Protocol for a systematic review of associations of bisphenol A exposure with declining semen quality in males to support derivation of a reference dose for mixture risk assessments for male reproductive health

Contents

1. Identification

Protocol for a systematic review of associations of bisphenol A exposure with declining semen quality in men to support derivation of a reference dose for mixture risk assessments for male reproductive health

2. Registration

Final protocol uploaded to Zenodo and attributed the following digital object identifier,

doi: **10.5281/zenodo.5083147**

3. Authors

Professor Andreas Kortenkamp, Centre for Pollution Research and Policy, Brunel University London*, andreas.kortenkamp@brunel.ac.uk

Dr Olwenn V Martin, Centre for Pollution Research and Policy, Brunel University London, olwenn.martin@brunel.ac.uk

Martin Scholze, Centre for Pollution Research and Policy, Brunel University London, martin.scholze@brunel.ac.uk

Dr Sibylle Ermler, Centre for Pollution Research and Policy, Brunel University London, sibylle.ermler@brunel.ac.uk

Dr Asma Baig, Centre for Pollution Research and Policy, Health and Societies, Brunel University London, asma.baig@brunel.ac.uk

Joanne McPhie, Library Services, Brunel University London, joanne.mcphie@brunel.ac.uk

* Halsbury 124, Kingston Lane, Uxbridge UB8 3PH, United Kingdom

Authors' contributions: Brunel University London (BUL) is a linked third party to Public Health England to contribute to the EU project Human Biomonitoring for the EU (HBM4EU; https://www.hbm4eu.eu). BUL's role in this project is to conduct case studies of mixture risk assessment. AK developed the concept for these case studies and selected bisphenol A as one of several substances to be evaluated. AK and OVM developed the systematic review protocol. AK, OVM, AB and SE piloted the screening of relevant studies, data extraction and risk-of-bias evaluations. MS provided statistical support in interpreting risk-of-bias questions for epidemiological studies during this piloting and advice regarding methods to derive reference doses for mixture risk assessment.

4. Amendments

Not applicable

5. Funding and sources of support

The time spent on this project by Andreas Kortenkamp, Sibylle Ermler and Olwenn V Martin is funded by the EU Horizon 2020 project HBM4EU. Martin Scholze is supported by BUL.

Role of the funder: The funder, the European Commission DG RTD has approved the conduct of mixture risk assessment case studies. There is no other role of the funder.

Competing interests: All contributors have no financial or non-financial competing interests to declare. A public annual declaration of interests of OVM can be found at https://echa.europa.eu/about-us/who-we-are/management-board/management-board-members. ICMJE Conflict of Interest Disclosure Forms for all team members can be fund in Supplementary Information – "SI File 1 – ICMJE COI Disclosure Forms".

6. Rationale

Current chemical regulations operate almost exclusively on a chemical-by-chemical basis. There is concern that this approach may not be sufficiently protective if two or more chemicals have the same toxic effect (Evans *et al.*, 2016). Methods for the routine consideration of mixture effects in chemical risk assessment and regulation are currently being elaborated. One aspect of this process is the identification of so-called priority mixtures – mixtures composed of chemicals that jointly affect the health outcomes of concern.

One health outcome of concern is the decline in male reproductive health. There are reports in the literature suggesting that bisphenol A contributes to these declines. Bisphenol A is an androgen receptor antagonist and has been shown to act together with other such antagonists when present in mixtures (Kortenkamp 2020). An adverse outcome pathway network for anti-androgenic chemicals was constructed to decide which chemicals to group together for mixture risk assessments with male reproductive health as the outcome (Kortenkamp 2020). This analysis also revealed that declines in semen quality are a critical endpoint to judge risks to male reproductive health.

Bisphenol A is a phenolic chemical that leaches out from polycarbonate plastic items. It has entered the food chain with widespread human exposure to this chemical.

To conduct a mixture risk assessment, it is necessary to estimate exposure levels to bisphenol A not associated with observable risks to semen quality in men of reproductive age, so-called reference doses. To our knowledge, previous systematic reviews of the association between bisphenol A have only considered epidemiological evidence (Bliatka et al. 2020; Bonde et al. 2017) and there is therefore a need to update and complement these works. To support this effort, a systematic review of associations between bisphenol A exposures and declining semen quality is required, both from human and from animal studies.

7. Objectives – PECO statement

The overall objective of this systematic review is to evaluate whether epidemiological evidence and data from animal studies can be used to derive a reference dose for bisphenol A that is protective against declines in semen quality.

This objective will be realised by achieving the following specific aims:

- Identify literature reporting on associations between bisphenol A exposure and semen quality in men of reproductive age.
- Identify literature reporting on effects of bisphenol A and semen quality in studies involving rats, mice and other mammalian species.

- Extract data on these associations from relevant studies to allow assessments of the quality of associations, first in a screening and evidence mapping stage, followed by a more exhaustive data extraction for the most relevant studies.
- Assess the internal validity (risk of bias) of relevant individual studies using pre-defined criteria.
- Synthesize the evidence using a quantitative approach, or meta-analysis if appropriate.
- Rate the confidence in the conclusions drawn from these quantitative analyses.

PECO statements are given in Tables 1 and 2.

Table 1. PECO statement for human studies

Question	Is exposure to bisphenol associated with declines in semen quality?
Populations	Men of reproductive age (between 18 and 40 years of age)
Exposures	Bisphenol A, measured as urinary levels in expectant mothers or at time points close to the collection of semen samples in adult men
Comparators	Men not exposed to bisphenol A, or men with bisphenol A levels in lower quartiles
Outcomes	Semen quality, as measured in terms of: • Total sperm count • Sperm concentration • Sperm motility • Sperm morphology • Sperm vitality

Table 2. PECO statement for animal studies

Question	Is exposure to bisphenol associated with declines in semen quality?
Populations	Laboratory mammalian species including rats, mice, rabbits, guinea pigs, dogs and monkeys
Exposures	Bisphenol A by oral gavage, via drinking water or the diet during gestation and postnatal life, when germ cell populations are established (gestational day 7 to postnatal day 8 in mice; gestational day 9 to postnatal day 10 in rats)
Comparators	Animals not exposed to bisphenol A
Outcomes	Semen quality, as measured in terms of: • Total sperm count • Sperm concentration • Sperm motility • Sperm morphology • Sperm vitality

Methods

This protocol was drafted with specific regards to the PRISMA-P (Preferred Reporting Items for Systematic review and Meta-Analysis Protocols) 2015 checklist (Shamseer *et al.*, 2015) and the recommendations for the conduct of systematic reviews in toxicology and environmental health research (COSTER) (Whaley et al. 2020) by using the generic protocol template available through the protocols.io platform (Supplementary Information "SI file 2 - COSTER protocols.io, Whaley 2020).

