it was possible for the research personnel to know what group animals were allocated to, or it is likely that they could have broken the blinding of allocation before assignment was complete and irrevocable, **OR** there is insufficient information provided about allocation to study groups (record "NR" as basis for answer).
		--	There is direct evidence that at the time of assigning study groups it was possible for the research personnel to know what group animals were allocated to, or it is likely that they could have broken the blinding of allocation before assignment was complete and irrevocable.
3	**Domain: Performance** Were experimental conditions identical across study groups?	++	There is direct evidence that same vehicle was used in control and experimental animals, **AND** there is direct evidence that caregivers and/or investigators were blinded to knowledge which intervention each animal received during the experiment, **AND** there is direct evidence that non-treatment-related experimental conditions were identical across study groups (*i.e.* the study report explicitly provides this level of detail).
		+	There is indirect evidence that the same vehicle was used in control and experimental animals, **OR** it is deemed that the vehicle used would not appreciably bias results. **AND** as described above caregivers and/or investigators are assumed to be blinded to knowledge which intervention each animal received during the experiment if authors did not report this information**,** **AND** as described above, identical non-treatment-related experimental conditions are assumed if authors did not report differences in housing or husbandry.
		-	There is indirect evidence that the vehicle differed between control and experimental animals, **OR** authors did not report the vehicle used (record "NR" as basis for answer), **OR** there is indirect evidence that that caregivers and/or investigators were not blinded to knowledge which intervention each animal received during the experiment, **OR** there is indirect evidence that non-treatment-related experimental conditions were not comparable between study groups.
		--	There is direct evidence from the study report that control animals were untreated, or treated with a different vehicle than experimental animals, **OR** there is direct evidence that caregivers and/or investigators were not blinded to knowledge which intervention each animal received during the experiment, **OR** there is direct evidence that non-treatment-related experimental conditions were not comparable between study groups.

4 Key question B	**Domain: Attrition** Were outcome data completely reported without attrition or exclusion from analysis?	++	There is direct evidence that loss of animals was adequately addressed and reasons were documented when animals were removed from a study. Acceptable handling of attrition includes: very little missing outcome data; reasons for missing animals unlikely to be related to outcome (or 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; missing outcomes is not enough to impact the effect estimate, **OR** missing data have been imputed using appropriate methods (ensuring that characteristics of animals are not significantly different from animals retained in the analysis).
		+	There is indirect evidence that loss of animals was adequately addressed and reasons were documented when animals were removed from a study, **OR** it is deemed that the proportion lost would not appreciably bias results. This would include reports of no statistical differences in characteristics of animals removed from the study from those remaining in the study.
		-	There is indirect evidence that loss of animals was unacceptably large and not adequately addressed, **OR** there is insufficient information provided about loss of animals (record "NR" as basis for answer). **Note:** Unexplained inconsistencies between materials and methods and results sections *(e.g.* inconsistencies in the numbers of animals in different groups) could be an example of indirect evidence;
		--	There is direct evidence that loss of animals was unacceptably and not adequately addressed. Unacceptable handling of attrition or exclusion includes: reason for loss is likely to be related to true outcome, with either imbalance in numbers or reasons for loss across study groups
5	**Domain: Detection** Can we be confident in the exposure characterisation?	++	There is direct evidence that exposure was consistently administered (*i.e.* with the same method and time-frame) across treatment groups. Consistent feed consumption would be an additional element to consider when rating this question.
		+	There is indirect evidence that exposure was consistently administered (*i.e.* with the same method and time-frame) across treatment groups. For dietary exposure studies, between-group differences in feed consumption would be an additional element to consider when rating this question.
		-	There is indirect evidence that exposure was not consistently administered (*i.e.* with the same method and time-frame) across treatment groups. For dietary exposure studies, between-group differences in feed consumption would be an additional element to consider when rating this question.
		--	There is direct evidence that exposure was not consistently administered (*i.e.* with the same method and time-frame) across treatment groups. For dietary exposure studies, between-group differences in feed consumption would be an additional element to consider when rating this question.
6 Key question C	**Domain: Detection** Can we be confident in the outcome	++	There is direct evidence that the outcome assessors were adequately blinded to the study group, and it is unlikely that they could have broken the blinding prior to reporting outcomes, **AND** that the outcome was assessed at the same length of time after initial exposure in all study groups.

	Assessment?	+	There is indirect evidence that the outcome assessors 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. For some outcomes, particularly histopathology assessment, outcome assessors are not blind to study group as they require comparison to the control to appropriately judge the outcome, but additional measures such as independent review by trained pathologists can minimize this potential bias, **AND** there is indirect evidence that the outcome was assessed at the same length of time after initial exposure in all study groups, **OR** it is deemed that assessment of the outcome at a different length of time after initial exposure among study groups would not appreciably bias results.
		-	There is indirect evidence that it was possible for outcome assessors to infer the study group prior to reporting outcomes without sufficient quality control measures **OR** that the outcome was assessed at a different length of time after initial exposure among study groups **OR** there is insufficient information provided about blinding of outcome assessors or length of time after exposure for outcome assessment (record "NR" as basis for answer).
		--	There is direct evidence for lack of adequate blinding of outcome assessors, including no blinding or incomplete blinding without quality control measures **OR** that the outcome was assessed at a different length of time after initial exposure among study groups.
7	**Domain: Selective reporting** Were all measured outcomes reported?	++	There is direct evidence that all of the study's measured outcomes (apical and intermediate) outlined in the protocol (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 (apical and intermediate) outlined in the protocol, **OR** 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 (apical and intermediate) outlined in the protocol, methods, abstract, and/or introduction (that are relevant for the evaluation) have not 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 (apical and intermediate) 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.

935

936

Appendix D – Methods for reporting the data from the included studies

937 It is anticipated that data will be synthesised for each relevant outcome, building on the measured results for each endpoint considered.

938

939 **Table D.1** Study summary table – human studies

Ref. ID (Author, year)	Health Outcome/ Endpoint	Study design	Subjects	Exposure	Results	Relevance to the sub-question	Internal validity

940

941 **Table D.2** Study summary table – animal studies

Ref. ID (Author, year)	Health Outcome/Endpoint	Animal species/Strain	No of animals/ group	Exposure (Route, period, duration of administration)	Treatment groups and BPA dose(s) (mg/kg bw per day)	HED	Results	Relevance to the sub-question	Internal validity

942

943