Published in final edited form as:

Environ Int. 2016 ; 92-93: 605–610. doi:10.1016/j.envint.2016.03.017.

Study Sensitivity: Evaluating the Ability to Detect Effects in Systematic Reviews of Chemical Exposures

Glinda S. Cooper¹, Ruth M. Lunn², Marlene Ågerstrand³, Barbara S. Glenn¹, Andrew D. Kraft¹, April M. Luke¹, and Jennifer M. Ratcliffe⁴

¹National Center for Environmental Assessment, Office of Research and Development, U.S. Environmental Protection Agency, Washington, DC, USA

²Office of the Report on Carcinogens, Division of the National Toxicology Program, National Institute of Environmental Health Sciences (NIEHS), National Institutes of Health (NIH), Research Triangle Park, North Carolina, USA

³Department of Environmental Science and Analytical Chemistry (ACES), Stockholm University, 106 91 Stockholm, Sweden

⁴Integrated Laboratory Systems (ILS), Research Triangle Park, North Carolina, USA

Abstract

A critical step in systematic reviews of potential health hazards is the structured evaluation of the strengths and weaknesses of the included studies; risk of bias is a term often used to represent this process, specifically with respect to the evaluation of systematic errors that can lead to inaccurate (biased) results (i.e. focusing on internal validity). Systematic review methods developed in the clinical medicine arena have been adapted for use in evaluating environmental health hazards; this expansion raises questions about the scope of risk of bias tools and the extent to which they capture the elements that can affect the interpretation of results from environmental and occupational epidemiology studies and in vivo animal toxicology studies, (the studies typically available for assessment of risk of chemicals). One such element, described here as “sensitivity”, is a measure of the ability of a study to detect a true effect or hazard. This concept is similar to the concept of the sensitivity of an assay; an insensitive study may fail to show a difference that truly exists, leading to a false conclusion of no effect. Factors relating to study sensitivity should be evaluated in a systematic manner with the same rigor as the evaluation of other elements within a risk of bias framework. We discuss the importance of this component for the interpretation of individual studies, examine approaches proposed or in use to address it, and describe how it relates to other evaluation components. The evaluation domains contained within a risk of bias tool can include, or can be modified to include, some features relating to study sensitivity; the explicit inclusion of these sensitivity criteria with the same rigor and at the same stage of study evaluation as other bias-related criteria can improve the evaluation process. In some cases, these and other features may be better addressed through a separate sensitivity domain. The combined evaluation

Corresponding Author: Glinda S. Cooper, address: U.S. Environmental Protection Agency, 1200 Pennsylvania Ave NW (8601P), Washington, DC 20460, courier address: One Potomac Yard (South Building) S-11342, 2777 S. Crystal Dr, Arlington, VA 22202, phone: 703-347-8636, cooper.glinda@epa.gov.

Competing Financial Interests: The authors have no competing financial interests.

of risk of bias and sensitivity can be used to identify the most informative studies, to evaluate the confidence of the findings from individual studies and to identify those study elements that may help to explain heterogeneity across the body of literature.

Keywords

systematic review; validity; bias; environmental health; chemical hazard assessment; study sensitivity

Introduction

Systematic reviews are designed to evaluate bodies of existing evidence regarding specific hypotheses using rigorous, transparent and unbiased methods or approaches (IOM, 2011). While initially developed to evaluate clinical trial data, often using a meta-analysis to summarize results, their application has been extended to the evaluation of observational and animal studies of environmental hazards (Woodruff and Sutton 2014; Rooney et al. 2014). The purpose of a hazard assessment is to identify and characterize chemical and other environmental hazards to provide the scientific basis, when warranted, for measures to protect public health. There are several challenges faced in adapting the tools used in clinical medicine to this field, however, including the need for an expanded focus on exposure measures, the greater heterogeneity of the type of studies addressing these questions (e.g., observational epidemiology, animal toxicology, in vitro studies) (Whaley et al., in press), and the greater heterogeneity within each type of study (e.g., among observational epidemiology studies, as noted in Sterne et al., 2014).

