Published in final edited form as:
Environ Int. 2020 August ; 141: 105736. doi:10.1016/j.envint.2020.105736.

A novel study evaluation strategy in the systematic review of animal toxicology studies for human health assessments of environmental chemicals

Laura Dishaw[a,*], **Erin Yost**[a], **Xabier Arzuaga**[b], **April Luke**[c,1], **Andrew Kraft**[b], **Teneille Walker**[b,1], **Kris Thayer**[a]

[a]US EPA, Center for Public Health and Environmental Assessment, Research Triangle Park, NC, United States

[b]US EPA, Center for Public Health and Environmental Assessment, Washington, DC, United States

[c]US EPA, Office of Emergency Management, Washington, DC, United States

Abstract

A key aspect of the systematic review process is study evaluation to understand the strengths and weaknesses of individual studies included in the review. The present manuscript describes the process currently being used by the Environmental Protection Agency's (EPA) Integrated Risk Information System (IRIS) Program to evaluate animal toxicity studies, illustrated by application to the recent systematic reviews of two phthalates: diisobutyl phthalate (DIBP) and diethyl phthalate (DEP). The IRIS Program uses a domain-based approach that was developed after careful consideration of tools used by others to evaluate experimental animal studies in toxicology and pre-clinical research. Standard practice is to have studies evaluated by at least two independent reviewers for aspects related to reporting quality, risk of bias/internal validity (e.g., randomization, blinding at outcome assessment, methods used to expose animals and assess outcomes, etc.), and sensitivity to identify factors that may limit the ability of a study to detect a true effect. To promote consistency across raters, prompting considerations and example responses are provided to reviewers, and a pilot phase is conducted. The evaluation process is performed separately for each outcome reported in a study, as the utility of a study may vary for different outcomes. Input from subject matter experts is used to identify chemical- and outcome-specific considerations (e.g., lifestage of exposure and outcome assessment when considering reproductive effects) to guide judgments within particular evaluation domains. For each evaluation domain, reviewers reach a consensus on a rating of *Good, Adequate, Deficient, or Critically Deficient*. These individual domain ratings are then used to determine the overall confidence in the study (*High Confidence, Medium Confidence, Low Confidence, or Deficient*). Study evaluation results, including the justifications for reviewer judgements, are documented and made publicly available

*Corresponding author. Dishaw.Laura@epa.gov (L. Dishaw).
[1]Formerly US EPA, Center for Public Health and Environmental Assessment, Washington, DC, United States.

Declaration of Competing Interest
The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

in EPA's version of Health Assessment Workspace Collaborative (HAWC), a free and open source web-based software application. (*The views expressed are those of the authors and do not necessarily represent the views or policies of the US EPA*).

Keywords

Systematic review; Study evaluation; Phthalates

1. Background and introduction

Systematic review is a process that uses explicit, pre-specified methods to gather, sort, evaluate, and synthesize evidence to inform a specific scientific question (IOM 2011). The purpose of systematic review is to increase transparency and objectivity through consistent application and documentation of methods, expert judgements, and decisions. Generically, systematic reviews involve the following steps: define the question to be asked, search and screen the literature for relevant data using clear screening criteria, evaluate the quality of the individual studies, summarize the study methods and findings of included studies, and synthesize the results to reach a conclusion (Moher et al., 2015).

Study evaluation is a critical step in the systematic review process (Cooper et al. 2016; Samuel et al. 2016), because it provides information about whether various aspects of the reporting, conduct, or design of an animal toxicology study may affect the reliability (i.e., the likelihood that the results of an outcome measurement will be replicated) or interpretability of the results. This step elucidates potential strengths and limitations of the available data that can inform subsequent steps of the systematic review process, including evidence synthesis and integration, as well as decisions for dose response analysis. Three key aspects of study evaluation are reporting, risk of bias, and sensitivity.

Reporting quality refers to how well the study authors communicated the details of the experiment (Higgins and Green 2011). Important aspects of reporting may vary depending on the design of the study and specific outcomes that are evaluated, but generally include information about the test animal (e.g., species/strain, sex, and housing conditions), exposure methods (e.g., route, purity), experimental design (e.g., exposure duration/frequency, sample size, life stage at exposure and outcome measurement), outcome evaluations (e.g., methods and assays used to evaluate the outcomes of interest), and sufficient presentation of results (e.g., quantitative vs qualitative results, variability estimates). Failure to include this information in the study report can make it difficult for the reviewer to adequately assess other aspects of the study evaluation (i.e., risk of bias and sensitivity). Thus, insufficient reporting can significantly impact the interpretability and usability of results.

Review of risk of bias is the assessment of systematic errors that may compromise the credibility of the study results. Many of the currently available tools to assess experimental animal studies (e.g., OHAT Risk of Bias Rating Tool (NIEHS, 2015); The Navigation Guide Systematic Review Methodology (Johnson et al. 2016; Koustas et al. 2014); CAMARADES check list (Macleod et al. 2004); SYRCLE Risk of Bias Tool (Hooijmans et al. 2014);

SciRAP in vivo Toxicity Tool (Beronius et al. 2018)) include elements the risk of bias ("internal validity") approach recommended by the Cochrane Collaboration to evaluate human trials. The Cochrane tool focuses on selection bias (randomization, allocation concealment), performance bias (blinding of participants and personnel during the study), detection bias (blinding at outcome assessment), attrition bias (incomplete outcome data), and selective reporting.

Study sensitivity is a measure of the ability of a study to detect a true effect for the outcome(s) of interest (Cooper et al. 2016). An insensitive study design biases the results towards the null, making it more likely to fail to detect a difference that truly exists and lead to a false conclusion of no effect. Study sensitivity is driven by features of the study design and outcome measures, test animals, and analysis and informs aspects of internal validity that are not typically encompassed by risk of bias-based study evaluation tools (Cooper et al. 2016). In general, factors affecting sensitivity should be evaluated at the outcome level (rather than study level) because they are likely to vary considerably depending on the outcome being evaluated. As a result, reviewers may need significant subject matter expertise to evaluate potential impacts on sensitivity.

This paper provides an overview of the study evaluation process that is being used by the Environmental Protection Agency's (EPA) Integrated Risk Information System (IRIS) Program to evaluate animal toxicity studies. The current iteration of the evaluation strategy is the result of extensive internal testing by EPA scientists that was used to develop and refine the process. Feedback was also received from external scientists familiar with study evaluation methods and tools. The approach has been positively reviewed as part of National Academy of Sciences (NAS) reports on implementation of systematic review by the IRIS Program (NRC, 2014; NASEM, 2018). The method was applied as part of the systematic reviews of animal evidence informing the potential health hazards of diisobutyl phthalate (DIBP; Yost et al. 2019) and diethyl phthalate (DEP; Weaver et al., *in press*) that are part of a larger evaluation published in this special issue on the health effects from exposure to phthalates. An overview of the method as well as specific examples of how this method has been applied for the DEP and DIBP assessments are presented, focusing on male reproductive outcomes.

2. Overview of the evaluation strategy

The strategy was designed to meet the needs of IRIS assessments, which encompass hazard identification and dose response analysis. Specifically, the goal was to create an approach that combined reporting quality, risk of bias, and sensitivity and could be applied to the wide variety of animal toxicology studies that may be included in an IRIS assessment evidence base. This is similar to the focus of other existing tools developed or optimized for use in environmental health. Namely, a focus on evaluation elements that improve the transparency and rigor of decisions in support of hazard identification and inform decisions for dose-response analysis.

An overview of the key elements of the process is provided here and in Fig. 1. The strategy was developed independently by EPA scientists and incorporates the key concepts of

reporting quality and risk of bias that are established in other study evaluation tools with the addition of sensitivity. The evaluation strategy provides a framework that guides reviewer responses with prompting questions and general considerations but includes flexibility when needed so that it can be applied to a wide range of study designs and outcomes. The evaluation strategy uses nine domains to examine strength and weaknesses:

1. reporting quality
2. allocation
3. observational bias/blinding
4. confounding and variable control
5. selective reporting and attrition
6. chemical administration and characterization
7. exposure timing, frequency, and duration
8. sensitivity and specificity
9. results presentation

These domains encompass the three key concepts of study evaluation that are evaluated separately in the individual domains: reporting quality (domain 1), risk of bias (domains 2–5), and sensitivity (domains 6–9). Core and prompting questions as well as general considerations were developed for each domain that are designed to aid consistency in applying the tool. The core and prompting questions focus on those aspects of study design and conduct that are known to be important to drawing conclusions about the reliability and interpretation of a study within the context of human health assessments of environmental chemicals, based on years of experience developing IRIS toxicological reviews (see Table 1). These questions guide the reviewer to evaluate information that is used to rate each domain as *Good, Adequate, Deficient, Not reported, or Critically Deficient* (see Fig. 1, Domain Judgements). Evaluations are performed by at least two subject matter experts and are performed at the level of the study, animal cohort, or outcome. The level of granularity depends on the focus of the domain and the potential impact the identified strengths or weaknesses could have on the reliability or interpretation of the study findings. For each domain, reviewers are provided guidance as to the granularity that is required.