8. Eligibility criteria

Criteria for the eligibility of epidemiological studies are listed in Table 3 overleaf.

Table 3. Eligibility criteria for human studies

		Inclusion criteria	Exclusion criteria
Populations	Men	Younger than 40 years, older than 18 years	Men with non-descended testes, hypospadias, and chronic diseases such as cancer, varicocele, or other known illnesses impacting on semen quality; men > 40 years; < 18 years
Exposures	Bisphenol A	Bisphenol A exposures measured as urinary concentrations	Bisphenol A in other body fluids or tissues, e.g. plasma or seminal fluid, exposure information derived from questionnaires or job exposure matrices
Comparators	**Exposure contrast** Bisphenol A at lower levels or in reference groups	Sufficient information reported to allow comparison/categorisation of exposures.	Insufficient information reported to allow comparison/categorisation of exposures.
Outcomes	Semen quality	• Total sperm count • Sperm concentration • Sperm motility • Sperm morphology • Sperm vitality	• Sperm DNA damage • Aneuploidies • Measures of in vitro fertilization success • Time to pregnancy
Design		• Case-control studies • Cohort studies • Cross-sectional studies	• Case reports • Reviews

Studies on men with non-descended testes, hypospadias, varicocele, chronic diseases such as cancer or other illnesses known to have a negative impact on semen quality will not be considered. Exposure assessments of bisphenol A other than as urinary concentrations will be excluded as measurements in e.g. plasma or seminal fluids are not deemed sufficiently reliable (matrix interferences) and do not currently permit the estimation of daily intakes that have resulted in these concentrations. by accepted and validated pharmacokinetic models Studies reporting on associations between bisphenol A and DNA damage in sperm, or aneuploidy will also not be considered, as these effects are not related to disruptions of male reproductive health by hormonal factors.

Criteria for the eligibility of experimental studies with laboratory rats or mice are listed in Table 4 overleaf.

Table 4. Eligibility criteria for animal studies

		Inclusion criteria	Exclusion criteria
Populations	Laboratory mammalian species including rats, mice, rabbits, guinea pigs, dogs and monkeys	Mammalian species	Non mammalian test species such as fish or amphibians
Exposures	Bisphenol A perinatally, e.g. at any time from gestational day 7 (mouse) or 9 (rat) to postnatal day 8 (mouse) or 10 (rat)	Administered by gavage, via drinking water or through the diet; at least 2 exposure doses.	Administered subcutaneously or intraperitoneally; only 1 exposure dose group; in juveniles or adults.
Comparators	Animals not exposed to bisphenol A	Control group (same species as exposure group(s))	No control group
Outcomes	Semen quality	• Total sperm count • Sperm concentration • Sperm motility • Sperm morphology • Sperm vitality	• Sperm DNA damage • Aneuploidies • Fertility and fertilization outcomes

9. Information sources

Searches for peer-reviewed articles will be conducted in the following bibliographic databases:

- PubMed
- Web of Science Core Collection
- Scopus
- The full text database ScienceDirect

Although grey literature is not strictly excluded, the search strategy does not include sources specifically targeting the grey literature.

10. Search strategy

The literature searches include epidemiological studies of associations between bisphenol A exposure and declines in semen quality, as well as experimental studies in mammalian species. Due to resource limitations, only reports in English will be considered. No restrictions will be placed on publication date.

A common search strategy was designed for epidemiological and animal experimental studies concurrently on the basis of two strands of search terms combined by the Boolean AND; one for synonyms of bisphenol A another for semen parameters (both lists combined by the Boolean OR).

For the search terms related to the chemical bisphenol A, two lists of search terms were piloted, the first using the full list of chemical synonyms for bisphenol as available through PubChem, the second by using the 'Find chemical synonyms' function of the SWIFTReview Software. We opted for the latter list of synonyms of BPA.

In PubMed, MeSH terms related to 'Sperm' or 'Semen' were used for semen parameters. For other databases, prefix and suffix wild cards were used to detect all terms with roots related to 'sperm',

'semen' and 'semin'. The inclusion of a more general search term related to fertility was later abandoned as it reduced the specificity of the search without any discernible benefits in terms of the sensitivity of the search.

Results of pilot searches for individual search terms and their combinations and the full search strings for all four literature databases are given in Supplementary Information "SI file 3 – Search Strategy".

11. Study records

Data management

Literature and all systematic review processes will be managed and coordinated with the support of the freely available online tool CADIMA established in a close collaboration between the Julius Kühn-Institut and the Collaboration for Environmental Evidence (https://www.cadima.info/index.php/area/evidenceSynthesisDatabase).

Relevance screening

The list of eligibility criteria will be applied to the merged reference list in two stages. In the first stage, only titles and abstracts will be checked for relevance to the study question. AK, OM, ES and AB will carry out the screening. As part of an initial consistency check, 200 studies will first be screened by all team members in parallel. Discrepancies detected during this pilot stage will be reviewed by all team members and the eligibility criteria will be clarified if necessary. Eligibility criteria will be applied to the remainder of the merged reference list by one team member. Clearly irrelevant studies will be excluded. The full text of the resulting list of included references after title/abstract screening will then be examined for inclusion in duplicate, i.e. by two team members independently. The reason for exclusion of studies after assessment of the full text will be recorded.

Multiple reports of the same research (e.g. multiple publications, conference abstracts etc.) will not be excluded but instead the methodological information from each of the reports shall be collated as part of the data extraction process as one unit of evidence.

Data collection process:

A data extraction template in Excel will be used to extract elements relevant for all studies, as reported in selected manuscripts. No data interpretation will take place at this stage.

12. Data items

From human studies we will extract data on bisphenol A exposures suitable for estimating the exposure-response relationship. From animal studies we will compile reported data on bisphenol A doses associated with no observed adverse effect levels (NOAEL) or effect doses associated with predetermined effect magnitudes (benchmark doses) or lowest doses associated with effects (lowest observed adverse effect levels (LOAELs).

The data extraction Excel templates contains the key data extraction elements to summarise study design, experimental model, methodology and results as described in the NTP OHAT 2019 Handbook for conducting a literature-based health assessment (Table 3, p 29) and reproduced overleaf. These were piloted by AB, AK, OM and SE in parallel (Supplementary Information – "SI file 4 – Data extraction pilot - human" and "SI file 5 – Data extraction pilot - animal").