A critical step in systematic reviews of potential health hazards is to conduct a systematic evaluation of the strengths and weaknesses of the included studies. The evaluation of internal validity assesses the extent to which a study can provide accurate (unbiased) evidence of a causal relationship between a given treatment or exposure and a given effect (e.g., does exposure to substance X cause effect A), if such a causal relationship exists. It is often discussed in terms of “risk of bias” or the degree to which specific types of systematic error may have been introduced into the design or execution of a study; these errors can result in a distortion of the results, such that the study does not provide an accurate answer to the research question (Higgins and Green, 2011).

Study sensitivity, as defined here in relation to experimental and observational environmental studies, is a measure of the ability of a study to detect a true effect. It addresses the question “Is this study able to detect a true effect or hazard, if present, that is due to the exposure?” It is analogous to the concept of sensitivity of an assay or a diagnostic test; an insensitive study will fail to detect a difference that truly exists, leading to a false conclusion of no effect; only a negative result from a highly sensitive study can be interpreted, with confidence, as evidence of no effect. Sensitivity can involve features of the study design and measures, study population, and analysis, and can be viewed as a component of internal validity. For the purpose of chemical hazard assessment, study sensitivity is a critical element to include in the evaluation of a study.

This concept of study sensitivity has not been a focus of risk of bias tools and is infrequently discussed with respect to its importance in evaluating “negative” results (of individual studies, or of a collection of studies). This paper aims to contribute to the discussion and development of recommendations for evaluating of studies by providing illustrative examples of study sensitivity and by highlighting the importance of this concept when evaluating epidemiological and toxicological studies. The thesis of this paper is that although the conceptualization of internal validity encompasses study sensitivity, this concept is not necessarily adequately addressed by methods used to operationalize the evaluation of internal validity. The specific consideration of study sensitivity should be included in an evaluation of the evidence to avoid a false conclusion that an exposure has no effect, when the lack of evidence may be due to the insensitivity of the studies to detect an effect.

Examples of Study Sensitivity

In this section, we explore in more detail some examples of study sensitivity for epidemiology and animal toxicology studies. Although some examples are unique to one of these types of studies, others are common to both.

Factors in epidemiology studies related to study sensitivity include, but are not limited to, evidence of substantial exposure (e.g., level, duration, frequency, or probability) during an etiologically relevant time window of exposure, and an adequate range of exposure levels or duration to evaluate exposure-response relationships. For example, the studies that found a positive association with trichloroethylene and kidney cancer were studies of highly exposed workers (NTP, 2015a). The validity of the outcome ascertainment methods, or the ability of the method to accurately differentiate between “diseased” and “non-diseased”, is also important. For cancers with high survival rates, use of cancer mortality, for example from death certificate data, may be an insensitive outcome measure. Similar issues apply to other kinds of outcomes. Adequate length of follow-up in cohort studies for a specific endpoint should also be considered, and will differ depending on the specific exposure, outcome, and potential mechanisms in play. For example, the optimal latency for mesothelioma in relation to asbestos exposure may differ from that of lung cancer, and the optimal latency for breast cancer may differ for an estrogenic chemical than for a chemical acting through a different mode of action. The number of exposed cases is especially important in the evaluation of rare cancers in cohort studies (such as nasopharyngeal cancer, which is linked to formaldehyde exposure) or rare exposures in population-based case-control studies. Although a systematic review may consider study size in the evaluation of precision of the study, this may not be possible when there are no observed cases in a cohort study, or no exposed cases or controls in a case-control study (Fu et al., 2011), or when it is not possible to develop a combined effect estimate across multiple studies. These factors should be evaluated in a systematic manner with the same rigor as the evaluation for the potential for biases and confounding.

