Annex to: Scientific Opinion on the Re-evaluation of the risks to public health related to the presence of bisphenol A (BPA) in foodstuffs. https://doi:10.2903/j.efsa.2023.6857

Annex A. Revised Bisphenol A (BPA) hazard assessment protocol

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) was 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 became available (initially foreseen in 2017).

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 required setting up a BPA hazard assessment protocol as a first step.

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

1.1. Background and Terms of Reference as provided by EFSA

This work aimed to ensure that the CEF Panel would be fully prepared to engage in a re-evaluation of the safety for consumers of BPA (to set a full tolerable daily intake (TDI)) when the two-year ongoing NTP CLARITY study report became available. 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 CLARITY study design covered 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 Scientific Assessment (PROMETHEUS) project (EFSA, 2015).

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

The Food Ingredients & Packaging (FIP) Unit ensured that the CEF Panel was 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 CLARITY study report became available in 2017.

Terms of reference

To ensure preparedness due to an upcoming evaluation in 2017, the FIP Unit was 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 also defined a priori how the new evidence would be appraised for relevance and reliability.

1.2. Background and Terms of Reference as provided by the EC

EFSA accepted a mandate upon request from the EC 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 was 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 defined *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 evaluating the confidence in the body evidence (Sections 6–8 and Appendix A.2)
- hazard characterisation (Section 9)
- uncertainty analysis (Section 10).

Protocol development was part of the 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: (i) planning (protocol development) and (ii) implementation of the protocol.

1.4. Development of the protocol

Following the receipt of the above-described mandates, the protocol was drafted and approved by the FIP Working Group on BPA assessment protocol[1] on 30 November 2017, following a public consultation (EFSA, 2017a). It was published as an EFSA supporting publication on 21 December 2017 (EFSA, 2017b).

During the testing and implementation of the protocol, amendments were applied that are reported in Section 11. The protocol presented in this annex is the final version as implemented in the re-evaluation of the risks to public health related to the presence of bisphenol A (BPA) in foodstuffs.

2. Problem formulation

2.1. Background information

The toxicity of BPA has been extensively characterised in previous risk assessments by EFSA (EFSA, 2007, 2008; EFSA CEF Panel, 2010, 2011, 2015, 2016) and international bodies such as FAO/WHO (2011) and US FDA (2013).

The 2015 EFSA BPA risk assessment (EFSA CEF Panel, 2015) estimated the degree of likelihood for the effects under consideration on the basis of the then available human and animal evidence. General toxicity and mammary gland proliferative changes were classified as 'Likely'. Reproductive/developmental, neurological/neurobehavioural/neuroendocrine, immune, cardiovascular, carcinogenic and metabolic effects were classified as 'As Likely As Not' (ALAN). Genotoxic effects were considered as being 'Unlikely'.

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 increased in the two-generation study in mice by Tyl et al. (2006, 2008), for which a $BMDL_{10}$ (Benchmark dose 10% lower confidence limit) of 8.96 mg/kg bw (body weight) per day was calculated, and 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

[1] https://ess.efsa.europa.eu/doi/doiweb/wg/682790

availability of this chemical for: (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; and (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 $BMDL_{10}$ 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 intraspecies differences (factor of 10), interspecies toxicodynamic differences (factor of 2.5) and uncertainties in the database on 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.

2.2. Objectives of the hazard assessment

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

More specifically, the evaluation covered:

-the adverse effects in humans associated with the exposure to BPA via any route;

-the adverse effects in animals after:

a) oral exposure[2] to BPA at doses equal or below the cut-off of 10 mg/kg bw per day (based on the benchmark dose lower confidence interval ($BMDL_{10}$) used by the EFSA CEF Panel to set the t-TDI in 2015); or

b) other exposure routes [subcutaneous (s.c.), intraperitoneal (i.p.), intravenous (i.v.), inhalation and intratesticular] at doses equal or below the cut-off of 10 mg/kg bw per day, when converted to oral dose, taking into account the interspecies kinetics differences (see Section 3.1.1.5 of the Scientific opinion). No cut-off was applied for dermal[3] studies.

With regard to point (b), if all the doses in one study converted from other routes to oral gave results above the oral cut-off of 10 mg/kg bw per day, the study was excluded from every step of the assessment.

-the human and animal toxicokinetics of BPA.

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

2.3. Target population

The target population of the hazard assessment was the EU general population including specific vulnerable groups (embryos, fetuses and infants).

2.4. Chemical of concern

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

2.5. Endpoints relevant to the hazard assessment

[2] Concentrations in feed and drinking water were converted in oral doses using default values (EFSA Scientific Committee, 2012). For studies with exposure during gestation/lactation, the WG decided to use the conversion factor for chronic studies because this reflected adult exposure.

[3] As there was very little toxicokinetic evidence available on dermal exposure, a conservative approach was used and no cut-off was established.

A categorisation system of health outcome categories (HOCs) similar to the one used in the EFSA opinion of 2015 was implemented as follows: General toxicity (e.g. liver and kidney), Immunotoxicity, Metabolic effects, Neurotoxicity and developmental neurotoxicity, Reproductive and developmental toxicity, Cardiotoxicity, Carcinogenicity and mammary gland proliferative effects and Genotoxicity. In addition, toxicokinetic aspects of BPA was examined.

If newly identified endpoints did not belong to any of the above, a new appropriate category was added.

Functionally interrelated endpoints from human and animal studies were grouped in clusters for assessing the weight of evidence (WoE).

The identification of the relevant endpoints for the weight of the evidence was identified based on the following criteria:

1) Endpoints identified as key in the 2015 EFSA Scientific opinion:

Human studies: endpoints assessed at least as ALAN in the WoE (Sections 3.3.4, 3.4.3, 3.5.4, 3.6.4 and 3.7.4 of the 2015 EFSA Scientific opinion), belonging to the HOCs developmental and reproductive effects, neurological, neurodevelopmental and neuroendocrine effects, immune effects, cardiovascular effects and metabolic effects.

Animal studies: endpoints from Section 3.2.5 of the 2015 opinion (not included in the uncertainty analysis tables) for the HOC general toxicity, and from Section 4.3.2 of the 2015 EFSA Scientific opinion (included in the uncertainty analysis tables) for the HOCs mammary gland proliferation, carcinogenicity, reproductive toxicity, neurotoxicity, immunotoxicity and metabolic effects.

2) Endpoints identified in the current assessment:

Human studies: endpoints belonging to relevant clusters, i.e. clusters composed of at least two studies, one of them showing a statistically significant effect for one of the endpoints measured;

Animal studies: endpoints identified as statistically significant in at least one Tier 1 or Tier 2 study.

In order to be considered and assessed in the WoE (see Section 8), the relevant endpoints identified in the animal studies needed also to be expressed quantitatively and studied in a relevant animal model. Moreover, Tier 3 studies containing relevant endpoints but with less than three doses (ctrl + 3 BPA doses) were excluded from the WoE.

All the endpoints assessed in the WoE were considered adverse[4] unless otherwise stated.

As regards the measurement of organ weights, these were reported as absolute and as relative weights. The following approach was applied in the consideration of absolute and relative organ weights in the risk assessment:

- Organs for which the weight was related to the body weight: heart, liver, kidney, spleen, lung, pancreas:

If both the absolute and relative organ weights were reported for the same study, both weights were considered as relevant endpoints. However, considering that a possible effect in absolute weight may be due to a change in body weight, only the relative organ weight was considered in the WoE.

If only the absolute organ weight was reported, this was considered a relevant endpoint and included in the WoE, taking into account that the change in absolute organ weight may be due to a change in body weight.

[4] Definition of 'adverse effects':
"Changes in the morphology, physiology, growth, development, reproduction or lifespan of an organism, system or (sub)population that results in an impairment of functional capacity, an impairment of the capacity to compensate for additional stress or an increase in susceptibility to other influences" (WHO/IPCS, 2009).
"Change in the morphology, physiology, growth, reproduction, development or lifespan of an organism that results in impairment of functional capacity to compensate for additional stress or increased susceptibility to the harmful effects of other environmental influences" (EFSA SC, 2019).

- Organs for which the weight was not related to the body weight: testis, uterus, ovaries, epididymis, bulbourethral glands, prostate, levator ani/bulbocavernosus muscle, brain, cerebrum and mammary gland:

If both the absolute and relative organ weights were reported for the same study, both weights were considered as relevant endpoints. However, considering that a possible effect in relative weight may be due to a change in body weight, only the absolute organ weight was considered in the WoE.

If only the relative weight was reported, the relative weight was considered a relevant endpoint. However, in the absence of the absolute organ weight, the study was considered not to bring additional information to the overall body of evidence related to the endpoint and the relative organ weight would not be taken into account in the WoE. The same approach was applied when it was unclear whether the organ weight was reported as absolute or relative.

2.6. Identification of the hazard assessment subquestions

This section illustrates the hazard identification and characterisation subquestions to be answered and the review approach, i.e. narrative vs. systematic[5], to follow for the new BPA re-evaluation (Table 1).

A full systematic process was applied to human and animal evidence of outcomes related to the exposure of BPA which could potentially provide a reference dose for setting a health-based guidance value (Table 1). Other types of evidence, such as cross-sectional studies, toxicokinetics and mode of action (MoA) studies were dealt with narratively.

As 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 was 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 were determined before combining the whole body of experimental evidence (see section on evaluation of the confidence in the body of evidence) to ensure comparability of the effects across different studies and species.

Additional subquestions referred 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 interspecies variability, if animal data are being used for deriving a health-based guidance value.

Table 1. Hazard assessment subquestions

Q#	Hazard assessment step	Hazard assessment subquestions	Approach
1	Hazard identification	Does exposure to BPA at any pre- and/or post-natal life stage cause general toxicity (e.g. liver and kidney), or reproductive and developmental, neurological and neurodevelopmental, immune, cardiovascular, metabolic, carcinogenic or mammary gland proliferation outcomes in humans?	Systematic
2	Hazard identification	Does BPA exposure at any pre- and/or post-natal life stage via the oral, s.c., i.p., i.v., intratesticular or inhalation route (at or below the oral dose of 10 mg BPA/kg bw per day) and via dermal routes (for which there was no cut-off dose) cause general toxicity (e.g. liver and kidney) or reproductive and developmental, neurological and neurodevelopmental, immune, cardiovascular, metabolic, mammary gland proliferation or carcinogenic outcomes in mammalian animals?	Systematic

[5] 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

3	Hazard identification	Is BPA genotoxic *in vitro* or *in vivo*?	Systematic
4	Hazard identification	Does exposure to BPA at any pre- and/or post-natal life stage cause any outcome not mentioned in Q1 in humans?	Systematic
5	Hazard identification	Does exposure to BPA at any pre- and/or post-natal life stage cause any outcome not mentioned in Q2 in mammalian animals?	Systematic
6	Hazard identification	What is the evidence on the mode of action (MoA) of BPA arising from *in vitro* studies at concentrations at or below 100 nM?	Narrative
7	Hazard identification	What is the evidence on BPA MoA arising from other MoA studies (not *in vitro*)?	Narrative
8	Hazard characterisation	What is BPA's toxicokinetic profile in humans?	Narrative
9	Hazard characterisation	What is BPA's toxicokinetic profile 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 subquestions 8 and 9
11	Hazard characterisation	What is BPA dose–response relationship for relevant outcomes in humans?	Informed by subquestions 1 and 4
12	Hazard characterisation	What is BPA dose–response relationship for relevant outcomes in experimental animals?	Informed by subquestions 2, 3 and 5

3. Methods for gathering the evidence

3.1. Time span of evidence search

The evaluation dealt 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) were not re-assessed in the re-evaluation.

The proposed ending date was 30 August 2018, but the publication of the BPA NTP CLARITY study report was delayed, therefore the actual end date was moved to 15 October 2018. The Grantees' studies, included in the NTP Clarity project, published after this date were also included in the assessment. For three out of 21 studies, only raw data were published on the NTP of the United States Food and Drug Administration (US FDA) website[6] and were used for the assessment, following a statistical analysis by EFSA (see Annex F for details).

For the genotoxicity evidence, initially the narrative approach was considered since no effect was identified in the previous risk assessment (EFSA 2015). However, evidence of genotoxic effect was identified *in vitro* in the Genotoxicity MoA studies and for this reason a systematic approach was considered more appropriate, and the time span of the evidence search was extended until 21st July 2021.

3.2. Information sources

Literature searches were conducted in the following bibliographic databases (see Appendix A.1):

- PubMed
- Web of Science™ Core Collection

[6] https://manticore.niehs.nih.gov/cebssearch/program/CLARITY-BPA

- Scopus
- TOXLINE + DART (TOXNET platform)

Furthermore, EFSA launched a call for data in order to gather study reports and other information which were not available in the specified bibliographic databases or were unpublished. Such studies underwent the same screening and appraisal procedures foreseen for studies gathered through bibliographic searches.

An open search strategy was 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.1

For the additional time span considered for the HOC Genotoxicity (from October 2018 until July 2021), only the first three bibliographic databases, considered more specific, were used.

3.3. Type of evidence

Only primary research studies were considered for the assessment. Reviews (both narrative and systematic), comments, letters to the editors, book chapters, poster and/or conference abstracts and PhD theses were excluded.

3.4. Management of the information

The evidence retrieved from each bibliographic database or obtained through the call for data was imported in the bibliographic reference management software EndNote X8 (EndNote™, www.endnote.com) and combined together. A first removal of duplicates was carried out at this step using the functionality available in the EndNote X8 reference manager software.

The EndNote file obtained from the merge of the records retrieved from the different sources of information was uploaded into an online systematic review tool, DistillerSR (Evidence Partners, Ottawa, Canada), for the subsequent steps of the review.

Following uploading of the records into DistillerSR, removal of duplicates was again 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 when available, the abstracts identified in the searches described in Section 3 and Appendix A.1 were 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) MoA?'

The screening of titles and abstract was performed by two reviewers working independently.

For the genotoxicity studies, for the additional time span considered in the literature search, the screening question was: 'Is the paper reporting information about exposure to BPA and genotoxicity?' and the screening of titles and abstract was performed by one reviewer.

The possibility of an 'unclear' reply was foreseen at this stage as the unavailability of full text may have hampered the possibility to take an informed decision. In case of an 'unclear' reply the paper was considered as meeting the inclusion criteria.

A check on a random sample of studies was performed during the initial stages of the work of screening in order to ensure that there were no misinterpretations of the selection criteria. If misinterpretations were identified from this check or from direct reporting of the reviewers, then the selection criteria were better explained.

The DistillerSR tool allowed the identification of potential disagreements between the two reviewers on study eligibility.

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

The EFSA proposal of initially screening papers on the basis of title and abstract only has been criticised during the protocol's public consultation, the main reason behind the criticism being that sometimes studies examining multiple chemicals do not discuss the chemicals with 'null' results in the abstract. The Working Group (WG) assessed the approximate impact of such a decision through a pilot test (see Appendix A.5 for details). On the basis of the results, it seemed reasonable to infer that the first screening step of papers on the basis of title and abstract only would not lead to an inappropriate exclusion of relevant 'null' studies, and hence would not compromise the overall assessment.

4.2. Examining full-text reports for eligibility of studies

For records passing the first screening based on titles and abstracts, the full text underwent a second screening against the inclusion criteria by means of two reviewers working independently.

This step also served for the first categorisation of the studies into the different HOCs identified in the subquestions in Table 1. For screening the additional genotoxicity studies, the categorisation was made into different subgroups of genotoxicity endpoints (genotoxicity, epigenetics and oxidative stress).

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

In case of disagreement, the two reviewers would discuss the paper in order to reach a common decision. If the disagreement persisted, the article was brought to the attention of the whole WG on BPA assessment for discussion and agreement on a final decision.

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

4.2.1. Availability of full text and language

Availability of the full text in English was a pre-requisite for an article was included in the assessment.

The reviewers were thus asked to reply to these questions:

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

If any answer was negative, the record would be excluded from the assessment. If both answers were 'Yes', the reviewers proceeded to the next question.

The EFSA proposal (driven by the available resources) of omitting non-English publications from the review has been criticised during the protocol's public consultation. As a result, the WG undertook a pilot test to assess the approximate impact of such a decision (see Appendix A.5 for detailed results). In brief, the search strings reported in Appendix A.1 for each database were used to gather references from 1 January 2013 until 25 August 2017.