Table 5. Key Data Extraction Elements to Summarise Study Design, Experimental Model, Methodology and Results (Reproduced from OHAT 2019)

HUMAN	
Funding	Funding source(s)
	Reporting of conflict of interest (COI) by authors (*reporting bias)
Subjects	Study population name/description
	Dates of study and sampling time frame
	Geography (country, region, state, etc.)
	Demographics (sex, race/ethnicity, age or lifestage at exposure and at outcome assessment)
	Number of subjects (target, enrolled, n per group in analysis, and participation/follow-up rates) (*missing data bias)
	Inclusion/exclusion criteria/recruitment strategy (*selection bias)
	Description of reference group (*selection bias)
Methods	Study design (e.g., prospective or retrospective cohort, nested case-control study, cross-sectional, population-based case-control study, intervention, case report, etc.)
	Length of follow-up (*information bias)
	Health outcome category, e.g., cardiovascular
	Health outcome, e.g., blood pressure (*reporting bias)
	Diagnostic or methods used to measure health outcome (*information bias)
	Confounders or modifying factors and how considered in analysis (e.g., included in final model, considered for inclusion but determined not needed (*confounding bias)
	Substance name and CAS number
	Exposure assessment (e.g., blood, urine, hair, air, drinking water, job classification, residence, administered treatment in controlled study, etc.) (*information bias)
	Methodological details for exposure assessment (e.g., HPLC-MS/MS, limit of detection) (*information bias)
	Statistical methods (*information bias)
Results	Exposure levels (e.g., mean, median, measures of variance as presented in paper, such as SD, SEM, 75th/90th/95th percentile, minimum/maximum); range of exposure levels, number of exposed cases
	Statistical findings (e.g., adjusted β, standardized mean difference, adjusted odds ratio, standardized mortality ratio, relative risk, etc.) or description of qualitative results. When possible, OHAT will convert measures of effect to a common metric with associated 95% confidence intervals (CI). Most often, measures of effect for continuous data are expressed as mean difference, standardized mean difference, and percent control response. Categorical data are typically expressed as odds ratio, relative risk (RR, also called risk ratio), or β values, depending on what metric is most commonly reported in the included studies and on OHAT's ability to obtain information for effect conversions from the study or through author query.
	If not presented in the study, statistical power can be assessed during data extraction using an approach that can detect a 10% to 20% change from response by control or referent group for continuous data, or a relative risk or odds ratio of 1.5 to 2 for categorical data, using the prevalence of exposure or prevalence of outcome in the control or referent group to determine sample size. For categorical data where the sample sizes of exposed and control or referent groups differ, the sample size of the exposed group will be used to determine the relative power category. Recommended sample sizes to achieve 80% power for a given effect size, i.e., 10% or 20% change from control, will be compared to sample sizes used in the study to categorize statistical power. Studies will be considered adequately powered when sample size for 80% power is met.
	Observations on dose response (e.g., trend analysis, description of whether dose-response shape appears to be monotonic, non-monotonic)

ANIMAL	
Funding	Funding source(s)
	Reporting of COI by authors (*reporting bias)
Animal Model	Sex
	Species
	Strain
	Source of animals
	Age or lifestage at start of dosing and at health outcome assessment
	Diet and husbandry information (e.g., diet name/source)
Treatment	Chemical name and CAS number
	Source of chemical
	Purity of chemical (*information bias)
	Dose levels or concentration (as presented and converted to mg/kg bw/d when possible)
	Other dose-related details, such as whether administered dose level was verified by measurement, information on internal dosimetry (*information bias)
	Vehicle used for exposed animals
	Route of administration (e.g., oral, inhalation, dermal, injection)
	Duration and frequency of dosing (e.g., hours, days, weeks when administration was ended, days per week)
Methods	Study design (e.g., single treatment, acute, subchronic (e.g., 90 days in a rodent), chronic, multigenerational, developmental, other)
	Guideline compliance (i.e., use of EPA, OECD, NTP or another guideline for study design, conducted under GLP guideline conditions, non-GLP but consistent with guideline study, non-guideline peer-reviewed publication)
	Number of animals per group (and dams per group in developmental studies) (*missing data bias)
	Randomization procedure, allocation concealment, blinding during outcome assessment (*selection bias)
	Method to control for litter effects in developmental studies (*information bias)
	Use of negative controls and whether controls were untreated, vehicle-treated, or both
	Report on data from positive controls – was expected response observed? (*information bias)
	Endpoint health category (e.g., reproductive)
	Endpoint (e.g., infertility)
	Diagnostic or method to measure endpoint (*information bias)
	Statistical methods (*information bias)
Results	Measures of effect at each dose or concentration level (e.g., mean, median, frequency, and measures of precision or variance) or description of qualitative results. When possible, OHAT will convert measures of effect to a common metric with associated 95% confidence intervals (CI). Most often, measures of effect for continuous data will be expressed as mean difference, standardized mean difference, and percent control response. Categorical data will be expressed as relative risk (RR, also called risk ratio).
	No Observed Effect Level (NOEL), Lowest Observed Effect Level (LOEL), benchmark dose (BMD) analysis, statistical significance of other dose levels, or other estimates of effect presented in paper. **Note:** The NOEL and LOEL are highly influenced by study design, do not give any quantitative information about the relationship between dose and response, and can be subject to author's interpretation (e.g., a statistically significant effect may not be considered biologically important). Also, a NOEL does not necessarily mean zero response. Ideally, the response rate at specific dose levels is used as the primary measure to characterize the response.
	If not presented in the study, statistical power can be assessed during data extraction using an approach that assesses the ability to detect a 10% to 20% change from control group's response for continuous data, or a relative risk or odds ratio of 1.5 to 2 for categorical data, using the outcome frequency in the control group to determine sample size. Recommended sample sizes to achieve 80% power for a given effect size, i.e., 10% or 20% change from control, will be compared to sample sizes used in the study to categorize statistical power. Studies will be considered adequately powered when sample size for 80% power is met.
	Observations on dose response (e.g., trend analysis, description of whether dose-response shape appears to be monotonic, non-monotonic)
	Data on internal concentration, toxicokinetics, or toxicodynamics (when reported)
Other	Documentation of author queries, use of digital rulers to estimate data values from figures, exposure unit, and statistical result conversions, etc.

As a result of the piloting of the template, the following items were added to the data extraction template in addition to the elements listed above for animal studies:

- Was the diet soy-free? (yes, no)
- Was consideration given to background contamination with bisphenol A, e.g. by avoiding polycarbonate plastic caging?
- Were positive controls for declines in semen quality included and did they show activity?

Revised data extraction templates can be found in Supplementary Information – "SI file 6 – Final data extraction template – human" and "SI file 7 – Revised data extraction template – animals").

The data extraction workload will be distributed between at least two project team members with some overlap to allow the evaluation of inter-rater reliability as a SR quality control measure.