Dilution of risk estimates comparing exposed and referent groups, with a reduction in sensitivity, can arise when there is a great deal of variation in the probability, frequency and level of exposure in the group defined as exposed. This issue was specifically noted in a

2006 NRC report on key scientific issues for assessing the health effects of trichloroethylene. The committee noted the potential for dilution of the risk estimate when effect estimates are calculated for “ever exposed” in studies with large numbers of individuals with low levels of exposure (e.g., based on average or cumulative exposure measures) (NRC 2006). Other examples of this type of low sensitivity can be found in studies of lung cancer and asbestos (Marsh et al., 2001), cause-specific mortality in a polyvinylchloride manufacturing cohort and a lead smelting cohort (Parodi et al., 2007), and pregnancy outcomes in relation to formaldehyde exposure among nursing personnel in surgical departments or sterilization units in general hospitals (Hemminki et al., 1982; Hemminki et al., 1985).

Sensitivity is also important in the consideration of animal studies. As with the epidemiology studies, important aspects include the exposure duration and levels, assessing a relevant time window of exposure, and the appropriate timing of endpoint assessment. For example, the most informative cancer bioassays are generally those that expose and observe animals for as long as possible without introducing end-of-life health complications (e.g., a 2-year bioassay), with shorter assays drawing into question the reliability of null findings. This may not always be the most sensitive study protocol, however. While studies of arsenic carcinogenicity in adult animals did not reveal substantial effects, more recent studies of exposure to arsenic or its metabolites suggested that gestational and early postnatal exposure may be a time of particular sensitivity in terms of carcinogenesis (IARC, 2012).

The reliability, specificity, and validity of the endpoint ascertainment in animal studies also requires systematic evaluation, some features of which are routinely covered during outcome assessment in risk-of-bias approaches (e.g., evaluating blinding of outcome assessors; ensuring consistent application of protocols across groups). Other features, such as whether the endpoint was measured at animal ages during which the endpoint being tested was sensitive to change (e.g., based on known biological maturation of the organ or function in question) or whether the specific endpoint evaluation protocol employed might be more or less sensitive for detecting changes in the endpoint being evaluated, may not be adequately addressed. Validation of the non-standard assays that are often the only data available for environmental health assessments through the use of positive and negative controls may be necessary to ensure that the assay can appropriately detect the effect under study. It is also important to consider the specificity of the assay protocols for measuring the outcome of interest. For example, while routine histopathology of all organs at necropsy may be capable of detecting overt damage to the tissue of interest, a more specific evaluation of the target tissue using stereological methods or histopathology from interim sacrifice data for age-related pathological lesions may be able to detect effects that would otherwise be missed.

For endocrine disrupting chemicals, sensitivity to endocrine disruption is highest during tissue development (UNEP-WHO, 2013). Thus, the sensitivity of a study would be reduced by a study design that does not include exposure during the developmental period, does not include the length of follow-up needed to assess latent developmental effects, or for some endpoints, such as pubertal development, examines effects at a time that is too late to detect effects on early maturation. Within the context of neurodevelopment, functional maturation of the nervous system continues through puberty, with an age-dependency to the

development and variability of different behavioral functions (Semple et al., 2013; Rice and Barone, 2000). Testing for effects of exposure prior to the full maturation of the specific behavior in question could make it difficult to detect effects. While endpoint timing sensitivity is an important aspect of study design to consider in a systematic fashion, it is important to recognize that, for some endpoints, knowledge of the underlying biology is poorly understood, and speculative inferences should be avoided.

Additional aspects of sensitivity are specific to experimental animal studies. An example includes the relative susceptibility (or resistance) of the animal model (e.g., species or strain or sex) to the effect under study. Similarly, the manner of administering the test article (e.g., bolus gavage versus dietary exposure) could alter the sensitivity at a specific dose due to differences in toxicokinetics. Although knowledge of differences in sensitivity for many of these factors may be unknown for the specific chemical or outcome being evaluated, it is important that these aspects are considered and recorded for subsequent comparisons or as sources of heterogeneity.