Overall, according to this pilot test, in the worst-case scenario less than 5% of the studies reaching full text screening were not published in English. This lent support to the idea that omitting non-English publications would only have a limited impact as the included English studies would account for about 95% of the overall evidence reaching full-text screening. EFSA nonetheless acknowledges that this exclusion may be a source of uncertainty and, as such, it was accounted for in the assessment's uncertainty analysis.

EFSA, in addition, offered the opportunity to authors of non-English publications to submit through an open call for data from their full-text articles, translated into English, for consideration by EFSA. Such translated studies were subject to the same inclusion criteria and appraisal process as all other literature.

4.2.2. Selection of the type of studies

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

- Is the paper a primary or a secondary study?

If the answer to the question was 'primary', the reviewers were prompted to reply to the following question.

If the answer to the question was 'secondary', the record was excluded from the assessment, but it would be used to check whether it contained additional references of primary studies that were not been captured by the literature search/call for data.

If the answer to the question was 'other', the record was excluded from the assessment.

4.2.3. Selection of the endpoints of interest

In the first instance the reviewers were 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 MoA of BPA (e.g. *in vitro*, cell cultures, specific molecular pathways). Primary studies that were not aimed at studying effects associated with exposure to BPA (e.g. human biomonitoring studies) were excluded at this step.

If the answer to the question was 'Yes', the reviewers were prompted to reply to the following question.

If the answer to the question was 'No', the record was excluded from the assessment.

The reviewers then classified 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;
- MoA (*in vitro*, non-mammalian animals, microbiota, etc.).

The effects considered relevant for the assessment were classified in the following HOCs: general toxicity (e.g. effects on liver and kidney), Immunotoxicity, Metabolic effects, Neurotoxicity and developmental neurotoxicity, Reproductive and developmental toxicity, Cardiotoxicity, Carcinogenicity and mammary gland proliferative effects, Genotoxicity or any other HOCs, with the addition of toxicokinetic studies. With regard to Genotoxicity, an additional screening of the relevance of the studies was done by experts in this field following the full-text screening.

One source could report on more than one outcome of interest and each outcome was assessed separately.

4.2.4. Selection of the exposure of interest

4.2.4.1. Human data

For human data, all types of exposure to BPA (alone or in mixtures) were considered, including the occupational exposure scenario.

4.2.4.2. Experimental animal studies

For experimental animal studies to be considered for the assessment, exposure to BPA (not given only as a part of a mixture) via any route was investigated.

For oral studies, at least one of the doses tested must be equal or below the oral cut-off value of 10 mg BPA/kg bw per day (based on the $BMDL_{10}$ 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 was on low-dose effects.

For other exposure routes (s.c., i.p., i.v., intratesticular and inhalation) at least one of the doses tested must be equal or below the cut-off of 10 mg BPA/kg bw per day, considering their doses converted into oral doses (see Tables 6 and 7 in Section 3.1.1.6 of the Scientific opinion). No cut-off was applied for dermal studies.

4.2.4.3. Mode of action (MoA) studies

Studies that investigate possible MoA of BPA must have been conducted using BPA alone, at concentrations that were considered to be in a toxicologically relevant range; hence *in vitro* studies were considered only if at least one of the concentrations tested was at or below 100 nM. In defining this cut-off concentration, we considered the concentration of unconjugated BPA in humans, as published by Thayer et al. (2015) [RefID 7183], at the exposure levels identified in the 2015 EFSA CEF Panel opinion, i.e. 1 nM, and a concentration that was subcytotoxic for many cell lines. In addition, a factor of 100 was applied to account for the amount possibly being absorbed by the experimental devices.

The studies concerning non-mammalian animal models (e.g. zebrafish) were collected but excluded from the MoA narrative assessment. For the *in vivo* studies including only mechanistic endpoints, the same cut-off doses considered for the standard endpoints were considered, when applicable.

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

Tables 2, 3 and 4 schematically list the criteria for including or excluding from the review human, animal and MoA studies, respectively.

Only studies with cohort and case–control designs were systematically appraised for humans.

Human studies with a cross-sectional design bear some limitations in relation to the scope of the BPA review that was to set up a causal dose–response relationship and, therefore, were presented in a narrative manner for informative purposes.

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

Table 2. Inclusion/exclusion criteria related to human studies

Subquestion 1: Does exposure to BPA at any pre- and/or post-natal life stage cause general toxicity (e.g. liver and kidney), or reproductive and developmental, neurological and neuro-developmental, immune, cardiovascular, metabolic, carcinogenic or mammary gland proliferation outcomes in humans? Subquestion 3: Is BPA genotoxic *in vitro* or *in vivo*? (systematic approach) Subquestion 4: Does exposure to BPA at any pre- and/or post-natal life stage cause any outcome not mentioned in Q1 in humans? Subquestion 8: What is the BPA's toxicokinetic profile in humans? (narrative approach) Subquestion 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) Cross-sectional studies (narrative approach)
	Out	Experimental animal studies *In vitro*/*in silico* studies
Population	In	All populations groups, all ages, males and females
	Out	–
Exposure/ intervention	In	All routes of exposure
	Out	Biomonitoring studies
Language	In	English
	Out	Other languages
Time	In	From 1 January 2013 (except those which were already included in the 2015 opinion) to 15 October 2018

	Out	Before 2013
Publication type	In	Primary research studies (i.e. studies generating new data)
	Out	Secondary studies[(a)] Expert opinions, editorials, and letters to the editor PhD theses Extended abstracts, conference proceedings

(a): They were used to obtain additional references of primary research studies.

Table 3. Inclusion/exclusion criteria related to experimental animal studies

Subquestion 2: Does BPA exposure at any pre- and/or post-natal life stage via the oral, s.c., i.p., i.v., inhalation or intratesticular route (at or below the oral dose of 10 mg BPA/kg bw per day) and via dermal routes (for which there was no cut-off dose) cause general toxicity (e.g. liver and kidney) or reproductive and developmental, neurological and neurodevelopmental, immune, cardiovascular, metabolic, mammary gland proliferation or carcinogenic outcomes in mammalian animals? Subquestion 3: Is BPA genotoxic *in vitro* or *in vivo*? (systematic approach) Subquestion 5: Does exposure to BPA at any pre- and/or post-natal life stage cause any outcome not mentioned in Q2 in mammalian animals? Subquestion 9: What is the BPA's toxicokinetic profile in experimental mammalian animal species/strains? (narrative approach) Subquestion 12: What is the BPA dose–response relationship for relevant outcomes in experimental animals?		
Study design	In	*In vivo* studies on animals not examining MoA Toxicokinetic studies (narrative approach)
	Out	Human data *In vitro/in silico* studies
Population	In	All mammalian animals
	Out	Non-mammalian animals
Exposure/ intervention	In	Oral, dermal, s.c., i.p., i.v., inhalation, intratesticular studies in which levels of BPA have been measured in biological samples (for toxicokinetic studies) For oral, s.c., i.p., i.v., inhalation or intratesticular studies at least one tested dose below the oral cut-off of 10 mg/kg bw per day, when converted to oral exposure; for dermal routes no cut-off dose was set. All *in vivo* genotoxicity studies with no cut-off dose
	Out	No negative control group Exposure routes other than oral, dermal, s.c., i.p., i.v., inhalation and intratesticular Mixtures with the exception that in a study arm BPA is used alone
Language	In	English
	Out	Other languages
Time	In	From 1 January 2013 (except those already included in the 2015 opinion) to 15 October 2018 For genotoxicity assessment, from 1 January 2013 (except those already included in the 2015 opinion) to 21 July 2021
	Out	Before 2013
Publication type	In	Primary research studies (i.e. studies generating new data)

	Out	Secondary studies[(a)] Expert opinions, editorials, and letters to the editor PhD theses Extended abstracts, conference proceedings

(a): They were used to obtain additional references of primary research studies.

Table 4. Inclusion/exclusion criteria related to MoA studies

Subquestion 3: Is BPA genotoxic[7] *in vitro* at any concentration? (systematic approach) Subquestion 6: What is the evidence of the MoA of BPA arising from *in vitro* studies at concentrations at or below 100 nM? (narrative approach) Subquestion 7: What is the evidence of the MoA of BPA arising from other studies (not *in vitro*)? (narrative approach)		
Study design	In	*In vitro*/*in silico* studies *In vivo* studies on MoA in humans, mammalian
	Out	Human data or *in vivo* studies not examining MoA, non-mammalian animal studies
Exposure/ intervention	In	At least one concentration at or 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[8] *In vitro* studies (except for *in vitro* genotoxicity studies) testing BPA only above 100 nM
Language	In	English
	Out	Other languages
Time	In	From 1 January 2013 (except those already included in the 2015 opinion) to 15 October 2018 For the genotoxicity assessment, from 1 January 2013 (except those already included in the 2015 opinion) to 21 July 2021
	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

5. Methods for collecting the data from the included studies

5.1. Data extraction

Pre-defined data extraction forms (see a draft example in Tables 5 and 6) were used for collecting the data from the individual studies undergoing a systematic review approach and validity appraisal.

For the human studies, the data extraction was performed using DistillerSR, and then converted to Microsoft Word. For the animal studies the data extraction was performed directly in Word.

For studies undergoing a narrative approach (i.e. cross-sectional studies and MoA *in vivo* and *in vitro* studies), the data extraction was outsourced to an external contractor (contracts number: Specific Agreements No. 2 and No. 3 implementing Framework Partnership Agreement GP/EFSA/FIP/2018/01 –

[7] Endpoints considered: gene mutation, recombination and gene conversion, sister chromatid exchanges, structural and numerical (aneuploidy) chromosome aberrations, DNA binding, DNA damage (comet assay), DNA repair (UDS) and DNA damage response (DDR).
[8] Only studies in which BPA was tested alone at least in one arm were considered for the assessment.

Lot 3; Specific Agreements No. 4, implementing Framework Partnership Agreement GP/EFSA/FIP/2018/01 – Lot 3). The outcome has been summarised in the external scientific reports 'Implementation of the evidence-based risk assessment for the re-evaluation of Bisphenol A: preparatory work on cross-sectional studies' (University of Hertfordshire, 2021a) and 'Implementation of the evidence-based risk assessment for the re-evaluation of Bisphenol A: preparatory work on Mode of Action studies in mammalian, human and/or *in vitro* models' (University of Hertfordshire, 2021b).

Table 5. Data extraction form for human cohort and case-control studies

Field	Type	Possible answers
RefID	Numeric	Automatically assigned by the systematic review software[9]
Reference	Text	Imported from bibliographic database
Study type	Categorical	Cohort Case–control
Study subjects	Text	Free text providing a description of the total number of participants, characteristics, age and country in which the study has been carried out
Sex	Categorical	Male Female Male and Female Not Reported
Exposure time	Categorical	Adulthood Childhood Pregnancy Pregnancy and Childhood Not Reported
Age (in case of exposure during childhood)	Text	Free text providing a description of the age at which the exposure in children was measured
Matrix analysed for BPA determination	Categorical	All the matrixes analysed
BPA exposure level	Text	Free text providing a description of the BPA exposure levels measured
Health outcome category	Categorical	General toxicity Immuno toxicity Metabolic and related endocrine effects Neurotoxicity Reproductive and developmental toxicity Carcinogenicity
Cluster	Categorical	Kidney toxicity Liver toxicity Asthma/allergy Infections other that respiratory tract Obesity Cardiometabolic Thyroid Type 2 diabetes mellitus (T2DM) Gestational diabetes mellitus Neurodevelopment Fetal and post-natal growth Prematurity Pre-eclampsia Male fertility Female fertility Pubertal/endocrine Prostate Lymphoid tissues

[9] The RefIDs assigned to Genotoxicity papers are identified in the opinion and annexes as RefID-G.

Endpoints	Categorical	All the endpoints measured
Direction	Categorical	Description of the direction of the measured results: Increase Decrease No change
Description of the results	Text	Free text providing a description of the observed results by endpoint

Table 6. Data extraction form for experimental animal studies for each HOC

Study identification	RefID number Author Year the study
Animal model	Species/(sub)strain/sex
Exposure	Period of exposure (pre-mating, mating, gestation, lactation, adult) Duration of exposure (e.g. GD0-GD20) Route of administration (diet, drinking water, gavage, s.c., i.p., i.v., dermal, inhalation, intratesticular) Dose regimen (dose level or concentration of BPA per group; documentation of details for dose conversion when conducted) Time of measurement
Results	Results per dose or concentration and at a specific time point (statistically significant increase, statistically significant decrease, no change)
Cluster (endpoint)	The cluster/s allocation of the relevant endpoint/s addressed in the study is/are reported

6. Internal validity of the studies

Internal validity relates to whether a study answers its research question 'correctly', that is, in a manner free from bias (Higgins and Green, 2011).[10] Risk of bias relates to the propensity of a study to be affected by systematic error. Biases can operate in either direction and can lead to underestimation or overestimation of the true intervention effect (Higgins and Green, 2011). In the current protocol risk of bias considered two aspects: (i) those that introduced a systematic difference between the control and the exposed group only (e.g. non-randomised allocation of animals to study groups); and (ii) those potentially affected to the same extent the control and exposed study groups (e.g. the reliability of the method used to test the outcome).

A structured approach was used to appraise the internal validity of human epidemiological and experimental animal studies, whereas for MoA studies a narrative approach was applied.

Internal validity of human and animal studies was evaluated by study design and by endpoint according to step 4 of the NTP Handbook for Conducting a Literature-Based Health Assessment Using OHAT Approach for Systematic Review and Evidence Integration (NTP-OHAT, 2015). The same questions considered by NTP and their related domains (e.g. selection, detection, attrition, etc.) were used to appraise the studies: in the tool some questions were identified as key based on the study design. Some of the elements reported in SciRAP (Science in Risk Assessment and Policy; Beronius et al., 2014, revised version at www.scirap.org) were inserted in the appraisal tool for experimental animal studies.

[10] This definition of internal validity partially overlaps with that of reliability as given in the 2017 EFSA Guidance on the use of the weight of evidence WoE (EFSA Scientific Committee, 2017b), with reliability being 'the extent to which the information comprising a piece or line of evidence is correct, i.e. how closely it represents the quantity, characteristic or event that it refers to. This includes both accuracy (degree of systematic error or bias) and precision (degree of random error)'. In the context of this protocol, internal validity covered the concept of accuracy but not that of precision.

For each internal validity question the response options were 'Definitely low risk of bias (RoB) (++)', 'Probably low RoB (+)', 'Probably high RoB (-)', 'Definitely High RoB (--)' (see Table 7).

The ratings of the key and non-key questions (++, +, -, --) were integrated to classify the studies in tiers from 1 to 3 corresponding to decreasing levels of internal validity.

All the studies were then considered jointly to evaluate the confidence in the overall body of evidence.

Table 7. Response options for each internal validity question (adapted from NTP-OHAT, 2015)

Rating	Response to the question	Description
++	Definitely Low risk of bias	There is direct evidence of low risk of bias practices
+	Probably Low risk of bias	There is indirect evidence of low risk of bias practices, or it is deemed that deviations from low risk of bias practices for these criteria during the study would not appreciably bias results. This includes consideration of direction and magnitude of bias
- /NR	Probably High risk of bias	There is indirect evidence of high risk of bias practices, or there is insufficient information provided about the relevant risk of bias practices (NR, not reported)
--	Definitely High risk of bias	There is direct evidence of high risk of bias practices

Each evaluation was performed by two reviewers in series (i.e. a first reviewer was initially tasked with the appraisal of a study and then the appraisal was forwarded to a second reviewer for validation). In case of disagreement, the reviewers discussed the appraisal and tried to reach consensus. If the disagreement persisted, the article was brought to the attention of the whole WG for consultation and agreement on a final decision.

The reviewer first appraised the study taking into consideration all the endpoints. If an endpoint triggered or required a different scoring compared with the other endpoints in that study, even if only in one question, the appraisal was cloned and conducted separately for that specific endpoint.

The evaluation of all studies was only based on the reported information/data. Due to limited resources, the study authors were not contacted for clarification or missing information related to the internal validity assessment.