13. Outcomes

The primary outcome will be estimates of bisphenol A exposures not associated with adverse effects on semen quality parameters, from both epidemiological studies and experimental studies with laboratory animals. For human epidemiological studies, the outcomes will be operationalised in terms of exposures associated with semen quality parameters not significantly different from referents. If most epidemiological studies are cross-sectional, the estimation of bisphenol A exposures not associated with adverse effects on semen parameters will rely on animal studies, due to difficulties with ruling out reverse causation in cross-sectional studies.

For animal studies, the outcomes will be operationalised in terms of NOAELs, LOAELs or benchmark doses (lower confidence limits). Studies that cover the period when germ cell populations are established during development (gestational day 7 to post-natal day 8 in the mouse; gestational day 9 to post-natal day 10 in the rat) will be prioritised as most relevant.

14. Study evaluation, risk of bias of individual studies

We will assess the internal validity (risk of bias) of individual studies by using separate criteria and considerations for human epidemiological and for animal studies.

Human epidemiological studies

Main concerns are risk of bias (understood as factors that affect the magnitude or direction of effect) and insensitivity (factors that limit the ability of a study to detect an effect that is there). We will examine studies of associations between BPA and semen quality using the procedures detailed by Radke *et al.* (2018). Accordingly, we will conduct evaluations of the following domains:

- Exposure measurement
- Outcome measurement
- Participant selection
- Confounding
- Analysis

We will judge each study for each outcome in each evaluation domain in terms of Good, Adequate, Poor, or Critically Deficient, following the criteria detailed in Radke et al. (2018). These ratings assess the utility of a study for purposes of hazard identification, as follows:

- **Good** represents a judgment that there was appropriate study conduct relating to the domain, and any minor deficiencies that were noted would not be expected to influence the study results.

- **Adequate** indicates a judgment that there were experimental limitations relating to the domain, but that those limitations are not likely to be severe or to have a substantive impact on the results.

- **Poor** denotes identified biases or deficiencies that are interpreted as likely to have had a substantial impact on the results or that prevent reliable interpretation of the study findings. *Not reported* indicates that the information necessary to evaluate the domain question was not available in the study. Generally, this term carries the same functional interpretation as Poor for the purposes of the study confidence classification.

- **Critically Deficient** reflects a judgment that the experimental conduct relating to the domain question introduced a flaw so serious that the study should not be used without exceptional justification (e.g., it is the only study of its kind and may highlight possible research gaps). This judgment should only be used if there is an interpretation that the limitation(s) would be the primary driver of any observed effect(s), or if it makes the study uninterpretable.

Details of the protocol used for the assessment of epidemiological studies are given in **Appendix 1**.

As in Radke *et al.* (2018), we will combine the identified strengths and limitations in each domain to reach a **study confidence classification** of High, Medium, Low, or Uninformative. The classifications, which reflect a consensus judgment between reviewers, are as follows:

- **High Confidence**: No notable deficiencies or concerns were identified; the potential for bias is unlikely or minimal, and the study used sensitive methodology. In general, although classifications are not decided by "scoring," high confidence studies would reflect judgments of good across all or most evaluation domains.

- **Medium Confidence**: Possible deficiencies or concerns were noted, but the limitations are unlikely to be of a substantive degree. Generally, medium confidence studies will include adequate or good judgments across most domains, with the impact of any identified limitation not being judged as severe.

- **Low Confidence**: Deficiencies or concerns were noted, and the potential for substantive bias or inadequate sensitivity could have a significant impact on the study results or their interpretation. Typically, low confidence studies would have a poor evaluation for one or more domains (unless the impact of the particular limitations on the results is judged as unlikely to be severe).

- **Uninformative**: Serious flaw(s) make the study results unusable for informing hazard identification. Studies with critical deficiencies in any evaluation domain will almost always be classified as uninformative (see explanation above). Studies with multiple poor judgments across domains may also be considered uninformative, particularly when there is a robust database of studies on the outcome(s) of interest or when the impact of the limitations is viewed as severe.

The risk of bias assessment tool was piloted by AK, SE and OVM and results of this piloting exercise can be found in Supplementary Information – "SI file 8 – Risk-of-bias pilot – human". The risk-of-bias tool, together with instructions how to rate each element of the protocol in terms of the risk categories can be found in Supplementary Information – "SI file 10 – Final risk-of-bias tool – human".

Experimental studies with animals

To appraise the **internal validity of studies**, we will use the internal validity appraisal protocol (risk of bias assessment) for BPA detailed in EFSA (2017, 2019) which was developed along the lines of the NTP OHAT Risk of Bias Tool, as described in the NTP OHAT 2019 Handbook for conducting a literature-based health assessment, p 33.

EFSA (2019) defined **key elements** for their BPA assessment, as follows:

- Can we be confident in the exposure characterisation, in terms of purity of test chemical, its consistent administration and absence of contaminations?
- Can we be confident in the outcome assessment, in terms of blinding of the assessors?
- Was the number of animals per dose group sufficient?

To assess specific quality issues related to studies of BPA and semen quality, we will use the following **additional key elements**:

- Use of polycarbonate-free caging
- Use of phytoestrogen-free chow
- Demonstration that a positive control was effective

It is known that BPA can leach from polycarbonate caging and obscure the effects of administered BPA. Similarly, the use of phytoestrogen-containing chow will introduce hormonal disturbances which may mask the effects of BPA on semen quality. Inclusion of a positive control (often ethinylestradiol or diethylstilbestrol) demonstrates the proficiency of the investigators to detect changes in semen quality and shows that the experimental system is sufficiently sensitive.

We will score each element using the NTP OHAT scores, as follows:

++ definitely low risk

+ probably low risk

~ probably high risk

~~ definitely high risk

A study that fails a key element will not be evaluated further. We will adopt the system in EFSA (2019) to rate each study in terms of three Tiers, with Tier 1 signifying the highest confidence, as follows:

Tier 1 (high confidence)

All three EFSA key elements and all three additional key elements are scored + or ++ **AND**

No more than 1 question not addressing these key elements (EFSA 2019, Table 2, p 8) is scored ~ or ~~

Tier 2 (medium confidence)

All combinations not covered in Tier 1 or Tier 3

Tier 3 (low confidence)

Any one of the three EFSA key elements and the additional key elements is scored ~ or ~~ **OR**

More than 50% of the questions not addressing these key elements (EFSA 2019, Table 2, p 8) are scored ~ or ~~

The risk of bias assessment tool was piloted by AK, SE, OVM and AB and results of this piloting exercise can be found in Supplementary Information – "SI file 9 – Risk-of-bias pilot – animal". The risk-of-bias tool, together with instructions how to rate each element of the protocol in terms of the risk categories can be found in Supplementary Information – "SI file 11 – Final risk-of-bias tool – animal".