The general considerations described for each domain are meant to provide broadly applicable guidelines for reviewers but may not cover all possible considerations that might be encountered within an evidence base. It is important to conduct a pilot phase in which reviewers are assigned a subset of studies to identify issues that are likely to arise during the full review and discuss any chemical-, exposure-, and outcome-specific considerations that will be applied during the study evaluation. Despite this preparatory step, there may be instances where an issue or consideration is not identified until through the review process has been partially completed. This may require reviewers to conduct a secondary review to apply the new considerations.

Reviewers should be assigned based on their knowledge of the health outcomes that are being evaluated in the studies. Depending on the complexity of the study, more than two reviewers may be needed (e.g., different subject matter experts for different health outcomes). In some cases, additional research on the part of the reviewer, including consulting outside scientists with extensive knowledge and experience in a particular area of toxicology, may be required to address specific concerns (e.g., adequacy of an experimental design to evaluate chemical-induced effects on function/development of specific organs/systems).

A written justification is recorded for each rating, including any identified limitations or concerns and the impact these may have on the interpretation of the study results. Documenting the rationale is a critical aspect of this method because it establishes a transparent record of reviewers' professional judgement of the strengths and weaknesses of the study for each of the reported outcomes, and the utility of the data for informing potential human health hazard(s). The identified strengths and weaknesses as described in the nine domain ratings are then used to inform the overall confidence for the outcome(s) of interest (rated as *High, Medium, or Low* confidence, or *Uninformative*; see Fig. 1, Overall study rating for an outcome).

After the independent evaluations are completed, they undergo a conflict resolution step where discrepancies in the individual domain and overall confidence ratings are discussed among reviewers to reach a final consensus judgement. If a consensus cannot be reached among the original reviewers, additional outside expertise may be consulted. The consensus judgments are recorded in EPA's version of Health Assessment Workspace Collaborative (HAWC), a free and open source web-based software application, and are available to the public (DIBP available at https://hawcprd.epa.gov/assessment/497/; DEP available at https://hawcprd.epa.gov/assessment/552/).

When concerns are raised due to deficiencies in the reporting of information that is highly influential to the study evaluation judgment, it may be useful to contact the corresponding author(s) and attempt to obtain the missing information. The decision on when to seek missing information takes into consideration whether obtaining this information is likely to affect the domain ratings, the overall confidence in the study, and whether the study is expected to be critical for informing hazard and/or dose response conclusions. Outreach to study authors is documented and considered unsuccessful if researchers do not respond to an email or phone request within an allotted amount of time (generally one month).

2.1. Reporting quality

This domain evaluates whether a report provides the information that is needed to proceed with study evaluation. Although reporting deficiencies may affect the ratings of other domains, the addition of a separate reporting quality domain allows the reviewer to determine whether there are any critical deficiencies in the reporting of the study that may warrant rating the study uninformative based on reporting quality alone and provides an overall indication of the extent to which reporting issues may be impacting the overall confidence rating versus issues related to the conduct or design of the study. In the latter

situation, reviewers are able to flag key pieces of information important to study evaluation that are missing but might be obtained by contacting the study authors.

The prompting questions in Table 1 describe two categories of information: critical and important. Critical information is considered the minimum necessary to move forward with the study evaluation and is comprised of the following:

- species
- test article name
- levels and duration of exposure
- route (e.g.,oral; inhalation)
- qualitative or quantitative results for *at least* one outcome of interest

If any piece of critical information is missing, the study is rated *Critically Deficient* for the reporting domain. This also results in an overall confidence rating of *Uninformative*, and the study is not considered further for hazard identification. The reporting quality domain, therefore, can serve as a triage for whether a full study evaluation is warranted.

Important information covers a much broader range of details about the study design and conduct (Table 1). This category of reporting information is used to distinguish between ratings of *Good, Adequate*, and *Deficient* for this domain. In general, studies rated *Good* in this domain describe all important information in the report. A rating of *Adequate* applies to studies where some important information is missing, but the omitted information is not expected to significantly impact the study evaluation. A study that is rated as *Deficient* is also missing some important information, but the omissions are expected to significantly impact the study evaluation. As discussed above, during the evaluation process attempts may be made to reach out to the corresponding authors in instances where missing important information is expected to significantly impact the study rating. If missing information is received from corresponding authors, it is documented and made available.

2.2. Allocation

This domain evaluates whether animals were randomly allocated to treatment groups, with each animal or litter having an equal chance of being assigned to any experimental group, or whether other steps were taken to normalize animals across treatment groups. Failure to randomize is considered a risk of bias and has been empirically demonstrated to increase effect sizes in animal studies (Hirst et al. 2014; Krauth et al. 2013). Factors to consider include:

- whether the allocation procedure was sufficiently described
- whether there were any indications that animals were assigned to groups in such a way that is known or expected to bias the interpretation of results.

Evaluation of this domain is performed at the level of the study, or for each cohort within a study when a study included multiple experiments.

In general, studies are considered *Good* for this domain if authors provide a description of the randomization method used to assign animals to groups (e.g. the use of a computerized randomization method), and *Adequate* if they indicate that randomization was used but do not describe the method. Alternatively, if a study reports that treatment groups were normalized by an important modifying factor (e.g. by equalizing body weight across groups) but do not indicate that randomization is used, this was also generally considered *Adequate*. A rating of *Deficient* for this domain indicates that animals were allocated in a manner that could bias the interpretation of results. In instances where a study does not provide any information on test animal allocation, it is marked as *Not Reported*, and it is assumed that randomization or normalization methods were not used. For this domain, *Not Reported* is interpreted as *Deficient*.

2.3. Observational Bias/Blinding

Assessment of this domain is performed at the level of the outcome and evaluates whether blinding or other appropriate measures are used to reduce observational bias during outcome evaluation. *Good* studies describe the measures used to mitigate the potential for observational bias to affect the results. These measures may include blinding investigators to the animal treatment group or using two or more investigators to independently measure the outcomes. A study is *Adequate* if they provide limited details on how the potential for observational bias is reduced or if these procedures are not explicitly stated but can be inferred (e.g., report indicates that methods followed those described in a separate document that specifies blinding or other methods of reducing observational bias).

If information on blinding or other steps to reduce observational bias is not available, this domain is marked as *Not Reported* and it is assumed that blinding was not performed. A *Not Reported* rating is interpreted as either *Adequate* or *Deficient* depending on the outcome being evaluated. For outcomes that use methods that limit the potential for observational bias (e.g., automated/computer driven data collection, standard laboratory kits, or simple, objective measures such as body or organ weights), concern is attenuated, and the domain is interpreted as *Adequate*. In contrast, for highly subjective measures where there is a strong potential for observational bias to impact the results (e.g., functional observational battery), the domain is interpreted as *Deficient*. In rare cases, a study may be rated *Critically Deficient* if there is strong evidence to suggest observational bias may have impacted the results.

It should be noted that the issues surrounding observational bias for histopathological evaluations are complex and largely dependent on the purpose of the evaluation. Here, histopathology is split into two broad categories: screening level and targeted evaluations. Screening level evaluations are defined as those that examine several tissues looking for a broad range of outcomes that are not prespecified. In contrast, targeted evaluations typically examine a much smaller set of tissues with prespecified outcomes that are known or hypothesized to be affected by the chemical of interest. Screening level evaluations can be used to inform subsequent targeted evaluations, and it is possible for both to be present within a single study. For screening level evaluations, best practices indicate that tissues should initially be read by pathologists with knowledge of the treatment group to facilitate

the separation of subtle treatment-related changes from normal variation. In these situations, concerns about observational bias are mitigated by a secondary evaluation that is blinded and/or performed by an independent pathologist or review board to ensure consistency in the diagnoses. For targeted evaluations, however, initial blinding is appropriate and recommended (Crissman et al. 2004).

2.4. Confounding and variable control

This domain evaluates whether the experiment was sufficiently controlled to attribute results to exposure to the compound of interest. The specific variables of concern can vary by experiment, outcome, or chemical. A judgment and rationale for this domain should be given for each cohort or experiment in the study, noting when the potential for confounding is restricted to specific outcomes. The rating reflects the extent to which the concerns are expected to affect the results of the study. In general, a *Good* study is one where, outside of the exposure of interest, variables that are likely to confound or modify results appear to be controlled for and consistent across experimental groups. A rating of *Adequate* is applied when there is some concern that variables likely to confound or modify results were uncontrolled or inconsistent across groups but are expected to have a minimal impact. A *Deficient* rating is used when there is notable concern that potentially confounding variables were uncontrolled or inconsistent across groups and are expected to substantially impact the results. In rare cases, a study may be rated *Critically Deficient* for this domain if confounding variables are known or presumed to be uncontrolled or inconsistent across groups and are expected to be a primary driver of the results.