Sensitivity concepts related to the animal model (e.g., age at endpoint assessment, species, sex) are considered separate from the issue of external validity (i.e. the generalizability of findings in these animal models to other models and/ or to humans). These concepts relate to the ability of a study to detect changes in a health outcome of interest using the selected model, and therefore whether the study findings accurately represent the true potential for toxicity (i.e. internal validity). Thus the sensitivity of the animal model – i.e., does it develop the endpoint under evaluation in the study, is a question of internal validity, but the relevance of an animal model to human disease, (is the effect seen in an animal model generalizable to humans?), is a question of external validity.

Approaches to Evaluating Study Sensitivity

The identification of sensitivity factors prior to evaluation of study findings will increase the objectivity of the systematic review. This process requires knowledge of the designs and other details of the studies, and research into what is known about the specific issues such as types of assays, animal models, work setting, and various exposure measures.

In some situations, it may be possible to use inclusion and exclusion criteria to address aspects of the sensitivity component prior to evaluating individual studies for internal, and subsequently external, validity. For a given question, specific criteria could be defined such that if these criteria were not met, the study would not be included. This approach can be incorporated into the procedures used to select studies using “PECO” criteria (i.e., criteria relating to features of population, exposure, comparator, and outcome) (NTP, 2015b; Woodruff and Sutton, 2014). Thus, one could define the population to be people enrolled in an occupational cohort followed for at least 20 years or animals exposed for at least two years in a bioassay if the issue of an adequate latency or exposure duration period, respectively, was deemed to be critical, and if these cut-points represented the minimum time needed for the development of the outcome or endpoint under consideration. This approach may be useful when there is a high degree of certainty concerning the aspects of a study design or measurement (including the underlying biological dependencies in exposed

Author Manuscript

animals) that would substantially reduce the sensitivity or make the studies not informative. Disadvantages of this approach are that it does not allow examination of relative sensitivity as an explanatory factor for variability in results, and it forces a dichotomization of continuous (or multi-level ordinal) attributes. Perhaps most importantly, this approach introduces the potential for exclusion of useful, or even crucial, information. Expanding on the example of carcinogenicity provided above, a cut-off of bioassays less than two years could erroneously exclude shorter-term studies that observed early development of cancers, which could otherwise contribute important information to the chemical assessment. The use of inclusion criteria to address critical aspects of the sensitivity component should carefully consider and address the potential for loss of relevant, informative data.

Author Manuscript

Another approach would be to modify the categories used to assess internal validity (sometimes referred to as risk of bias domains) to ensure that all elements relating to study sensitivity are routinely included. However, because some attributes of study sensitivity are not issues of bias per se, the use of the term “risk of bias” may be confusing to a reader without additional explanation. Specification of these details will result in a systematic collection and evaluation of the relevant information across all of the identified studies. An example of this approach is shown in **Table 1** for observational epidemiology. Note that other choices for categorization are possible, and Table 1 is not an exhaustive list of elements. **Table 2** summarizes some examples of the application of this approach for animal toxicology studies, incorporating specific questions pertaining to specific features of a study. Both the animal toxicology and observational epidemiology examples are intended to be applied for each outcome assessed in a study, rather than to the study as a whole (e.g., a study may be interpreted as insensitive for one set of endpoints, but highly sensitive for others. As noted for the epidemiology studies, different choices for the placement of these items into categories are possible. For both disciplines, an explicit consideration of the sensitivity of studies can increase the transparency of an analysis of heterogeneity among studies during synthesis of study results.

