This draft document is being disseminated to obtain public comment. It does not represent and should not be construed to represent final NTP determination or policy.

Appendix 2: Guidance for Assessing Risk of Bias in the BPA-Obesity Systematic Review

This risk of bias rating assessment tool for human and animal studies was developed based on the most recent guidance from the Agency for Healthcare Research and Quality (Viswanathan et al. 2012), Cochrane Handbook (Higgins and Green 2011), CLARITY Group at McMaster University (CLARITY Group at McMaster University), consultation with technical advisors (NTP 2013), staff at other federal agencies, and other sources (Downs and Black 1998; Dwan et al. 2010; Genaidy et al. 2007; Johnson et al. 2013; Koustas et al. 2013; Krauth et al. 2013; Shamliyan et al. 2010; Shamliyan et al. 2011; Wells et al.). For each study, risk of bias is assessed at the outcome level because certain aspects of study design and conduct may increase risk of bias for some outcomes and not others within the same study.

How this tool is structured:

• 14 Risk of Bias Questions • Applicability of each item to animal studies, controlled exposure human studies, and observational human studies • Each rated by 4 possible answers (see below) • The document presents the complete set of instructions. However, during the evaluations we will use web-based forms to collect risk of bias ratings and the reviewer will only see the question and instruction text that is relevant to the study under review (i.e., text related to evaluating human studies will not appear during the evaluation of an animal study, text that is only relevant to evaluating a human controlled trial will not appear during evaluation of a cross-sectional study)	**Study Type Abbreviations:** **EA:** Experimental Animal **HCT:** Human Controlled Trial[1] **Co:** Cohort **CaCo:** Case-Control **CrSe:** Cross-sectional **CaS:** Case Series/Case report

General Question Format:

- Definition of the general category of bias
- Clarifying text to explain what study aspects are relevant
- Summary of available empirical information about the direction and magnitude of the bias
- Support for including item based on risk of bias guidance or other internal validity tools.

[1] Human controlled trial study design used here refers to studies in humans with a controlled exposure including randomized controlled trials and non-randomized experimental studies.

General Answer Format:

++ ***Definitely Low*** *risk of bias*:
There is direct evidence of low risk of bias practices in the form of an explicit statement from the study report or through contacting the authors

\+ ***Probably Low*** *risk of bias*:
Low risk of bias practice can be inferred from study report ("indirect evidence") **OR** it is deemed by the risk of bias evaluator that deviations from definitely low risk of bias practices would not appreciably bias results, including consideration of direction and magnitude of bias.

– ***Probably High*** *risk of bias*:
There is indirect evidence of high risk of bias practices **OR** there is insufficient information provided about relevant risk of bias practices to infer.

–– ***Definitely High*** *risk of bias*:
There is direct evidence of high risk of bias practices

Risk of bias versus study applicability: Timing of exposure and health outcome assessment

Risk of bias evaluates internal validity: "Are the results of the study credible?" The issue of timing and duration of exposure in relation to health outcome assessment in most cases is an issue of applicability: "Did the study design address the topic of the evaluation?" However, there may be instances where it is best considered as part of risk of bias. For example, if there are differences in the duration of follow-up across study groups, this would be a source of bias considered under detection bias "Can we be confident in the outcome assessment?" If the duration of follow-up was not optimal for the development of the outcome of interest (e.g., short duration of time between exposure and health outcome assessment for chronic disease), then it would be considered under applicability. Ideally, windows of exposure and health outcome assessment that not considered relevant to an evaluation would be considered in determining study eligibility criteria in Step 1.

SELECTION BIAS

Selection bias refers to systematic differences between baseline characteristics of the groups that are compared (Higgins and Green 2011).

1. Was administered dose or exposure level adequately randomized?

Randomization requires that each human subject or animal had an equal chance of being assigned to any study group including controls (e.g., use of random number table or computer generated randomization).

A lack of randomization will bias results away from the null towards larger effect sizes. This effect has been empirically assessed in both controlled human trials (reviewed in Higgins and Green 2011) and experimental animals (reviewed in Krauth et al. 2013).

This item is widely recommended to assess risk of bias for controlled human trials (Guyatt et al. 2011; Higgins and Green 2011; IOM 2011; Viswanathan et al. 2012) and is included in most risk of bias instruments for animal studies (reviewed in Krauth et al. 2013).

We recognize that given reporting practices for animal studies it is unlikely that the method of randomization will be explicitly reported in most studies. Thus, we will assume in cases where the randomization method is unknown (i.e., not reported and cannot be obtained through author query) that randomization was not undertaken and classify such studies as "probably high risk of bias".

Applies to: HCT, EA

Definitely Low risk of bias:

HCT: There is direct evidence that subjects were allocated to any study group including controls using a method with a random component. Acceptable methods of randomization include: referring to a random number table, using a computer random number generator, coin tossing, shuffling cards or envelopes, throwing dice, or drawing of lots (Higgins and Green 2011). Restricted randomization (e.g., blocked randomization) to ensure particular allocation ratios will be considered low risk of bias. Similarly, stratified randomization and minimization approaches that attempt to minimize imbalance between groups on important factors prognostic factors (e.g., body weight) will be considered acceptable.

EA: There is direct evidence that animals were allocated to any study group including controls using a method with a random component. Acceptable methods of randomization include: referring to a random number table, using a computer random number generator, coin tossing, shuffling cards or envelopes, throwing dice, or drawing of lots (Higgins and Green 2011). Restricted randomization (e.g., blocked randomization) to ensure particular allocation ratios will be considered low risk of bias. Similarly, stratified randomization and minimization approaches that attempt to minimize imbalance between groups on important factors prognostic factors (e.g., body weight) will be considered acceptable. This type of approach is used by NTP and included in OECD guidelines for toxicology protocols, i.e., random number generator with body weight as a covariate such that body weight is consistent across study groups. Discrimination criteria applied prior to randomization across study groups (e.g., only female rats displaying normal estrus cycles in the prior 3 months were included; rats were then randomly assigned to study groups using a random number table) will also be considered acceptable. Investigator-selection of animals from a cage is not considered random allocation because animals may not have an equal chance of being selected, e.g., investigator selecting animals with this method may inadvertently choose healthier, easier to catch, or less aggressive animals. Use of concurrent controls is required as an indication that randomization covered all study groups.

Probably Low risk of bias:

HCT: There is indirect evidence that subjects were allocated to study groups using a method with a random component (i.e., authors state that allocation was random, without description of the method used) **OR** it is deemed that allocation without a clearly random component during the study would not appreciably bias results. For example, approaches such as biased coin or urn randomization, replacement randomization, mixed randomization, and maximal randomization may require consultation with a statistician to determine risk of bias rating (Higgins and Green 2011).

EA: There is indirect evidence that animals were allocated to study groups using a method with a random component (i.e., authors state that allocation was random, without description of the

method used) **OR** it is deemed that allocation without a clearly random component during the study would not appreciably bias results. For example, approaches such as biased coin or urn randomization, replacement randomization, mixed randomization, and maximal randomization may require consultation with a statistician to determine risk-of-bias rating (Higgins and Green 2011). Use of concurrent controls is required as an indication that randomization covered all study groups.

Probably High risk of bias:

HCT: There is indirect evidence that subjects were allocated to study groups using a method with a non-random component **OR** there is insufficient information provided about how subjects were allocated to study groups. Non-random allocation methods may be systematic, but have the potential to allow participants or researchers to anticipate the allocation to study groups. Such "quasi-random" methods include alternation, assignment based on date of birth, case record number, or date of presentation to study (Higgins and Green 2011).

EA: There is indirect evidence that animals were allocated to study groups using a method with a non-random component **OR** there is insufficient information provided about how subjects were allocated to study groups. Non-random allocation methods may be systematic, but have the potential to allow researchers to anticipate the allocation of animals to study groups (Higgins and Green 2011). Such "quasi-random" methods include investigator-selection of animals from a cage, alternation, assignment based on shipment receipt date, date of birth, or animal number. A study reporting lack of concurrent controls is another indication that randomization to all study groups was not conducted.

Definitely High risk of bias:

HCT: There is direct evidence that subjects were allocated to study groups using a non-random method including judgment of the clinician, preference of the participant, the results of a laboratory test or a series of tests, or availability of the intervention (Higgins and Green 2011).

EA: There is direct evidence that animals were allocated to study groups using a non-random method including judgment of the investigator, the results of a laboratory test or a series of tests (Higgins and Green 2011). A study reporting lack of concurrent controls is another indication that randomization to all study groups was not conducted.

2. Was allocation to study groups adequately concealed?

Allocation concealment requires that research personnel not know which administered dose or exposure level is assigned at the start of a study. Human studies also require that allocation be concealed from human subjects prior to entering the study.

A lack of allocation concealment will bias results away from the null towards larger effect sizes. This effect has been empirically assessed in both controlled human trials [(Pildal et al. 2007; Schulz et al. 2002; Schulz et al. 1995); see also studies reviewed in (Higgins and Green 2011)] and in a systematic review of animal studies by Macleod et al. (2008) to evaluate the efficacy of NXY-059 in experimental focal cerebral ischemia (see also studies reviewed in Krauth et al. 2013).

This item is widely recommended to assess risk of bias for controlled human trials (Guyatt et al. 2011; Higgins and Green 2011; IOM 2011; Viswanathan et al. 2012) and included in some risk of bias instruments for animal studies (reviewed in Krauth et al. 2013).

Note there are separate risk of bias issues affected by concealment or blinding that are important for selection, performance and detection bias: 1) a question under performance bias addresses blinding of

research personnel and human subjects to study groups during the study; and 2) a question under detection bias addresses blinding during outcome assessment.

Applies to: HCT, EA

Definitely Low risk of bias:

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

EA: There is direct evidence that at the time of assigning study groups the research personnel did not know what group animals were allocated to, and it is unlikely that they could have broken the blinding of allocation until after assignment was complete and irrevocable. Methods used to ensure allocation concealment include sequentially numbered treatment containers of identical appearance or equivalent methods.

Probably Low risk of bias:

HCT: There is indirect evidence that the research personnel and subjects did not know what study group subjects were allocated to **OR** it is deemed that lack of adequate allocation concealment would not appreciably bias results.

EA: There is indirect evidence that at the time of assigning study groups the research personnel did not know what group animals were allocated to **OR** it is deemed that lack of adequate allocation concealment would not appreciably bias results.

Probably High risk of bias:

HCT: There is indirect evidence that at the time of recruitment it was possible for the research personnel and subjects to know what study group subjects were allocated to, or it is likely that they could have broken the blinding of allocation before recruitment was complete and irrevocable **OR** there is insufficient information provided about allocation to study groups.