2.5. Selective reporting and attrition

This domain evaluates whether results are reported on all prespecified outcomes of interest to the assessment and whether any apparent or explicit animal loss is addressed by the study authors. Aspects to consider include whether all study animals are accounted for in the results. Authors should provide clear explanations and/or rationales for discrepancies between the sample sizes at the beginning and end of the study (e.g., at the start of the study, n = 10/treatment but at outcome assessment is n = 5–10) and whether expected comparisons or certain groups with animal loss were appropriately addressed in the analyses. Evaluation of this domain is performed at the level of the cohort or experiment.

In general, *Good* studies present results quantitatively or qualitatively for all prespecified outcomes (including those that are inferred as well as explicitly stated), exposure groups and evaluation timepoints either in the main study report or supplemental materials. If results omissions or animal attrition are identified, the authors provide a reasonable (in the judgment of the evaluator) explanation and these are not expected to impact the interpretation of the results. *Adequate* studies report quantitative or qualitative results for most prespecified outcomes, exposure groups and evaluation timepoints; omissions and/or attrition (e.g., loss of some animals in an exposure group) are not explained or otherwise addressed but are not expected to significantly impact the interpretation of the results. A *Deficient* rating is applied to studies where quantitative or qualitative results are missing for prespecified outcomes, exposure groups and evaluation timepoints, and/or there is high animal attrition. In these cases, the omissions and/or attrition are expected to significantly

impact the interpretation of the results, regardless of whether it is explained by the authors. *Critically Deficient* studies are limited to those where extensive results omissions and/or animal attrition are identified that prevents comparisons of results across treatment groups.

2.6. Chemical administration and characterization

This domain evaluates the confidence in an exposure and factors that may cause actual exposure levels to deviate from those reported in a manuscript. Depending on the chemical being assessed, this may include factors such as the following:

- the stability and composition of the test chemical (e.g. purity, iso-meric composition)
- exposure generation and analytic verification methods (including whether the tested levels and spacing between exposure groups is resolvable using current methods)
- details of exposure methods (e.g. gavage volume, inhalation chamber type).

In some cases, exposure biomarkers in blood, urine, or tissues of treated animals can mitigate concerns regarding the accuracy of the dosing. Evaluation of this metric is performed at the level of the cohort or experiment.

In general, *Good* studies provide information about the chemical source and purity, and analytically verify the concentration of the test article. Additionally, there are no concerns about the composition, stability, or purity of the administered chemical, or the specific methods of administration. An *Adequate* rating is appropriate if some uncertainties in the chemical administration and characterization are identified but are expected to have minimal impact on interpretation of the results. For example, source and vendor-reported purity are presented but not independently verified by the study authors, or the purity of the test article is sub-optimal but not concerning. A study rated as *Deficient* in this domain has uncertainties in the exposure characterization that are expected to substantially impact the results.

Notably, the specific circumstances that would result in a *Deficient* rating for this domain are largely dependent on what is known about the chemical production and impurities. Examples may include studies where:

- the original source of the test article is not reported (e.g., material was provided by a collaborator)
- there are concerns about impurities (e.g., low purity reported, chemical synthesized in house, but purity was not verified), or
- the administration methods are not appropriate for the chemical of interest.

If uncertainties in the exposure characterization and/or administration are identified and there is reasonable certainty that the results are largely attributable to these factors rather than exposure to the chemical of interest (e.g., impurities are identified that are expected to be a primary driver of the results), a rating of *Critically Deficient* is appropriate.

2.7. Exposure timing, frequency, and duration

This domain evaluates whether the design of an exposure is appropriate for evaluating an outcome of interest. Developmental exposures, for instance, have greater relevance when they are designed to cover the full developmental windows that are critical to a system of interest. Similarly, the duration of exposure should be long enough for the expected outcome to develop. Additionally, it may be more complicated to interpret the results of studies that expose animals infrequently or sporadically. Evaluation of this metric was performed at the level of the outcome.

In general, a rating of *Good* is applied to outcomes where the timing, duration, and frequency of the exposure is sensitive and fully covers any known critical window(s) of sensitivity. Outcomes may be rated *Adequate* if the timing, duration, and frequency of the exposure is sensitive but covers only a portion of a known critical window of sensitivity. A *Deficient* rating is applied if the duration and/or frequency of the exposure is known or suspected to be insensitive for the outcome(s) of interest, or if most of a known critical window of sensitivity is not covered by the exposure. These limitations are expected to bias the results towards the null. In rare cases, an outcome may be rated *Critically Deficient* if the exposure design was determined to be inappropriate for evaluating the outcome(s) of interest. The rationale that is documented as part of the study evaluation process should clearly indicate the specific concern(s) that were identified. Notably, this critical deficiency is based on the expected insensitivity of the design; therefore, if the study is otherwise well conducted, the utility of this study will depend on whether effects were observed (i.e., if an exposure related effect is reported, it suggests that the design was, in fact, sensitive for that chemical, and the study could be adjusted to a higher rating).

2.8. Sensitivity and specificity

*T*his domain evaluates the ability of the outcome evaluation methods to reliably measure the effect(s) of interest for the assessment, which can differ widely between different assays and protocols. This includes both overestimates or underestimates of the true effect, as well as a higher or lower probability for detecting the outcome being assessed. Considerations for this domain are highly variable depending on the outcome(s) of interest. To the extent possible, outcome-specific considerations for this domain should be established by reviewers during the pilot phase, considering the most current knowledge of the methods and best practices for the outcomes being evaluated. In some cases, it is useful to consult external scientists with extensive expertise in the outcomes and methods. Although this may require a significant amount of effort early in the process, careful consideration and documentation of the factors that would affect the evaluation judgements at an early stage will improve the quality and consistency of the reviewer judgements.

Some general factors to consider include the following:

- the timing of the outcome evaluation relative to the exposure
- the specificity and validity of the assay or protocol for evaluating an outcome of interest
- whether there are serious concerns about the sample sizes

Notably, small sample size alone is not sufficient to conclude that a study is *Critically Deficient* for this domain. Evaluation of this domain is performed at the level of the outcome, and the strengths and/or weaknesses that were identified during the course of the evaluation should be clearly described in the rating rationale.

2.9. Results presentation

This domain evaluates whether results are presented in a way that allows for an informed interpretation of the data, including any concerns about the way that data are compared or presented. Evaluation of this domain is performed at the level of the outcome. As with the sensitivity and specificity domain, the specific considerations used to evaluate the usability and transparency of the data are dependent on the outcomes of interest and should be refined accordingly. When possible, the considerations for this domain should be developed in parallel with those for the outcome sensitivity and specificity domain.

Examples of potential concerns include:

- Limitations in the presentation of data (e.g., presentation of only relative organ weights when absolute weights are known to be more reliable; developmental toxicity data averaged across pups in a treatment group, when litter responses are more appropriate)
- Providing only a qualitative description of the results
- Pooling data when responses are known or expected to differ substantially (e.g., across sexes or ages)
- Failing to report on or address overt toxicity when exposure levels are known or expected to be highly toxic
- Incomplete data presentation (e.g., presentation of mean without variance data; concurrent control data are not presented)

In general, the outcomes were considered *Good* if there was a full quantitative presentation of results (e.g., means and standard error or standard deviation for continuous data; incidence data for categorical data) or *Adequate* if some details were missing (e.g., means were presented without a measure of variability; qualitative description of no effect). A rating of *Deficient* was applied in cases where authors reported a treatment-related effect, but only qualitative results were provided.

Although the magnitude or direction of effect is not generally considered when evaluating this domain, as described above, an exception is made for studies that provide only a qualitative description of the results. The general considerations above acknowledge that word or page limitations in journal articles may not allow for detailed reporting of negative findings. Note, that while publication bias is not explicitly evaluated as part of this study evaluation approach, this potential source of bias may be included as part of the broader assessment process on a case-by-case basis depending on the assessment-specific uncertainties.

2.10. Overall confidence

After evaluating the study for issues related to reporting quality, risk of bias, and sensitivity, reviewers must consider all the strengths and weaknesses that were identified during the study evaluation and rate their overall confidence for each of the outcomes of interest. As with the previous domains, reviewers should clearly document the rationale for their judgement, including a summary of the specific concerns, if any. Notably, the relative weight accorded to strengths and weaknesses should be based on their expected impact on reliability and validity of the study results and should not necessarily represent an 'average' of the individual metric ratings. In some cases, concerns identified in a single metric may be judged to be so critical that it drives down the overall rating of an otherwise well-conducted study.

In general, a rating of *High Confidence* is applied when no notable concerns are identified during the study evaluation (e.g., most or all domains rated *Good*). *Medium Confidence* indicates that although some concerns are identified, these are expected to have minimal impact on the interpretation of the results. In most cases this will include outcomes where most domains are rated *Adequate* or *Good* but may be appropriate if a domain is rated as *Deficient* if the identified concerns are not expected to strongly impact the magnitude or direction of the results. The *Low Confidence* rating is used when the concerns that have been identified are expected to significantly impact the study results or their interpretation (e.g., generally, *Deficient* ratings for one or more domains). Notably, the specific concerns leading to this confidence judgment are carried forward to later steps of the systematic review to facilitate the appropriate comparison and synthesis of evidence across sets of related studies. An overall confidence of *Uninformative* is reserved only for instances where serious flaw(s) are identified that make the study results unusable for informing hazard identification (e.g., generally, *Critically Deficient* rating in any domain or many *Deficient* ratings). *Uninformative* results are not considered further for the systematic review but may be used to highlight potential data gaps. In addition to the overall confidence rating, the primary limitations that are identified for the individual studies (e.g., the specific type of bias for which a concern was determined) are carried forward and considered during the health outcome-specific evidence syntheses.