The experts reviewing the studies were selected in compliance with an EFSA standard operating procedure (SOP 06_S, EFSA was quality certified according to the ISO9001:2015 system). This SOP set out clear rules for identifying and appointing independent experts in working groups. In the BPA case these areas of expertise were essential: general toxicology, genotoxicity, mammary gland proliferation and carcinogenicity, neurotoxicology, developmental/reproductive toxicology, immunotoxicology, endocrinology, epidemiology, toxicokinetics, pathology, biostatistics, uncertainty analysis and risk assessment.

With regard to the HOC Genotoxicity, a specific internal validity approach was applied, as described in the Section 2.3.5 of the Scientific opinion.

6.1. Internal validity appraisal for human studies

The seven questions that addressed the internal validity of human studies are presented in Table 8. Five of them are considered key. Whenever one of the elements to be appraised for internal validity was not reported, this was by default judged as 'Probably high RoB'. However, when there was 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' was given.

The instructions on how to rate each internal validity aspect can be found in Appendices B.1 and B.2.

Table 8. Internal validity appraisal tool for human data (case–control and cohort study design)

(adapted from NTP-OHAT, 2015)

#	Key Q	Question	Domain	Rating (++, +, -, --)
1	A	Did selection of study participants result in appropriate comparison groups?	Selection	
2		Were outcome data completely reported without attrition or exclusion of experimental units from analysis?	Attrition	
3	B	Can we be confident in the exposure characterisation?	Detection	
4	C	Can we be confident in the outcome assessment?	Detection	
5	D	Did the study design or analysis account for important confounding and modifying variables?	Confounding	
6		Were all measured outcomes reported?	Selective reporting	
7	E	Do the statistical methods seem appropriate?	Other sources of bias	

The ratings of the key and non-key questions (++, +, -, --) were integrated to classify the studies in tiers from 1 to 3 corresponding to decreasing levels of internal validity.

Tier 1:

- All the key questions are scored +/++

  AND

- No more than one non-key questions was scored -

  AND

- No non-key question was scored --

Tier 3:

- Any key question scored -/--

  OR

- Any non-key question was scored --

Tier 2:

- All the other combinations, not falling under Tier 1 or 3

A tiered approach was applied for the appraisal of the studies.

The appraisal started from Question 3 (Can we be confident in the exposure characterisation?), if the paper fell in Tier 3 due to a judgement of 'Probably high' or 'Definitely high' then the evaluation was stopped at this point.

6.2. Internal validity appraisal for experimental animal studies

The eight questions that addressed the internal validity of experimental animal studies are presented in Table 9. Three subquestions are considered key.

In general, whenever one of the elements to be appraised for internal validity was not reported, this was by default judged as 'Probably high RoB'. However, when there was 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' was given.

The instructions on how to rate each internal validity aspect can be found in Appendix A.2.3.

Table 9. Internal validity tool for experimental animal studies (adapted from NTP-OHAT, 2015)

#	Key Q	Question	Domain	Rating (++, +, -, --)
1		Was the administered dose or exposure level adequately randomised?	Selection	
2		Was the allocation to study group adequately concealed	Selection	
3		Were the experimental conditions identical across study groups?	Performance	
4		Were outcome data completely reported without attrition or exclusion from analysis?	Attrition	
5		Can we be confident in the exposure characterisation?	Detection	
	A	Sub-question: Did the test compound contain any impurities?		
6		Can we be confident in the outcome assessment?	Detection	
	B	Sub-question: Were the outcome assessors adequately blinded to the study group?		
7		Were all measured outcomes reported?	Selective reporting	
8		Were the statistical methods and the number of animals per dose group appropriate?	Other sources of bias	
	C	Sub-question: Was the number of animals per dose group appropriate?		

The ratings of the key and non-key questions (++, +, -, --) were integrated to classify the studies in tiers from 1 to 3 corresponding to decreasing levels of internal validity.

Tier 1:

- All the key subquestions are scored + /++

  AND

- No more than one question was scored - / - -

Tier 3:

- Any key subquestion was scored – /- -

  OR

- More than four questions are scored – / - -

Tier 2:

- All the other combinations not falling under Tier 1 or 3

A tiered approach was applied for the appraisal of the studies:

The appraisal started with Question 5 (Can we be confident in the exposure characterisation?). Three elements were considered to answer the main question including the key subquestion on the purity of the test compound. If the study was considered at high risk of bias for the key subquestion then the evaluation was stopped at this point and the study was classified as a Tier 3 otherwise the appraisal was continued.

The next question was Question 6 (Can we be confident in the outcome assessment?). Eight elements were considered to answer the main question including the key subquestion on the blinding of the outcome assessor. If the study was considered at high risk of bias for the key subquestion then the

evaluation was stopped at this point and the study was classified as a Tier 3, otherwise the appraisal was continued.

The next question was Question 8 (Were the statistical methods and the number of animals per dose group appropriate?). Two elements were considered to answer the main question including the key subquestion on the appropriateness of the number of animals per dose group. If the study was considered at high risk of bias for the key subquestion then the evaluation was stopped at this point and the study was classified as a Tier 3 otherwise the appraisal was continued with the other questions starting from Question 1.

7. External validity

In this protocol the external validity refers to the relevance for human health of measuring a given endpoint in a given animal model.

The assessors were asked to consider whether the specific endpoints measured in a specific animal model would be relevant to humans. Thus, animal models differing from humans in terms of target anatomical or pathophysiological features for the chemical under investigation would not be considered relevant.

Appendix A.3 reports the criteria, as taken from the SciRAP tool (www.scirap.org; Beronius et al., 2014) which was used to evaluate the relevance of the animal model and of the endpoint to human. Animal models can at maximum be considered as indirectly relevant for human health.

The outcome of this assessment (directly relevant, indirectly relevant or not relevant) was considered for the evaluation of the confidence in the body of evidence.

Each evaluation was performed by two reviewers in series (i.e. a first reviewer was initially be tasked with the appraisal of a study and then the appraisal was forwarded to a second reviewer for validation). In case of disagreement, the reviewers discussed the appraisal and tried to reach consensus. If the disagreement persisted, the article was brought to the attention of the whole WG for consultation and agreement on a final decision. The appraisal of external validity was performed on the same groups of endpoints considered for the appraisal of internal validity. The evaluation of the studies was only based on the reported information/data. Due to limited resources the study authors were not contacted for clarification or missing information related to the external validity assessment.

8. Weighing the body of evidence

8.1. Evaluation of the confidence in the body of evidence

Following the appraisal of the individual human and experimental animal studies for internal and external validity (the latter only for the animal studies), the experts evaluated the confidence in the overall body of evidence by clusters of endpoints for each HOC.

Within the WoE it was judged whether an endpoint could be considered as apical (e.g. breast cancer) or as intermediate (e.g. mammary gland proliferation/hyperplasia) (Guyatt et al., 2011). An apical endpoint means an observable outcome in a whole organism, such as a clinical sign or pathologic state, which was indicative of a disease state that could 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 MoA or a biomarker of effect (see e.g. OECD, 2013) and are experimentally quantifiable.

Data were extracted in summary tables in MS Word or Excel with appropriate information for all the studies containing the relevant endpoints (see 0and 0above) and grouped by HOC and by cluster. For human studies, relevant clusters were identified for the WoE (see Section 2.5). For the animal experimental studies, relevant endpoints were identified for the WoE (see Section 2.5). Data were extracted from all the studies in relevant clusters or containing a relevant endpoint, irrespective of their tier allocation and evidence of effect.

A cluster is a group of biologically functionally interrelated endpoints, i.e. addressing biological pathways known to lead to a certain toxicity or disease state, or some established clinical or other related research measures or biomarkers of the core cluster element (e.g. for the human studies, the cluster of obesity groups together the endpoints of body mass index, fat mass, waist circumference, leptin, adiponectin, etc.).

Overall, this collective assembling of the results of various experimental or epidemiological studies on a certain cluster enable easy visualisation of the consistency in qualitative terms of BPA effects across e.g. different studies and/or exposure periods and levels, and/or animal species, taking into account the study validity.

With regard to the HOC Genotoxicity, a specific WoE approach was applied, as described in the Section 2.3.5 of the Scientific opinion.

8.1.1. Collective evaluation of endpoints groups (clusters)

Confidence ratings in the overall body of evidence were reached by assessing the weight of relevant clusters (in human) or of clusters of relevant endpoints (in animals) per different exposure categories. These categories consist of: 'Exposure during pregnancy', 'Exposure during childhood' and 'Exposure during adulthood' for the epidemiological studies, and 'Developmental exposure (pre-natal and/or post-natal until weaning)', 'Developmental and adult exposure (pre-natal and post-natal in pups until adulthood)', 'Growth phase/young age', 'Adult exposure (after puberty)', 'Indirect (germline) exposure' for the animal studies. The latter took into account the ranges of ages for the different animal species reported in Table 10.

When the exposure was through the dams or sires (F0 generation), the effects of BPA were measured on the F1, F2 and/or F3 generations.

Results on these generations were separated in the WoE table as their exposure was different. In the studies F1 offspring were exposed as embryos in the uterus of the exposed F0 dams, whereas the F2 offspring in such studies were exposed only as germline cells in the F1 embryos (Skinner, 2008).

Post-natal or adult exposure of the F0 generation resulted in the F1 generation germline cells being exposed.

The F3 generation was the first generation not directly exposed from F0 dams given the chemical while pregnant and was said to be a true transgenerational effect (meaning not having any form of direct exposure).

Any effects in F3 required a germline transmission and a permanent reprogramming of the germline. Thus, the F1 generation that were exposed from F0 dams or sires were under the time period 'Developmental' (pre-natal and/or post-natal until weaning) or 'Developmental and adult' (pre-natal and/or post-natal in pups until adulthood).

The F2 generation that were exposed from F0 dams or sires were under the time period called 'Indirect' (germline) exposure.

The F3 generation was also included in the exposure period 'Indirect' (germinal) exposure, although the effect on the germline was likely to be epigenetic rather than genetic.

Transversal endpoints refer to endpoints relevant in more than one HOC and evaluated collectively.

Table 10. Lifespan, gestational period and weaning of animal species

Species	Life span	Gestation period	Weaning
Mice	18 months	19–20 days	21 days
Rat	2 years	21–23days	21 days
Dog	>2 years[(a)]	63 days[(a)]	6–8 weeks[(a)]
Monkey	<15 years – median lifespan in the wild;	5.5 months in macaques; 146 to 180 days in	4 months gradually

	>25 years – median lifespan in captivity	rhesus cynomolgus: 153 to 179 days)	
Sheep	>2 years[(a)]	137–152 days[(a)]	4–12 weeks[(a)]
Rabbit	>2 years[(a)]	29–31 days	28 days

(a): Retrieved via web search.

Biological plausibility was a fundamental concept for this appraisal; indeed, concordance of results would increase the confidence in the body of evidence for a certain effect. In case of non-concordance in the results concerning the same biological pathway, in principle priority was given to the evidence arising from 'apical' endpoints (i.e. overt effect or disease state) (Guyatt et al., 2011). This was because, in all study types, the apical endpoints were 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). In some cases, intermediate endpoints may be as decisive as apical endpoints.

In vitro and mechanistic data may be useful to support the evidence for the existence of an intermediate effect in qualitative terms, so they were considered to derive the conclusions on the hazard identification, but not for weighting the evidence (i.e. the likelihood of an effect).

8.2. Assessment of the likelihood of a health effect

The likelihood of a health effect in the overall body of evidence was evaluated using a modified version of step 5 of the NTP Handbook for Conducting a Literature-Based Health Assessment Using OHAT Approach for Systematic Review and Evidence Integration (NTP-OHAT, 2015) and of the VKM risk assessment of energy drinks and caffeine (VKM, 2019).

The studies addressing endpoints belonging to a specific cluster were grouped according to study design features. As detailed in the NTP-OHAT (2015) an initial confidence rating of human and animal studies should be assigned on the basis of the study design and its intrinsic ability to potentially set up an association between exposure to a substance and a subsequent effect. The following four descriptors were used to determine this initial level of confidence:

- Controlled exposure conditions.
- Exposure preceding the effect onset.
- Outcome being assessed at individual level.
- Presence of an appropriate comparison group.

For experimental animal studies, the initial confidence was rated high, as this design ensured the fulfilment of all the four key study features listed above.

For cohort and case-control human studies the initial confidence was rated moderate as the 'controlled exposure conditions' feature was not fulfilled. Considerations on whether the exposure precedes the outcome were carried out at internal validity level, thus, resulting in considering this aspect as fulfilled.

The studies grouped for a given cluster and exposure period were then evaluated for elements that would downgrade or upgrade the likelihood of a health effect at study level and then overall at body of evidence level. Similar considerations as those described in the NTP-OHAT tool (NTP-OHAT, 2015) and in Balshem et al. (2011) were applied in order to decide on downgrading or upgrading the likelihood of a health effect and therefore the reader should refer to these publications for more detailed explanations. In brief, on a cluster basis the following elements were considered for downgrading the initial ratings of the confidence in the body of evidence:

- internal validity;
- external validity (for animal studies only);
- unexplained inconsistency;
- imprecision (for human studies only).

Elements that conversely were considered for upgrading the confidence in the body of evidence are:

- effect size (in human studies only);
- dose–response;
- consistency across study design type/dissimilar populations/animal models or species (at body of evidence level only);
- residual confounding (this applies mainly to human observational studies. If a study reports an effect or association despite the presence of residual confounding, confidence in the association was increased).

Downgrading the confidence were reserved to cases in which there were at least serious limitations affecting the study or the body of evidence.

As reported in the NTP-OHAT (2015), 'if a decision to downgrade was borderline for two domains, the body of evidence was downgraded once in a single domain…'. The body of evidence was not downgraded/upgraded twice for the same reason if it was applicable to more than one domain of the body of evidence.

The likelihood of a health effect was assessed per exposure period considering the internal and external validity (for animal studies only) of the different studies and the rationale underlying the ratings is documented in Table 11 (as modified from the NTP-OHAT, 2015, evidence profile table).

After the potential downgrading and upgrading of the evidence, five judgements on the final likelihood of a health effect at the body of evidence level were possible:

- Very Likely: There is very high confidence in the body of evidence for an association between exposure to the substance and health effect/s (e.g. there is much evidence showing consistent effect/s).
- Likely: There is high confidence in the body of evidence for an association between exposure to the substance and health effect/s (e.g. there is evidence showing consistent effects).
- As Likely As Not (ALAN): There is low confidence in the body of evidence for an association between exposure to the substance and health effect/s (e.g. there is evidence showing inconsistent effects).
- Not Likely: There is very low confidence in the body of evidence for an association between exposure to the substance and health effect/s (e.g. there is evidence showing consistent no effects).
- Inadequate evidence: There is insufficient evidence available to assess if the exposure to the substance is associated with and health effect/s or data are missing.

The likelihood levels were scored by the experts taking into account all the elements above described.

The Tier 1 and Tier 2 studies were considered in the assessment firstly as evidence for deriving the likelihood of the effects; if these showed inconsistent results, then also the Tier 3 studies were taken into account. Tier 3 studies including relevant endpoints, but with less than three BPA tested doses, were excluded from the WoE.

An evidence was judged 'Likely' when, for instance, three Tier 1, together with one Tier 2, or one Tier 3 studies, were available, all showing effects in the same direction (good quality evidence, consistent results); or if one Tier 1 study showing clear effects observed in both genders, with a clear dose-response, was available.

An evidence was judged 'Inadequate' when, for instance, only Tier 3 studies or only one Tier 1 or 2 single-dose studies were available. An evidence was judged 'Not Likely' when, for instance, two Tier 1 and one Tier 2 studies were available, all showing no effects (good quality evidence, no effects); or if two Tier 1 studies were available showing inconsistent effects at different doses (good quality evidence, inconsistent effects); or if no dose-response in one Tier 1 and one Tier 2 study (good quality evidence, no dose-response).