Regarding the **external validity** of the endpoint measured in animal studies for human health, we will assume that measures of declining semen quality in animals are directly relevant to human health.

15. Data synthesis

A systematic narrative synthesis will be provided with information presented in the text and tables to summarise and explain the characteristics and findings of the included studies. This will only consider human epidemiological studies rated as high or medium confidence, and experimental animal studies rated as high confidence (Tier 1). Summary tables with characteristics of eligible studies will follow the same template as that described in EFSA (2019) and reproduced below.

Table 6. Study summary table – Human data (reproduced from EFSA 2019)

Reference	Funding source	Endpoint	Study design	Subjects	Exposure	Results	Internal validity

Table 7. Study summary table – Experimental animal data (reproduced from EFSA 2019)

Reference	Funding source	Endpoint	Species/ strain	No animals/ group	Treatment (Route, period, duration)	Dose(s) (mg/ kg bw per day)	Human equivalent dose	Results	Internal validity	External validity

To enable quantitative comparisons between bisphenol A exposures in human studies and experimental studies with animals, we will convert urinary bisphenol A levels into daily intakes for humans by employing the toxicokinetic model detailed in Koch et al. (2012).

16. Meta-biases

It is not planned to analyse meta-biases.

17. Confidence in cumulative evidence, evidence synthesis

The evidence synthesis will be guided by two separate questions, one of a qualitative, the other of a quantitative nature:

Qualitative – how reliable is the evidence linking BPA with declines in semen quality, from both human and animal studies? This will require methods for **weighing evidence** from different lines of evidence (here: human and animal studies) (EFSA 2017a).

Quantitative – can we identify ranges of BPA exposures not associated with declines in semen quality? This will require methods for **integrating evidence** (EFSA 2017a). The integration will be achieved by comparison of each line of evidence to derive a NOAEL or benchmark dose. Quantitative comparisons will then determine a point of departure for the most sensitive species, taking account of biological relevance, reliability and consistency of data.

In the **qualitative** evaluation, we will synthesise the human and animal study evidence separately. The quantitative synthesis will proceed by a comparison of exposure estimates from human and animal studies.

Following the principles developed in the EFSA (2017a) Guidance on the use of weight of evidence in scientific assessments, as implemented in the EFSA Bisphenol A hazard assessment protocol (EFSA 2017b), in each **qualitative** synthesis (human studies and animal studies) we will consider aspects of an association that may suggest causation, according to the Bradford Hill criteria: consistency, exposure–response relationship, strength of association, temporal relationship, biological plausibility, and coherence.

We will distinguish between *conflicting evidence* (unexplained positive and negative results in similarly exposed human populations or in similar animal models) and *differing results* (mixed results attributable to differences between human populations, animal models, or exposure conditions).

Human studies: qualitative evidence synthesis

We will synthesise human studies by adopting the framework developed by Radke *et al.* (2018), which is compatible with the OHAT (2019) scheme, as follows:

Robust evidence from human studies

A set of *high* or *medium* confidence independent studies reporting an association between BPA exposure and declines in semen quality, with reasonable confidence that alternative explanations, including chance, bias, and confounding, can be ruled out across studies. The set of studies is primarily consistent, with reasonable explanations when results differ; and an exposure response gradient is demonstrated. Additional supporting evidence, such as associations with biologically related endpoints in human studies (coherence) or large estimates of risk or severity of the response, may increase confidence but are not required.

Mechanistic evidence from exposed humans or human cells, if available, may add support informing considerations such as exposure response, temporality, coherence, and MOA, thus raising the level of certainty to *robust* for a set of studies that otherwise would be described as *moderate*.

Moderate evidence from human studies

A smaller number of studies (at least one *high* or *medium confidence* study with supporting evidence), or with some heterogeneous results, that do not reach the degree of confidence required for *robust.* For multiple studies, there is primarily consistent evidence of an association, but there may be some uncertainty due to potential chance, bias or confounding.

If only a single study is available, there is a large magnitude or severity of the effect, or a dose-response gradient, or other supporting evidence, and there are no serious residual methodological uncertainties. Supporting evidence could include associations with related endpoints, including mechanistic evidence from exposed humans or human cells, if available, based on considerations such as exposure response, temporality, coherence, and MOA.

Slight evidence from human studies

One or more studies reporting an association between exposure and the health outcome, where considerable uncertainty exists. In general, the evidence is limited to a set of consistent *low* confidence studies, or higher confidence studies with unexplained heterogeneity. Supporting coherent evidence is sparse. Strong biological support from mechanistic evidence in exposed humans or human cells may also be independently interpreted as s*light*. This also includes scenarios where there are serious residual uncertainties across studies (these uncertainties typically relate to exposure characterization or outcome ascertainment, including temporality) in a set of largely consistent medium or high confidence studies. This category serves primarily to encourage additional study where evidence does exist that might provide some support for an association, but for which the evidence does not reach the degree of confidence required for *moderate*.

Indeterminate evidence in human studies

No studies available in humans or situations when the evidence is highly inconsistent and primarily of *low* confidence. In addition, this may include situations where higher confidence studies exist, but unexplained heterogeneity exists and there are additional outstanding concerns such as effect estimates of low magnitude, uninterpretable patterns with respect to exposure levels, or uncertainties or methodological limitations that result in an inability to discern effects from exposure. A set of largely null studies could be concluded to be *indeterminate* if the evidence does not reach the level required for *compelling evidence of no effect*.

Compelling evidence of a lack of association

Several *high* confidence studies showing null results (for example, an odds ratio of 1.0), ruling out alternative explanations including chance, bias, and confounding with reasonable confidence. Each of the studies should have used an optimal outcome and exposure assessment and adequate sample size (specifically for higher exposure groups and for susceptible populations). The set as a whole should include the full range of levels of exposures that human beings are known to encounter, an evaluation of an exposure response gradient, and an examination of at-risk populations and lifestages.

Animal studies: qualitative evidence synthesis

Animal studies will be synthesised according to the framework used by Radke *et al.* (2018), modified in line with EFSA (2019) as follows:

Robust evidence from animal studies

The set of experiments rated as falling into **Tier 1** (see above, 14. Experimental studies with animals) includes consistent findings of adverse or toxicologically significant effects across multiple laboratories and species, where the experiments can reasonably rule out the potential for nonspecific effects (e.g., resulting from toxicity) to have resulted in the findings. Any inconsistent evidence (evidence that cannot be reasonably explained by the respective study design or differences in animal model) is from a set of experiments of lower confidence (**Tier 2** or **Tier 3**). At least two of the following additional factors in the set of experiments support a causal association: coherent effects across multiple related endpoints (may include mechanistic endpoints); an unusual magnitude of effect, or severity; a strong dose response relationship; consistent observations across animal strains and species. Alternatively, mechanistic data in animals or animal cells that address the above considerations or that provide experimental support for a MOA that defines a causal relationship with reasonable confidence may

raise the level of certainty to *robust* for evidence that otherwise would be described as *moderate* or, exceptionally, *slight, or indeterminate*.