Note: Inadequate methods include using an open random allocation schedule (e.g., a list of random numbers), assignment envelopes used without appropriate safeguards (e.g., if envelopes were unsealed or non-opaque or not sequentially numbered), alternation or rotation; date of birth; case record number; or any other explicitly unconcealed procedure. For example, if the use of assignment envelopes is described, but it remains unclear whether envelopes were sequentially numbered, opaque and sealed.

EA: There is indirect evidence that at the time of assigning study groups it was possible for the research personnel to know what group animals were allocated to, or it is likely that they could have broken the blinding of allocation before assignment was complete and irrevocable **OR** there is insufficient information provided about allocation to study groups.

Definitely High risk of bias:

HCT: There is direct evidence that at the time of recruitment it was possible for the research personnel and subjects to know what study group subjects were allocated to, or it is likely that they could have broken the blinding of allocation before recruitment was complete and irrevocable.

EA: There is direct evidence that at the time of assigning study groups it was possible for the research personnel to know what group animals were allocated to, or it is likely that they could have broken the blinding of allocation before assignment was complete and irrevocable.

3. Were the comparison groups appropriate?

Comparison group appropriateness refers to having similar baseline characteristics between groups aside from the exposures and outcomes under study.

Assessment of appropriate selection of comparison groups is a widely used element of tools to assess study quality for observational human studies (CLARITY Group at McMaster University 2013; Downs and Black 1998; Shamliyan et al. 2010; Viswanathan et al. 2012; Wells et al.). This addresses whether exposed and unexposed subjects were recruited from the same populations in cohort studies and consideration of appropriate selection of cases and controls in case-control studies.

The direction of the bias (towards or away from the null) will differ based on the nature of differences between comparison groups and may be difficult to predict.

For example, in occupational cohorts workers have lower rates of disease and mortality than the general population – the healthy worker effect – because the severely ill and chronically disabled are commonly excluded from employment (Gerstman 1998). Therefore, comparing workers to an inherently less healthy group (general population or workers with less physically demanding work) will bias the estimate of disease risk towards the null (Rothman 1986). Conversely if cases of disease identified from a screening program were compared to controls from the general population, the effect estimate could be overestimated as those being screened may inherently have a higher risk (e.g., family history) so the better comparison group would be subjects screened as not having disease (Szklo and Nieto 2007).

Applies to: Co, CaCo, CrSe

Definitely Low risk of bias:

Co, CrSe: There is direct evidence that subjects (both exposed and non-exposed) were similar (e.g., recruited from the same eligible population, recruited with the same method of ascertainment using the same inclusion and exclusion criteria, and were of similar age and health status), recruited within the same time frame, and had the similar participation/response rates.

CaCo: There is direct evidence that cases and controls were similar (e.g., recruited from the same eligible population including being of similar age, gender, ethnicity, and eligibility criteria other than outcome of interest as appropriate), recruited within the same time frame, and controls are described as having no history of the outcome. Note: A study will be considered low risk of bias if baseline characteristics of groups differed but these differences were considered as potential confounding or stratification variables (see question #4).

Probably Low risk of bias:

Co, CrSe: There is indirect evidence that subjects (both exposed and non-exposed) were similar (e.g., recruited from the same eligible population, recruited with the same method of ascertainment using the same inclusion and exclusion criteria, and were of similar age and health status), recruited within the same time frame, and had the similar participation/response rates **OR** differences between groups would not appreciably bias results.

CaCo: There is indirect evidence that cases and controls were similar (e.g., recruited from the same eligible population, recruited with the same method of ascertainment using the same inclusion

and exclusion criteria, and were of similar age), recruited within the same time frame, and controls are described as having no history of the outcome **OR** differences between cases and controls would not appreciably bias results.

Probably High risk of bias:

Co, CrSe: There is indirect evidence that subjects (both exposed and non-exposed) were not similar, recruited within very different time frames, or had the very different participation/response rates **OR** there is insufficient information provided about the comparison group including a different rate of non-response without an explanation.

CaCo: There is direct evidence that controls were drawn from a very dissimilar population than cases or recruited within very different time frames **OR** there is insufficient information provided about the appropriateness of controls including rate of response reported for cases only.

Definitely High risk of bias:

Co, CrSe: There is direct evidence that subjects (both exposed and non-exposed) were not similar, recruited within very different time frames, or had very different participation/response rates.

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

CONFOUNDING BIAS

The heading of confounding bias brings together consideration of systematic differences between risk factors and other characteristics of the groups that are compared that may reduce or increase the observed effect (IOM 2011). While epidemiologists working on environmental health questions may commonly consider confounding as a separate area of study quality, other methods such as the AHRQ risk of bias tool (Viswanathan *et al.* 2012) consider these risk of bias issues under multiple domains such as selection bias and performance bias. We have developed a separate category for confounding bias for this risk of bias OHAT tool given the importance of human observational studies for addressing environmental health questions and the importance of evaluating confounding for these studies.

4. Did the study design or analysis account for important confounding and modifying variables?

Interpretation of study findings may be distorted by failure to consider the extent to which systematic differences in baseline characteristics risk factors, prognostic variables[2], or co-occurring exposures among comparison groups may reduce or increase the observed effect (IOM 2011). Appropriate methods to account for these differences would include multivariable analysis, stratification, matching of cases and controls, or other approaches.

[2] "Risk" factors are those which as associated with causing a condition (like smoking for lung cancer or being born premature for chronic lung disease). 'Prognostic' factors are those which, in people who have the condition, influence the outcome (like resectability of tumor for lung cancer or duration of intubation for CLD). Risk factors are determined by looking at things that influence new cases ('incident' ones), while prognostic factors can only be determined by following up people who already have the disease (http://blogs.bmj.com/adc-archimedes/2009/03/09/risk-vs-prognostic-factors/) .

Note: a parallel question under detection bias addresses reliability of the measurement of confounding or modifying variables.

Human text: This item is commonly included in tools used to assess the quality of observational studies (CLARITY Group at McMaster University 2013; Shamliyan et al. 2010) and recent guidance from AHRQ also recommends consideration for controlled human trials ("randomized clinical trials") (Viswanathan et al. 2012).

The direction of the bias (towards or away from the null) will differ based on the nature of differences between comparison groups. Generally, confounding results in effect sizes that are overestimated. However, confounding factors can lead to an underestimation of the effect of a treatment or exposure, particularly in observational studies. In other words, if the confounding variables were not present, the measured effect would have been even larger (IOM 2011).

We developed a list of potential confounders for the BPA and obesity project by using information from NHANES biomonitoring data (Calafat et al. 2008) and statistics on overweight and obesity (CDC 2012) to develop a directed acyclic graph (DAG). Factors that could potentially cause or help predict BPA exposure level and obesity status, based on what is known in the literature, were considered "key" potential confounders for assessing risk of bias. **Sex, age, race/ethnicity, and socioeconomic status** were all associated with BPA and obesity. More specifically, in NHANES 2003-2004 least square geometric mean urinary concentrations of total BPA differed by race/ethnicity (non-Hispanic whites and non-Hispanic blacks > Mexican Americans), age (children>adolescents>adults), sex (women>men), and household income (low household income>high household income) (Calafat et al. 2008). Smoking status in adults was not significantly associated with BPA concentrations. With respect to obesity, prevalence rates are higher in adults compared to children and non-Hispanic blacks have higher age-adjusted rates of obesity compared with Mexican Americans, all Hispanics, and non-Hispanic whites (CDC 2012). The relationship between obesity and socioeconomic status differs by sex and race/ethnicity. In addition, consumption of canned and packaged foods are associated with higher BPA exposure (Braun et al. 2011; Cao et al. 2011; Rudel et al. 2011) and consumption of energy dense, low-nutrient food is associated with obesity (IOM 2006). Many canned food products are lined with epoxy resins that contain BPA but paper and paperboard products used in contact with food can also contain BPA (Ozaki et al. 2006; Ozaki et al. 2004). So, **consumption of canned or packaged food and drink ("processed" food) that is also energy dense and low-nutrient (e.g., soda)** could also be an important potential confounder to consider. Caloric intake and television watching were not associated with BPA concentrations in children (Trasande et al. 2012)

Animal text: This item is not typically included in study quality tools for animal studies (Krauth et al. 2013), but is considered applicable in this current tool because recent guidance from AHRQ recommends consideration for the analogous human study design, randomized clinical trials (Viswanathan et al. 2012). In the context of an animal study, this element would include consideration of covariates such as body weight, litter size, or other outcome specific covariates.

Applies to: HCT, EA, Co, CaCo, CrSe, CaS

Definitely Low risk of bias:

HCT, Co, CrSe, CaS: There is direct evidence that appropriate adjustments or explicit considerations were made for primary covariates and confounders in the final analyses through the use of statistical models to reduce research-specific bias including standardization, case matching, adjustment in multivariate model, stratification, propensity scoring, or other methods were

appropriately justified. Acceptable consideration of appropriate adjustment factors includes cases when the factor is not included in the final adjustment model because the author conducted analyses that indicated it did not need to be included.

EA: There is direct evidence that appropriate adjustments were made for body weight, litter size in studies of offspring (especially when the outcome measure is growth-related and assessed prior to weaning) or any other relevant covariates.

CaCo: There is direct evidence that appropriate adjustments were made for primary covariates and confounders in the final analyses through the use of statistical models to reduce research specific bias including standardization, matching of cases and controls, adjustment in multivariate model, stratification, propensity scoring, or other methods were appropriately justified.

Probably Low risk of bias:

HCT, Co, CaCo, CrSe, CaS: There is indirect evidence that appropriate adjustments were made for most primary covariates and confounders **OR** it is deemed that not considering or only considering a partial list of covariates or confounders in the final analyses would not appreciably bias results.

EA: There is indirect evidence that appropriate adjustments were made for body weight, litter size in studies of offspring (especially when the outcome measure is growth-related and assessed prior to weaning), or any other relevant covariates **OR** it is deemed that not considering or only considering a partial list of covariates or confounders in the final analyses would not appreciably bias results.

Probably High risk of bias:

HCT, Co, CrSe, CaS: There is indirect evidence that the distribution of primary covariates and known confounders differed between the groups and was not appropriately adjusted for in the final analyses **OR** there is insufficient information provided about the distribution of known confounders.

EA: There is indirect evidence that appropriate adjustments were not made for body weight, litter size in studies of offspring (especially when the outcome measure is growth-related and assessed prior to weaning), or any other relevant covariates **OR** there is insufficient information provided about analysis of relevant covariates.