An evidence was judged 'ALAN' in cases considered between 'Likely' and the 'Not Likely', so, for instance, when one Tier 1 study that showed no effects in both sexes and one Tier 2 study that showed

effects only in one sex were available (good quality evidence, unexplained inconsistency in the results); or if one Tier 1 with a monotonic dose response (MDR), one Tier 1 single-dose with no effects and one Tier 3 with no effects were available (good and medium/low quality evidence, inconsistent results).

The final likelihood at cluster level was determined by the highest likelihood in the exposure periods considered in the cluster.

Table 11. Template for grading confidence in the body of evidence on the likelihood of a health effect per cluster in human studies (adapted from NTP-OHAT, 2015 and VKM Risk assessment of energy drinks and caffeine, VKM, 2019)

RefID - Endpoints	Elements downgrading the confidence rating			Elements upgrading confidence rating			Final rating for health effect (single paper rating)
	Internal validity (single paper rating)	**Unexplained inconsistency (single paper rating)**	**Imprecision (single paper rating)**	**Effect size (single paper rating)**	**Dose-response (single paper rating)**	**Residual confounding (single paper rating)**	
RefID n – Citation 1, endpoint 1, endpoint 2, endpoint 3 RefID n – Citation 2, endpoint 2, endpoint 4, endpoint 5 RefID n – Citation 3, endpoint 1, endpoint 2, endpoint 5	Tier 1; or Tier 2; or Tier 3	Very serious Serious Not serious concerns	Very serious Serious Not serious concerns	Large Not large	Yes No NA	Yes No NA	Very Likely Likely ALAN Not Likely Inadequate Evidence
	Internal validity (body of evidence rating)	**Unexplained inconsistency (body of evidence rating)**	**Imprecision (body of evidence rating)**	**Effect size (body of evidence rating)**	**Dose-response (body of evidence rating)**	**Consistency (body of evidence rating)**	**Residual confounding (body of evidence rating)**	**Final rating for health effect (body of evidence rating)**
Overall cluster judgement	Very serious Serious Not serious concerns	Very serious Serious Not serious concerns	Very serious Serious Not serious concerns	Large Not large	Yes No NA	Yes No NA	Yes No NA	Very likely Likely ALAN Not likely Inadequate evidence

Table 12. Template for grading confidence in the body of evidence on the likelihood of a health effect per cluster in animal studies (adapted from NTP-OHAT, 2015 and VKM Risk assessment of energy drinks and caffeine (VKM, 2019))

Cluster /Exposure time	RefID	Species	Endpoint measured	Internal validity	External validity (animal model)	External validity (Endpoints)	Is there an effect?	Is there a MDR or a NMDR or is a single-dose study?	Unexplained inconsistency of the results between different studies (per endpoint)	Likelihood of effect (per endpoint), rationale	Unexplained inconsistency of the results between different studies (per cluster)	Likelihood of effect (per cluster), rationale
RefID n – Citation 1												
RefID n – Citation 2												

To answer the questions in the table 12, the experts followed the indications reported below:

- Is there an effect?
  If yes: to indicate the direction of the effect (↑,↓), the corresponding doses (µg/kg bw per day), the duration of exposure until measurement, and the sex of the animal tested (f/m).
  If no: to mention 'no effect' (~), the corresponding doses (µg/kg bw per day) and the sex of the animal tested (f/m).

- Is there an MDR or a Non-Monotonic dose response (NMDR) or is it a single-dose study?
  y, MDR if at least two adjacent doses show a monotonic effect.
  y, NMDR or if at least two adjacent doses show a non-monotonic effect (see explanations in Section 2.3.3. of the Scientific opinion).
  n (n, no effects or n, no dose–response).
  ? when the statistically significant effect is present only at the lowest or at the highest dose tested.
  Single-dose study.

- Unexplained inconsistency of the results between different studies (per endpoint):
  y (in case of unexplained inconsistency).
  n (in case of explained inconsistency, or in case there is no inconsistency, neither explained or unexplained). If there is explained inconsistency, the explanation can be given in column K, close to the likelihood.
  na, in case there is only one study and therefore it is not possible to have a unexplained inconsistency between studies. Also in case there are more endpoints but all in the same study.

- Likelihood of effect (per endpoint):
  To indicate the likelihood for endpoint (Inadequate evidence, Not Likely, ALAN, Likely, Very Likely), giving a rationale.
  If needed, to add also the explanation for an inconsistency.
  If only one sex is tested, to indicate the likelihood is only valid for this sex.
  If there is a different likelihood for both sexes, to indicate for which sex the likelihood is valid.
  To indicate if one sex was not tested → for example: f (nt), and provide a rationale for the answer.

- Unexplained inconsistency of the results between different studies (per cluster):
  y (in case of unexplained inconsistency).
  n (in case of explained inconsistency). If there is explained inconsistency the explanation can be given in column M.
  na, in case there is only one study and therefore it is not possible to have a unexplained inconsistency between studies. Also in case there are more endpoints but all in the same study.

- Likelihood of effect (per cluster):
  To indicate the likelihood for cluster (inadequate evidence, Not Likely, ALAN, Likely, Very Likely), giving a rationale.
  If needed, to add also the explanation for an inconsistency.
  If only one sex is tested, to indicate the likelihood is only valid for this sex.
  If there is a different likelihood for both sexes, to indicate for which sex the likelihood is valid.
  To indicate if one sex was not tested → for example: f (nt), and provide a rationale for the answer.

8.3. Integration of human and animal evidence for the final assessment of the likelihood of a health effect

Error! Reference source not found. provides an overview of the process of integration of the human a

nd animal evidence for the final assessment of the likelihood of a health effect at cluster level.

Human evidence	**Very likely**	VL	VL	VL	VL	VL
	Likely	L	L	L	L	VL
	ALAN	ALAN	ALAN	ALAN	L	VL
	Not Likely	Not Likely	Not Likely	ALAN	L	VL
	Inadequate evidence	Non-classifiable	Not Likely	ALAN	L	VL
		Inadequate evidence	**Not Likely**	**ALAN**	**Likely**	**Very Likely**
		Animal evidence				

Figure 1. Integration of human and animal evidence for the final assessment of the likelihood of a health effect

The highest level of evidence for a health effect among the different exposure periods within a cluster was considered as the likelihood of a health effect for the cluster.

The level of evidence for a health effect at cluster level resulting from the human evidence stream was combined with that deriving from the animal evidence stream to reach a single hazard identification conclusion (using a process adapted from step 7 of the NTP-OHAT Handbook, NTP-OHAT, 2015).

Differently from the NTP process, this protocol foresaw five hazard identification conclusion categories, i.e. 'Very Likely', 'Likely', 'As Likely As Not (ALAN)', 'Not Likely' and 'Not classifiable', expressing the level of likelihood of a health effect associated with the exposure to BPA for the cluster under consideration.

These verbal terms were chosen to align the hazard identification conclusions of the next BPA evaluation with the expressions used in the 2015 EFSA opinion on BPA (EFSA CEF Panel, 2015). To date, 'As Likely As Not' means a level of likelihood in which it was about equally likely that BPA caused, or did not cause, the effect.

In general, the human or animal stream with the highest level of evidence would drive the conclusions on the likelihood of the effect. If for a specific cluster in one stream, no corresponding cluster was available in the other stream, and so an integration of the likelihood from the two lines of evidence would not be possible, then the likelihood of the only available evidence (e.g. human or animal) was considered for the next steps of the assessment.

9. Method for performing hazard characterisation

By hazard characterisation it was intended that the analysis of the dose-response relationship and the identification of a reference point [lower confidence limit of the benchmark dose (BMDL) associated with a specified change in response, the so-called benchmark response (BMR)] would be the basis for a new TDI.

Analysis of the data was performed according to the EFSA Guidance on the use of the benchmark dose (BMD) approach in the risk assessment (EFSA Scientific Committee, 2017a)[11].

[11] The CEP Panel was aware that an updated EFSA Scientific Committee guidance on the use of the benchmark dose approach in risk assessment was under development during the BPA re-evaluation reported in this opinion; it was in fact published in October 2022 (EFSA Scientific Committee, 2022). The 2018 EFSA cross-cutting guidance lifecycle document (EFSA, 2018) states

In humans, exposure could only be estimated by the sum of urinary conjugated and unconjugated BPA concentrations. These could not be directly related to an internal/systemic concentration of unconjugated BPA, which was considered the toxicologically relevant fraction for a reference point. Therefore, human studies were not brought forward for BMD analysis.

Dose–response analysis was performed for 'Very Likely' and 'Likely' effects using experimental animal studies showing adverse effects relevant to humans. An effect was 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 IPCS, 2009). Given the broad number of endpoints examined, the adversity of a specific effect and a critical effect size were evaluated case by case based on expert judgement. A justification was provided.

Dose–response analysis was carried out only for experimental animal studies that have been assigned relatively high internal and external validity (Tier 3 studies were excluded) and included the number of doses sufficient to estimate the parameters of the BMD model. Single-dose studies were not brought forwards for BMD analysis.

Studies supporting 'Very Likely', 'Likely' and ALAN effects were collectively considered in an uncertainty analysis, to define the need for an additional uncertainty factor in deriving a TDI to address uncertainties in the database (see Appendix D of the Scientific opinion).

If information needed to perform the BMD analysis was missing in a study, e.g. the real number of animals or individual data due to a litter effect, the author of the study could be contacted in order to retrieve them.

A cut-off value of 10 for the ratio of the lowest non-zero dose and the BMDL was used to bring a study forward for selection of the Reference Point (RP). If this criterion was not fulfilled (ratio > 10), the BMDL was extrapolated too far outside the dose range and the study design was considered not suitable to evaluate the relevant effect sizes. However, the study was considered in the uncertainty analysis.

If several dose–response analyses were performed, the lowest RP, after conversion to HED, was selected.

For the human hazard characterisation, data on the toxicokinetics [Absorption, Distribution, Metabolism and Excretion (ADME) and PBPK modelling] supported the extrapolation of results from experimental animal studies to humans. This information was also important to determine which uncertainty factors had to be applied when establishing the health-based guidance value. It should be noted that the default factor of 4 for interspecies kinetic differences was already taken into consideration by the chemical-specific approach in which the ratio of AUCs in animals to the AUC in humans was used to adjust the external doses in animals to the external doses in humans. The remaining uncertainty factor should cover interspecies differences in toxicodynamics (default factor was 2.5) and inter-individual variability in both toxicokinetics and toxicodynamics (the default factor being 10). An additional uncertainty factor could also be necessary to cover for uncertainty in the database (see Appendix D of the Scientific opinion).

10. Uncertainty analysis

The evaluation of the uncertainties linked to the evaluation of the new evidence on BPA and the setting of a full TDI was performed in accordance with the EFSA Guidance on Uncertainty Analysis in Scientific Assessment (EFSA Scientific Committee, 2018). The detailed description of the methodology used for the uncertainty analysis is presented in Appendix D of the Scientific opinion.

11. Amendments to the protocol

The protocol was approved by the FIP WG on BPA assessment protocol[12] on 30 November 2017 and it

that 'a cross-cutting guidance document has to be applied to all new risk assessment projects that start 6 months after its publication and the completion of the dissemination, communication and capacity building activities'. Therefore, the 2017 EFSA Scientific Committee guidance on BMD approach (EFSA Scientific Committee, 2017a) was used in this opinion.

[12] https://ess.efsa.europa.eu/doi/doiweb/wg/682790

was originally published as an EFSA supporting publication on 21 December 2017 (EFSA, 2017b).

During the testing and implementation of the protocol, the amendments reported below were introduced:

All the verbs were converted from future to the past tense.

Section 2.2. Objectives of the hazard assessment: A clarification on the routes of exposure and their cut-off doses was added.

Section 2.5. Endpoints relevant to the hazard assessment: An explanation of the categorisation system of the health outcome categories (HOCs) applied in the opinion and of the criteria used for the selection of the relevant endpoints and their grouping into clusters was provided. Moreover, a clarification on the consideration of absolute and relative organ weights in the risk assessment and references on the definition of adversity were included.

Section 2.6. Identification of the hazard assessment subquestions: Examples of other types of evidence were included. A more precise description of the health outcomes categories, and of the exposure routes and their cut-off doses was reported in Table 1.

Section 3.1. Time-span of the literature search: A clarification about the timespan of the evidence search, most of all related to the additional search for Genotoxicity, was added.

Section 3.2. Information sources: The information sources used for the literature search for the HOC Genotoxicity was added.

Section 3.3. Type of evidence: A clarification about the use of the Reviews was included.

Section 4.1. Screening of titles and abstracts: The screening question for selecting at title and abstract level the literature for the HOC Genotoxicity for the additional time span considered was added.

Clarification: The title and abstract screening process for all the HOC apart for Genotoxicity, which was done internally, was outsourced to a contractor, and therefore not carried out by the WG experts, as could have been perceived when reading the initial protocol: 'A check on a random sample of studies was performed by the WG during the initial stages of the work…'. It was therefore considered more appropriate to move this sentence up.

Section 4.2. Examining full-text reports for eligibility of studies: A clarification about the categorisation of the additional genotoxicity studies was included.

Section 4.2.3. Selection of the endpoints of interest. A clarification of the HOCs considered in the selection of the endpoints of interest was added.

Section 4.2.4.2. Selection of exposure of interest: A clarification on the routes of exposure and their cut-off doses was added. Moreover, a clarification on the treatment of the non-mammalian animal studies was added.

Section 4.2.5. A clarification on the treatment of the non-mammalian animal studies was added in in Table 4 (inclusion/exclusion criteria related to MoA studies). A more correct naming of the HOCs and a clarification of the routes of exposure and their cut-off doses were reported in Table 2 and 3. A specification of the time-span considered was reported in Table 2, 3 and 4. A correction of the approach used for the genotoxicity assessment was reported in Table 2 and 4. A footnote with the genotoxic endpoints considered was added in Table 4.

Section 5.1. Data extraction: For studies undergoing a systematic approach, the data extraction was not outsourced to a contractor. Due to the high number of publications retrieved and their complexity, for animal studies data extraction was not implemented using DistillerSR. Moreover, due to the high number of publications retrieved, data extraction was not performed by one reviewer and then

systematically checked for quality/consistency by a second reviewer. Due to the high number of relevant endpoints identified (more than 300 for which all the data from all the studies – Tiers 1, 2 and 3 – should have been extracted) and resource limitation, the data extraction for the production of the summary graphs was considered unfeasible. It was then decided to extract the data in summary tables with appropriate information for all the studies containing the relevant endpoints and grouped by HOC, cluster and period of exposure. For hazard identification, gathering the results of the various studies into a certain cluster enabled easy visualisation of the consistency in qualitative terms of BPA effects across e.g. different studies and/or exposure periods and levels, and/or animal species, taking into account study validity. Tables 5 and 6 were revised accordingly. In Table 5 the correct naming of the HOCs and of the clusters considered was introduced.

For studies undergoing a narrative approach (i.e. cross-sectional studies and MoA in vivo and in vitro studies), the data extraction was outsourced to an external contractor. This information was added in the text.

Section 6. Internal validity assessment: Due to the high workload and resources limitations, it was decided to appraise the studies (both for internal and external validity) in series and not in parallel as originally planned. Moreover, the appraisal was not conducted separately for each study by endpoint; the endpoints measured in a study were first gathered in groups of endpoints for which a unique appraisal (both for internal and external validity) was considered appropriate and then appraised together. If some of the appraisal questions would be answered differently for one or more endpoint in the same study, a new assessment should have been made for this/these endpoint(s).

A revision of the areas of expertise for the WG was added in Section 6: genotoxicity, mammary gland proliferation and carcinogenicity and uncertainty analysis were added.

The indication of the specific approach implemented for the HOC Genotoxicity was included.

Section. 6.1 Internal validity assessment of human studies: Considering the high workload, resource limitations and the fact that a valid exposure characterisation was the most common problem affecting human epidemiological studies investigating the relation between BPA exposure and adverse outcomes[13], it was decided to apply a tiered approach for the appraisal of the human studies. For the appraisal started with Question 3 'Can we be confident in the exposure characterisation?', if the study fell in Tier 3 due to a judgement of 'Probably high' or 'Definitely high' then the evaluation was stopped at this point.

Section 6.2. Internal validity assessment of animal studies: Considering the high workload, resource limitations and the most common problems affecting animal studies investigating the relation between BPA exposure and adverse outcomes, it was decided to apply a tiered approach for the appraisal of the animal studies.