Moderate evidence from animal studies

A set of evidence that does not reach the degree of certainty required for *robust*, but which includes at least one **Tier 1** confidence study and information strengthening the likelihood of a causal association. Although the results are largely consistent, notable uncertainties remain. However, while inconsistent evidence and/or evidence indicating nonspecific effects (e.g., toxicity) may exist, it is not sufficient to reduce or discount the level of concern regarding the positive findings from the supportive experiments or it is from a set of experiments of lower confidence. The set of experiments supporting the effect provide additional information supporting a causal association, such as consistent effects across laboratories or species; an unusual magnitude of effect, or severity; a strong dose response relationship; and/or consistent observations across animal strains. Mechanistic data in animals or animal cells that address the above considerations or that provide information supporting an association between exposure and effect with reasonable confidence may raise the level of certainty to *moderate* for evidence that otherwise would be described as *slight*.

Slight evidence from animal studies

Scenarios in which there is a signal of a possible effect, but the evidence is conflicting or weak. Most commonly, this includes situations where only *low* confidence experiments are available and supporting coherent evidence is sparse. It also applies when one *medium* or *high* confidence experiment is available without additional information strengthening the likelihood of a causal association (e.g., corroboration within the same study or from other studies). Lastly, this includes scenarios in which there is evidence that would typically be characterized as *moderate*, but inconsistent evidence (evidence that cannot be reasonably explained by the respective study design or differences in animal model) from a set of experiments of higher confidence (may include mechanistic evidence) exists. Strong biological support from mechanistic studies in exposed animals or animal cells may also be independently interpreted as *slight*. Notably, to encourage additional research, it is important to describe situations for which evidence does exist that might provide some support for an association but is insufficient for a conclusion of *moderate*.

Indeterminate evidence from animal studies

No animal studies are available, the evidence is highly inconsistent and primarily of *low* confidence. In addition, this may include situations where higher confidence studies exist, but there is unexplained heterogeneity and additional concerns such as small effect sizes (given what is known about the endpoint) or a lack of dose-dependence. A set of largely null studies could be concluded to be *indeterminate* if the evidence does not reach the level required for *compelling evidence of no effect*.

Compelling evidence of no effect

A set of *high* confidence experiments that demonstrate a lack of biologically significant effects across multiple species, both sexes, and a broad range of exposure levels. The data are compelling in that the experiments have examined the range of scenarios across which health effects in animals could be observed, and an alternative explanation (e.g., inadequately controlled features of the studies' experimental designs; inadequate sample sizes) for the observed lack of effects is not available. The experiments were designed to specifically test for effects of interest, including suitable exposure

timing and duration, post-exposure latency, and endpoint evaluation procedures, and to address potentially susceptible populations and lifestages.

Quantitative synthesis: integrating evidence

To achieve integration of evidence to arrive at quantitative assessment allowing us to derive a reference dose with respect to endpoint semen quality we will follow the procedure sketched out in EFSA (2017). Briefly, qualitative comparisons will be made for each line of evidence (per animal species, and human) where it is possible to derive a point of departure (NOAEL, LOAEL or benchmark dose). If necessary, NOAELs will be extrapolated from LOAELs by using standard assessment factors (AF = 3). These comparisons will be based on high quality studies (high or medium confidence human studies, Tier 1 animal studies). We will then select the study (or studies) for the most sensitive species and derive a reference dose.

References

Bliatka D *et al.* (2020) The effects of postnatal exposure of endocrine disruptors on testicular function: a systematic review and a meta-analysis. *Hormones-International Journal of Endocrinology and Metabolism* 19 (2): 157–169. doi: 10.1007/s42000-019-0170-0

Bonde JP *et al.* (2017) The epidemiologic evidence linking prenatal and postnatal exposure to endocrine disrupting chemicals with male reproductive disorders: a systematic review and meta-analysis. *Human Reproduction Update* 23(1): 104–125. doi: 10.1093/humupd/dmw036

EFSA Scientific Committee (2017a) Scientific Opinion on the guidance on the use of the weight of evidence approach in scientific assessments. *EFSA Journal* 15(8):4971, 69 pp. doi: 10.2903/j.efsa.2017.4971

EFSA (European Food Safety Authority) *et al.* (2017b) Bisphenol A (BPA) hazard assessment protocol. *EFSA supporting publication* 2017:EN-1354. 76 pp. doi:10.2903/sp.efsa.2017.EN-1354

EFSA (European Food Safety Authority) *et al.* (2019) Testing the study appraisal methodology from the 2017 Bisphenol A (BPA) hazard assessment protocol. EFSA supporting publication 2019:EN-1732. 100 pp. doi:10.2903/sp.efsa.2019.EN-1732

Evans RM *et al.* (2016) Should the scope of human mixture risk assessment span legislative/regulatory silos for chemicals?, *The Science of the total environment*, 543(Pt A): 757–764. doi: 10.1016/j.scitotenv.2015.10.162.

Kortenkamp A (2020) Which chemicals should be grouped together for mixture risk assessments of male reproductive disorders?, *Molecular and Cellular Endocrinology* 449:110581 doi: 10.1016/j.mce.2019.110581

NTP OHAT (2019) Handbook for conducting a literature-based health assessment using OHAT approach for systematic review and evidence integration. Available from https://ntp.niehs.nih.gov/ntp/ohat/pubs/handbookmarch2019_508.pdf

Radke EG *et al.* (2018) Phthalate exposure and male reproductive outcomes: A systematic review of the human epidemiological evidence. *Environ International* 121: 764-793

Shamseer L *et al.* (2015) Preferred reporting items for systematic review and meta-analysis protocols (prisma-p) 2015: Elaboration and explanation. *BMJ (Online)*. doi: 10.1136/bmj.g7647.