CaCo: There is indirect evidence that the distribution of primary covariates and known confounders differed between cases and controls and was not investigated further **OR** there is insufficient information provided about the distribution of known confounders in cases and controls.

Definitely High risk of bias:

HCT, Co, CrSe, CaS: There is direct evidence that the distribution of primary covariates and known confounders differed between the groups, confounding was demonstrated, and was not appropriately adjusted for in the final analyses.

EA: There is direct evidence that appropriate adjustments were not made for body weight, litter size in studies of offspring (especially when the outcome measure is growth-related and assessed prior to weaning), or any other relevant covariates.

CaCo: There is direct evidence that the distribution of primary covariates and known confounders differed between cases and controls, confounding was demonstrated, but was not appropriately adjusted for in the final analyses.

PERFORMANCE BIAS

Human text: Performance bias refers to systematic differences in the care provided to participants and protocol deviation. Examples include contamination of the control group with the exposure or intervention, unbalanced provision of additional interventions or co-interventions, difference in co-interventions, and inadequate blinding of providers and participants (Viswanathan et al. 2012).

Animal text: Performance bias refers to systematic differences in the care provided to animals and protocol deviation. Examples include contamination of the control group with the exposure or intervention, unbalanced provision of additional interventions or co-interventions, and inadequate blinding of research personnel to the animal's study group (Sena et al. 2007).

5. Did researchers adjust or control for other exposures that are anticipated to bias results?

Human introductory text. This risk of bias element is often included in tools or guidance developed to assess the quality of observational studies (CLARITY Group at McMaster University 2013; Shamliyan et al. 2010; Viswanathan et al. 2012). Recent guidance from AHRQ also considers it relevant to randomized clinical trials (Viswanathan et al. 2012).

The direction of the bias (towards or away from the null) will differ based on the nature of unintended exposure. For example, in a human study if the exposed group lives at a Superfund site they may be exposed to high levels of other environmental contaminants that, if not accounted for, may bias results away from the null (towards larger effects sizes).

It is understood in environmental health that people are exposed to complex mixtures of environmental contaminants and other types of exposures that make it difficult to establish chemical-specific associations. Thus, we will not penalize studies if other exposures are not adjusted or controlled for in most cases. For some projects exceptions may include studies where levels of other chemicals aside from the chemical of interest are likely to be high, such as in occupational cohorts or contaminated regions (e.g., Superfund sites). For some health outcomes, consideration of additional therapies, including medications, may also be appropriate.

Animal introductory text. This risk of bias element is often included in tools or guidance developed to assess the quality of human observational studies (CLARITY Group at McMaster University 2013; Shamliyan et al. 2010; Viswanathan et al. 2012). Recent guidance from AHRQ also considers it relevant to human randomized clinical trials (Viswanathan et al. 2012). Experimental animal studies also need to consider the impact of inadvertent chemical or biological co-exposures. For example, if exposure is to bisphenol A or other chemical with estrogenic properties, husbandry practices that raise the background level of estrogenicity may make the model system less sensitive to detect low-dose effects of BPA, including use of a diet high in phytoestrogens (Muhlhauser et al. 2009; Thigpen et al. 2007). In this case, the direction of the bias would be towards the null (towards smaller effect sizes). Infectious agents and non-treatment related co-morbidity should also be monitored as potential sources of bias.

The direction of the bias will depend on the nature of differences between the groups. For example, certain types of infections may be related to outcomes of interest (Baker 1998; GV-SOLAS 1999; NRC 1991). *Helicobacter hepaticus* is a bacterial carcinogen and can cause hepatitis, hepatocellular carcinoma, and proliferative typhlocolitis in rodents (Hailey et al. 1998; Kusters et al. 2006). If the infection occurs in control animals, then the bias for an effect on the liver may be towards the null (smaller effect size). If the infection occurs in treated animals, then the bias for an effect on the liver may be away from the null (larger effect size).

Applies to: HCT, EA, Co, CaCo, CrSe, CaS

Definitely Low risk of bias:

HCT: There is direct evidence that other exposures anticipated to bias results were not present or were appropriately adjusted for.

Co, CaCo, CrSe, CaS: There is direct evidence that other exposures anticipated to bias results were not present or were appropriately adjusted for. For occupational studies or studies of contaminated sites, other chemical exposures known to be associated with those settings were appropriately considered.

EA: There is direct evidence that other exposures anticipated to bias results were not present or were appropriately adjusted for. For estrogenic exposures or endpoints anticipated to be affected by estrogenic or endocrine pathways, this would include if animals were fed a phytoestrogen-free or low phytoestrogen diet.

Probably Low risk of bias:

HCT, Co, CaCo, CrSe, CaS: There is indirect evidence that other co-exposures anticipated to bias results were not present or were appropriately adjusted for **OR** it is deemed that co-exposures present would not appreciably bias results. Note, as discussed above, this includes insufficient information provided on co-exposures in general population studies.

EA: There is indirect evidence that other exposures anticipated to bias results were not present or were appropriately adjusted for **OR** it is deemed that co-exposures present would not appreciably bias results.

Probably High risk of bias:

HCT: There is indirect evidence that the control group may have received the treatment or there was an unbalanced provision of additional co-exposures which were not appropriately adjusted for.

EA: There is indirect evidence that the control group may have received the treatment or there was an unbalanced provision of additional co-exposures which were not appropriately adjusted for. For estrogenic exposures or endpoints anticipated to be affected by estrogenic or endocrine pathways, this would include if animals were likely fed a diet that did not minimize or eliminate phytoestrogen content (or phytoestrogen content of diet was not reported).

Co, CrSe, CaS: There is indirect evidence that there was an unbalanced provision of additional co-exposures across the primary study groups, which were not appropriately adjusted for **OR** there is insufficient information provided about co-exposures in occupational studies or studies of contaminated sites where high exposures to other chemical exposures would have been reasonably anticipated.

CaCo: There is indirect evidence that there was an unbalanced provision of additional co-exposures across cases and controls, which were not appropriately adjusted for **OR** there is insufficient information provided about co-exposures in occupational studies or studies of contaminated sites where high exposures to other chemical exposures would have been reasonably anticipated.

Definitely High risk of bias:

HCT: There is direct evidence that the control group received the treatment or there was an unbalanced provision of additional co-exposures which were not appropriately adjusted for.

EA: There is direct evidence that the control group received the treatment or there was an unbalanced provision of additional co-exposures which were not appropriately adjusted for. For estrogenic exposures or endpoints anticipated to be affected by estrogenic or endocrine pathways, this would include that animals were fed a diet that did not minimize or eliminate phytoestrogen content.

Co, CrSe, CaS: There is direct evidence that there was an unbalanced provision of additional co-exposures across the primary study groups, which were not appropriately adjusted for.

CaCo: There is direct evidence that there was an unbalanced provision of additional co-exposures across cases and controls, which were not appropriately adjusted for.

6. Were experimental conditions identical across study groups?

Housing conditions and husbandry practices should be identical across control and experimental groups because these variables may impact the outcome of interest (Duke et al. 2001; Gerdin et al. 2012). Identical conditions include use of the same vehicle in control and experimental animals. This risk of bias element is included in some tools used to assess animal studies (Krauth et al. 2013).

We recognize that given reporting practices it is unlikely that similarity of conditions will be explicitly reported in most studies. Thus, we will assume unless stated otherwise that experimental conditions (other than use of appropriate vehicle for control animals) were identical across groups which will result in most studies considered "probably low risk of bias". In the short-term this risk of bias item is unlikely to be informative for the purposes of discriminating between studies of higher quality and studies of lower quality. However, in the long-term, especially if reporting standards improve, collecting this information may generate data that will allow us to empirically assess evidence of bias or to remove this risk of bias question from consideration. Note, that the use of appropriate vehicle for control animals will remain an important risk of bias consideration.

Applies to: EA

Definitely Low risk of bias:

EA: There is direct evidence that non-treatment-related experimental conditions were identical across study groups (i.e., the study report explicitly provides this level of detail) and the same vehicle was used in control and experimental animals.

Probably Low risk of bias:

EA: There is indirect evidence that the same vehicle was used in control and experimental animals **OR** it is deemed that the vehicle used would not appreciably bias results. As described above, identical non-treatment-related experimental conditions are assumed if authors did not report differences in housing or husbandry.

Probably High risk of bias:

EA: There is indirect evidence that the vehicle differed between control and experimental animals **OR** authors did not report the vehicle used.

Definitely High risk of bias:

EA: There is direct evidence from the study report that non-treatment-related experimental conditions were not comparable between study groups or control animals were untreated, or treated with a different vehicle than experimental animals.

7. Did deviations from the study protocol impact the results?

Failure of the study to maintain fidelity to the protocol is recommended as an important consideration when assessing performance bias (IOM 2011; Viswanathan et al. 2012). However, it will likely be difficult to assess with confidence for most studies, particularly when the methods section of a publication is all that is available. In some instances the protocol is meant to be "fluid" and the protocol explicitly allows for modification based on need; such fluidity does not mean the interventions are implemented incorrectly. The deviation may not result in a risk of bias, or if it does the direction of the bias (towards or away from the null) will differ based on the deviation from the protocol.

We recognize that given reporting practices it is unlikely that deviations from the protocol will be explicitly reported in most studies. Thus, we will assume unless stated otherwise that no deviations occurred which will result in most studies considered "probably low risk of bias". In the short-term this risk of bias item is unlikely to be informative for the purposes of discriminating between studies of higher quality and studies of lower quality. However, in the long-term, especially if reporting standards improve, collecting this information may generate data that will allow us to empirically assess evidence of bias or to remove this risk of bias question from consideration.

Animal introductory text. One of the more common deviations from protocol that can occur in toxicity studies is when a dose level is decreased based on evidence of mortality or severe toxicity. Documentation of this change could be reflected as an amended protocol. However, depending upon how the author addresses this change it may or may not impact results. For example, when this occurs in NTP studies, the usual analysis would be conducted on the dose groups remaining after the toxic dose level is dropped. A similar situation arises when a dose group has to be euthanized due to overt toxicity. Other deviations such as inconsistencies or mistakes in following the protocol suggest a greater risk of bias (e.g., animals receiving the wrong treatment).

Applies to: HCT, EA, Co, CaCo, CrSe, CaS

Definitely Low risk of bias:

HCT, Co, CaCo, CrSe, CaS, EA: There is direct evidence that there were no deviations from the protocol (i.e., the study report explicitly provides this level of detail).

Probably Low risk of bias:

HCT, Co, CaCo, CrSe, CaS, EA: There is indirect evidence that there were no deviations from the protocol (i.e., authors did not report any deviations) **OR** deviations from the protocol are described and it is deemed that they would not appreciably bias results.