Author Manuscript

Tables 1 and 2 also include examples of how some risk of bias tools currently in use address the sensitivity elements discussed above. The three tools for epidemiology studies are from the Agency for Healthcare Research and Quality Evidence-based Practice Centers (AHRQ EPC) (Viswanathan et al., 2012), the National Toxicology Program Office of Health Assessment and Translation (NTP OHAT (NTP, 2015b), and the Risk of Bias for Non-Randomized Studies – Interventions (ROBINS-I) (Sterne et al., 2014). Some examples show the explicit incorporation of the consideration into the assessment tool, some show the possibility of inclusion (e.g., into an “other sources of bias” category), others are addressed in applicability (discussed below), and others are not addressed. The three tools for animal toxicology studies are from NTP OHAT, Science in Risk Assessment and Policy (SciRAP) (Molander et al., 2014; Beronius et al., 2014), and the Systematic Review Centre for Laboratory Animal Experimentation (SYRCLE) (Hooijmans et al., 2014). As with the epidemiology elements, there are a variety of approaches to address these elements. SciRAP includes these example questions in a variety of places (e.g., in the relevance criteria, and in Tier I and Tier II reliability criteria). Although the inclusion of a category for “other sources

of bias” could prove to be useful, its value will depend on the extent to which these elements are expressly specified during protocol development and reviewer training.

Some of these elements, such as relevance of the exposure period, timing of endpoint evaluation, adequacy of follow-up period, could be considered under applicability or indirectness in the “upgrading” and “downgrading” of confidence of a body of evidence (NTP, 2015b). Elements relating to study sensitivity do affect the internal validity of a study (i.e., the likelihood the study will produce accurate evidence of a causal relationship, if such a relationship exists). Thus we believe it is important to systematically evaluate these elements as part of the evaluation of potential biases of individual studies rather than during evaluations of bodies of evidence (which could overlook such differences in sensitivity across studies). This evaluation should consider not just differences in these elements between groups (e.g., differences in length of follow-up between an exposed and a non-exposed group), but also the absolute value of these elements and how that applies to knowledge of underlying biological processes (e.g., is the length of follow-up appropriate for the outcome under study?)

Alternatively (or in addition to this approach), a separate domain could be developed to address the sensitivity component. A 4-level rating scheme, similar to the risk of bias domains, (e.g., low, modest, serious, or critical risk of insensitivity) could be used, and this evaluation can then be incorporated into the overall evaluation of the study and/or in the confidence of its results. A separate sensitivity domain could be especially useful for elements that may not be thought of as a “bias” *per se*. If a separate sensitivity domain is included in an evaluation, it is important to avoid “double counting” a particular strength or limitation. There also may be situations in which use of an overall assessment of sensitivity is less informative than use of the individual elements (e.g., exposure level) as a stratification variable to facilitate examination of heterogeneity (or consistency) of results.

These approaches are not mutually exclusive and can be combined. For example, it may be useful to devise inclusion and exclusion criteria for some attributes, to modify one component of a risk of bias assessment to incorporate some sensitivity elements, and also to examine an additional set of elements relating to study sensitivity as another evaluation domain.

Recommendations for a comprehensive approach for study evaluation

We have described the concept of the sensitivity of a study, and described examples of approaches to address this concept in a systematic review of environmental health studies. As noted previously, some, but not all, of the attributes relating to study sensitivity could be addressed in domains included in a risk of bias evaluation, if the domains or categories of evaluation are designed to do so. Other attributes are not “biases” *per se*, and so would not easily fit into a risk of bias framework without additional modification (e.g., an additional evaluation category or domain) or changes in terminology. Sensitivity can affect the extent to which the results or inferences accurately reflect the true results for the research question under review; thus, assuring that a risk of bias assessment that adequately addresses the elements affecting study sensitivity fits into the broader goal of evaluating internal validity.

We are not advocating for one approach, or specifying a set of questions to be included in an evaluation of bias. Rather, we are advocating for consistent and thoughtful consideration of these issues in the development of the specific methods used for an evaluation of given topic.

Although sensitivity can be included (in part) in the evaluation of risk of bias, there may be benefits to developing a separate component for the evaluation of sensitivity, or identifying and evaluating specific elements within this component. As noted previously, the discussion of heterogeneity among studies can be aided by an explicit consideration of the sensitivity of individual studies. More importantly, it may be necessary to separately evaluate the sensitivity component because it captures aspects that are missing from the risk of bias evaluation. For example, the relationship of the outcome under study to the particulars of the exposure levels, duration, range and timing must be carefully evaluated to determine whether it was adequately sensitive. An insensitive study may not be able to address the question that is the focus for the systematic review for assessments of health risks of chemical exposures.