Out of the eight questions comprised in the critical appraisal tool for animal studies, three subquestions were finally considered as key.

The appraisal started with Question 5: 'Can we be confident in the exposure characterisation?' Three elements had to be considered to answer the main question: one key subquestion on the purity of the test compound and the two other elements. If the study was considered at high risk of bias for the key subquestion then the evaluation was stopped at this point and the study was classified as Tier 3, otherwise the appraisal was continued.

The next question was Question 6: 'Can we be confident in the outcome assessment?' Eight elements had to be considered to answer the main question: one key subquestion on the blinding of the outcome assessor and the seven other elements. If the study was considered at high risk of bias for the key subquestion then the evaluation was stopped at this point and the study was classified as Tier 3,

[13] Taking into account BPA's toxicokinetics it was decided to consider as appropriate the frequency of the measurements when there were at least three measurements per year in 24-hour urine samples.

otherwise the appraisal was continued.

The next question was Question 8: 'Were the statistical methods and the number of animals per dose group appropriate?' Two elements had to be considered to answer the main question: one key subquestion on the appropriateness of the number of animals per dose group and the other element. If the study was considered at high risk of bias for the key subquestion then the evaluation was stopped at this point and the study was classified as Tier 3, otherwise the appraisal was continued with the other questions starting from Question 1.

Table 9 was revised accordingly.

Section 7. External validity: Considering external validity, as animal models are proxies for human health, it was decided to consider them at the maximum as indirectly relevant for human health.

Due to the high workload and resources limitations, it was decided to appraise the studies (both for internal and external validity) in series and not in parallel as originally planned. Moreover, the appraisal was not conducted separately for each study by endpoint; the endpoints measured in a study were first gathered in groups of endpoints for which a unique appraisal (both for internal and external validity) was considered appropriate and then appraised together. If some of the appraisal questions would be answered differently for one or more endpoint in the same study, a new assessment should have been made for this/these endpoint(s).

Section 8. Weighting the body of evidence:

Due to the high number of relevant endpoints identified, the fact that some of them were biologically related and due to resource limitation for hazard identification, the evaluation of the confidence in the overall body of evidence was performed by clusters of endpoints and not by endpoint as originally planned. The highest level of evidence for a health effect among the different exposure periods within a cluster was considered as the likelihood of a health effect for the cluster.

For the epidemiological studies, relevant clusters were identified for the WoE (see Section 2.5). For the animal studies, due to the high number of endpoints measured and resource limitation, it was decided to extract the data only for relevant endpoints (see Section 2.5). Subsequently data were extracted from all the studies containing a relevant endpoint, irrespective of their Tier allocation and evidence of effect.

Due to the high number of relevant endpoints identified and resource limitation for hazard identification, the doses in the experimental animal studies were not converted into the corresponding HEDs for the WoE assessment. The conversion of doses to an HED was limited to selected studies within hazard characterisation and uncertainty analysis.

Considering that the studies were grouped in clusters, adapting the methodologies from NTP-OHAT (NTP-OHAT, 2015) and VKM (VKM, 2019), the assessment of the likelihood of a health effect was derived for both the streams of human and animal studies without first passing an evaluation of the confidence in the body of evidence. First, the likelihood of a health effect was evaluated by study, measuring endpoints in the same cluster. Then at the level of the whole body of evidence for the cluster.

Due to the high number of clusters identified and resource limitation the identification of clusters with no health effect was not performed. Instead, confidence was only expressed in terms of likelihood of a health effect.

Due to the high number of clusters identified and resource limitations the likelihood of a health effect was not assessed independently by two reviewers.

The likelihood of a health effect was expressed for both the streams of human and animal studies as Very Likely, Likely, As Likely As Not (ALAN), Not likely or Inadequate evidence, and not as High, Moderate, Low or Inadequate as originally planned.

A description of the exposure categories considered for the WoE and a clarification of the way to perform the assessment was added in Section 8.1.1 and 8.2.

Some details on the assessment of the likelihood of a health effect were added in Section 8.2. Tables 10 and 11 were revised accordingly.

The indication of the approach implemented for the HOC Genotoxicity was included.

The integration scheme of the likelihoods from the human and animal streams was revised (see Figure 1 in Section 8.3) and some clarifications on how to do the integration were added in Section 8.3.

Section 9. Method for performing hazard characterisation: Clarifications on the consideration of studies for the BMD analysis and the hazard characterisation and on the possibility to ask for data to the authors for BMD analyses were added.

Section 10. Uncertainty analysis: A detailed description of the methodology used for the uncertainty analysis has been included in Section 2.3.4, 2.3.5 and in Appendix D of the Scientific opinion. Consequently, most text in Section 10 on the uncertainty analysis became obsolete.

Section 11. Amendments to the protocol: The previous Section 11 ('Plans for updating the protocol and for dealing with newly available evidence') was deleted because it was no longer applicable. All the revisions implemented in the Protocol used in the re-evaluation of BPA are reported in the new Section 11.

Appendix A.1 - Search strings used for each database:

The search terms and databases used for the retrieve of the literature on Genotoxicity were added.

Appendix A.2 - Guidelines for the assessment of internal validity:

More specifications on the assessment of the internal validity were added and the questions were re-ordered listing as first the revised key ones.

Appendix A.3 - Guidelines for the assessment of external validity:

Further indications on how to answer to the questions related to the animal model were added.

References

Balshem H, Helfand M, Schünemann HJ, Oxman AD, Kunz R, Brozek J, Vist GE, Falck-Ytter Y, Meerpohl J, Norris S and Guyatt GH, 2011. GRADE guidelines: 3. Rating the quality of evidence. Journal of Clinical Epidemiology, 64(4), 401–406. doi: 10.1016/j.jclinepi.2010.07.015

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. doi: 10.1002/jat.2991

EFSA (European Food Safety Authority), 2007. Opinion of the Scientific Panel on Food Additives, Flavourings, Processing Aids and Materials in Contact with Food (AFC) related to 2,2-bis(4-Hydroxyphenyl)propane. EFSA Journal 2007;428, 75 pp. doi: 10.2903/j.efsa.2007.428

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. EFSA Journal 2008;759, 10 pp.

EFSA (European Food Safety Authority), 2010. Application of systematic review methodology to food and feed safety assessments to support decision making. EFSA Journal 2010;8(6):1637, 90 pp. doi: 10.2903/j.efsa.2010.1637

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

EFSA (European Food Safety Authority), Gundert-Remy U, Barizzone F, Croera C, Putzu C and Castoldi AF, 2017a. Report on the public consultation on the draft EFSA Bisphenol A (BPA) hazard assessment protocol. EFSA Supporting Publications 2017a;14(12), 1355E pp. doi: 10.2903/sp.efsa.2017.EN-1355

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 A, 2017b. Bisphenol A (BPA) hazard assessment protocol. EFSA Supporting Publications 2017;14(12):1354E, 75 pp. doi: 10.2903/sp.efsa.2017.EN-1354

EFSA CEF Panel (EFSA Panel on Food Contact Materials, Enzymes, Flavourings and Processing Aids), 2010. Scientific Opinion on Bisphenol A: evaluation of a study investigating its neurodevelopmental toxicity, review of recent scientific literature on its toxicity and advice on the Danish risk assessment of Bisphenol A. EFSA Journal 2010;8(9):1829, 116 pp. doi: 10.2903/j.efsa.2010.1829

EFSA CEF Panel (EFSA Panel on Food Contact Materials, Enzymes, Flavourings and Processing Aids), 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 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. EFSA Journal 2015;13(1):3978, 1040 pp. doi: 10.2903/j.efsa.2015.3978

EFSA CEF Panel (EFSA Panel on Food Contact Materials, Enzymes, Flavourings and Processing Aids), Silano V, Bolognesi C, Castle L, Cravedi JP, Engel KH, Fowler P, Franz R, Grob K, Gürtler R, Kärenlampi S, Mennes W, Milana M, Penninks A, Smith A, Tavares Poças M, Tlustos C, Wölfle D, Zorn H, Zugravu C, Anderson S, Germolec D, Pieters R, Castoldi A and Husøy T, 2016. 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):e04580, 22 pp. doi: 10.2903/j.efsa.2016.4580

EFSA Scientific Committee, 2012. Guidance on selected default values to be used by the EFSA Scientific Committee, Scientific Panels and Units in the absence of actual measured data. EFSA Journal 2012;10(3):2579, 32 pp. doi: 10.2903/j.efsa.2012.2579

EFSA Scientific Committee, Hardy A, Benford D, Halldorsson T, Jeger MJ, Knutsen KH, More S, Mortensen A, Naegeli H, Noteborn H, Ockleford C, Ricci A, Rychen G, Silano V, Solecki R, Turck D,

Aerts M, Bodin L, Davis A, Edler L, Gundert-Remy U, Sand S, Slob W, Bottex B, Abrahantes J, Marques D, Kass G and Schlatter J, 2017a. Update: use of the benchmark dose approach in risk assessment. EFSA Journal 2017;15(1):4658, 41 pp. doi: 10.2903/j.efsa.2017.4658

EFSA Scientific Committee, Hardy A, Benford D, Halldorsson T, Jeger MJ, Knutsen HK, More S, Naegeli H, Noteborn H, Ockleford C, Ricci A, Rychen G, Schlatter J, Silano V, Solecki R, Turck D, Benfenati E, Chaudhry Q, Craig P, Frampton G, Greiner M, Hart A, Hogstrand C, Lambre C, Luttik R, Makowski D, Siani A, Wahlstroem H, Aguilera J, Dorne J, Fernandez Dumont A, Hempen M, Valtuena Martínez S, Martino L, Smeraldi C, Terron A, Georgiadis N and Younes M, 2017b. Guidance on the use of the weight of evidence approach in scientific assessments. EFSA Journal 2017;15(8):4971, 69 pp. doi: 10.2903/j.efsa.2017.4971

EFSA Scientific Committee, Benford D, Halldorsson T, Jeger MJ, Knutsen HK, More S, Naegeli H, Noteborn H, Ockleford C, Ricci A, Rychen G, Schlatter J, Silano V, Solecki R, Turck D, Younes M, Craig P, Hart A, Von Goetz N, Koutsoumanis K, Mortensen A, Ossendorp B, Martino L, Merten C, Mosbach-Schulz O and Hardy A, 2018. Guidance on uncertainty analysis in scientific assessments. EFSA Journal 2018;16(1):5123, 39 pp. doi: 10.2903/j.efsa.2018.5123

EFSA Scientific Committee, 2019. Guidance on harmonised methodologies for human health, animal health and ecological risk assessment of combined exposure to multiple chemicals. EFSA Journal 2019;17(3):5634, 77 pp. https://doi.org/10.2903/j.efsa.2019.5634ISSN:1831-4732

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: final report, including report of stakeholder meeting on bisphenol A. Available online: https://apps.who.int/iris/bitstream/handle/10665/44624/97892141564274_eng.pdf

Guyatt G, Oxman AD, Akl EA, Kunz R, Vist G, Brozek J, Norris S, Falck-Ytter Y, Glasziou P, DeBeer H, Jaeschke R, Rind D, Meerpohl J, Dahm P and Schünemann HJ, 2011. GRADE guidelines: 1. Introduction – GRADE evidence profiles and summary of findings tables. Journal of Clinical Epidemiology, 64(4), 383–394. doi: 10.1016/j.jclinepi.2010.04.026

Higgins J and Green S, 2011. Cochrane Handbook for Systematic Reviews of Interventions Version 5.1. 0 [updated March 2011]. The Cochrane Collaboration, 2011. Available from www.cochrane-handbook.org.

Holson RR, Freshwater L, Maurissen JPJ, Moser VC and Phang W, 2008. Statistical issues and techniques appropriate for developmental neurotoxicity testing. A report from the ILSI Research Foundation/Risk Science Institute expert working group on neurodevelopmental endpoints. Neurotoxicology and Teratology, 30(4), 326–348. doi: 10.1016/j.ntt.2007.06.001Krewski D, Westphal M, Al-Zoughool M, Croteau MC and Andersen ME, 2011. New directions in toxicity testing. Annual Review of Public Health 2011;32, 161–178 pp. doi: 10.1146/annurev-publhealth-031210-101153

Li AA, Baum MJ, McIntosh LJ, Day M, Liu F and Gray Jr LE, 2008. Building a scientific framework for studying hormonal effects on behavior and on the development of the sexually dimorphic nervous system. Neurotoxicology, 29(3), 504–519. doi: 10.1016/j.neuro.2008.02.015

NTP-OHAT (National Toxicology Program – Oral Health Assessment Tool), 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.

OECD (Organisation for Economic Co-operation and Development), 2002a. Guidance notes for analysis and evaluation of chronic toxicity and carcinogenicity studies. Available online: https://www.oecd-ilibrary.org/docserver/9789264078499-en.pdf?expires=1631206791&id=id&accname=guest&checksum=72F2D35087337CC7E285B2627F239A85

OECD (Organisation for Economic Co-operation and Development), 2002b. Series on Testing and Assessment Number 32 and OECD Series on Pesticides Number 10. Guidance Notes for Analysis and Evaluation of Repeat-Dose Toxicity Studies. Available online:

http://www.oecd.org/env/ehs/testing/seriesontestingandassessmenttestingforhumanhealth.htm

OECD (Organisation for Economic Co-operation and Development), 2007. Test No. 440: Uterotrophic Bioassay in Rodents. Available online: http://www.oecd-ilibrary.org/environment/oecd-guidelines-for-the-testing-of-chemicals-section-4-health-effects_20745788

OECD (Organisation for Economic Co-operation and Development), 2008. Series on testing and assessment. Number 43. Guidance document on mammalian reproductive toxicity testing and assessment. Available online: http://www.oecd.org/env/ehs/testing/seriesontestingandassessmenttestingforhumanhealth.htm

OECD (Organisation for Economic Co-operation and Development), 2013. Guidance document on developing and assessing adverse outcome pathways. Available online: http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=env/jm/mono%282013%296&doclanguage=en

Skinner MK, 2008. What is an epigenetic transgenerational phenotype?: F3 or F2. Reproductive Toxicology, 25(1), 2–6. doi: 10.1016/j.reprotox.2007.09.001

Thayer KA, Doerge DR, Hunt D, Schurman SH, Twaddle NC, Churchwell MI, Garantziotis S, Kissling GE, Easterling MR, Bucher JR and Birnbaum LS, 2015. Pharmacokinetics of bisphenol A in humans following a single oral administration. Environment International, 83, 107–115. doi: 10.1016/j.envint.2015.06.008 [RefID 7183].

Tyl R, Myers C and Marr M, 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). Research Triangle Park: RTI International Center for Life Sciences and Toxicology.