Probably High risk of bias:

HCT, Co, CaCo, CrSe, CaS, EA: There is indirect evidence that there were large deviations from the protocol as outlined in the methods or study report.

Definitely High risk of bias:

HCT, Co, CaCo, CrSe, CaS, EA: There is direct evidence that there were large deviations from the protocol as outlined in the methods or study report.

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

Blinding requires that research personnel do not know which administered dose or exposure level the human subject or animal is being given (i.e., study group). Human studies also require blinding of the human subjects when possible.

Human introductory text. If research personnel or human subjects are not blinded to the study groups it could affect the actual outcomes of the participants due to differential behaviors across intervention groups. During the course of a study blinding of participants and research personnel is a recommended risk of bias element in the most recent Cochrane guidance for assessing randomized clinical trials (Higgins and Green 2011).

There is no empirical evidence of bias due to failure to blind during the course of a study is currently available. However, 'blind' or 'double-blind' study descriptions usually includes blinding of research personnel, human subjects, or both. Without distinguishing between the different stages of blinding during the conduct of a study, lack of blinding in randomized trials has been empirically shown to be associated with larger estimations of intervention effects (on average a 9% increase in an odds ratio) (Pildal et al. 2007). Schulz et al. (1995) analyzed 250 controlled trials and found that studies that were not double-blinded had a 17% larger estimation of treatment effect, on average. If additional investigations or co-interventions occur differentially across intervention groups, bias can also be introduced by not blinding research personnel or human subjects.

For some exposures, it is not possible to entirely blind research personnel and subjects during the course of the study (an exercise intervention or patients receiving surgery). However, adherence to a strict study protocol to minimize differential behaviors by research personnel and human subjects can reduce the risk of bias. In practice, successful blinding cannot be ensured, as it can be compromised for most interventions. In some cases the treatment may have side effects possibly allowing the participant to detect which intervention they received, unless the study compares interventions with similar side effects or uses an active placebo (Boutron et al. 2006).

Animal introductory text. Lack of blinding of research personnel could bias the results by affecting the actual outcomes of the animals in the study. This may be due to differences in handling of animals (e.g., stress-related effects) or monitoring for health outcomes. For example, an investigator may be more likely to take measures to ensure that animals in experimental groups receive the appropriate dose volume compared to animals in the control group. Lack of blinding might also lead to bias caused by additional investigations or co-interventions regardless of the type of outcomes, if these occur differentially across intervention groups (Higgins and Green 2011).

This item is recommended to assess performance bias for controlled human trials (Higgins and Green 2011) and animal studies (reviewed in Krauth et al. 2013), although empirical evidence of bias due to lack of blinding of research personnel during the course of the study is not currently available. Rosenthal and Lawson (1964) reported that rats that experimenters had been told were "bright" performed better than rats labeled "dull" in Skinner box learning tests, despite the fact that they were the same rats. The study design did not allow clear separation between experimenter bias introduced during handling or training from bias at outcome assessment. As discussed under detection bias, lack of blinding at outcome assessment in animal studies is associated with larger measures of the effect (Bebarta et al. 2003; Sena et al. 2007; Vesterinen et al. 2010).

In animal studies, blinding of study group during the course of the study is often not possible for animal welfare considerations and the need to determine if treated animals are affected relative to controls in a treatment or dose-dependent manner (examples include clinical observations and histopathologic

assessment of non-neoplastic lesions). However, animal studies are in general more tightly controlled than human studies and additional measures may be taken to reduce the risk of bias, such as the generation and use of standard operating procedures, training, and randomized husbandry or handling practices (e.g., placement in the animal room, necropsy order, etc.).

We recognize that given animal welfare practices it is unlikely that blinding of research personnel during the course of a study can be fully achieved. Given the lack of empirical evidence to directly assess this potential source of bias, animal studies that do not report blinding and studies that report practices designed to reduce potential risk of bias such as randomized necropsy order will be categorized as "probably low risk of bias". In the short-term this risk of bias item is unlikely to be informative for the purposes of discriminating between studies of higher quality and studies of lower quality. However, in the long-term collecting this information may generate data that will allow us to empirically assess evidence of bias.

Note there are separate risk of bias issues affected by concealment or blinding that are important for selection, performance and detection bias: 1) a question under selection bias addresses allocation concealment of research personnel and human subjects; and 2) a question under detection bias addresses blinding during outcome assessment.

Applies to: HCT, EA

Definitely Low risk of bias:

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

EA: There is direct or indirect evidence that the research personnel were adequately blinded to study group, and it is unlikely that they could have broken the blinding during the study. Methods used to ensure blinding include central allocation, sequentially numbered drug containers of identical appearance; sequentially numbered animal cages; or equivalent methods.

Probably Low risk of bias:

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

EA: Blinding was not reported **OR** blinding was not possible but research personnel took steps to minimize potential bias, such as randomized necropsy order.

Probably High risk of bias:

HCT: There is indirect evidence that it was possible for research personnel or subjects to infer the study group, **OR** there is insufficient information provided about blinding of study group. Inadequate methods include using an open random allocation schedule (e.g., a list of random numbers), assignment envelopes used without appropriate safeguards (e.g., if envelopes were unsealed or non-opaque or not sequentially numbered), alternation or rotation; date of birth; case record number; or any other explicitly unconcealed procedure. For example, if the use of assignment envelopes is described, but it remains unclear whether envelopes were sequentially numbered, opaque and sealed.

EA: There is indirect evidence that the research personnel were not adequately blinded to study group and did not take steps to minimize potential bias.

Definitely High risk of bias:

HCT: There is direct evidence for lack of adequate blinding of the study group including no blinding or incomplete blinding of research personnel and subjects. For some treatments, such as behavioral interventions, allocation to study groups cannot be concealed.

EA: There is direct evidence that the research personnel were not adequately blinded to study group and did not take steps to minimize potential bias.

ATTRITION/EXCLUSION BIAS

Attrition or exclusion bias refers to systematic differences in the loss or exclusion from analyses of participants or animals from the study and how they were accounted for in the results (Viswanathan et al. 2012).

9. Were outcome data incomplete due to attrition or exclusion from analysis?

Incomplete outcome data includes loss due to attrition (nonresponse, dropout, or loss to follow-up) or exclusion from analyses. The degree of bias resulting from incomplete outcome data depends on the reasons that outcomes are missing, the amount and distribution of missing data across groups, and the potential association between outcome values and likelihood of missing data (Higgins and Green 2011). The risk of bias from incomplete outcome data can be reduced if study authors address the problem in their analyses (e.g., intention to treat analysis and imputation).

Human introductory text. Differential or overall attrition because of nonresponse, dropping out, loss to follow-up, and exclusion of participants can introduce bias when missing outcome data are related to both exposure/treatment and outcome. Those who drop out of the study or who are lost to follow-up may be systematically different from those who remain in the study. Attrition or exclusion bias can potentially change the collective (group) characteristics of the relevant groups and their observed outcomes in ways that affect study results by confounding and spurious associations (Viswanathan et al. 2012). This risk of bias item is recommended to assess controlled human trials (Higgins and Green 2011), observational human studies (Viswanathan et al. 2012) and animal studies (Krauth et al. 2013). However, concern over bias from incomplete outcome data is mainly theoretical and most studies that have looked at whether aspects of missing data are associated with magnitude of effect estimates have not found clear evidence of bias (reviewed in Higgins and Green 2011).

Animal introductory text. Attrition or exclusion because of illness, death, or other reasons can introduce bias when missing outcome data are related to both exposure and outcome. Attrition bias can potentially change the collective (group) characteristics of the relevant groups and their observed outcomes in ways that affect study results by confounding and spurious associations (Viswanathan et al. 2012). This risk of bias item is recommended to assess controlled human trials (Higgins and Green 2011), observational human studies (Viswanathan et al. 2012) and animal studies (Krauth et al. 2013). However, concern over bias from incomplete outcome data is mainly theoretical and most studies that have looked at whether aspects of missing data are associated with magnitude of effect estimates have not found clear evidence of bias (reviewed in Higgins and Green 2011).

Applies to: HCT, EA, Co, CaCo, CrSe

Definitely Low risk of bias:

HCT: There is direct evidence that there was no loss of subjects during the study and outcome data were complete **OR** loss of subjects (i.e., incomplete outcome data) was adequately addressed and reasons were documented when human subjects were removed from a study. Review authors should be confident that the participants included in the analysis are exactly those who were randomized into the trial. Acceptable handling of subject attrition includes: very little missing outcome data (less than 10% in each group (Genaidy et al. 2007)); reasons for missing subjects unlikely to be related to outcome (for survival data, censoring unlikely to be introducing bias); missing outcome data balanced in numbers across study groups, with similar reasons for missing data across groups **OR** analyses (such as intention-to-treat analysis) in which missing data have been imputed using appropriate methods(insuring that the characteristics of subjects lost to follow up or with unavailable records are described in an identical way and are not significantly different from those of the study participants).

NOTE: participants randomized but subsequently found not to be eligible need not always be considered as having missing outcome data (Higgins and Green 2011).

EA: There is direct evidence that loss of animals was adequately addressed and reasons were documented when animals were removed from a study. Acceptable handling of attrition includes: very little missing outcome data; reasons for missing animals unlikely to be related to outcome (or for survival data, censoring unlikely to be introducing bias); missing outcome data balanced in numbers across study groups, with similar reasons for missing data across groups; missing outcomes is not enough to impact the effect estimate **OR** missing data have been imputed using appropriate methods (insuring that characteristics of animals are not significantly different from animals retained in the analysis).

Co: There is direct evidence that loss of subjects (i.e., incomplete outcome data) was adequately addressed and reasons were documented when human subjects were removed from a study. Acceptable handling of subject attrition includes: very little missing outcome data; reasons for missing subjects unlikely to be related to outcome (for survival data, censoring unlikely to be introducing bias); missing outcome data balanced in numbers across study groups, with similar reasons for missing data across groups; **OR** missing data have been imputed using appropriate methods, **AND** characteristics of subjects lost to follow up or with unavailable records are described in identical way and are not significantly different from those of the study participants.

CaCo, CrSe: There is direct evidence that exclusion of subjects from analyses was adequately addressed, and reasons were documented when subjects were removed from the study or excluded from analyses.

Probably Low risk of bias:

HCT: There is indirect evidence that loss of subjects (i.e., incomplete outcome data) was adequately addressed and reasons were documented when human subjects were removed from a study **OR** it is deemed that the proportion lost to follow-up would not appreciably bias results (less than 20% in each group (Genaidy et al. 2007)). This would include reports of no statistical differences in characteristics of subjects lost to follow up or with unavailable records from those of the study participants. Generally, the higher the ratio of participants with missing data to participants with events, the greater potential there is for bias. For studies with a long duration of follow-up, some withdrawals for such reasons are inevitable.