The importance of considering study elements specifically related to study sensitivity has been emphasized by several National Research Council (NRC) committees. In a 2014 report of the evaluation of potential cancer hazards of formaldehyde, the criteria used to systematically assess (ranks as strong, moderate or weak) the human epidemiology studies included consideration of aspects relating to what is often included in risk of bias evaluations, and aspects relating to what we have defined above as sensitivity; i.e., a strong study was defined as having a large population, long-duration of exposure, sufficient follow-up for cancer latency, and a substantial proportion of the exposed population that was probably highly exposed to formaldehyde. Similarly, the styrene NRC committee also considered elements such as large cohorts and high and varied exposure as attributes of informative studies. These are aspects of study sensitivity that would affect the interpretation of the results, and so should be included in a systematic evaluation of factors affecting internal validity of each study under review.

Just as a highly sensitive test (or study, or set of studies) is needed to have confidence that a negative result is really negative, a highly specific test (or study, or set of studies) is needed to have confidence that a positive result is really positive. Currently available risk of bias evaluations, such as those described above, generally emphasize elements of specificity, i.e., the confidence that a given positive result is in fact positive. The consideration of study sensitivity in addition to the evaluation of other aspects of internal validity can be used to identify the most informative studies to evaluate the confidence of the findings from individual studies and identify those study elements (e.g., sensitivity, risk of bias) that may help to explain heterogeneity across the body of literature. Systematic reviews that do not evaluate sensitivity may arrive at a different conclusion regarding the hazards of a chemical exposure than if study sensitivity was evaluated. Insensitive studies that have low risk of bias based on other attributes examined (e.g., blinding of outcome assessors, randomization procedures) may not be informative for addressing questions relating to hazard identification and informing public health decisions. Systematic reviews that do not evaluate study sensitivity may inappropriately combine results from studies, give undue weight to the results of insensitive studies, or erroneously interpret evidence as being conflicting.

Addressing the concept of study sensitivity is an important piece in the application of systematic review methods to environmental health and chemical hazard assessments.

Acknowledgements

The authors thank Teneille Walker (National Center for Environmental Assessment, US Environmental Protection Agency) and Kristina Thayer (Office of Health Assessment and Translation, National Toxicology Program, National Institute of Environmental Health Sciences) for their thoughtful reviews of this manuscript. The views expressed are those of the authors and do not necessarily reflect the policies of the US Environmental Protection Agency.