Vandenberg L *et al.* (2016) A proposed framework for the systematic review and integrated assessement (SYRINA) of endocrine disrupting chemicals. *Environmental Health* 15: 74, doi: 10.1186/s12940-016-0156-6

Whaley P *et al.* (2020a) Recommendations for the conduct of systematic reviews in toxicology and

environmental health research (COSTER). *Environment International* 143: 105926. doi: 10.1016/j.envint.2020.105926

Whaley P (2020b) Generic Protocol for Environmental Health Systematic Reviews Based on COSTER Recommendations. *protocols.io* https://dx.doi.org/10.17504/protocols.io.biktkcwn

Supplementary Information

SI file 1 – ICMJE COI Disclosure Forms

SI file 2 – COSTER protocols.io

SI file 3 - Search strategy

SI file 4 – Data extraction pilot – human

SI file 5 – Data extraction pilot - animal

SI file 6 – Final data extraction – human

Si file 7 – Revised data extraction - animal

SI file 8 – Risk-of-bias pilot – human

SI file 9 – Risk-of-bias pilot - animal

SI file 10 – Final risk-of-bias – human

SI file 11 – Final risk-of-bias – animal

Appendix 1. Risk-of-bias evaluation of epidemiological studies (adapted from Radke et al. 2018)

For the **exposure measurement** domain, we will use the following guiding questions, reproduced from Radke *et al.* (2018):

- Does the exposure measure capture the variability in exposure among the participants, considering intensity, frequency, and duration of exposure?
- Does the exposure measure reflect a relevant time window? If not, can the relationship between measures in this time and the relevant time window be estimated reliably?
- Was the exposure measurement likely to be affected by a knowledge of the outcome?
- Was the exposure measurement likely to be affected by the presence of the outcome (i.e., reverse causality)?

For case-control studies of occupational exposures:

- Is exposure based on a comprehensive job history describing tasks, setting, time period, and use of specific materials?

For biomarkers of exposure, general population:

- Is a standard chemical analytical assay used? What are the intra- and inter-assay coefficients of variation? Is the assay likely to be affected by contamination? Are values less than the limit of detection dealt with adequately?
- What exposure time period is reflected by the biomarker? If the half-life is short, what is the correlation between serial measurements of exposure?

For **rating the exposure measures**, we will use the following descriptors (adapted from Radke et al. 2018):

Good

- Three or more urine samples within etiologically relevant time period (±1−3 mo) and analysis includes a summed variable, or similar results seen with each of the metabolites from the parent compound **and**
- High proportion (>50%) above the LOD **and**
- Discussion of laboratory QC procedures or no discussion of laboratory QC procedures but analysis by an experienced laboratory

Adequate

- Two or more urine samples within etiologically relevant period (±1−3 mo) and analysis includes a summed variable, or similar results seen with each of the metabolites from the parent compound **and**
- High proportion (>50%) above the LOD

Poor

- One urine sample within etiologically relevant period (±1−3 mo) and high proportion (>50%) above the LOD

Critically deficient

- Measures in urine likely to be affected by differential misclassification (e.g., after disease diagnosis) **or**
- Low proportion (<50%) above the LOD

In addition, we will rate the **timing of exposure measurements,** by taking account of the specifics of spermatogenesis: All the cell divisions that produce mature sperm, occur after puberty. A man's semen quality is influenced by the *in utero* environment and by events in childhood, before puberty. In adulthood, spermatogenesis takes 74 days to complete and requires an additional 12 days of maturation as sperm migrate through the epididymis. For short-lived chemicals such as BPA, exposure measurements during the critical period of spermatogenesis *in utero*, before puberty and at least 86 days before taking semen samples would therefore be **ideal**. If the study question relates to adult life, then measurements during the period of spermatogenesis in adulthood (86 days before semen sampling) are **good**, for assessing prenatal exposures, measures corresponding to pregnancy are **good**. Measurements occurring after a man's fertility problem was recognised are **critically deficient**.

For the **outcome measurement** domain, semen quality measurements, we will use the following guiding questions (from Radke *et al.* 2018):

For all studies:

- Is outcome ascertainment likely to be affected by knowledge of, or presence of, exposure (e.g., consider access to health care, if based on self-reported history of diagnosis)?

For case-control studies:

- Is the comparison group without the outcome (e.g., controls in a case-control study) based on objective criteria with little or no likelihood of inclusion of people with the disease?

For rating the semen quality measures, we will use the following criteria (Radke *et al.* 2018):

Good

- One or more samples collected with instructions regarding abstinence time; abstinence data collected and discussed and/or addressed in analysis.
- Analyzed within two hours of collection (if collected off site, should be kept at body temperature until transfer to the laboratory)
- Analyzed by a single laboratory according to WHO standards with analysis and quality control procedures described in enough detail to provide assurance of quality standards.
- Manual ascertainment of morphology (CASA acceptable for other parameters); presents analyses of concentration (or counts), motility, and morphology; definition/criteria for motility and morphology are specified
- If all three parameters are not presented, an explanation is provided (e.g. manual ascertainment of morphology was not performed).

Adequate

Same as *Good*, except:

- Analysis and quality control procedures are mentioned, but described with limited detail **or**

- Sample collected at home and shipped overnight for analysis and otherwise meets Adequate criteria (should be considered critically deficient for motility) **or**
- Use of multiple labs for sample analyses, but documented use of the same protocol.

Poor

- One sample collected without consideration of abstinence time in the analysis (unless variability in abstinence length is < 2 days) or without mention of abstinence time **or**
- Analysis more than two hours after collection (issue for motility) **or**
- Analysis by multiple labs without documented use of the same protocol **or**
- Analysis methods and quality control not discussed **or**
- Does not present analysis of concentration (or counts), motility, **and** morphology, and does not provide explanation.

Critically deficient

- Use of local clinically oriented labs without discussion of procedures and quality control.

For the **participant selection** domain, we will use the following guiding questions:

For longitudinal cohorts:

- Did participants volunteer for the cohort based on knowledge of exposure and/or preclinical disease symptoms? Was entry into the cohort or continuation in the cohort related to exposure and outcome?

For occupational cohorts:

- Did entry into the cohort begin with the start of the exposure?
- Was follow-up or outcome assessment incomplete, and if so, was follow-up related to both exposure and outcome status?
- Could exposure produce symptoms that would result in a change in work assignment/work status ("healthy worker survivor effect")?

For case-control studies:

- Were controls representative of population and time periods from which cases were drawn?
- Are hospital controls selected from a group whose reason for admission is independent of exposure?
- Could recruitment strategies, eligibility criteria, or participation rates result in differential participation relating to both disease and exposure?

For population-based surveys:

- Was recruitment based on advertisement to people with knowledge of exposure, outcome, and hypothesis?