2. Application of the evaluation strategy: examples from systematic reviews of DIBP and DEP

The core and prompting questions for each domain presented in Table 1 were applied to the animal toxicity studies identified in the systematic reviews of DIBP and DEP. Heat maps summarizing the study evaluation results by domain for DIBP and DEP are presented in Fig. 2; interactive versions of these heat maps are publicly available in HAWC (DIBP available at https://hawcprd.epa.gov/summary/visual/100500053/; DEP available at https://hawcprd.epa.gov/summary/visual/100000097/).

Both the DIBP and DEP systematic reviews included studies related to six broad hazard categories: male reproductive, female reproductive, developmental, liver, kidney, and cancer. The ratings shown in the heat maps represent composite ratings across all outcomes in

a study, although there were instances where certain outcomes within the same study were rated differently due to outcome-specific considerations; details on outcome-specific ratings can be found in the interactive visuals in HAWC. The examples given in the sections below for outcome-specific considerations focus on male reproductive outcomes. Male reproductive outcomes were reported in 16 studies for DIBP and 15 studies for DEP, and consisted of effects on testosterone, morphological development (e.g. anogenital distance, nipple retention, preputial separation, cryptorchidism, hypospadias), reproductive organ weights, histological changes, sperm effects, and fertility. As described previously, phthalate-induced effects in the developing male reproductive system are mediated via androgen-dependent and -independent MOAs (Arzuaga et al. 2019; NRC 2008; U.S. EPA 2009). This biological understanding was considered when evaluating the sensitivity of various study designs for detecting male reproductive effects.

3.1. Reporting quality

Most DIBP and DEP studies were rated as *Good* or *Adequate* for reporting quality, with few rated *Deficient* in this domain. No studies were rated *Critically Deficient*.

Examples of reporting quality concerns that resulted in a rating of *Adequate* for this domain:

- Chemical purity not reported
- Strain of animals not reported
- When evaluating developmental outcomes where the specific age at exposure is important, age of the animals is not reported, but can be inferred based on the reported body weights

Examples of reporting quality concerns that resulted in a rating of *Deficient* for this domain:

- Sample size is not reported anywhere in the paper. This impacts the ability to evaluate the sensitivity and specificity of the endpoint evaluations, as well as the ability to evaluate whether there was attrition of animals over the course of the experiment.
- Study report is missing information on endpoint evaluation methods

3.2. Allocation

Most DIBP and DEP studies were rated as *Adequate* or *Not Reported* (interpreted as *Deficient*) for this domain. The *Adequate* ratings largely reflect studies where study authors reported that animals were randomly allocated to groups but do not provide details on their randomization procedure, or studies where animals were allocated to groups using only normalization procedures (e.g. normalizing body weights across treatment groups without indication of randomization). No information was provided on animal allocation in 7 out of 19 DIBP studies and 9 out of 34 DEP informative studies (i.e., high, medium, or low confidence); these studies received a rating of *Not Reported* (interpreted as *Deficient*). Few studies received a *Deficient* rating. An example of this rating is a study evaluating DEP by Manservisi et al. (2015), which indicated that the control group included 5 pairs of sisters;

this was considered a risk of bias concern, since the control animals were not independent of one another.

3.3. Observational Bias/Blinding

In the systematic reviews of DIBP and DEP animal studies, most studies did not report whether blinding or other measures to reduce observational bias were used during outcome evaluation. When authors were contacted for clarification, they often indicated that some outcomes were read blinded, or that outcome measurements were independently evaluated by multiple researchers to reach a consensus judgement (i.e., for screening level histopathology outcomes). In these cases, the study evaluation was updated to *Adequate* or *Good* for this domain.

3.4. Confounding and variable control

In all DIBP and DEP studies, it was noted whether the vehicle was the same across the control and treatment groups. Since phthalate diesters are moderately lipophilic, oil is generally used as the vehicle control in gavage studies for DIBP and DEP. Olive oil and corn oil were both considered by reviewers to be appropriate vehicles. Vehicle control studies have found that olive oil and peanut oil treatments may increase male body weight relative to untreated controls (Yamasaki et al. 2001), but this was not considered a concern in the phthalate exposure studies if the same vehicle and gavage volume was used in control and treatment groups.

Most studies provided limited details on the housing conditions used for the animals, which could provide additional insight into whether there were any uncontrolled variables. However, none of the DIBP or DEP studies provided any indication that housing conditions were different between exposure groups, so this was not considered to be a risk of bias in any study.

In DIBP and DEP dietary exposure studies, decreased food consumption was frequently observed in phthalate-treated animals dosed with relatively high doses of phthalates (> 1000 mg/kg-day) and was identified as a confounder for body weight gain in these animals. Reduced food consumption also may bias towards the null by reducing the ingested dose of chemical; however, in these studies, bias towards the null was not considered a significant concern as the dose levels were high enough that the dose was still expected to elicit a toxicological response.

3.5. Selective reporting and attrition

For DIBP and DEP animal studies, it was noted whether the sample sizes reported in the methods were consistent with those reported in the results. There were several instances in which sample sizes were reported in the results but not in the methods; these studies were generally considered *Adequate* in this domain, unless there was reason to suspect that attrition had occurred. Some studies reported sample sizes as a range (e.g. n = 18–20 dams per treatment group), which was generally considered to be *Adequate* or *Good* if the potential differences in sample sizes across the different treatment groups was small enough that it was not expected to interfere with the interpretation of the results. Studies were

rated as *Deficient* in some cases where sample size was not clearly reported and a large amount of variation across the reported outcomes or treatment groups was observed without explanation from the authors, making it difficult to ascertain what the initial sample size was and whether attrition had occurred. No studies were rated *Critically Deficient* for this domain.

3.6. Chemical administration and characterization

All DIBP and DEP studies exposed animals via oral gavage or diet. A primary consideration for rating this domain was whether the authors reported the source and purity of the test chemical. Studies that independently verified the purity or stability of the test chemical were generally considered *Good* in this domain, whereas those that only provided the manufacturer's reported purity were considered *Adequate*. If a study did not provide any information on purity and this information was not inferable (e.g., obtained based on CAS number from the manufacturer's website), it was rated *Deficient*. For gestational exposures, it was noted whether and how often the doses were adjusted to account for dam body weight gain; if this was not reported, or if there was an indication that doses were not adjusted to account for dam weight gain, this would result in a lower rating in this domain. For dietary exposures, it was noted whether the amount of food consumption was monitored; if there was no indication that the authors monitored the amount of food consumption, the rating for this domain was decreased.

For DEP, specific concerns were raised over a series of studies that exposed animals to low doses of phthalates (e.g. ≤2 mg/kg-day) without verifying the nominal concentrations. This was considered a potential concern because phthalates may be present in environmental media due to their wide use in plastics and other commercial products. For instance, one dose range-finding study for DEP observed elevated levels of DEP metabolites in the urine of control animals, suggesting an unknown source of background exposure (Teitelbaum et al. 2016). Low-dose dietary studies were rated *Deficient* in this domain if they did not take measures to verify the concentrations of phthalates ingested by the test animals. Such concerns could be mitigated by including higher dose group(s) in addition to the low dose group, which would allow for better evaluation of dose-response.

Although the DIBP and DEP systematic reviews were limited to oral exposure studies, it should be noted that some additional considerations should be applied for inhalation exposure studies. In a *Good* study, animals are exposed in a dynamic chamber and concentrations in the exposure chambers are monitored regularly using reliable analytical methods. An inhalation study may be rated *Adequate* for this domain if the actual exposure concentrations are missing or verified with less reliable methods. Studies that use static inhalation chambers are generally considered *Deficient*. These are described in greater detail by Whalan et al. (2019).

3.7. Exposure timing, frequency, and duration

For male reproductive effects following gestational exposure to phthalates, the critical window of exposure will vary depending upon the mode of action (MOA) for a given outcome. There is considerable evidence that in utero exposure to certain phthalates can

disrupt male reproductive development through multiple distinct MOAs: decreased testicular production of androgens, which are integral to male sexual development; decreased insulin-like-3 hormone, which regulates transabdominal testicular descent; and disrupted seminiferous cord formation, Sertoli cells, and germ cell development via an unknown MOA (Arzuaga et al. 2019; CHAP, 2014; NRC 2008; U.S. EPA 2009). Outcomes associated primarily with the androgen-dependent MOA include decreased testosterone production, decreased male reproductive organ weights, hypospadias, decreased anogenital distance, and female-like nipple retention. Outcomes associated primarily with androgen-independent MOAs include cryptorchidism and germ cell toxicity.