EA: There is indirect evidence that loss of animals was adequately addressed and reasons were documented when animals were removed from a study **OR** it is deemed that the proportion of

animals lost would not appreciably bias results. This would include reports of no statistical differences in characteristics of animals removed from the study from those remaining in the study.

Co: There is indirect evidence that loss of subjects (i.e., incomplete outcome data) was adequately addressed and reasons were documented when human subjects were removed from a study **OR** it is deemed that the proportion lost to follow-up would not appreciably bias results. This would include reports of no statistical differences in characteristics of subjects lost to follow up or with unavailable records from those of the study participants. Generally, the higher the ratio of participants with missing data to participants with events, the greater potential there is for bias. For studies with a long duration of follow-up, some withdrawals for such reasons are inevitable.

CaCo, CrSe: There is indirect evidence that exclusion of subjects from analyses was adequately addressed, and reasons were documented when subjects were removed from the study or excluded from analyses.

Probably High risk of bias:

HCT: There is indirect evidence that loss of subjects (i.e., incomplete outcome data) was unacceptably large (greater than 20% in each group (Genaidy et al. 2007)) and not adequately addressed **OR** there is insufficient information provided about numbers of subjects lost to follow-up.

EA: There is indirect evidence that loss of animals was unacceptably large and not adequately addressed **OR** there is insufficient information provided about loss of animals.

Co: There is indirect evidence that loss of subjects (i.e., incomplete outcome data) was unacceptably large and not adequately addressed **OR** there is insufficient information provided about numbers of subjects lost to follow-up.

CaCo, CrSe: There is indirect evidence that exclusion of subjects from analyses was not adequately addressed, **OR** there is insufficient information provided about why subjects were removed from the study or excluded from analyses.

Definitely High risk of bias:

HCT, Co: There is direct evidence that loss of subjects (i.e., incomplete outcome data) was unacceptably large and not adequately addressed. Unacceptable handling of subject attrition includes: reason for missing outcome data likely to be related to true outcome, with either imbalance in numbers or reasons for missing data across study groups; or potentially inappropriate application of imputation.

EA: There is direct evidence that loss of animals was unacceptably large and not adequately addressed. Unacceptable handling of attrition includes: reason for loss is likely to be related to true outcome, with either imbalance in numbers or reasons for loss across study groups.

CaCo, CrSe: There is direct evidence that exclusion of subjects from analyses was not adequately addressed. Unacceptable handling of subject exclusion from analyses includes: reason for exclusion likely to be related to true outcome, with either imbalance in numbers or reasons for exclusion across study groups.

DETECTION BIAS

Detection bias refers to systematic differences between experimental and control groups with regards to how outcomes and exposures are assessed (Higgins and Green 2011) and also considers validity and reliability of methods used to assess outcomes and exposures (Viswanathan et al. 2012).

10. Were the outcome assessors blinded to study group or exposure level?

Blinding requires that outcome assessors do not know the study group or exposure level of the human subject or animal when the outcome was assessed.

Human introductory text. If outcome assessors are not blinded to the study group or exposure level it could bias the outcome assessment, so this is a recommended risk of bias element for controlled trials and observational studies (Higgins and Green 2011; Viswanathan et al. 2012).

Without distinguishing between the different stages of blinding during the conduct of a study, lack of blinding in randomized trials has been empirically shown to be associated with larger estimations of intervention effects (on average a 9% increase in an odds ratio) (Pildal et al. 2007). Schulz et al. (1995) analyzed 250 controlled trials and found that studies that were not double-blinded had a 17% larger estimation of treatment effect, on average. In trials with more subjective outcomes, more bias has been observed with lack of blinding (Wood et al. 2008), indicating that blinding outcome assessors could be more important for these effects.

For some exposures, it is not possible to entirely blind outcome assessors, particularly if subjects are self-reporting outcomes. However, adherence to a strict study protocol can reduce the risk of bias. In practice, successful blinding cannot be ensured, as it can be compromised for most interventions. In some cases the treatment may have side effects possibly allowing the participant to detect which intervention they received, unless the study compares interventions with similar side effects or uses an active placebo (Boutron et al. 2006).

Animal introductory text. If outcome assessors are not blinded to the study group or exposure level it could bias the outcome assessment, so this is a recommended risk of bias element animal studies (reviewed in Krauth et al. 2013) and human controlled trials and observational studies (Higgins and Green 2011; Viswanathan et al. 2012).

There is empirical evidence that lack of blinding at outcome assessment in animal studies is associated with larger measures of the effect (Bebarta et al. 2003; Sena et al. 2007; Vesterinen et al. 2010). In animal studies, blinding of study group at outcome assessment may not be possible because of the need to determine if treated animals are affected relative to controls in a treatment or dose-dependent manner (examples include clinical observations and histopathologic assessment of non-neoplastic lesions). However, animal studies are in general more tightly controlled than human studies and additional measures may be taken to reduce the risk of bias.

Note there are separate risk of bias issues affected by concealment or blinding that are important for selection, performance and detection bias: 1) a question under selection bias addresses allocation concealment of research personnel and human subjects ; and 2) a question under performance bias addresses blinding of research personnel and human subjects to the study group during the study.

Applies to: HCT, EA, Co, CaCo, CrSe, CaS

Definitely Low risk of bias:

HCT: There is direct evidence that the outcome assessors (including study subjects, if outcomes were self-reported) were adequately blinded to the study group, and it is unlikely that they could have broken the blinding prior to reporting outcomes.

EA: There is direct evidence that the outcome assessors were adequately blinded to the study group, and it is unlikely that they could have broken the blinding prior to reporting outcomes.

Co, CrSe, CaS: There is direct evidence that the outcome assessors (including study subjects, if outcomes were self-reported) were adequately blinded to the exposure level, and it is unlikely that they could have broken the blinding prior to reporting outcomes.

CaCo: There is direct evidence that the outcome assessors (including study subjects, if outcomes were self-reported) were adequately blinded to the exposure level when reporting outcomes.

Probably Low risk of bias:

HCT: There is indirect evidence that the outcome assessors (including study subjects, if outcomes were self-reported) were adequately blinded to the study group, and it is unlikely that they could have broken the blinding prior to reporting outcomes, **OR** it is deemed that lack of adequate blinding of outcome assessors would not appreciably bias results, which may vary by outcome (i.e., blinding is especially important for subjective measures).

EA: There is indirect evidence that the outcome assessors were adequately blinded to the study group, and it is unlikely that they could have broken the blinding prior to reporting outcomes **OR** it is deemed that lack of adequate blinding of outcome assessors would not appreciably bias results, which may vary by outcome (i.e., blinding is especially important for subjective measures). For some outcomes, particularly pathology assessment, outcome assessors are not blind to study group as they require comparison to the control to appropriately judge the outcome, but additional measures such as multiple levels of independent review by trained pathologists can minimize this potential bias.

Co, CrSe, CaS: There is indirect evidence that the outcome assessors were adequately blinded to the exposure level, and it is unlikely that they could have broken the blinding prior to reporting outcomes **OR** it is deemed that lack of adequate blinding of outcome assessors would not appreciably bias results (including that subjects self-reporting outcomes were likely not aware of reported links between the exposure and outcome lack of blinding is unlikely to bias a particular outcome).

CaCo: There is direct evidence that the outcome assessors were adequately blinded to the exposure level when reporting outcomes **OR** it is deemed that lack of adequate blinding of outcome assessors would not appreciably bias results (including that subjects self-reporting outcomes were likely not aware of reported links between the exposure and outcome or lack of blinding is unlikely to bias a particular outcome).

Probably High risk of bias:

HCT: There is indirect evidence that it was possible for outcome assessors (including study subjects if outcomes were self-reported) to infer the study group prior to reporting outcomes, **OR** there is insufficient information provided about blinding of outcome assessors.

EA: There is indirect evidence that it was possible for outcome assessors to infer the study group prior to reporting outcomes without sufficient quality control measures **OR** there is insufficient information provided about blinding of outcome assessors.

Co, CrSe, CaS: There is indirect evidence that it was possible for outcome assessors to infer the exposure level prior to reporting outcomes (including that subjects self-reporting outcomes were likely aware of reported links between the exposure and outcome) **OR** there is insufficient information provided about blinding of outcome assessors.

CaCo: There is indirect evidence that it was possible for outcome assessors to infer the exposure level prior to reporting outcomes (including that subjects self-reporting outcomes were likely aware of reported links between the exposure and outcome) **OR** there is insufficient information provided about blinding of outcome assessors.

Definitely High risk of bias:

HCT: There is direct evidence for lack of adequate blinding of outcome assessors (including study subjects if outcomes were self-reported), including no blinding or incomplete blinding.

EA: There is direct evidence for lack of adequate blinding of outcome assessors, including no blinding or incomplete blinding without quality control measures.

Co, CrSe, CaS: There is direct evidence that outcome assessors were aware of the exposure level prior to reporting outcomes (including that subjects self-reporting outcomes were aware of reported links between the exposure and outcome).

CaCo: There is direct evidence that outcome assessors to were aware of the exposure level prior to reporting outcomes (including that subjects self-reporting outcomes were aware of reported links between the exposure and outcome).

11. Were confounding variables assessed consistently across groups using valid and reliable measures?

Consistent application of valid, reliable, and sensitive methods of assessing important confounding or modifying variables is required across study groups.

This item is included in this current risk of bias tool because it is recommended in recent guidance from AHRQ on assessing risk of bias for observational human studies (Viswanathan et al. 2012) and a similar item is recommended by the CLARITY Group at McMaster University (2013) for assessment of cohort studies ("confidence in the assessment of the presence or absence of prognostic factors").

The requirement for assessing the confounding variables with valid and reliable measures is directly linked to the relative importance of the confounding variable considered under selection bias (i.e., independent of study design, if a confounder needed to be accounted for in design or analyses, then measurement of that variable had to be reliable).

Empirical evidence of bias due to this factor is not currently available. The direction of the bias (towards or away from the null) will differ based on the nature of any inconsistent assessment of confounding across groups and limitations in the validity and reliability of the measurement.

Note, a parallel question under selection bias addresses whether design or analysis account for confounding.

Applies to: HCT, EA, Co, CaCo, CrSe, CaS

Definitely Low risk of bias:

HCT, EA, Co, CaCo, CrSe, CaS: There is direct evidence that primary covariates and confounders were assessed using valid and reliable measurements.