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 evaluation of Bisphenol A administered in the feed to CD-1® Swiss mice. Report from RTI International Center for life Sciences and Toxicology, Research Triangle (2006). Two-generation reproductive toxicity study of dietary bisphenol A in CD-1 (Swiss) mice. Toxicological Sciences, 104(2), 362–384. doi: 10.1093/toxsci/kfn084

University of Hertfordshire, 2021a. Implementation of the evidence-based risk assessment for the re-evaluation of Bisphenol A: preparatory work on cross-sectional studies. EFSA supporting publication 2021:EN-6997. 77 pp. doi:10.2903/sp.efsa.2021.EN-6997

University of Hertfordshire, 2021b. Implementation of the evidence-based risk assessment for the re-evaluation of Bisphenol A: preparatory work on Mode of Action studies. EFSA supporting publication 2021:EN-6995. 410 pp. doi:10.2903/sp.efsa.2021.EN-6995

VKM (Norwegian Scientific Committee for Food and Environment), 2019. Protocol for the risk assessment of energy drinks and caffeine. Available online: https://vkm.no/download/18.30da4543166cf237d7c51f66/1541418430780/Energidrikk%20protokoll.pdf

WHO/IPCS (World Health Organization & International Programme on Chemical Safety), 2009. Environmental Health Criteria 240. Principles and Methods for the Risk Assessment of Chemicals in Food. Available online: https://apps.who.int/iris/bitstream/handle/10665/44065/WHO_EHC_240_eng.pdf

Abbreviations

ADME	Absorption, Distribution, Metabolism and Excretion
ALAN	As Likely As Not
AMU	Assessment and Methodological support Unit
AUC	Area under the curve
BMD	Benchmark dose
BMDL	Benchmark dose (lower confidence limit)
BMDL10	Benchmark dose (10% lower confidence limit)
BPA	Bisphenol A
bw	body weight
CEF Panel	Panel on Food Contact Materials, Enzymes, Flavourings and Processing Aids
EC	European Commission
EFSA	European Food Safety Authority
FAO	Food and Agriculture Organisation of the United Nations
FDA	(US) Food and Drug Administration
FIP	Food Ingredients and Packaging
GLP	Good laboratory practices
HED	Human equivalent dose
HEDF	Human equivalent dose factor
i.p.	Intraperitoneal
i.v.	Intravenous

MoA	Mode of action
MDR	Monotonic dose response
NCTR	(US) National Center for Toxicological Research
NTP	(US) National Toxicology Program
NMDR	Non-Monotonic dose response
OECD	Organisation for Economic Co-operation and Development
OHAT	(US) Office of Health Assessment and Translation
PBPK	Physiologically-based pharmacokinetic (model)
PROMETHEUS	Promoting Methods for Evidence Use in Science
RoB	Risk of bias
s.c.	Subcutaneous
TDI	Tolerable daily intake
t-TDI	Temporary tolerable daily intake
WHO	World Health Organisation
WoE	Weight of evidence

Appendices

Appendix A.1 - Search strings used for each database

Information sources

Information source	Platform	Dates
PubMed	National Library of Medicine	1 January 2013 to 15 October 2018
Scopus	Scopus	1 January 2013 to 15 October 2018
Web of Science Core Collection. Science Citation Expanded Index	Web of Science	1 January 2013 to 15 October 2018
Web of Science Core Collection. Emerging Sources Citation Index (ESCI)	Web of Science	1 January 2013 to 15 October 2018
Web of Science Core Collection. Current Chemical Reactions (CCR-EXPANDED)	Web of Science	1 January 2013 to 15 October 2018
Web of Science Core Collection. Index Chemicus (IC)	Web of Science	1 January 2013 to 15 October 2018
DART	TOXNET	1 January 2013 to 15 October 2018
TOXLINE	TOXNET	1 January 2013 to 15 October 2018

Search strategies

PubMed

Search	Query
#4	Search #1 NOT #2 Filters: Publication date from 1 January 2013
#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=2013–2018
#1	TS=("bisphenol A" OR BPA OR "80 05 7" OR "201 245 8") Indexes=SCI-EXPANDED, ESCI, CCR-EXPANDED, IC Timespan=2013–2018

DART

Search	Query
# 1	(80-05-7 [rn] OR "bisphenol a" OR bpa OR "80 05 7" OR "201 245 8") AND 2013:2018 [yr]

TOXLINE

Search	Query
# 1	(80-05-7 [rn] OR "bisphenol a" OR bpa OR "80 05 7" OR "201 245 8") AND 2013:2018 [yr]

Search terms for genotoxicity

Information source	Platform	Dates
PubMed	National Library of Medicine	15 October 2018 to 21 July 2021
Scopus	Scopus	15 October 2018 to 21 July 2021
Web of Science Core Collection	Web of Science	15 October 2018 to 21 July 2021

Controlled vocabulary	Free text terms
"Aneugens"[Mesh]	Ames assay*
"Chromosomes"[Mesh]	Ames test
"Comet Assay"[Mesh]	Ames tests
"Germ-Line Mutation"[Mesh]	Aneugen*
"Mutagenicity Tests"[Mesh]	Clastogen*
"Mutagenesis"[Mesh]	Chromatid
"Mutagenicity Tests"[Mesh]	Chromosom*
"Mutagens"[Mesh]	COMET assay*
"Mutation"[Mesh]	COMET test
"Sister Chromatid Exchange"[Mesh]	COMET test*
	DNA
	Genotox*
	Micronucl*
	Mutagen*
	Mutagenicity
	Mutation*
	Sister Chromatid Exchange
	Strand break

PubMed

Search	Query	Results
#24	Search: #22 AND #23 Sort by: Most Recent	311
#23	Search: (english[Language]) AND (("2018"[Date - Publication] : "3000"[Date - Publication])) Sort by: Most Recent	4,659,868
#22	Search: #21 AND #19 Sort by: Most Recent	1,036
#21	Search: "Aneugens"[Mesh] OR "Chromosomes"[Mesh] OR "Comet Assay"[Mesh] OR "Germ-Line Mutation"[Mesh] OR "Mutagenicity Tests"[Mesh] OR "Mutagenesis"[Mesh] OR "Mutagenicity Tests"[Mesh] OR "Mutagens"[Mesh] OR "Mutation"[Mesh] OR "Sister Chromatid	2,271,169

	Exchange"[Mesh] OR ((Ames[tiab] OR COMET[tiab]) AND (assay*[tiab] OR test[tiab] OR tests[tiab])) OR aneugen*[tiab] OR Clastogen*[tiab] OR Chromatid[tiab] OR Chromosom*[tiab] OR DNA [tiab] OR Genotox*[tiab] OR Micronucl*[tiab] OR Mutagen*[tiab] OR Mutagenicit*[tiab] OR Mutation*[tiab] OR Sister Chromatid Exchange[tiab] OR Strand break*[tiab] Sort by: Most Recent	
#20	Search: "Aneugens"[Mesh] OR "Chromosomes"[Mesh] OR "Comet Assay"[Mesh] OR "Germ-Line Mutation"[Mesh] OR "Mutagenicity Tests"[Mesh] OR "Mutagenesis"[Mesh] OR "Mutagenicity Tests"[Mesh] OR "Mutagens"[Mesh] OR "Mutation"[Mesh] OR "Sister Chromatid Exchange"[Mesh] OR ((Ames[tiab] OR COMET[tiab]) AND (assay*[tiab] OR test[tiab] OR tests[tiab])) OR aneugen*[tiab] OR Clastogen*[tiab] OR Chromatid[tiab] OR Chromosom *[tiab] OR DNA [tiab] OR Genotox*[tiab] OR Micronucl*[tiab] OR Mutagen*[tiab] OR Mutagenicit*[tiab] OR Mutation*[tiab] OR Sister Chromatid Exchange[tiab] OR Strand break*[tiab] Sort by: Most Recent	2,178,491
#19	Search: "bisphenol A" [Supplementary Concept] OR "bisphenol A"[tiab] OR BPA[tiab] OR "80 05 7"[tiab] OR "201 245 8"[tiab]	13,224

Scopus

Search	Query	Results
#1	(CASREGNUMBER (80-05-7)) OR (TITLE-ABS-KEY ("bisphenol a" OR bpa OR "80 05 7" OR "201 245 8"))) AND (TITLE-ABS-KEY (((ames OR comet) W/5 (assay* OR test OR tests)) OR aneugen* OR clastogen* OR chromatid OR chromosom* OR dna OR genotox* OR micronucl* OR mutagen* OR mutagenicit* OR mutation* OR "Sister Chromatid Exchange" OR "Strand break*")) AND (LIMIT-TO (PUBYEAR , 2021) OR LIMIT-TO (PUBYEAR , 2020) OR LIMIT-TO (PUBYEAR , 2019) OR LIMIT-TO (PUBYEAR , 2018)) AND (LIMIT-TO (LANGUAGE , "English"))	571

Web of Science Core Collection

Set	Query	Results
# 6	#5 AND #2 Indexes=SCI-EXPANDED, ESCI, CCR-EXPANDED, IC Timespan=2018–2021	491
# 5	(TI=(((Ames OR COMET) NEAR/5 (assay* OR test OR tests)) OR aneugen* OR Clastogen* OR Chromatid OR Chromosom * OR DNA OR Genotox* OR Micronucl* OR Mutagen* OR Mutagenicit* OR Mutation* OR "Sister Chromatid Exchange" OR "Strand break*") OR AB=(((Ames OR COMET) NEAR/5 (assay* OR test OR tests)) OR aneugen* OR Clastogen* OR Chromatid OR Chromosom * OR DNA OR Genotox* OR Micronucl* OR Mutagen* OR Mutagenicit* OR Mutation* OR "Sister Chromatid Exchange" OR "Strand break*") OR AK=(((Ames OR COMET) NEAR/5 (assay* OR test OR tests)) OR aneugen* OR Clastogen* OR Chromatid OR Chromosom * OR DNA OR Genotox* OR Micronucl* OR Mutagen* OR Mutagenicit* OR Mutation* OR "Sister Chromatid Exchange" OR "Strand break*")) AND LANGUAGE: (English) Indexes=SCI-EXPANDED, ESCI, CCR-EXPANDED, IC Timespan=2018–2021	307,868
# 2	TS=("bisphenol A" OR BPA OR "80 05 7" OR "201 245 8") Indexes=SCI-EXPANDED, ESCI, CCR-EXPANDED, IC Timespan=1975–2021	35,017

Appendix A.2 - Guidelines for the assessment of internal validity

A.2.1. Human case–control studies[14]

Question no. 3 – Domain: Detection

Key Question B: **Can we be confident in the exposure characterisation?**

Subquestion 1: Consistency of the exposure assessment

Was the exposure consistently assessed (i.e. under the same method and time frame) across groups?

Judgement	Explanation for expert's judgement
A	There is direct evidence that exposure was consistently assessed (i.e. under the same method and time frame) across groups.
B	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.
C	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).
D	There is direct evidence that exposure was not consistently assessed (i.e. under different methods and time frames) across groups.

Subquestion 2: Validity of the methods used for exposure assessment

Was the exposure 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.)

Judgement	Explanation for expert's judgement
A	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 Using less established methods which are validated against well established methods. Note: taking into account BPA's toxicokinetics the frequency of the measurements is appropriate when there are at least three measurements per year in 24-hour urine samples.

[14] These guidelines were slightly adapted from NTP-OHAT (2015).

B	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 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). Note: taking into account BPA's toxicokinetics the frequency of the measurements is appropriate when there are at least three measurements per year in 24-hour urine samples.
C	There is indirect evidence that the exposure was assessed using poorly validated methods that directly measure exposure. OR 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). OR There is insufficient information provided about the method used for exposure assessment (record 'NR' as basis for answer). Note: taking into account BPA's toxicokinetics the frequency of the measurements is appropriate when there are at least three measurements per year in 24-hour urine samples.
D	There is direct evidence that the exposure was assessed using poorly validated methods. Note: taking into account BPA's toxicokinetics the frequency of the measurements is appropriate when there are at least three measurements per year in 24-hour urine samples.

Question no. 1 – Domain: Selection

Key Question A: Did selection of study participants result in appropriate comparison groups?

Appraisal	Explanation for expert's appraisal
++	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), were recruited within the same time frame 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, ethnicity), were recruited within the same time frame 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 (record 'NR' as basis for answer).

- -	There is direct evidence that controls were drawn from a very dissimilar population than cases or recruited within very different time frames.

Question no. 2 – Domain: Attrition

Were outcome data completely reported without attrition or exclusion from analysis?

Appraisal	**Explanation for expert's appraisal**
++	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: reasons for exclusion likely to be related to the outcome, with either imbalance in numbers or reasons for exclusion across study groups.

Question no. 4 – Domain: Detection

Key Question C: Can we be confident in the outcome assessment?

Subquestion 1: Blinding of the outcome assessors

Were the outcome assessors (including study subjects, if outcomes were self-reported) adequately blinded to the exposure level?

Judgement	**Explanation for expert's judgement**
A	There is direct evidence that the outcome assessors (including study subjects, if outcomes were self-reported) were adequately blinded to the exposure level.
B	There is indirect evidence that the outcome assessors (including study subjects, if outcomes were self-reported) 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 was unlikely to bias a particular outcome).
C	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).
D	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).

Subquestion 2: Validity of the methods used to assess the outcome

Was the outcome assessed in cases (i.e. case definition) and controls using well established methods?

Judgement	**Explanation for expert's judgement**
A	There is direct evidence that the outcome was assessed in cases (i.e. case definition) and controls using well established methods.
B	There is indirect evidence that the outcome was assessed in cases (i.e. case definition) and controls using well established methods. OR It is deemed that the outcome assessment methods used would not appreciably bias results.
C	There is indirect evidence that the outcome was assessed in cases (i.e. case definition) and controls using non-acceptable methods. OR There is insufficient information provided about how cases were identified (record 'NR' as basis for answer).
D	There is direct evidence that the outcome was assessed in cases (i.e. case definition) and controls using non-acceptable methods.

Subquestion 3: Time window between exposure and outcome assessment

Was the time window between exposure and outcome assessment appropriate for the endpoint of interest?

Judgement	**Explanation for expert's judgement**
A	There is direct evidence that the time window between exposure and outcome assessment was appropriate for the endpoint of interest.
B	There is indirect evidence that the time window between exposure and outcome assessment was appropriate for the endpoint of interest.
C	There is indirect evidence that the time window between exposure and outcome assessment was not appropriate for the endpoint of interest. OR There is insufficient information provided on the time window between exposure and outcome assessment (record 'NR' as basis for answer).
D	There is direct evidence that the time window between exposure and outcome assessment was not appropriate for the endpoint of interest.

Question no. 5 – Domain: Confounding

Key Question D: Did the study design or analysis account for important confounding and modifying variables?

Subquestion 1: Appropriateness of the adjustments

Were appropriate adjustments made for primary covariates and confounders in the final analyses?

Judgement	Explanation for expert's judgement
A	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). Acceptable considerations 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.
B	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.
C	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).
D	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.

Subquestion 2: Validity of the methods used to measure confounders

Were primary covariates and confounders assessed using valid and reliable measurements?

Judgement	Explanation for expert's judgement
A	There is direct evidence that primary covariates and confounders were assessed using valid and reliable measurements.
B	There is evidence (direct or indirect) that primary covariates and confounders were assessed using valid and reliable measurements (i.e. the authors justified the validity of the measures from previously published research). OR It is deemed that the measures used would not appreciably bias results.
C	There is indirect evidence that primary covariates and confounders were assessed using measurements of unknown validity. OR There is insufficient information provided about the measurement of primary covariates and confounders (record 'NR' as basis for answer).
D	There is direct evidence that primary covariates and confounders were assessed using non-valid measurements.

Question no. 6 – Domain: Selective reporting

Were all measured outcomes reported?

Appraisal	Explanation for expert's appraisal
++	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. This would include outcomes reported with insufficient detail such as only reporting that results were statistically significant (or not). 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.
-	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. Note: 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 that were not pre-specified or reporting outcomes not pre-specified, or that unplanned analyses were included that would appreciably bias results.

Question no. 7 – Domain: Other sources of bias

Key Question E: Do the statistical methods seem appropriate?

Appraisal	Explanation for expert's appraisal
++	The statistical methods have been described with enough detail and 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 summarised 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. OR There is insufficient information provided about the statistical methods (record 'NR' as basis for answer).
- -	The statistical methods have been described in detail, and there is direct evidence that statistical methods are inappropriate, unusual or unfamiliar.

A.2.2. Human cohort studies[15]

Question no. 3 – Domain: Detection

Key Question B: Can we be confident in the exposure characterisation?

Subquestion 1: Consistency of the exposure assessment

Was the exposure consistently assessed (i.e. under the same method and time frame) across groups?

Judgement	Explanation for expert's judgement
A	There is direct evidence that exposure was consistently assessed (i.e. under the same method and time frame) across groups.
B	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.
C	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).
D	There is direct evidence that exposure was not consistently assessed (i.e. under different methods and time frames) across groups.

Subquestion 2: Validity of the methods used for exposure assessment

Was the exposure 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.)

Judgement	Explanation for expert's judgement
A	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 Using less established methods which are validated against well established methods. Note: taking into account BPA's toxicokinetics the frequency of the measurements is appropriate when there are at least three measurements per year in 24-hour urine samples.
B	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

[15] These guidelines were slightly adapted from NTP-OHAT (2015).