Male sexual differentiation occurs during the masculinization programming window in late gestation [approximately between gestation days (GD) 14–18 in the rat], when a surge in testosterone production triggers the growth and development of the male reproductive system (Evans and Ganjam, 2011; Foster and Gray 2013; Sharpe 2010). Testosterone production peaks at approximately GD 14 in the rat (Scott et al. 2009). Experimental designs that expose animals for only part of this window (e.g. 1 or 2 days) have been shown to be less sensitive for evaluating androgen-dependent effects of phthalates (e.g. see (Ema et al. 2000; Scott et al. 2008)). Therefore, gestational exposure studies aimed at evaluating androgen-dependent male reproductive outcomes (including testosterone production, male reproductive organ weight, hypospadias, anogenital distance, and nipple retention) were considered less sensitive for this domain if the exposure did not span this entire male programming window.

For studies aimed at evaluating testicular histopathology or effects on sperm following gestational exposure to phthalates, the critical window of exposure is difficult to determine *a priori* since the windows of sensitivity for different cellular targets within the testis vary according to developmental stage (Creasy and Chapin, 2018). For instance, in rats, the most sensitive developmental stages during gestation for effects on Sertoli and germ cell numbers do not overlap with the masculinization programming window. Scott et al (2008) observed that DBP exposures from GD 11–20, 13–20, and 19–20 resulted in significant decreases in Sertoli cell number, whereas exposures from GD 13–15 and 15–17 had no effect. Ferrara et al. (2006) observed the strongest decline in germ cell numbers following DBP exposure from GD 13–20, whereas animals treated from GD 19–20 did not experience any changes in germ cell number.

Puberty is another sensitive life stage characterized by rapid changes in hormone levels and reproductive organ growth (Stoker et al. 2000). Sexually mature animals are found to be less sensitive than prepubertal animals to the androgen-dependent and -independent effects of phthalates (Albert and Jégou 2014). Therefore, studies exposing adult animals to phthalates is considered to have reduced sensitivity in this domain. More generally, concerns were raised if a study did not describe the life stage of the animals in sufficient detail to allow for outcome interpretation.

Notably, exposure designs may be sensitive for evaluating some outcomes but not others. For instance, phthalate exposure studies that target the male programming window (e.g. exposure from GD 14–18) may be insensitive for identifying reproductive effects in females

or effects in other organ systems. Although there does not appear to be a "critical window" of sensitivity for female reproductive effects, female reproductive tract malformations were induced following phthalate exposures that encompassed the major window of organogenesis (GD 8–13), whereas exposure from GD 14–19 produced no effects on the developing female reproductive tract (Hannas et al. 2013). Therefore, this domain should be addressed on a case by case basis based on what is known about the etiology of each specific outcome. Longer exposures would generally be considered more sensitive than shorter exposures.

3.8. Sensitivity and specificity

For all outcomes in the DIBP and DEP systematic reviews, the sensitivity and specificity was generally inferred to be *Adequate* or *Good* if a study evaluated outcomes in a manner consistent with guideline recommendations or used standard/well-regarded assays for outcome evaluation. If a study used a novel protocol, the inclusion of positive controls was considered useful for this evaluation. Most male reproductive outcomes reported in the literature databases for DIBP and DEP are considered standard indicators of male reproductive toxicity that are routinely evaluated in studies aimed at assessing reproductive effects (Mangelsdorf et al. 2003; U.S. EPA 1996). Some specific considerations for the most widely reported male reproductive outcomes are as follows.

- *Testosterone:* Serum testosterone, testicular testosterone production, and testicular testosterone levels were all considered to be acceptable measurements of testosterone levels, as these methods have similar sensitivity and reliability (Gray et al. 2016). Standard quantification assays were considered acceptable, including radio-immunoassay or enzyme-linked immunosorbent assay. Timing of assessment was also considered, as phthalate studies that include a recovery period may observe no effect (Auharek et al. 2010; Ferrara et al. 2006).
- *Male reproductive organ weights:* It was not considered a significant concern for organ weight measurements if the procedural methods were not described in detail, since it is a relatively straightforward measurement.
- *Histopathology:* Concerns were raised if the testis was preserved in formalin, which is known to introduce artifacts to the testis (Foley 2001); however, for other reproductive organs (e.g. epididymis), formalin was considered acceptable. Preservation in Bouin's solution or modified Davidson's solution were considered acceptable methods for the testis, and whole-body perfusion (not used in any of the DIBP or DEP studies) is considered the gold standard (Foley 2001). It was also noted whether the evaluation was performed by an accredited pathologist, how many sections were evaluated, and how the lesions were classified (e.g. whether standardized terminology was used).

An additional consideration is whether there was adequate sampling. When evaluating offspring effects in gestational exposure studies, there is a general understanding that it may not be practical to include all pups in each analysis; however, if only a subset of litters was included in an analysis, it was noted as a potential concern for study sensitivity.

3.9. Results presentation

For DIBP and DEP, a common concern in this domain was the presentation of testis weight as an organ: body weight ratio (relative organ weight). Relative testis weights without corresponding data on absolute testis weights were reported in 2 out of 9 DIBP studies and 4 out of 10 DEP studies that evaluated male reproductive organ weights. Relative testis weight has been found to be a potentially unreliable metric for testicular toxicity because testis and body weight are not proportional (Bailey et al. 2004), so studies that reported only relative testis weight were considered to have reduced sensitivity and were rated as *Deficient* in this domain. Another concern encountered in several studies was the presentation of offspring data from developmental studies as means of individual animals rather than as litter means, which has the potential to overestimate the statistical significance of experimental findings (Haseman et al. 2001). Studies that reported offspring data as means of individual animals were therefore rated as *Deficient* in this domain. This was noted only in 1 DIBP study that evaluated male reproductive effects in F1 males, but also in 8 DEP studies that evaluated other types of outcomes in F1 or F2 offspring. Other deficiencies in results presentations encountered in the DIBP and DEP systematic reviews were the presentation of testosterone data as a percentage of control (noted in 2 studies) or as means without a measure of variance (noted in 1 study), and qualitative description of histopathological lesions with no quantitative information provided on incidence or severity (noted in 1 study that evaluated testicular histopathology and in 12 studies that evaluated histopathology in other organs, e.g. liver, kidney, and mammary glands). One DEP study was rated as *Critically Deficient* in this domain because it presented con-flicting results for testosterone across the figures and tables and presented data for other outcomes as dose-response curves on a log scale (which is implausible for the control dose of 0 mg/kg-day) (Hu et al. 2018).

3.10. Overall confidence

In general, the overall confidence ratings follow the general considerations discussed earlier (see 2.10). However, In some cases, concerns identified in a single metric may be judged to be so critical that it drives down the overall rating of an otherwise well-conducted study. For instance, the DIBP study by Zhu et al. (2010) was rated as *Low Confidence*, with the study evaluation indicating significant concerns (i.e., rated *Not Reported - interpreted as Deficient*) because the sample sizes and allocation methods were not reported in addition to less critical limitations (i.e., rated *Adequate* in most other domains). For DEP, low dose studies by two laboratory groups were rated as *Low Confidence* primarily due to concerns about exposure characterization, although these studies also had additional concerns (e.g. pooling data across litters in developmental studies, qualitative reporting of data) that contributed to the *Low Confidence* ratings (Hu et al., 2016; Manservisi et al. 2015; Mapuskar et al. 2007; Pereira et al. 2006; 2007a,b; 2008a,b; Pereira et al. 2007c; Pereira and Rao 2006a,b; 2007; Sonde et al. 2000). Additionally, the DEP study by Hu et al. (2018) was found to be *Uninformative* due to its *Critically Deficient* rating in the results presentation domain, as described above. Examples of other considerations that may warrant an *Uninformative* rating include the use of wild-caught animals (i.e., due to the high potential for confounding variables, such as health status, prior exposures, age at exposure/outcome assessment, etc), or severe inconsistencies between methods and results that make interpretation of results impossible.

4. Discussion

One of the key features of the IRIS program study evaluation method is that reviewers have the flexibility to apply expert judgment and subject matter expertise when evaluating studies. The core and prompting questions under the nine domains provide a general framework for developing the outcome-, assessment-, and chemical-specific considerations but are not intended to supersede application of professional judgement. One critique of this flexible approach is that it may reduce the consistency of the evaluations. This concern was addressed by the iterative nature of the review process. A pilot study is first conducted which allows reviewers the opportunity to identify issues that may be specific to the evidence base at hand and discuss how they should be evaluated before conducting their independent evaluations on the full set of studies. The use of at least two reviewers per study also helps to ensure consistency and rigor in the process. Transparency is achieved by requiring reviewers to document the rationale for each of the domains and the overall confidence rating. The final judgements are publicly shared as part of the systematic review documentation. Here, the HAWC platform was used to document and share the specific rationales that were applied to animal toxicology studies for the DIBP and DEP systematic reviews; however, other platforms or systems could be used.