Probably Low risk of bias:

HCT, EA, Co, CaCo, CrSe, CaS: There is indirect evidence primary covariates and confounders were assessed using valid and reliable measurements **OR** it is deemed that the measures used would not appreciably bias results (i.e., the authors justified the validity of the measures from previously published research).

Probably High risk of bias:

HCT, EA, Co, CaCo, CrSe, CaS: There is indirect evidence that primary covariates and confounders were assessed using measurements of unknown validity **OR** there is insufficient information provided about the measures used.

Definitely High risk of bias:

HCT, EA, Co, CaCo, CrSe, CaS: There is direct evidence that primary covariates and confounders were assessed using non valid measurements.

12. Can we be confident in the exposure characterization?

Confidence requires valid, reliable, and sensitive methods to measure exposure applied consistently across groups.

Detection bias can be minimized by using valid and reliable exposure measures applied consistently across groups consistently assessed (i.e., under the same method and time-frame). For example, studies relying on indirect measures of exposure (e.g., self-report) may be rated as having a higher risk of bias than studies that directly measure exposure (e.g., measurement of the chemical in air or measurement of the chemical in blood, plasma, urine, etc.).

For controlled exposure studies (i.e., experimental human and animal studies), independent verification of purity would be considered best practice because the identity and purity as listed on the bottle can be inaccurate. In NTP's experience, about 3% are the wrong chemical and inaccuracy rises to 10% if you include cases where the purity is not as stated on the bottle (unpublished, personal communication Brad Collins, NTP chemist). It is also possible that impurities may be more toxic than the compound of interest. This occurred during an NTP study of PCB 118 where analysis revealed the presence of 0.622% of the much more potent PCB 126, resulting in the study being continued as a mixture study [(NTP 2006), see page 13].

Human text: Assessment of exposure is a widely used element of tools to assess study quality for observational human studies (CLARITY Group at McMaster University 2013; Downs and Black 1998; Shamliyan et al. 2010; Viswanathan et al. 2012; Wells et al.). Key factors to assess the quality of exposure characterization in observational human studies are understanding potential for exposure misclassification, whether there is an adequate level of exposure variability to detect an effect, and the extent of reliance on imputed exposure levels.

The direction of the bias (towards or away from the null) will differ based on the nature of differences between comparison groups and may be difficult to predict. Non-differential misclassification of exposure will generally bias results towards the null, but differential misclassification can bias towards or away from the null, making it difficult to predict the direction of effect (Szklo and Nieto 2007). Noncompliance with the allocated treatment could introduce differential misclassification if compliance was unequal across study groups. Adherence to a strict study protocol that includes measures to assure or assess compliance can reduce the risk of bias.

Applies to: HCT, EA, Co, CaCo, CrSe, CaS

Definitely Low risk of bias:

HCT: There is direct or indirect evidence that the test material is confirmed as ≥99% pure (or impurities have been characterized and not considered to be of serious concern), and that the concentration, stability, and homogeneity of stock material and formulation have been verified as appropriate (**Note:** ≥99% purity value is considered achievable based on current advertised purity from Sigma-Aldrich); **AND FOR INTERNAL DOSIMETRY STUDIES** there is direct evidence that most data points for the aglycone, conjugated and/or total BPA are ***above*** the level of quantitation (LOQ) for the assay; **AND** the study utilized spiked samples to confirm assay performance and the stability of BPA and conjugated BPA in biological samples was appropriately addressed; **AND** studies took measures to assess potential BPA contamination that might have occurred during sample collection and analysis, including method blanks. Note: Use of method blanks is necessary to identify potential sources of contamination in blood and urine but cannot rule out all possible sources of contamination (Ye et al. 2012). The risk of contamination for blood-based measurements is likely higher than for urinary measurements in part because sterile plastic blood collection containers can increase the number of sources of contamination and because of higher levels of protein and lipid levels in blood versus urine. Preferred practices include (1) measurement of aglycone AND conjugated or total BPA for blood measurements, and (2) use of isotopically labeled BPA dosing material (e.g., deuterated) to avoid issues of contamination, although we will not "downgrade" if a study did not follow these preferred practices.

EA: There is direct or indirect evidence that the test material is confirmed as ≥99% pure (or impurities have been characterized and not considered to be of serious concern), and that the concentration, stability, and homogeneity of stock material and formulation have been verified as appropriate (**Note:** ≥99% purity value is considered achievable based on current advertised purity from Sigma-Aldrich); **AND** the study provides information about consumption through measurement of the dosing medium and dose intake quantity, e.g., feed or water consumption; **AND FOR INTERNAL DOSIMETRY STUDIES** there is direct evidence that most data points for the aglycone, conjugated and/or total BPA are ***above*** the level of quantitation (LOQ) for the assay; **AND** the study utilized spiked samples to confirm assay performance and the stability of BPA and conjugated BPA in biological samples was appropriately addressed; **AND** studies took measures to assess potential BPA contamination that might have occurred during sample collection and analysis including method blanks. Note: Use of method blanks is necessary to identify potential sources of contamination in blood and urine but cannot rule out all possible sources of contamination (Ye et al. 2012). The risk of contamination for blood-based measurements is likely higher than for urinary measurements in part because sterile plastic blood collection containers can increase the number of sources of contamination and because of higher levels of protein and lipid levels in blood versus urine. Preferred practices include (1) measurement of aglycone AND conjugated or total BPA for blood measurements, and (2) use use of isotopically labeled BPA dosing material (e.g., deuterated) is ideal to avoid issues of contamination, although we will not "downgrade" if a study did not follow these preferred practices.

Co, CaCo, CrSe, CaS: There is direct evidence that most data points for the aglycone, conjugated and/or total BPA are ***above*** the level of quantitation (LOQ) for the assay; **AND** the study utilized spiked samples to confirm assay performance and the stability of BPA and conjugated BPA in biological samples was appropriately addressed; **AND** studies took measures to assess potential BPA contamination that might have occurred during sample collection and analysis including

method blanks. Note: Use of method blanks is necessary to identify potential sources of contamination in blood and urine but cannot rule out all possible sources of contamination (Ye et al. 2012). The risk of contamination for blood-based measurements is likely higher than for urinary measurements in part because sterile plastic blood collection containers can increase the number of sources of contamination and because of higher levels of protein and lipid levels in blood versus urine. Preferred practices include (1) measurement of aglycone AND conjugated or total BPA for blood measurements, and (2) inclusion of multiple measurements of BPA because a single sample from an individual does not appear to be strong predictor of a subject's exposure category. Mahalingaiah et al. {, 2008 #300} analyzed samples from at least six repeat urinary BPA measurements from eight subjects. The sensitivity, specificity, and positive predictive value of a single urine sample to predict the highest BPA tertile were 0.64, 0.76, and 0.63, respectively. The positive predictive value increased to 0.85 when two samples were used to predict those individuals in the highest BPA tertile. Use of a single measurement in large sample size studies such as NHANES is less of a issue because the number of participants offsets potential concern for differential exposure misclassification. We will not downgrade if a study did not follow these preferred practices.

Probably Low risk of bias:

HCT: There is direct or indirect evidence that purity was ≥98%, (or impurities have been characterized and not considered to be of serious concern i.e., purity was independently confirmed by lab, purity is reported in paper or obtained through author query, or purity not reported but the source is listed and the supplier of the chemical provides documentation of the purity of the chemical; **AND FOR INTERNAL DOSIMETRY STUDIES** there is indirect evidence that most data points for the aglycone, conjugated and/or total BPA are ***above*** the level of quantitation (LOQ) for the assay, i.e., the central estimate (median, mean, geometric mean) is ***above*** the LOQ but results for individual data values are not presented or the presentation of variance estimates do not permit assessment of whether most data points are likely ***above*** the LOQ; **AND** the study utilized spiked samples to confirm assay performance and the stability of BPA and conjugated BPA in biological samples was appropriately addressed; **AND** studies took measures to assess potential BPA contamination that might have occurred during sample collection and analysis including method blanks.

EA: There is direct or indirect evidence that purity was ≥98% pure (or impurities have been characterized and not considered to be of serious concern),, i.e., purity was independently confirmed by lab, purity is reported in paper or obtained through author query, or purity not reported but the source is listed and the supplier of the chemical provides documentation of the purity of the chemical; **BUT** the study does not provide information about consumption through measurement of the dosing medium and dose intake quantity, e.g., feed or water consumption; **AND FOR INTERNAL DOSIMETRY STUDIES** there is indirect evidence that most data points for the aglycone, conjugated and/or total BPA are ***above*** the level of quantitation (LOQ) for the assay, i.e., the central estimate (median, mean, geometric mean) is ***above*** the LOQ but results for individual data values are not presented or the presentation of variance estimates do not permit assessment of whether most data points are likely ***above*** the LOQ; **AND** the study utilized spiked samples to confirm assay performance and the stability of BPA and conjugated BPA in biological samples has been appropriately addressed; **AND** studies took measures to assess potential BPA contamination that might have occurred during sample collection and analysis including method blanks.

Co, CaCo, CrSe, CaS: There is indirect evidence that most data points for the aglycone, conjugated and/or total BPA are ***above*** the level of quantitation (LOQ) for the assay, i.e., the central

estimate (median, mean, geometric mean) is ***above*** the LOQ but results for individual data values are not presented or the presentation of variance estimates do not permit assessment of whether most data points are likely ***above*** the LOQ; **AND** the study utilized spiked samples to confirm assay performance and the stability of BPA and conjugated BPA in biological samples has been appropriately addressed; **AND** studies took measures to assess potential BPA contamination that might have occurred during sample collection and analysis including method blanks; OR use of questionnaire items where results of biomonitoring studies support the use of the questionnaire item(s) as an indicator of relative level of exposure; **OR** job description for occupational studies where levels in the work environment or results of biomonitoring studies support the use of job description as an indicator of relative level of exposure.