	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). Note: taking into account BPA's toxicokinetics the frequency of the measurements is appropriate when there are at least three measurements per year in 24-hour urine samples.
C	There is indirect evidence that the exposure was assessed using poorly validated methods that directly measure exposure. OR 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). OR There is insufficient information provided about the method used for exposure assessment (record 'NR' as basis for answer). Note: taking into account BPA's toxicokinetics the frequency of the measurements is appropriate when there are at least three measurements per year in 24-hour urine samples.
D	There is direct evidence that the exposure was assessed using poorly validated methods. Note: taking into account BPA's toxicokinetics the frequency of the measurements is appropriate when there are at least three measurements per year in 24-hour urine samples.

Question no. 1 – Domain: Selection

Key Question A: Did selection of study participants result in appropriate comparison groups?

Appraisal	Explanation for expert's appraisal
++	There is direct evidence that subjects (both exposed and non-exposed) 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 and health status), were recruited within the same time frame and had 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, with the same method of ascertainment using the same inclusion and exclusion criteria, and were of similar age and health status), were recruited within the same time frame and had similar participation/response rates. OR Differences between exposure groups would not appreciably bias the results.
-	There is indirect evidence that subjects (both exposed and non-exposed) were not similar, recruited within very different time frames and had very different participation/response rates. OR There is insufficient information provided about the comparison groups including a different rate of response reported 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 and had very different

	participation/response rates.

Question no. 2 – Domain: Attrition

Were outcome data completely reported without attrition or exclusion from analysis?

Appraisal	Explanation for expert's appraisal
++	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. 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 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. 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. OR There is insufficient information provided about numbers of subjects lost to follow-up (record 'NR' as basis for answer).
- -	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.

Question no. 4 – Domain: Detection

Key Question C: Can we be confident in the outcome assessment?

Subquestion 1: Blinding of the outcome assessors

Were the outcome assessors (including study subjects, if outcomes were self-reported) adequately blinded to the exposure level?

Judgement	Explanation for expert's judgement
A	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.
B	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.
C	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).
D	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.

Subquestion 2: Validity of the methods used to assess the outcome

Was the outcome assessed using well established methods?

Judgement	Explanation for expert's judgement
A	There is direct evidence that the outcome was assessed using well established methods.
B	There is indirect evidence that the outcome was assessed using well established methods. OR It is deemed that the outcome assessment methods used would not appreciably bias results.
C	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 methods (*record 'NR' as basis for answer*).
D	There is direct evidence that the outcome was assessed using a non-acceptable method.

Subquestion 3: Time window between exposure and outcome assessment

Was the time window between exposure and outcome assessment appropriate for the endpoint of interest?

Judgement	Explanation for expert's judgement
A	There is direct evidence that the time window between exposure and outcome assessment was appropriate for the endpoint of interest.
B	There is indirect evidence that the time window between exposure and outcome assessment was appropriate for the endpoint of interest.
C	There is indirect evidence that the time window between exposure and outcome assessment was not appropriate for the endpoint of interest. OR There is insufficient information provided on the time window between exposure and outcome assessment (record 'NR' as basis for answer).
D	There is direct evidence that the time window between exposure and outcome assessment was not appropriate for the endpoint of interest.

Question no. 5 – Domain: Confounding

Key Question D: Did the study design or analysis account for important confounding and modifying variables?

Subquestion 1: Appropriateness of the adjustments

Were appropriate adjustments made for primary covariates and confounders in the final analyses?

Judgement	Explanation for expert's judgement
A	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, adjustment in multivariate model, stratification, propensity scoring). Acceptable considerations 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.
B	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.
C	There is indirect evidence that the distribution of primary covariates and known confounders (including other exposures) differed between the exposure groups and was not appropriately adjusted for in the final analysis. OR There is insufficient information provided about the distribution of known confounders (including other exposures) (record 'NR' as basis for answer).
D	There is direct evidence that the distribution of primary covariates and known confounders (including other exposures) differed between exposure groups, confounding was demonstrated and was not appropriately adjusted for in the final analyses.

Subquestion 2: Validity of the methods used to measure confounders

Were primary covariates and confounders assessed using valid and reliable measurements?

Judgement	Explanation for expert's judgement
A	There is direct evidence that primary covariates and confounders were assessed using valid and reliable measurements.
B	There is evidence (direct or indirect) that primary covariates and confounders were assessed using valid and reliable measurements (i.e. the authors justified the validity of the measures from previously published research). OR It is deemed that the measures used would not appreciably bias results.
C	There is indirect evidence that primary covariates and confounders were assessed using measurements of unknown validity. OR There is insufficient information provided about the measurement of primary covariates and confounders (record 'NR' as basis for answer).
D	There is direct evidence that primary covariates and confounders were assessed using non-valid measurements.

Question no. 6 – Domain: Selective reporting

Were all measured outcomes reported?

Appraisal	Explanation for expert's appraisal
++	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. This would include outcomes reported with insufficient detail such as only reporting that results were statistically significant (or not). 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.
-	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.

	Note: 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 that were not pre-specified or reporting outcomes not pre-specified, or that unplanned analyses were included that would appreciably bias results.

Question no. 7 – Domain: Other sources of bias

Key Question E: Do the statistical methods seem appropriate?

Appraisal	Explanation for expert's appraisal
++	The statistical methods have been described with enough detail and 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 summarised 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. OR There is insufficient information provided about the statistical methods (record 'NR' as basis for answer).
- -	The statistical methods have been described in detail, and there is direct evidence that statistical methods are inappropriate, unusual or unfamiliar.

A.2.3. Animal experimental studies[16]

Question no. 5 – Domain: Detection

Can we be confident in the exposure characterisation?

(Please note that potentially three elements have to be considered for answering the main question: one key subquestion on the purity of the test compound and, if the reply to that subquestion is not - or --, two other elements)

Key sub-question: Purity of the test compound

Did the test compound contain any impurities?

Judgement	Explanation for expert's judgement Note that: The judgement can be adjusted depending on the relevant available information in the paper, e.g. if purity is equal to 97% but there is clear indication about the 3% of impurities, then it could become ++. If BPA is purchased from Sigma-Aldrich, at least + should be given, since this company sells compounds with purity of either 97% or 99%. If the name of another company is reported, the experts should quickly check if in its website a minimal % of purity of the BPA they sell is reported. If this key subquestion is scored with - or - -, then the appraisal stops and the study is allocated to Tier 3.
++	There is direct evidence that the test compound was unlikely to contain any impurities that may have significantly affected its toxicity. The test compound has been clearly identified and characterised and is of sufficient purity. In the absence of indications about the impurities a purity >99% is considered ++.
+	There is indirect evidence that the test compound was unlikely to contain any impurities that may have significantly affected its toxicity. E.g. 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. In the absence of indications about the impurities a purity ≥97% is considered +.
-	There is indirect evidence that the test compound was likely to contain any impurities that may significantly have affected its toxicity. E.g. 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 likely that impurities are present that would significantly affect the results of the study. In the absence of indications about the impurities a purity <97% is considered -. OR There is no information provided about the test compound (record 'NR' as basis for answer). Please note that in such cases, no additional raw data will be requested to the authors, because the consideration of raw data only for specific studies could bias the overall results.
--	There is direct evidence that the test compound contains impurities that can affect study results. In the absence of indications about the impurities a purity <95% is considered --.

[16] These guidelines were adapted from NTP-OHAT (2015) and SciRAP (Beronius et al., 2014).

Element a: Consistency of the exposure assessment

Was the exposure consistently administered (with the same method and time frame i.e. treatment day and/or time of day) across treatment groups?

Judgement	Explanation for expert's judgement Note that: For dietary exposure studies, between-group differences in feed consumption would be an additional element to consider when dealing with this element.
A	There is direct evidence that exposure was consistently administered (i.e. with the same method and time frame) across treatment groups.
B	There is indirect evidence that exposure was consistently administered (i.e. with the same method and time frame) across treatment groups. OR There is insufficient information provided about exposure administration (record 'NR' as basis for answer) but it is considered that this does not have an impact on the validity of the study.
C	There is indirect evidence that exposure was not consistently administered (i.e. with the same method and time frames) across groups. OR There is insufficient information provided about exposure administration (record 'NR' as basis for answer) and it is suspected that this has an impact on the validity of the study.
D	There is direct evidence that exposure was not consistently administered (i.e. with the same method and time frames) across groups.

Element b: Test system contamination

Did the test system contain any relevant contaminant?

Judgement	Explanation for expert's judgement Note that: 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 oestrogenic 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, 2007). Ideally, feed and water should be tested for the presence of relevant contaminants and phytoestrogens. Similarly, the bedding material should be considered, especially if endocrine activity/disruption is being investigated, since it may contain naturally occurring oestrogenic or anti-oestrogenic substances. e.g. corn cob appears to be anti-oestrogenic and affects cyclicity in rats (OECD, 2007). Specifically, phytoestrogen content should be minimised in the bedding material in these cases. A full report of possible relevant 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.
A	There is direct evidence that the test system is unlikely to contain relevant contaminants that could affect the study results. No relevant contaminants that could have influenced study results are suspected (some materials such as feed, water, bedding) may have been analysed and/or selected in order to avoid any contamination.

B	There is indirect evidence that the test system is unlikely to contain relevant contaminants that could affect the study results. No relevant contaminants that could have influenced study results are suspected (some materials such as feed, water, bedding) may have been analysed and/or selected in order to avoid any contamination. OR There is insufficient information provided about the test system (record 'NR' as basis for answer) but it is considered that this does not have an impact on the validity of the study.
C	There is indirect evidence that the test system is likely to contain relevant contaminants that could affect the study results. Relevant contaminants that could have influenced study results are suspected. OR There is insufficient information provided about the test system (record 'NR' as basis for answer) and it is suspected that this has an impact on the validity of the study.
D	There is direct evidence that the test system is likely to contain relevant contaminants that could affect the study results. e.g. a bedding material known to contain oestrogenic or anti-oestrogenic substances was used in a study investigating endocrine endpoints.

Question no. 6 – Domain: Detection

Can we be confident in the outcome assessment?

(Please note that potentially eight elements have to be considered for answering the main question: one key subquestion on the blinding of the outcome assessor and, if the reply to that subquestion is not - or --, seven other elements)

Key sub-question: Blinding of the Outcome Assessors

Were the outcome assessors adequately blinded to the study group?

Judgement	Explanation for expert's judgement Note that: For manually scored endpoints, blinding is more important, whereas for non-manually scored endpoints is less important, therefore a paper may need to be cloned because of different scoring in this subquestion. If this key subquestion is scored with - or --, then the appraisal stops and the study is allocated to Tier 3.
++	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.
+	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 with the control to appropriately judge the outcome, but additional measures such as independent review by trained pathologists can minimise this potential

	bias. OR There is insufficient information provided about blinding of outcome assessors (record 'NR' as basis for answer) but it is considered that this does not have an impact on the validity of the study.
-	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 There is insufficient information provided about blinding of outcome assessors (record 'NR' as basis for answer) and it is suspected that this has an impact on the validity of the study.
--	There is direct evidence that there was a lack of adequate blinding of outcome assessors, including no blinding or incomplete blinding without quality control measures.

Element a: Consistency of the length of time between exposure and assessment

Was the outcome assessed at the same length of time (i.e. day and/or time of day) after initial exposure in all study groups? (remember to take into consideration the endpoints assignments)

Judgement	**Explanation for expert's judgement**
A	There is direct evidence that the outcome was assessed at the same length of time after initial exposure in all study groups.
B	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. OR There is insufficient information provided about consistency of length of time after exposure for outcome assessment (*record 'NR' as basis for answer*) but it is considered that this does not have an impact on the validity of the study.
C	There is indirect evidence that the outcome was assessed at a different length of time after initial exposure among study groups. OR There is insufficient information provided about consistency of length of time after exposure for outcome assessment (record 'NR' as basis for answer) and it is suspected that this has an impact on the validity of the study.
D	There is direct evidence that the outcome was assessed at a different length of time after initial exposure among study groups.

Element b: Reliability and sensitivity of the animal model

Was a reliable and sensitive animal model used for investigating the test compound and selected endpoints?

Judgement	Explanation for expert's judgement

	Note that: 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.
A	There is direct evidence that a reliable and sensitive animal model was used for investigating the test compound and selected endpoints.
B	There is indirect evidence that a reliable and sensitive animal model was used for investigating the test compound and selected endpoints. The model was not fully described but it may be inferred from other information that it was reliable and sensitive for investigating the test compound and selected endpoints. OR The animal model used is not suspected to be insensitive or unreliable. OR There is insufficient information provided about the animal model (record 'NR' as basis for answer) but it is considered that this does not have an impact on the validity of the study.
C	There is indirect evidence that an unreliable or insensitive animal model was used for investigating the test compound and selected endpoints. The model was not fully described but it may be inferred from other information that it was unreliable and/or insensitive for investigating the test compound and selected endpoints. OR There is insufficient information provided about the animal model (record 'NR' as basis for answer) and it is suspected that this has an impact on the validity of the study.
D	There is direct evidence that an unreliable or insensitive animal model was used for investigating the test compound and selected endpoints.

Element c: Housing conditions

Were housing conditions appropriate for the study type and animal model?

Judgement	Explanation for expert's judgement Note that: 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)
A	There is direct evidence that housing conditions (temperature, relative humidity, light–dark cycle) were appropriate for the study type and animal model. Housing conditions have been fully described and were in line with standard recommendations relevant to the study type and animal model.
B	There is indirect evidence that housing conditions (temperature, relative humidity, light–dark cycle) were appropriate for the study type and animal model. Housing conditions were not fully described but it may be inferred from other information that they were in line with standard recommendations relevant

	to the study type and animal model. OR It is deemed that the deviation from standard recommendations would not appreciably bias results OR There is insufficient information provided about hosing conditions (record 'NR' as basis for answer) but it is considered that this does not have an impact on the validity of the study.
C	There is indirect evidence that housing conditions (temperature, relative humidity, light–dark cycle) were not appropriate for the study type and animal model. Housing conditions were not fully described but it may be inferred from other information that they were not in line with standard recommendations relevant to the study type and animal model. OR There is insufficient information provided about hosing conditions (record 'NR' as basis for answer) and it is suspected that this has an impact on the validity of the study.
D	There is direct evidence that housing conditions (temperature, relative humidity, light–dark cycle) were not appropriate for the study type and animal model. Housing conditions were not in line with standard recommendations relevant to the study type and animal model.

Element d: Appropriateness of the number of animals

Was the number of animals per sex in each cage appropriate for the study type and animal model?

Judgement	Explanation for expert's judgement Note that: The number of animals housed together may have an effect on behaviour 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. 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.
A	There is direct evidence that the number of animals per sex in each cage was appropriate for the study type and animal model.
B	There is indirect evidence that the number of animals per sex in each cage was appropriate for the study type and animal model The number of animals per sex in each cage was not reported but it may be inferred from other information that they were in line with standard recommendations relevant to the study type and animal model. OR It is deemed that the deviation from standard recommendations would not appreciably bias results.

	OR There is insufficient information provided about number of animals per sex in each cage (record 'NR' as basis for answer) but it is considered that this does not have an impact on the validity of the study.
C	There is indirect evidence that the number of animals per sex in each cage was not appropriate for the study type and animal model. The number of animals per sex in each cage was not reported but it may be inferred from other information that they were not line with standard recommendations relevant to the study type and animal model. OR There is insufficient information provided about number of animals per sex in each cage (record 'NR' as basis for answer) and it is suspected that this has an impact on the validity of the study.
D	There is direct evidence that the number of animals per sex in each cage was not appropriate for the study type and animal model.

Element e: Timing and duration of administration of the test compound

Was the timing and duration of administration of the test compound appropriate?

Judgement	Explanation for expert's judgement Note that: 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 behaviour the potential of the administration to produce acute effects on behavioural measures should be considered, especially where the test substance is administered directly to offspring daily (OECD, 2008).
A	There is direct evidence that 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.
B	There is indirect evidence that 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. The timing and duration of administration of the test compound was not reported but it may be inferred from other information that it 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. OR It is deemed that the deviation from standard recommendations would not appreciably bias results. OR There is insufficient information provided about timing and duration of administration of the test compound (record 'NR' as basis for answer) but it is considered that this does not have an impact on the validity of the study.
C	There is indirect evidence that the timing and duration of administration of the test compound is not line with general recommendations for the study type.