Another key feature of the present method is that most judgements are not made at the level of the study. In many instances, study level evaluations are likely to over- or underestimate issues that are identified for an individual outcome because the issues do not affect all components of the study equally. Similarly, the overall confidence rating(s) allow for weighting of domain ratings based on the severity of the identified strengths and weaknesses within the context of the chemical or outcomes being evaluated. Not all identified limitations are equal in terms of their expected impact on the interpretation of the results and, in some cases, a single limitation could ultimately drive the overall confidence rating.

Lastly, sensitivity is an important aspect of this evaluation method that is often addressed separately from the study evaluation in other approaches. Studies encountered in environmental health assessments typically report heterogeneous results; the relative informativeness of one study versus another for assessing hazard can be evaluated using information on issues related to both risk of bias and sensitivity. Therefore, it was important for us to incorporate both in our review of individual studies so the study ratings reflect whether reported results are reasonably expected to reflect the true, underlying cause-effect relationship between exposure and the outcomes of interest.

There are, however, disadvantages to this approach. Most notable are the significant time, skill, and resources required to complete the study evaluation process. Internal testing found that it can take 1–3 h per reviewer to evaluate a single study, depending on the complexity of the study design and the number of outcomes that are being evaluated. Other considerable time investments are the pilot phase and conflict resolution steps and the potential for an iterative review process. Reviewers must have extensive training and knowledge of the field of study to adequately identify and assess the potential impact of various aspects of the study design and conduct. If the level of expertise is not available within the pool of reviewers they may need to reach out to outside experts to develop the appropriate

considerations during the pilot phase. The subjectivity of the judgements for some domains (e.g., reliability of specific exposure and/or outcome evaluation methods) relative to others (e.g., randomization) introduces the potential for inconsistencies across reviewers and assessments that must be resolved during the pilot phase and conflict resolution steps.

The DIBP and DEP systematic reviews illustrate how the method works in practice and the final judgements from these reviews will be used to generate examples that serve as a reference and training tool for future applications of the method. Although the examples discussed here focus on male reproductive outcomes, reviewers applied the method to over 50 studies covering a wide range of study designs, health effects, and outcomes. In general, some deficiencies were noted across multiple studies; however, specific deficiencies tended to vary across studies, and no single domain stood out as a primary driver of the overall confidence ratings. The incorporation of study evaluation ratings into the systematic reviews of DIBP and DEP resulted in more informed hazard evaluations, as it provided a transparent way to explore potential sources of inconsistency when synthesizing data across studies. For instance, for DEP, a series of *Low Confidence* studies reported profound histopathological and biochemical changes in the liver following exposure to relatively low dose levels; these studies described hepatic lesions qualitatively with no information on the incidence and severity of lesions, and had other concerns including lack of validation of the low-dose exposure levels. Conversely, *High* or *Medium Confidence* studies testing higher dose levels observed little to no evidence of effects on hepatic histopathology or biochemistry. The confidence ratings were considered when synthesizing data across studies, and it was concluded that the evidence for effects on hepatic histopathology and biochemistry was limited (Weaver et al., in preparation). The DIBP and DEP reviewers did not require an additional reviewer during the conflict resolution step to reach a final consensus judgement; however, they found it helpful to consult with other scientists to gain insights when identifying outcome, chemical and assessment-specific considerations during the pilot phase.

It is important to note that the present method may undergo further refinement as insights are gained during implementation across different study designs, outcomes, and assessment needs. Feedback from reviewers during method development and testing was critical for improving clarity in the core and prompting questions and additional adjustments may be made as the process is applied to more assessments. Future assessments can build on existing study evaluations such as those done for DIBP and DEP. As more study evaluations are completed, there is likely to be some overlap in the types of chemical-, exposure-, and outcome-specific considerations. Completed reviews can serve as a useful starting point for developing assessment-specific considerations that are adapted or updated as needed to reflect recent advances in scientific understanding or unique circumstances of future assessments.

Acknowledgements

The authors would like to acknowledge Susan Euling, Nagalakshmi Keshava, Anuradha Mudipalli, Lily Wang, and James A. Weaver for their work on the study evaluations for DIBP and DEP. We would also like to thank Geniece Lehman and Ingrid Druwe for their thoughtful comments on previous drafts of this manuscript.

Abbreviations:

DEP	Diethyl Phthalate
DIBP	Diisobutyl Phthalate
EPA	Environmental Protection Agency
GD	Gestational Day
HAWC	Health Assessment Workspace Collaborative
IRIS	Integrated Risk Information Systems
MOA	Mode of Action
NAS	National Academy of Sciences

References

Albert O, Jégou B, 2014. A critical assessment of the endocrine susceptibility of the human testis to phthalates from fetal life to adulthood. Hum. Reprod. Update20, 231–249. [PubMed: 24077978]

Arzuaga X, Walker T, Yost EE, Radke EG, Hotchkiss AK, 2019. Use of the Adverse Outcome Pathway (AOP) framework to evaluate species concordance and human relevance of Dibutyl phthalate (DBP)-induced male reproductive toxicity. Reprod. Toxicol

Auharek SA, de Franca LR, McKinnell C, Jobling MS, Scott HM, Sharpe RM, 2010. Prenatal plus postnatal exposure to di(n-butyl) phthalate and/or flutamide markedly reduces final sertoli cell number in the rat. Endocrinology151, 2868–2875. [PubMed: 20392824]

Bailey SA, Zidell RH, Perry RW, 2004. Relationships between organ weight and body/brain weight in the rat: What is the best analytical endpoint?Toxicol. Pathol32, 448–466. [PubMed: 15204968]

Beronius A, Molander L, Zilliacus J, Rudén C, Hanberg A, 2018. Testing and re-fining the Science in Risk Assessment and Policy (SciRAP) web-based platform for evaluating the reliability and relevance of in vivo toxicity studies. J. Appl. Toxicol38, 1460–1470. [PubMed: 29806706]

CHAP, 2014. Chronic Hazard Advisory Panel on phthalates and phthalate alternatives (with appendices). Bethesda, MD: U.S. Consumer Product Safety Commission, Directorate for Health Sciences.

Cooper GS, Lunn RM, Ågerstrand M, Glenn BS, Kraft AD, Luke AM, Ratcliffe JM, 2016. Study sensitivity: Evaluating the ability to detect effects in systematic reviews of chemical exposures. Environ. Int92–93, 605–610.

Creasy DM, Chapin RE, 2018. Chapter 17 - Male Reproductive System edêds. London, UK; San Diego, CA, USA: Academic Press of Elsevier.

Crissman JW, Goodman DG, Hildebrandt PK, Maronpot RR, Prater DA, Riley JH, Seaman WJ, Thake DC, 2004. Best practices guideline: Toxicologic histopathology. Toxicol. Pathol32, 126–131. [PubMed: 14713558]

Ema M, Miyawaki E, Kawashima K, 2000. Critical period for adverse effects on development of reproductive system in male offspring of rats given di-n-butyl phthalate during late pregnancy. Toxicol. Lett111, 271–278. [PubMed: 10643872]

Evans TJ, Ganjam VK (Eds.), 2011. Chapter 2 - Reproductive anatomy and physiology. Academic Press, San Diego, CA.

Ferrara D, Hallmark N, Scott H, Brown R, McKinnell C, Mahood IK, Sharpe RM, 2006. Acute and long-term effects of in utero exposure of rats to di(n-butyl) phthalate on testicular germ cell development and proliferation. Endocrinology147, 5352–5362. [PubMed: 16916955]

Foley GL, 2001. Overview of male reproductive pathology. Toxicol. Pathol29, 49–63. [PubMed: 11215684]

Foster P, Gray LE (Eds.), 2013. Toxic responses of the reproductive system. New York, NY, McGraw-Hill Education.

Gray LE, Furr J, Tatum-Gibbs KR, Lambright C, Sampson H, Hannas BR, Wilson VS, Hotchkiss A, Foster PM, 2016. Establishing the 'Biological Relevance' of dipentyl phthalate reductions in fetal rat testosterone production and plasma and testis testosterone levels. Toxicol. Sci149, 178–191. [PubMed: 26454885]

Hannas BR, Howdeshell KL, Furr J, Earl Gray L Jr., 2013. In utero phthalate effects in the female rat: A model for MRKH syndrome. Toxicol. Lett223:315–321. [PubMed: 23542816]

Haseman JK, Bailer AJ, Kodell RL, Morris R, Portier K, 2001. Statistical issues in the analysis of low-dose endocrine disruptor data. Toxicol. Sci61, 201–210. [PubMed: 11353128]

Higgins JPT, Green S, 2011. Cochrane handbook for systematic reviews of interventions. Version 5.1.0 (Updated March 2011). The Cochrane Collaboration.

Hirst JA, Howick J, Aronson JK, Roberts N, Perera R, Koshiaris C, Heneghan C, 2014. The need for randomization in animal trials: an overview of systematic reviews. PLoS ONE9, e98856. [PubMed: 24906117]

Hooijmans CR, Rovers MM, De Vries RB, Leenaars M, Ritskes-Hoitinga M, Langendam MW, 2014. SYRCLEs risk of bias tool for animal studies. BMC Med. Res. Method14, 43.