For rating the participant selection, we will use the following specific criteria for semen quality studies (Radke *et al.* 2018):

Good

Selection at setting other than infertility clinic (e.g. population-based, occupational) **and all of the following**:

- Minimal concern for selection bias based on description of recruitment process.
- Exclusion and inclusion criteria specified and would not induce bias.
- Participation rate is reported at all steps of study (e.g., initial enrollment, follow-up, selection into analysis sample). If participation rate is not high, there is appropriate rationale for why it is unlikely to be related to exposure (e.g., comparison between participants and nonparticipants or other available information indicates differential selection is not likely.
- Minimal concern that selection of comparison population introduced selection bias.

Adequate

Selection at any setting (including infertility clinic) and **all of the following**:

- Enough of a description of the recruitment process to be comfortable that there is no serious risk of bias.
- Inclusion and exclusion criteria specified and would not induce bias.
- Participation rate is incompletely reported but available information indicates participation is unlikely to be related to exposure.

Poor

- Little information on recruitment process, selection strategy, sampling framework, and/or participation **or**
- Aspects of the recruitment process, selection strategy, sampling framework, or participation raise the potential for bias (e.g., healthy worker effect, survivor bias) **or**
- Study is limited to men with normal or only slight oligozoospermia (i.e. excludes men with moderate to severe oligozoospermia). Exclusion of only men with azoospermia does not meet this criterion.

Critically deficient

- Population selected in such a way that selection bias is likely for example, if cases taken from clinic where exposed workers are treated, or from a high-exposure geographic area, while controls taken from a different clinic where exposure would be lower. This may include clinic or center-based sample of volunteers with known male fertility problems, with comparison to people who have not gone through a similar recruitment-selection process.

For the **confounding domain**, we will use the following guiding questions:

Is confounding adequately addressed by considerations in…

- … participant selection (matching or restriction)?
- … accurate information on potential confounders, and statistical adjustment procedures?
- … lack of association between confounder and outcome, or confounder and exposure in the study?

- … information from other sources?

Is the assessment of confounders based on a thoughtful review of published literature, potential relationships (e.g., as can be gained through directed acyclic graphing), minimizing potential overcontrol (e.g., inclusion of a variable on the pathway between exposure and outcome)?

Key confounders to be considered in semen quality studies are:

- Age
- Abstinence time
- Smoking history
- Body mass index
- Chronic disease status (e.g., diabetes, kidney disease).
- Possible: Alcohol use and stress also could be considered but are not as well established as risk factors compared to the variables listed above.

Following Radke et al. (2018) we will **rate** the confounding domain as follows:

Good

Contains all of the following:

- Conveys thoughtful discussion of strategy for identifying confounders including analysis of different functional forms of key covariates, if warranted. This may include justification for inclusion or exclusion of variables (based on a priori biological considerations, statistical analysis or results in the published literature) with the recognition that not all "risk factors" are confounders;
- Inclusion in model not based solely on statistical significance criteria (e.g., $p < 0.05$ from stepwise regression)
- Does not include variables on the causal pathway (intermediaries) in the model
- Shows progression of adjustment with different variables, presents other data relevant to potential for confounding (e.g., distribution of variables by exposure category) **or** discusses likelihood that confounding or residual confounding could explain the magnitude of observed effect;
- Descriptive information for relevant population characteristics/potential confounders presented (with amount of missing data noted);

Adequate

Contains **one** of these:

- Conveys some discussion of strategy for identifying confounders;

or

- Shows progression of adjustment with different variables, presents other data relevant to potential for confounding (e.g., distribution of variables by exposure category) **or** discusses likelihood that confounding or residual confounding could explain the magnitude of observed effect;

And all of the following:

- Inclusion in model not based solely on statistical significance criteria (e.g., $p < 0.05$ from stepwise regression)

- Does not include variables on the causal pathway (intermediaries) in the model
- Descriptive information for relevant population characteristics/potential confounders presented

Poor

- Strategy of evaluating confounding is unclear or is not recommended (e.g., based on statistical significance criteria only); **or**
- Descriptive information on population characteristics or potential confounders not presented; **or**
- Some residual confounding is likely, given the observed effect and likely measurement misclassification for a confounder present in the study.

Critically deficient

- Established intermediary(ies) included as confounder(s); **or**
- Confounding is present and not accounted for, indicating a strong bias (i.e., key variable is associated with the outcome and exposure in study and was not accounted for in the analysis), and lack of consideration of the confounder could explain the results observed with the exposure.

For the **analysis domain**, we will use the following guiding questions:

- Are missing outcome, exposure, and covariate data recognized, and if necessary, accounted for in the analysis?
- Does the analysis appropriately consider variable distributions and modeling assumptions?
- Does the analysis appropriately consider subgroups of interest (e.g., based on variability in exposure level or duration, or susceptibility)?
- Is an appropriate analysis used for the study design?
- Is effect modification considered, based on considerations developed a priori?
- Does the study include additional analyses addressing potential biases or limitations (i.e., sensitivity analyses)?

We will adopt the criteria by Radke *et al.* (2018) to rate the analysis domain:

Good

- Analyzes outcomes as either continuous or categorical (with justified cutpoints)
- Quantitative results presented (effect estimates and confidence limits or variability in estimates) (i.e., not presented only as a *p*-value or "significant"/"not significant")
- Descriptive information about **outcome and exposure** provided (where applicable).
  - Amount of missing data noted and addressed appropriately (discussion of selection issues—missing at random versus differential—i.e., related to exposure and outcome; if amount of missing data is large, multiple imputations or examinations by sensitivity analyses are necessary)
  - Where applicable, **for exposure**, includes LOD (and percentage less than LOD) and discussion of:
    - Choice of cut-points (if analyzed categorically)

  - Decision to use log transformation; interpretation of log-transformed variables

- Includes analyses that address robustness of findings, e.g., examination of shape of exposure-response (explicit consideration of nonlinear possibilities; quadratic, spline, or threshold/ceiling effects included, when feasible); relevant sensitivity analyses; effect modification examined based only on *a priori* rationale with sufficient numbers
- No deficiencies in analysis evident. Discussion of some details may be absent (e.g., examination of outliers).

Adequate

Same as Good, except:

- Descriptive information about **exposure** provided (where applicable), but may be incomplete; might not have discussed missing data, or cut-points, or shape of distribution **or**
- Includes analyses that address robustness of findings (examples in Very Good), but some important analyses are not performed

Poor

- Descriptive information about **exposure levels not provided** (where applicable); **or**
- Effect estimate and p-value presented, without standard error or confidence interval which are impossible to estimate **or**
- Results presented only as statistically “significant”/”not significant”;

Critically deficient

- Results of analyses of effect modification examination without clear a priori rationale and without providing main/principal effects (e.g., presentation only of statistically significant interactions that were not hypothesis driven) **or**
- Analysis methods are not appropriate for design or data of the study.