	The timing and duration of administration of the test compound was not reported but it may be inferred from other information that it is not in line with general recommendations for the study type and/or it is likely to interfere with the measurements conducted, and cover sensitive periods of development, where relevant. OR There is insufficient information provided about timing and duration of administration of the test compound (record 'NR' as basis for answer) and it is suspected that this has an impact on the validity of the study.
D	There is direct evidence that the timing and duration of administration of the test compound is not line with general recommendations for the study type.

Element f: Reliability and sensitivity of the test methods

Were reliable and sensitive test methods used for investigating the selected endpoints?

Judgement	Explanation for expert's judgement Note that: 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.
A	There is direct evidence that reliable and sensitive test methods were used for investigating the selected endpoint.
B	There is indirect evidence that reliable and sensitive test methods were used for investigating the selected endpoint. OR It is deemed that the deviation from standard recommendations would not appreciably bias results. OR There is insufficient information provided about reliability and sensitivity of the test methods used for investigating the selected endpoint (record 'NR' as basis for answer) but it is considered that this does not have an impact on the validity of the study.
C	There is indirect evidence that reliable and sensitive test methods were not used for investigating the selected endpoint. OR There is insufficient information provided about reliability and sensitivity of the test methods used for investigating the selected endpoint (record 'NR' as basis for answer) and it is suspected that this has an impact on the validity of the study.
D	There is direct evidence that reliable and sensitive test methods were not used for investigating the selected endpoint.

Element g: Timing of the outcome assessment

Were the measurements collected at suitable time points in order to generate sensitive, valid and reliable data?

Judgement	Explanation for expert's judgement Note that: 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 behavioural 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. Data should be collected so that the time of day does not influence measurements. For example, responses in behavioural testing in nocturnal animals like mice and rats are likely to produce different behaviour during the day than during the night. For such reasons reversed lighting conditions may be applied to test nocturnal animals during the day.
A	There is direct evidence that measurements do not seem to have been collected at unsuitable time points in order to generate sensitive, valid and reliable data. 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.
B	There is indirect evidence that measurements do not seem to have been collected at unsuitable time points in order to generate sensitive, valid and reliable data. The timing of tests and measurements where not fully described but it may be inferred from other information that they were appropriate to detect sensitive effects and there are no related aspects that are likely to influence the reliability of the results. OR It is deemed that the deviation from standard recommendations would not appreciably bias results. OR There is insufficient information provided about time points for collection of measurements (record 'NR' as basis for answer) but it is considered that this does not have an impact on the validity of the study.
C	There is indirect evidence that measurements seem to have been collected at unsuitable time points in order to generate sensitive, valid and reliable data. The timing of tests and measurements were not fully described but it may be inferred from other information that they were not appropriate to detect sensitive effects and/or there are related aspects that are likely to influence the reliability of the results. OR There is insufficient information provided about time points for collection of measurements (record 'NR' as basis for answer) and it is suspected that this has an impact on the validity of the study.
D	There is direct evidence that measurements seem to have been collected at unsuitable time points in order to generate sensitive, valid and reliable data. The timing of tests and measurements were not appropriate to detect sensitive effects and/or there are related aspects that are likely to influence the reliability of the results.

Question no. 8 – Domain: Other sources of bias

Were the statistical methods and the number of animals per dose group appropriate?

(Please note that potentially two subquestions have to be considered for answering the main question: one key subquestion on the number of animals and, if the reply to that subquestion is not - or --, another subquestion on the appropriateness of the statistical methods.)

Key sub-question: Number of animals per group**Was the number of animals per dose group appropriate?**

Appraisal	Explanation for expert's appraisal

	Note that: 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. 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. You are requested to consider if the number of animals is sufficient for examining a specific endpoint. If for some endpoints, the number of animals is sufficient and for others not, then the you should not degrade the whole paper, but clone for specific endpoints accordingly (separate the endpoints with sufficient no. of animals from the endpoints with too few animals). The number of animals per endpoint does not necessarily need to be mentioned in the Materials and Methods section, if it is mentioned in the Results section (often reported in the figures or tables included in this section). If the number of animals is not sufficient to perform a reliable statistical analysis, the score in this key subquestion should be - or - - and subsequently the appraisal stops and the paper is allocated to Tier 3.
++	There is direct evidence that 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. Information on the animals were clearly reported and were appropriate.
+	There is indirect evidence that 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. Information on the animals were not clearly reported but it may be inferred from other information that they were appropriate. OR There is insufficient information provided about the number of animals per dose group (record 'NR' as basis for answer) but it is considered that this does not have an impact on the validity of the study.
-	There is indirect evidence that an insufficient number of animals was included in the different treatment groups and loss of animals during the study is likely to have substantially affected statistical power. Information on the animals were not clearly reported but it may be inferred from other information that they were not appropriate. OR There is insufficient information provided about the number of animals per dose group (record 'NR' as basis for answer) and it is suspected that this has an impact on the validity of the study.
- -	There is direct evidence that an insufficient number of animals was included in the different treatment groups and loss of animals during the study is likely to have substantially affected statistical power. Information on the animals were clearly reported and were not appropriate.

Element a: Appropriateness of the statistical analyses

Where the statistical methods appropriate?

Appraisal	Explanation for expert's appraisal Note that:

	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, 2002a) 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 pre-natal conditions can have considerable influence on the statistical 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.
++	There is direct evidence that the statistical methods do not seem inappropriate, unusual or unfamiliar. Statistical methods were clearly reported and were appropriate.
+	There is indirect evidence that the statistical methods do not seem inappropriate, unusual or unfamiliar. Statistical methods were not clearly reported but it may be inferred from other information that they were appropriate. OR There is insufficient information provided about statistical methods (record 'NR' as basis for answer) but it is considered that this does not have an impact on the validity of the study.
-	There is indirect evidence that the statistical methods seem inappropriate, unusual or unfamiliar. Statistical methods were not clearly reported but it may be inferred from other information that they were not appropriate. OR There is insufficient information provided about statistical methods (record 'NR' as basis for answer) and it is suspected that this has an impact on the validity of the study.
- -	There is direct evidence that the statistical methods seem inappropriate, unusual or unfamiliar OR an insufficient. Statistical methods were clearly reported and were not appropriate.

Question no. 1 – Domain: Selection

Was the administered dose or exposure level adequately randomised?

Appraisal	Explanation for expert's appraisal Note that: The randomisation process in this question has to be considered up to the study unit (dams or pups) of interest. So, for instance in a developmental study where also the pups are then assigned to different level of exposure the whole randomisation process from mothers to pups needs to be considered.
++	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). Notes:

	If the study reports that it was performed according to good laboratory practices (GLP), randomisation cannot be considered as done unless specified in other parts of the manuscript. 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 minimise 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. 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). OR There is insufficient information provided about how subjects were allocated to study groups (record 'NR' as basis for answer) but it is considered that this does not have an impact on the validity of the study.
-	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), OR There is insufficient information provided about how subjects were allocated to study groups (record 'NR' as basis for answer) and it is suspected that this has an impact on the validity of the study. 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 judgement 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.

Question no. 2 – Domain: Selection

Was the allocation to study group adequately concealed?

Appraisal	Explanation for expert's appraisal Note that: This question refers to the dosing and treatment part of the experiment (i.e. that the people taking care of the animals may know or not which ones are

	the controls and which ones are the dosed) and not to the outcome measurement. The question that refers specifically to the blinding of the outcome assessor is reported n this form under key subquestion of Question no. 6 – Domain: Detection. Acceptable methods used to ensure allocation concealment include sequentially numbered treatment containers of identical appearance or equivalent methods.
++	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.
+	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. OR There is insufficient information provided about allocation to study groups (record 'NR' as basis for answer) but it is considered that this does not have an impact on the validity of the study.
-	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) and it is suspected that this has an impact on the validity of the study.
- -	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.

Question no. 3 – Domain: Performance

Were the experimental conditions identical across study groups?

(Please note that the following two elements have to be considered for answering the main question.)

Element a: Vehicle

Was the same vehicle used in control and experimental animals?

Judgement	Explanation for expert's judgement
A	There is direct evidence that the same vehicle was used in control and experimental animals.
B	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. OR

	Authors did not report the vehicle used (record 'NR' as basis for answer) but it is considered that this does not have an impact on the validity of the study.
C	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) and it is suspected that this has an impact on the validity of the study.
D	There is direct evidence from the study report that control animals were untreated, or treated with a different vehicle than experimental animals.

Element b: Non-treatment related conditions

Were non-treatment related experimental conditions identical across study groups?

Judgement	**Explanation for expert's judgement**
A	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).
B	Identical non-treatment-related experimental conditions are assumed if authors did not report differences in housing or husbandry.
C	There is indirect evidence that non-treatment-related experimental conditions were not comparable between study groups.
D	There is direct evidence that non-treatment-related experimental conditions were not comparable between study groups.

Question no. 4 – Domain: Attrition

Were outcome data completely reported without attrition or exclusion from analysis?

Appraisal	Explanation for expert's appraisal Note that: Acceptable handling of attrition includes: very little missing outcome data; reasons for missing animals 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; missing outcomes is not enough to impact the effect estimate. 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. If the loss of the animals is not acceptable, the score should be at least - and if the acceptability of loss is different for different endpoints, then you should clone the appraisal forms in Distiller accordingly.
++	There is direct evidence that loss of animals was adequately addressed and reasons were documented when animals were removed from a study. 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. OR There is insufficient information provided about loss of animals (record 'NR' as basis for answer) but it is considered that this does not have an impact on the validity of 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) and it is suspected that this would have an impact on the validity of the study. 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 large and not adequately addressed.

Question no. 7 – Domain: Selective reporting

Were all measured outcomes reported?

Appraisal	Explanation for expert's appraisal
++	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 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. This would include outcomes reported with insufficient detail such as only reporting that results were statistically significant (or not). 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). OR There is insufficient information provided about selective outcome reporting (record 'NR' as basis for answer) but it is considered that this does not have an impact on the validity of the study.
-	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 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) and it is suspected that this has an impact on the validity of the study.
- -	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 that were not pre-specified or reporting outcomes not pre-specified, or that unplanned analyses were included that would appreciably bias results.

Appendix A.3 - Guidelines for the assessment of external validity

As explained in Section 7, in this protocol external validity refers to the relevance of both the animal model and of the endpoint investigated to human health. The guidance reported below has been extracted from the guidance for evaluating relevance of *in vivo* toxicity studies present in the SciRAP tool.[17]

A.3.1. The animal model

Guidance: Is the animal model relevant for human health outcomes?

Consider the motivation behind the choice of animal model (species and strain) given, i.e. why one species or strain was preferred above another. Take into account factors such as species and strain differences in kinetics, metabolism, receptors, etc. Note that the default assumption is that effects observed in the animal model are indirectly relevant for human health as by definition an animal is only a model for humans and can never be directly relevant.

Indirectly relevant: there is no clear evidence that the animal model is irrelevant for the hazard or risk assessment being conducted. However, there may be a suspicion of species and/or strain differences affecting the sensitivity of the model.

Not relevant: there is evidence that the animal model is not relevant for the hazard or risk assessment being conducted.

A.3.2. The endpoint studied

Guidance: Is the endpoint relevant for human health outcomes?

Consider the rationale given for the selection of endpoints. Note that several endpoints may have been investigated. The study should be evaluated based on each individual endpoint, i.e. the reliability and relevance of a study may have to be evaluated several times based on different endpoints. Also consider that even if an endpoint is not (directly) relevant to humans it may be related to another relevant endpoint that was not measured in the study.

Directly relevant: the study addresses the endpoint of interest for the hazard or risk assessment being conducted.

Indirectly relevant: the study addresses a related endpoint to the one of direct interest for the hazard or risk assessment being conducted.

Not relevant: the study addresses an endpoint that is not relevant for the specific hazard or risk assessment being conducted.

[17] Available from www.scirap.org ; also see Beronius et al. (2014)

Appendix A.4 - Impact assessment of excluding non-English studies

The WG undertook a pilot test to assess the approximate impact of omitting non-English publications from the review. The search strings reported for each database in Appendix A.1 were used to gather references from 1 January 2013 until 25 August 2017. After deduplication, 10,814 references were retrieved. The languages of the publications are recorded in Table A.1.

Table A.1: Languages of the references retrieved during the pilot test

Language	No. of references	Proportion (%)
Not reported	687	**6.35**
Chinese	362	**3.35**
Czech	4	**0.04**
Dutch	1	**0.01**
English	9619	**88.95**
English; Dutch	1	**0.01**
English; French	3	**0.03**
English; German	5	**0.05**
English; Hungarian	1	**0.01**
English; Indonesian	1	**0.01**
English; Italian	1	**0.01**
English; Spanish	8	**0.07**
French	25	**0.23**
German	16	**0.15**
Hungarian	1	**0.01**
Japanese	16	**0.15**
Japanese; English	1	**0.01**
Korean	18	**0.17**
Persian	10	**0.09**
Polish	16	**0.15**
Portuguese	6	**0.06**
Russian	3	**0.03**
Slovak	1	**0.01**
Slovenian	1	**0.01**
Spanish	5	**0.05**
Turkish	2	**0.02**
Total	10814	**100.00**

A random sample of 1,100 references out of 10,814 was tested against inclusion/exclusion criteria at 'Title and Abstract' and 'Full-Text' level. Here, 197 papers (18%) reached 'Full-Text' screening. The languages of these publications are reported in Table A.2. For 25 publications the language was not specified in the bibliographic database: however, at full-text examination 22 out of 25 were in English.

Table A.2: Languages of the publications reaching 'Full-Text' screening during pilot test

Language	No. of references	Proportion (%)
Not reported	3 (25 – 22)	**1.52**
Chinese	3	**1.52**
English	188 (166 + 22)	**95.43**
French	2	**1.02**
Polish	1	**0.51**
Total	**197**	**100.00**

Overall, according to this pilot test in the worst-case scenario less than 5% (9/197) of the studies reaching 'Full-Text' screening were not published in English. This lent support to the idea that omitting non-English publications would only have a limited impact as the included English studies would be about 95% of the overall evidence reaching 'Full-Text' screening.

Appendix A.5 - Impact assessment of possibly missing studies with 'null' results on BPA not reported in either the title or the abstract

The EFSA's proposal to first screen papers on the basis of title and abstract was criticised during the public consultation, the main reason being that sometimes studies examining multiple chemicals may not mention the chemicals with 'null' results in the title or abstract. Hence, an initial screening based on title and abstract might lead to the inappropriate exclusion of such studies and in turn compromise the whole assessment.

EFSA's proposed approach to screen the output of the search was in line with internationally recognised guidelines (Higgins and Green, 2011; NTP-OHAT, 2015). Nevertheless, to address this criticism the WG decided to estimate the approximate impact of such a decision.

A pilot test against inclusion/exclusion criteria at 'Title and Abstract' and 'Full-Text' level was performed. The search strings reported for each database in Appendix A.1 were used to gather references from 1 January 2013 until 25 August 2017.

After deduplication, 10,814 references were retrieved and approximately a 10% random sample of references was tested against the inclusion/exclusion criteria at 'Title and Abstract' and 'Full-Text' level. Out of 1100 references, 197 (18%) reached the 'Full-Text' screening stage.

The 903 references that did not reach 'Full-Text' level were classified on the basis of the presence of BPA-related terms in the title or abstract (namely 'bisphenol', 'Bisphenol', 'BPA', 'bpa', '80 05 7', '80057', '201 245 8' and '2012458'). Overall, 441 references did not report any BPA-related terms in the title or abstract.

Full text (directly available through EFSA electronic literature databases) was retrieved for 364 out of 441 references. A text mining exercise was conducted and 319 out of 364 full texts did contain some BPA-related words. These 319 full texts were screened against the inclusion/exclusion criteria for their potential relevance for the assessment. Overall, 25 full texts were found as potentially relevant for BPA hazard assessment: 22/25 were secondary studies, while the remaining three full texts were related to MoA studies. No primary studies directly relevant to human and animal studies were missed.

On the basis of the above analysis it seems reasonable to conclude that an initial screening of papers just on the basis of title and abstract would not compromise the overall assessment. Clear instructions were provided to the reviewers in order to correctly identify secondary studies.