Hu J, Raikhel V, Gopalakrishnan K, Fernandez-Hernandez H, Lambertini L, Manservisi F, Falcioni L, Bua L, Belpoggi F, L Teitelbaum S, Chen J, 2016. Effect of postnatal low-dose exposure to environmental chemicals on the gut microbiome in a rodent model4:26.

Hu G, Li J, Shan Y, Li X, Zhu Q, Li H, Wang Y, Chen X, Lian Q, Ge RS, 2018. In utero combined di-(2-ethylhexyl) phthalate and diethyl phthalate exposure cumulatively impairs rat fetal Leydig cell development. Toxicology395, 23–33. [PubMed: 29325824]

IOM, 2011. Introduction edêds. Washington, DC: The National Academies Press.

Johnson PI, Koustas E, Vesterinen HM, Sutton P, Atchley DS, Kim AN, Campbell M, Donald JM, Sen S, Bero L, Zeise L, Woodruff TJ, 2016. Application of the Navigation Guide systematic review methodology to the evidence for developmental and reproductive toxicity of triclosan. Environ. Int92–93, 716–728.

Koustas E, Lam J, Sutton P, Johnson PI, Atchley DS, Sen S, Robinson KA, Axelrad DA, Woodruff TJ, 2014. The Navigation Guide - Evidence-based medicine meets environmental health: Systematic review of nonhuman evidence for PFOA effects on fetal growth. Environ. Health Perspect122, 1015–1027. [PubMed: 24968374]

Krauth D, Woodruff TJ, Bero L, 2013. Instruments for assessing risk of bias and other methodological criteria of published animal studies: a systematic review. Environ. Health Perspect121, 985–992. [PubMed: 23771496]

Macleod MR, O'Collins T, Howells DW, Donnan GA, 2004. Pooling of animal experimental data reveals influence of study design and publication bias. Stroke35, 1203–1208. [PubMed: 15060322]

Mangelsdorf I, Buschmann J, Orthen B, 2003. Some aspects relating to the evaluation of the effects of chemicals on male fertility. Regul. Toxicol. Pharm37, 356–369.

Manservisi F, Gopalakrishnan K, Tibaldi E, Hysi A, Iezzi M, Lambertini L, Teitelbaum S, Chen J, Belpoggi F, 2015. Effect of maternal exposure to endocrine disrupting chemicals on reproduction and mammary gland development in female Sprague-Dawley rats. Reprod. Toxicol54, 110–119. [PubMed: 25554385]

Mapuskar K, Pereira C, Rao CV, 2007. Dose-dependent sub-chronic toxicity of diethyl phthalate in female Swiss mice. Pestic. Biochem. Physiol87, 156–163.

Moher D, Shamseer L, Clarke M, Ghersi D, Liberati A, Petticrew M, Shekelle P, Stewart LA, Group P-P, 2015. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. Syst Rev4:1. [PubMed: 25554246]

NASEM, 2018. Progress toward transforming the Integrated Risk Information System (IRIS) program. A 2018 evaluation eds. Washington, DC: The National Academies Press.

NIEHS, 2015. OHAT Risk of Bias Tool. NTP.

NRC, 2008. Phthalates and cumulative risk assessment: The task ahead. Phthalates and cumulative risk assessment: The task ahead. National Academies Press, Washington, DC.

NRC, 2014. Review of EPA's Integrated Risk Information System (IRIS) process. Washington, DC: National Academies Press.

Pereira C, Mapuskar K, Rao CV, 2006. Chronic toxicity of diethyl phthalate in male Wistar rats–A dose-response study. Regul. Toxicol. Pharm45, 169–177.

Pereira C, Mapuskar K, Rao CV, 2007a. Chronic toxicity of diethyl phthalate–A three generation lactational and gestational exposure study on male Wistar rats. Environ. Toxicol. Pharmacol23, 319–327. [PubMed: 21783775]

Pereira C, Mapuskar K, Rao CV, 2007b. Reproductive failure associated with chronic interactive mixture toxicity of diethyl phthalate and Clophen A60 after gestational and lactational exposure over two generations in Wistar rats. Toxicol Int14, 111–122.

Pereira C, Mapuskar K, Rao CV, 2008a. Chronic mixture toxicity study of Clophen A60 and diethyl phthalate in male rats. Toxicol. Environ. Chem90, 349–359.

Pereira C, Mapuskar K, Rao CV, 2008b. Effect of diethyl phthalate on rat testicular antioxidant system: A dose-dependent toxicity study. Pestic. Biochem. Physiol90, 52–57.

Pereira C, Mapuskar K, Vaman Rao C, 2007c. A two generation chronic mixture toxicity study of Clophen A60 and diethyl phthalate after gestational and lactational exposure in female Wistar rats. Pestic. Biochem. Physiol88, 156–166.

Pereira C, Rao CV, 2006. Chronic toxicity of diethyl phthalate and polychlorinated biphenyls in rats–a sex related biochemical interaction study13, 53–60.

Pereira C, Rao CV, 2006b. Combined and individual administration of diethyl phthalate and polychlorinated biphenyls and its toxicity in female Wistar rats. Environ. Toxicol. Pharmacol21, 93–102. [PubMed: 21783644]

Pereira C, Rao CV, 2007. Toxicity study of maternal transfer of polychlorinated biphenyls and diethyl phthalate to 21-day-old male and female weanling pups of Wistar rats. Ecotoxicol. Environ. Saf68, 118–125. [PubMed: 16814384]

Samuel GO, Hoffmann S, Wright RA, Lalu MM, Patlewicz G, Becker RA, Degeorge GL, Fergusson D, Hartung T, Lewis RJ, Stephens ML, 2016. Guidance on assessing the methodological and reporting quality of toxicologically relevant studies: A scoping review. Environ. Int92–93, 630–646.

Scott HM, Hutchison GR, Jobling MS, Mckinnell C, Drake AJ, Sharpe RM, 2008. Relationship between androgen action in the Male Programming Window, Fetal Sertoli Cell Number, and Adult Testis Size in the Rat. Endocrinology149, 5280–5287. [PubMed: 18566125]

Scott HM, Mason JI, Sharpe RM, 2009. Steroidogenesis in the fetal testis and its susceptibility to disruption by exogenous compounds. Endocr. Rev30, 883–925. [PubMed: 19887492]

Sharpe RM, 2010. Environmental/lifestyle effects on spermatogenesis. Philos. Trans. R. Soc. Lond. B Biol. Sci365, 1697–1712. [PubMed: 20403879]

Sonde V, D'souza A, Tarapore R, Pereira L, Khare MP, Sinkar P, Krishnan S, Rao CV, 2000. Simultaneous administration of diethylphthalate and ethyl alcohol and its toxicity in male Sprague-Dawley rats. Toxicology147:23–31. [PubMed: 10837929]

Stoker TE, Parks LG, Gray LE, Cooper RL, 2000. Endocrine-disrupting chemicals: prepubertal exposures and effects on sexual maturation and thyroid function in the male rat. A focus on the EDSTAC recommendations. Endocrine Disrupter Screening and Testing Advisory Committee. Crit. Rev. Toxicol30, 197–252. [PubMed: 10759431]

Teitelbaum SL, Li Q, Lambertini L, Belpoggi F, Manservisi F, Falcioni L, Bua L, Silva MJ, Ye X, Calafat AM, Chen J, 2016. Paired serum and urine concentrations of biomarkers of diethyl phthalate, methyl paraben, and triclosan in rats. Environ. Health Perspect124, 39–45. [PubMed: 26047088]

U.S. EPA, 1996. Guidelines for reproductive toxicity risk assessment. Fed Reg61:56274–56322.

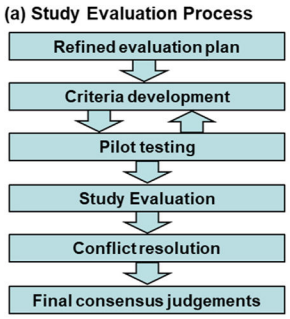
U.S. EPA, 2009. An approach to using toxicogenomic data in U.S. EPA human health risk assessments: A dibutyl phthalate case study (Final Report). Washington, DC.

Whalan JE, Stanek J, Woodall G, Reinhart P, Galizia A, Glenn B, Kraft A, SL M, Jarabek AM, 2019. The evaluation of inhalation studies for exposure quality: A case study with formaldehyde312:167–172.