Probably High risk of bias:

HCT, EA: Neither the source or purity of the chemical was reported in the study and information on purity could not be obtained through author query/vendor documentation; **AND FOR INTERNAL DOSIMETRY STUDIES** there is direct or indirect evidence that most data points for the aglycone, conjugated and/or total BPA are ***above*** the level of quantitation (LOQ) for the assay **BUT** no steps were taken to assess potential BPA contamination that might have occurred during sample collection and analysis; **OR** there is indirect or direct evidence that most individual data points for the aglycone, conjugated and/or total BPA are ***below*** the level of quantitation (LOQ) for the assay; **OR** method to measure BPA used ELISA which is less accepted as providing quantitatively accurate values and because of potential uncharacterized antibody cross-reactivity with conjugates and endogenous components of sample matrices (Chapin et al. 2008; Vandenberg et al. 2007)

Co, CaCo, CrSe, CaS: There is direct or indirect evidence that most data points for the aglycone, conjugated and/or total BPA are ***above*** the level of quantitation (LOQ) for the assay **BUT** no steps were taken to assess potential BPA contamination that might have occurred during sample collection and analysis; **OR** there is indirect or direct evidence that most individual data points for the aglycone, conjugated and/or total BPA are ***below*** the level of quantitation (LOQ) for the assay; **OR** method to measure BPA used ELISA which leads to concern because of uncharacterized antibody cross-reactivity with conjugates and endogenous components of sample matrices (Chapin et al. 2008; Vandenberg et al. 2007); **OR** use of questionnaire items that are not supported by results of biomonitoring studies; **OR** job description for occupational studies that are not supported by information on levels in the work environment or results of biomonitoring studies

Definitely High risk of bias:

HCT, EA: There is indirect or direct evidence that purity was <98%; **AND FOR INTERNAL DOSIMETRY STUDIES** there is direct evidence of uncontrolled contamination.

Co, CaCo, CrSe, CaS: There is direct evidence of uncontrolled contamination; **OR** not reporting of methods used to assess exposure and this information could not be obtained through author query; **OR** self-report exposure.

13. Can we be confident in the outcome assessment?

Confidence requires valid, reliable, and sensitive methods to assess the outcome applied consistently across groups.

Detection bias can be minimized by using valid and reliable methods to assess outcome. For example, studies relying on self-report of outcome may be rated as having a higher risk of bias than studies with clinically observed outcomes (Viswanathan et al. 2012). Differential assessment of outcomes is an additional source of bias and this element is a widely used element of tools to assess study quality for observational human studies (Downs and Black 1998; Genaidy et al. 2007; Shamliyan et al. 2010; Viswanathan et al. 2012).

Note: For case-control studies, confirmation that the control subjects are free of the outcome is considered under a separate risk of bias question, "Were the comparison groups appropriate?"

Applies to: HCT, EA, Co, CaCo, CrSe, CaS

Definitely Low risk of bias:

HCT, Co: There is direct evidence that the outcome was assessed using well-established methods, the "gold standard" or with validity and reliability >0.70 (Genaidy et al. 2007) and subjects had been followed for the same length of time in all study groups. Acceptable assessment methods will depend on the outcome, but examples of such methods may include: objectively measured with diagnostic methods, measured by trained interviewers, obtained from registries (Shamliyan et al. 2010).

EA: There is direct evidence that the outcome was assessed using well-established methods (the gold standard) assessed at the same length of time after initial exposure in all study groups.

CaCo: There is direct evidence that the outcome was assessed in cases using well-established methods (the gold standard) and subjects had been followed for the same length of time in all study groups.

CrSe, CaS: There is direct evidence that the outcome was assessed using well-established methods (the gold standard).

Probably Low risk of bias:

HCT, Co: There is indirect evidence that the outcome was assessed using acceptable methods [i.e., deemed valid and reliable but not the gold standard or with validity and reliability ≥0.40 (Genaidy et al. 2007)] and subjects had been followed for the same length of time in all study groups **OR** it is deemed that the outcome assessment methods used would not appreciably bias results. Acceptable, but not ideal assessment methods will depend on the outcome, but examples of such methods may include proxy reporting of outcomes and mining of data collected for other purposes.

EA: There is indirect evidence that the outcome was assessed using acceptable methods (i.e., deemed valid and reliable but not the gold standard) assessed at the same length of time after initial exposure in all study groups **OR** it is deemed that the outcome assessment methods used would not appreciably bias results.

CaCo: There is indirect evidence that the outcome was assessed in cases (i.e., case definition) using acceptable methods and subjects had been followed for the same length of time in all study groups **OR** it is deemed that the outcome assessment methods used would not appreciably bias results.

CrSe, CaS: There is indirect evidence that the outcome was assessed using acceptable methods **OR** it is deemed that the outcome assessment methods used would not appreciably bias results.

Probably High risk of bias:

HCT, Co: There is indirect evidence that the outcome assessment method is an insensitive instrument, the authors did not validate the methods used, or the length of follow up differed by study group **OR** there is insufficient information provided about validation of outcome assessment method.

EA: There is indirect evidence that the outcome assessment method is an insensitive instrument, the authors did not validate the methods used, or the length of time after initial exposure differed by study group **OR** there is insufficient information provided about validation of outcome assessment method.

CaCo: There is indirect evidence that the outcome was assessed in cases using an insensitive instrument or was not adequately validated **OR** there is insufficient information provided about how cases were identified.

CrSe, CaS: There is indirect evidence that the outcome assessment method is an insensitive instrument or was not adequately validated **OR** there is insufficient information provided about validation of outcome assessment method.

Definitely High risk of bias:

HCT, Co: There is direct evidence that the outcome assessment method is an insensitive instrument, or the length of follow up differed by study group.

EA: There is direct evidence that the outcome assessment method is an insensitive instrument or the length of time after initial exposure differed by study group.

CaCo: There is direct evidence that the outcome was assessed in cases using an insensitive instrument.

CrSe, CaS: There is direct evidence that the outcome assessment method is an insensitive instrument.

SELECTIVE REPORTING BIAS

Selective reporting bias refers to selective inclusion of outcomes in the publication of the study on the basis of the results (Higgins and Green 2011; Hutton and Williamson 2000).

14. Were all measured outcomes reported?

Selective reporting of results is a recommended element of assessing risk of bias (Guyatt et al. 2011; Higgins et al. 2011; IOM 2011; Viswanathan et al. 2012). Selective reporting is present if pre-specified outcomes are not reported or incompletely reported. It is likely widespread and difficult to assess with confidence for most studies unless the study protocol is available. Selective reporting bias can be assessed by comparing the "methods" and "results" section of the paper, and by considering outcomes measured in the context of knowledge in the field. Abstracts of presentations relating to the study may contain information about outcomes not subsequently mentioned in publications. Selective reporting bias should be suspected if the study does not report outcomes in the results section that would have been expected based on the methods, or if a composite score is present without the individual component outcomes (Guyatt et al. 2011). It may be useful to pay attention to author affiliations and

funding source which can contribute to selective outcome reporting when results are not consistent with expectations or value to the research objectives.

Applies to: HCT, EA, Co, CaCo, CrSe, CaS

Definitely Low risk of bias:

HCT, EA, Co, CaCo, CrSe, CaS: There is direct evidence that all of the study's measured outcomes (primary and secondary) outlined in the protocol, methods, abstract, and/or introduction (that are relevant for the evaluation) have been reported. This would include outcomes reported with sufficient detail to be included in meta-analysis or fully tabulated during data extraction.

Probably Low risk of bias:

HCT, EA, Co, CaCo, CrSe, CaS: There is indirect evidence that all of the study's measured outcomes (primary and secondary) outlined in the protocol, methods, abstract, and/or introduction (that are relevant for the evaluation) have been reported **OR** analyses that had not been planned at the outset of the study (i.e., retrospective unplanned subgroup analyses) are clearly indicated as such and it is deemed that the omitted analyses were not appropriate and selective reporting would not appreciably bias results. This would include outcomes reported with insufficient detail such as only reporting that results were statistically significant (or not).

Probably High risk of bias:

HCT, EA, Co, CaCo, CrSe, CaS: There is indirect evidence that all of the study's measured outcomes (primary and secondary) outlined in the protocol, methods, abstract, and/or introduction (that are relevant for the evaluation) have been reported **OR** there is insufficient information provided about selective outcome reporting.

Definitely High risk of bias:

HCT, EA, Co, CaCo, CrSe, CaS: There is direct evidence that all of the study's measured outcomes (primary and secondary) outlined in the protocol, methods, abstract, and/or introduction (that are relevant for the evaluation) have not been reported. In addition to not reporting outcomes, this would include reporting outcomes based on composite score without individual outcome components or outcomes reported using measurements, analysis methods or subsets of the data (e.g., subscales) that were not pre-specified or reporting outcomes not pre-specified (unless clear justification for their reporting is provided, such as an unexpected effect).

OTHER

15. Were there any other potential threats to internal validity (e.g., inappropriate statistical methods)?

On a project specific basis, additional questions for other potential threats to internal validity can be added and applied to study designs as appropriate.

Example:

Failure to statistically or experimentally adjust for litter in an animal study with a developmental outcome. The direction of the bias is away from the null towards a larger effect size (Haseman et al. 2001).

REFERENCES

Baker DG. 1998. Natural pathogens of laboratory mice, rats, and rabbits and their effects on research. Clinical Microbiology Reviews 11(2):231-266.

Bebarta V, Luyten D, Heard K. 2003. Emergency medicine animal research: does use of randomization and blinding affect the results? Acad Emerg Med 10(6):684-687.

Boutron I, Estellat C, Guittet L, Dechartres A, Sackett DL, Hrobjartsson A, et al. 2006. Methods of blinding in reports of randomized controlled trials assessing pharmacologic treatments: a systematic review. PLoS Med 3(10):e425.

Braun JM, Kalkbrenner AE, Calafat AM, Bernert JT, Ye X, Silva MJ, et al. 2011. Variability and predictors of urinary bisphenol A concentrations during pregnancy. Environ Health Perspect 119(1):131-137.

Calafat AM, Ye X, Wong LY, Reidy JA, Needham LL. 2008. Exposure of the U.S. population to bisphenol A and 4-tertiary-octylphenol: 2003-2004. Environ Health Perspect 116(1):39-44.

Cao XL, Perez-Locas C, Dufresne G, Clement G, Popovic S, Beraldin F, et al. 2011. Concentrations of bisphenol A in the composite food samples from the 2008 Canadian total diet study in Quebec City and dietary intake estimates. Food Addit Contam Part A Chem Anal Control Expo Risk Assess 28(6):791-798.

CDC (Centers for Disease Control and Prevention). 2012. Overweight and Obesity: Data and Statistics. http://www.cdc.gov/obesity/data/index.html [accessed 18 December 2012].

Chapin RE, Adams J, Boekelheide K, Gray LE, Jr., Hayward SW, Lees PS, et al. 2008. NTP-CERHR Expert Panel Report on the Reproductive and Developmental Toxicity of Bisphenol A. Birth Defects Res B Dev Reprod Toxicol 83(3):157-395.

CLARITY Group at McMaster University. 2013. Tools to assess risk of bias in cohort and case control studies; randomized controlled trials; and longitudinal symptom research studies aimed at the general population. http://www.evidencepartners.com/resources/ [accessed 19 January 2013].