Yamasaki K, Sawaki M, Noda S, Takatuki M, 2001. Effects of olive, corn, sesame or peanut oil on the body weights and reproductive organ weights of immature male and female rats. Exp. Anim50, 173–177. [PubMed: 11381622]

Yost EE, Euling SY, Weaver JA, Beverly BEJ, Keshava N, Mudipalli A, Arzuaga X, Blessinger T, Dishaw L, Hotchkiss A, Makris SL, 2019. Hazards of diisobutyl phthalate (DIBP) exposure: A systematic review of animal toxicology studies. Environ. Int125, 579–594. [PubMed: 30591249]

Zhu XB, Tay TW, Andriana BB, Alam MS, Choi EK, Tsunekawa N, Kanai Y, Kurohmaru M, 2010. Effects of di-iso-butyl phthalate on testes of prepubertal rats and mice. Okajimas Folia Anat. Jpn86, 129–136. [PubMed: 20560449]



(b) Individual Evaluation Domains

Animal Studies
Reporting Quality
Allocation
Observational bias/blinding
Confounding and variable control
Selective reporting and attrition
Chemical administration and characterization
Exposure timing, frequency, and duration
Sensitivity and specificity
Results presentation

(c) Domain Judgements

Judgment	Interpretation
++ Good	Appropriate study conduct relating to the domain and minor deficiencies not expected to influence results.
+ Adequate	A study that may have some limitations relating to the domain, but they are not likely to be severe or to have a notable impact on results.
- Deficient	Identified biases or deficiencies interpreted as likely to have had a notable impact on the results or prevent reliable interpretation of study findings.
NR Not Reported	The necessary information to evaluate the domain (i.e., observational bias/blinding or allocation) was not reported in the study . Reviewers should interpret the expected impact that the missing information will have on the interpretation of the findings (i.e., interpreted as adequate or deficient)
-- Critically Deficient	A serious flaw identified that makes the observed effect(s) uninterpretable. Studies with a critical deficiency are considered "uninformative" overall.

(d) Overall Confidence Rating

Rating	Interpretation
High	No notable deficiencies or concerns identified; potential for bias unlikely or minimal; sensitive methodology.
Medium	Possible deficiencies or concerns noted, but resulting bias or lack of sensitivity is unlikely to be of a notable degree.
Low	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.
Uninformative	Serious flaw(s) makes study results unusable for hazard identification but may be used to highlight possible research gaps

Fig. 1.

Overview of the study evaluation approach. (a) An overview of the evaluation process. (b) Individual evaluation domains for animal studies. Definitions for the (c) domain and (d) overall confidence ratings.

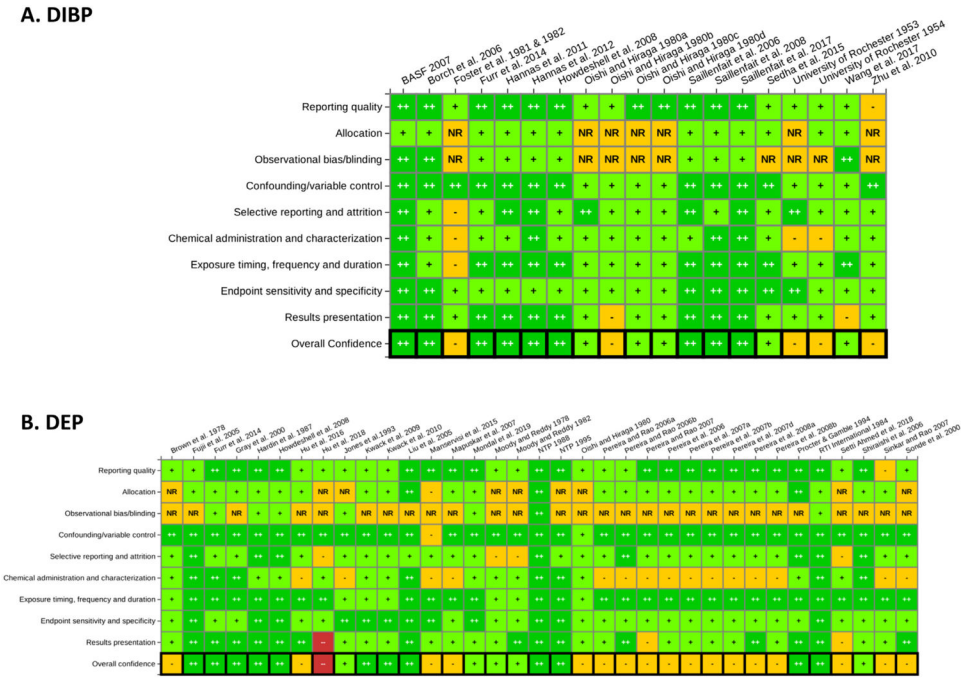


Fig. 2.

Summary of study evaluation ratings by domain for (A) DIBP and (B) DEP. The ratings presented in for each study in these heat maps represent the composite ratings across outcomes within each study evaluation domain. There were instances where certain outcomes were rated differently within the same study because of outcome specific considerations (see individual study evaluations in HAWC for details). Links to interactive graphics in HAWC can be found at the following URLs for DIBP (https://hawcprd.epa.gov/summary/visual/100500053/) and DEP (https://hawcprd.epa.gov/summary/visual/100000097/). ++ = Good/High Confidence; + = Adequate/Medium Confidence, − = Deficient/Low Confidence; – = Critically Deficient/Uninformative; NR = Not Reported.

Table 1

Study evaluation domains with core and prompting questions.

Key concept	Domain - Core Question	Prompting questions
Reporting Quality	**1. Reporting quality** - Does the study report information for evaluating the design and conduct of the study for the outcome(s) of interest?	Does the study report the following? • Critical information necessary to perform study evaluation: – Species; test article name; levels and duration of exposure; route (e.g., oral; inhalation); qualitative or quantitative results for at least one outcome of interest • Important information for evaluating the study methods: – Test animal: strain, sex, source, and general husbandry procedures – Exposure methods: source, purity, method of administration – Experimental design: frequency of exposure, animal age and lifestage during exposure and at outcome evaluation – Outcome evaluation methods: assays or procedures used to measure the outcomes of interest
Risk of Bias	**2. Allocation** - Were animals assigned to experimental groups using a method that minimizes selection bias?	For each study: • Did each animal or litter have an equal chance of being assigned to any experimental group (i.e., random allocation)? • Is the allocation method described? • Aside from randomization, were any steps taken to balance variables across experimental groups during allocation?
	3. Observational bias/blinding - Did the study implement measures to reduce observational bias?	For each outcome or grouping of outcomes in a study: • Does the study report blinding or other methods/procedures for reducing observational bias? • If not, did the study use a design or approach for which such procedures can be inferred? • What is the expected impact of failure to implement (or report implementation) of these methods/procedures on results?
	4. Confounding - Are variables with the potential to confound or modify results controlled for and consistent across all experimental groups?	For each study: • Are there differences across the treatment groups (e.g., co-exposures, vehicle, diet, palatability, husbandry, health status, etc.) that could bias the results? • If differences are identified, to what extent are they expected to impact the results?
	5. Selective reporting and attrition - Did the study report results for all prespecified outcomes and tested animals?	For each study: *Selective reporting bias:* • Are results presented for all outcomes described in the methods? *Attrition bias:* • Are all animals accounted for in the results? • If there are discrepancies, do authors provide an explanation (e.g., death or unscheduled sacrifice during the study)?

Key concept	Domain - Core Question	Prompting questions
		• If results omissions and/or attrition are identified, what is the expected impact on the interpretation of the results?
Sensitivity	**6. Chemical administration and characterization** - Did the study adequately characterize exposure to the chemical of interest and the exposure administration methods?	For each study: • Does the study report the source and purity and/or composition (e.g., identity and percent distribution of different isomers) of the chemical? If not, can the purity and/or composition be obtained from the supplier (e.g., as reported on the website) • Was independent analytical verification of the test article purity and composition performed? • Did the authors take steps to ensure the reported exposure levels were accurate? – For inhalation studies: were target concentrations confirmed using reliable analytical measurements in chamber air? – For oral studies: if necessary, based on consideration of chemical-specific knowledge (e.g., instability in solution; volatility) and/or exposure design (e.g., the frequency and duration of exposure), were chemical concentrations in the dosing solutions or diet analytically confirmed? • Are there concerns about the methods used to administer the chemical (e.g., inhalation chamber type, gavage volume, etc.)?
	7. Exposure timing, frequency, and duration - Was the timing, frequency, and duration of exposure sensitive for the outcome(s) of interest?	For each outcome or grouping of outcomes in a study: • Does the exposure period include the critical window of sensitivity (if known)? • Was the duration and frequency of exposure sensitive for detecting the outcome of interest?
	8. Sensitivity and specificity - Are the procedures sensitive and specific for evaluating the outcome(s) of interest?	For each outcome or grouping of outcomes in a study: • Are there concerns regarding the specificity and validity of the protocols? • Are there serious concerns regarding the sample size? • Are there concerns regarding the timing of the outcome assessment?
	9. Results presentation - Are the results presented in a way that makes the data usable and transparent?	For each outcome or grouping of outcomes in a study: • Does the level of detail allow for an informed interpretation of the results? • Are the data compared or presented in a way that is inappropriate or misleading?
Overall Confidence	**10. Overall Confidence** - Considering the identified strengths and limitations, what is the overall confidence rating for the outcome(s) of interest?	For each outcome or grouping of outcomes in a study: • Were concerns (i.e., limitations or uncertainties) related to the reporting quality, risk of bias, or sensitivity identified? • If yes, what is their expected impact on the overall reliability and validity of the study results, including (when possible) interpretations of impacts on the magnitude or direction of the reported effects?