Downs SH, Black N. 1998. The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. Journal of epidemiology and community health 52(6):377-384.

Duke JL, Zammit TG, Lawson DM. 2001. The effects of routine cage-changing on cardiovascular and behavioral parameters in male Sprague-Dawley rats. Contemp Top Lab Anim Sci 40(1):17-20.

Dwan K, Gamble C, Kolamunnage-Dona R, Mohammed S, Powell C, Williamson PR. 2010. Assessing the potential for outcome reporting bias in a review: a tutorial. Trials 11:52.

Genaidy AM, Lemasters GK, Lockey J, Succop P, Deddens J, Sobeih T, et al. 2007. An epidemiological appraisal instrument - a tool for evaluation of epidemiological studies. Ergonomics 50(6):920-960.

Gerdin AK, Igosheva N, Roberson LA, Ismail O, Karp N, Sanderson M, et al. 2012. Experimental and husbandry procedures as potential modifiers of the results of phenotyping tests. Physiol Behav 106(5):602-611.

Gerstman BB. 1998. Epidemiology Kept Simple. New York, NY: John Wiley and Sons, Inc.

Guyatt GH, Oxman AD, Vist G, Kunz R, Brozek J, Alonso-Coello P, et al. 2011. GRADE guidelines: 4. Rating the quality of evidence-study limitations (risk of bias). Journal of Clinical Epidemiology 64(4):407-415.

GV-SOLAS. 1999. Implications of infectious agents on results of animal experiments. Report of the Working Group on Hygiene of the Gesellschaft fur Versuchstierkunde--Society for Laboratory Animal Science (GV-SOLAS). Laboratory Animals 33 Suppl 1:S39-87.

Hailey JR, Haseman JK, Bucher JR, Radovsky AE, Malarkey DE, Miller RT, et al. 1998. Impact of Helicobacter hepaticus infection in B6C3F1 mice from twelve National Toxicology Program two-year carcinogenesis studies. Toxicol Pathol 26(5):602-611.

Haseman JK, Bailer AJ, Kodell RL, Morris R, Portier K. 2001. Statistical issues in the analysis of low-dose endocrine disruptor data. Toxicol Sci 61(2):201-210.

Higgins J, Green S. 2011. Cochrane Handbook for Systematic Reviews of Interventions. Version 5.1.0 (updated March 2011). http://handbook.cochrane.org/ (accessed 3 February 2013).

Higgins JP, Altman DG, Gotzsche PC, Juni P, Moher D, Oxman AD, et al. 2011. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. BMJ 343:d5928.

Hutton JL, Williamson PR. 2000. Bias in meta-analysis due to outcome variable selection within studies. Journal of the Royal Statistical Society Series C-Applied Statistics 49:359-370.

IOM (Institute of Medicine). 2006. Committee on Food Marketing and the Diets of Children and Youth: Food Marketing to Children and Youth: Threat or Opportunity? [JM McGinnis, J Gootman and VI Kraak, editors]. Washington, DC:Institute of Medicine of the National Academies. Available at http://www.nap.edu/catalog.php?record_id=11514http://www.nap.edu/catalog.php?record_id=11514 (accessed 5 March 2013).

IOM (Institute of Medicine). 2011. Finding What Works in Health Care: Standards for Systematic Reviews. http://www.nap.edu/openbook.php?record_id=13059&page=R1 [accessed 13 January 2013].

Johnson PI, Sutton P, Atchley D, Koustas E, Lam J, Robinson K, et al. 2013. Applying the Navigation Guide: Case Study #1. The Impact of Developmental Exposure to Perfluorooctanoic Acid (PFOA) On Fetal Growth. A Systematic Review of the Human Evidence - Protocol.

Koustas E, Lam J, Sutton P, Johnson PI, Atchley D, Robinson K, et al. 2013. Applying the Navigation Guide: Case Study #1. The Impact of Developmental Exposure to Perfluorooctanoic Acid (PFOA) On Fetal Growth. A Systematic Review of the Non-Human Evidence - Protocol.

Krauth D, Woodrull T, Bero L. 2013. A systematic review of quality assessment instruments for published animal studies (submitted).

Kusters JG, van Vliet AH, Kuipers EJ. 2006. Pathogenesis of Helicobacter pylori infection. Clin Microbiol Rev 19(3):449-490.

Macleod MR, van der Worp HB, Sena ES, Howells DW, Dirnagl U, Donnan GA. 2008. Evidence for the efficacy of NXY-059 in experimental focal cerebral ischaemia is confounded by study quality. Stroke; a journal of cerebral circulation 39(10):2824-2829.

Muhlhauser A, Susiarjo M, Rubio C, Griswold J, Gorence G, Hassold T, et al. 2009. Bisphenol A effects on the growing mouse oocyte are influenced by diet. Biol Reprod 80(5):1066-1071.

NRC (National Research Council). 1991. Infectious Diseases of Mice and Rats. Washington, D.C.: Committee on Infectious Diseases of Mice and Rats, Institute of Laboratory Animal Resources, Commission on Life Sciences, National Research Council, The National Academies Press.

NTP (National Toxicology Program). 2006. Toxicology and Carcinogenesis Studies of a Binary Mixture of 3,3',4,4',5-Pentachlorobiphenyl (PCB 126) (CAS No. 57465-28-8) and 2,3',4,4',5-Pentachlorobiphenyl (PCB 118) (CAS No. 31508-00-6) in Female Harlan Sprague-Dawley Rats (Gavage Studies). http://ntp.niehs.nih.gov/?objectid=D16D6C59-F1F6-975E-7D23D1519B8CD7A5 [accessed 28 January 2013].

NTP (National Toxicology Program). 2013. Draft OHAT Approach for Systematic Review and Evidence Integration for Literature-based Health Assessments – February 2013. http://ntp.niehs.nih.gov/go/38138 [accessed 26 January 2013].

Ozaki A, Kawasaki C, Kawamura Y, Tanamoto K. 2006. [Migration of bisphenol A and benzophenones from paper and paperboard products used in contact with food]. Shokuhin Eiseigaku Zasshi 47(3):99-104.

Ozaki A, Yamaguchi Y, Fujita T, Kuroda K, Endo G. 2004. Chemical analysis and genotoxicological safety assessment of paper and paperboard used for food packaging. Food Chem Toxicol 42(8):1323-1337.

Pildal J, Hrobjartsson A, Jorgensen KJ, Hilden J, Altman DG, Gotzsche PC. 2007. Impact of allocation concealment on conclusions drawn from meta-analyses of randomized trials. Int J Epidemiol 36(4):847-857.

Rosenthal R, Lawson R. 1964. A Longitudinal Study of the Effects of Experimenter Bias on the Operant Learning of Laboratory Rats. Journal of psychiatric research 69:61-72.

Rothman KJ. 1986. Modern Epidemiology. Boston, MA: Little, Brown and Company.

Rudel RA, Gray JM, Engel CL, Rawsthorne TW, Dodson RE, Ackerman JM, et al. 2011. Food packaging and bisphenol A and bis(2-ethyhexyl) phthalate exposure: findings from a dietary intervention. Environ Health Perspect 119(7):914-920.

Schulz KF, Altman DG, Moher D. 2002. Allocation concealment in clinical trials. JAMA 288(19):2406-2407; author reply 2408-2409.

Schulz KF, Chalmers I, Hayes RJ, Altman DG. 1995. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. JAMA 273(5):408-412.

Sena E, van der Worp HB, Howells D, Macleod M. 2007. How can we improve the pre-clinical development of drugs for stroke? Trends in neurosciences 30(9):433-439.

Shamliyan T, Kane RL, Dickinson S. 2010. A systematic review of tools used to assess the quality of observational studies that examine incidence or prevalence and risk factors for diseases. Journal of Clinical Epidemiology 63(10):1061-1070.

Shamliyan TA, Kane RL, Ansari MT, Raman G, Berkman ND, Grant M, et al. 2011. Development of quality criteria to evaluate nontherapeutic studies of incidence, prevalence or risk factors of chronic diseases: Pilot study of new checklists. Available at http://www.ncbi.nlm.nih.gov/books/NBK53272/ [accessed March 6, 2012]. Rockville (MD): Agency for Healthcare Research and Quality (US); 2011 Jan. Report No.: 11-EHC008-EF. AHRQ Methods for Effective Health Care.

Szklo M, Nieto FJ. 2007. Epidemiology: Beyond the basics (2nd edition). 2nd ed. Sudbury, MA: Jones and Bartlett Publishers.

Thigpen JE, Setchell KD, Padilla-Banks E, Haseman JK, Saunders HE, Caviness GF, et al. 2007. Variations in phytoestrogen content between different mill dates of the same diet produces significant differences in the time of vaginal opening in CD-1 mice and F344 rats but not in CD Sprague-Dawley rats. Environ Health Perspect 115(12):1717-1726.

Trasande L, Attina TM, Blustein J. 2012. Association between urinary bisphenol A concentration and obesity prevalence in children and adolescents. JAMA 308(11):1113-1121.

Vandenberg LN, Hauser R, Marcus M, Olea N, Welshons WV. 2007. Human exposure to bisphenol A (BPA). ReproductiveToxicology 24(2):139-177.

Vesterinen HM, Sena ES, ffrench-Constant C, Williams A, Chandran S, Macleod MR. 2010. Improving the translational hit of experimental treatments in multiple sclerosis. Mult Scler 16(9):1044-1055.

Viswanathan M, Ansari M, Berkman ND, Chang S, Hartling L, McPheeters LM, et al. 2012. Assessing the risk of bias of individual studies when comparing medical interventions (March 8, 2012). Agency for Healthcare Research and Quality Methods Guide for Comparative Effectiveness Reviews. March 2012. AHRQ Publication No. 12-EHC047-EF. Available at: www.effectivehealthcare.ahrq.gov/, or direct link at http://www.effectivehealthcare.ahrq.gov/index.cfm/search-for-guides-reviews-and-reports/?pageaction=displayproduct&productid=998 [accessed 3 January 2013].

Wells G, Shea B, O'Connell D, Peterson J, Welch V, Losos M, et al. The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses. Available at http://www.ohri.ca/programs/clinical_epidemiology/oxford.asp [accessed January 18 2012].

Wood L, Egger M, Gluud LL, Schulz KF, Juni P, Altman DG, et al. 2008. Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: meta-epidemiological study. BMJ 336(7644):601-605.

Ye X, Zhou X, Wong LY, Calafat AM. 2012. Concentrations of bisphenol a and seven other phenols in pooled sera from 3-11 year old children: 2001-2002 national health and nutrition examination survey. Environ Sci Technol 46(22):12664-12671.