References

- Beronius A, Molander M, Rudén C, Hanberg A. Facilitating the use of non-standard in vivo studies in health risk assessment of chemicals: a proposal to improve evaluation criteria and reporting. *J Appl Toxicol*. 2014; 34:607–617. [PubMed: 24481642]
- Fu R, Gartlehner G, Grant M, Shamliyan T, Sedrakyan A, Wilt TJ, Griffith L, Oremus M, Raina P, Ismaila A, Santaguida P, Lau J, Trikalinos TA. Conducting quantitative synthesis when comparing medical interventions: AHRQ and the Effective Health Care Program. *J Clin Epidemiol*. 2011; 64:1187–1197. [PubMed: 21477993]
- Hemminki K, Mutanen P, Salonenmi I, Niemi ML, Vainio H. Spontaneous abortions in hospital staff engaged in sterilizing instruments with chemical agents. *Br Med J*. 1982; 28:1461–1463.
- Hemminki K, Kyyronen P, Lindbohm ML. Spontaneous abortions and malformations in the offspring of nurses exposed to anesthetic gases, cytostatic drugs, and other potential hazards in hospitals based on registered information of outcome. *J Epidemiol Commun Health*. 1985; 39:141–147.
- Higgins, JPT.; Green, S., editors. *Cochrane Handbook for Systematic Reviews of Interventions* Version 5.1.0. The Cochrane Collaboration; 2011. Chapter 8: Assessing risk of bias in included studies. [updated March 2011] available from www.cochrane-handbook.org
- Hooijmans CR, Rovers MM, de Vries RB, Leenaars M, Ritskes-Hoitinga M, Langendam MW. SYRCLE's risk of bias tool for animal studies. *BMC Med Res Methodol*. 2014; 14:43. DOI: 10.1186/1471-2288-14-43. [PubMed: 24667063]
- IARC (International Agency for Research on Cancer). *IARC Monographs on the Evaluation of Carcinogenic Risks to Humans. Arsenic, Metals, Fibres and Dusts. Vol. 100C. International Agency for Research on Cancer; Lyon, France: 2012. Arsenic.; p. 39-325.*
- IOM (Institute of Medicine). *Finding What Works in Health Care: Standards for Systematic Reviews*. The National Academies Press; Washington, DC: 2011.
- Marsh GM, Youk AO, Stone RA, Buchanich JM, Gula MJ, Smith TJ, Quinn MM. Historical cohort study of US man-made vitreous fiber production workers: I. 1992 fiberglass cohort follow-up: Initial findings. *J. Occup. Environ. Med*. 2001; 43(9):741–756. [PubMed: 11561358]
- Molander L, Ågerstrand M, Beronius A, Hanberg A, Rudén C. Science in Risk Assessment and Policy (SciRAP): An Online Resource for Evaluating and Reporting In Vivo (Eco) Toxicity Studies. *Human and Ecological Risk Assessment*. 2014; 21:753–762.
- NRC (National Research Council). *Assessing the Human Health Risks of Trichloroethylene: Key Scientific Issues*, National Research Council. National Academies Press; Washington, D.C.: 2006. p. 379
- NRC (National Research Council). *Review of the Formaldehyde Assessment in the National Toxicology Program 12th Report on Carcinogens*. National Research Council; National Academies Press; Washington, DC: 2014a. p. 232
- NRC (National Research Council). *Review of Styrene Assessment in the National Toxicology Program 12th Report on Carcinogens*. National Research Council; National Academies Press; Washington, DC: 2014b. p. 178
- NTP (National Toxicology Program). *Report on Carcinogens Monograph on Trichloroethylene* – January 2015. Office of the Report on Carcinogens; RTP, NC: 2015a. Available: <http://ntp.niehs.nih.gov/go/37899>

- NTP (National Toxicology Program). Handbook for Conducting a Literature-Based Health Assessment Using OHAT Approach for Systematic Review and Evidence Integration - January 9 2015. Office of Health Assessment and Translation; RTP, NC: 2015b. Available: <http://ntp.niehs.nih.gov/go/38673>
- Parodi S, Gennaro V, Ceppi M, Cocco P. Comparison Bias and Dilution Effect in Occupational Cohort Studies. *Int J Occup Environ Health*. 2007; 13:143–152. [PubMed: 17718170]
- Rice D, Barone S Jr. Critical periods of vulnerability for the developing nervous system: evidence from humans and animal models. *Environ Health Perspect*. 2000; 108(Suppl 3):511–533. [PubMed: 10852851]
- Rooney AA, Boyles AL, Wolfe MS, Bucher JR, Thayer KA. Systematic Review and Evidence Integration for Literature-Based Environmental Health Science Assessments. *Environ Health Perspect*. 2014; 122:711–718. [PubMed: 24755067]
- Schneider K, Schwarz M, Burkholder I, Kopp-Schneider A, Edler L, Kinsner-Ovaskainen A, Hartung T, Hoffmann S. “ToxRTool”, a new tool to assess the reliability of toxicological data. *Toxicology Letters*. 2009; 189:138–144. [PubMed: 19477248]
- Semple BD1, Blomgren K, Gimlin K, Ferriero DM, Noble-Haesslein LJ. Brain development in rodents and humans: Identifying benchmarks of maturation and vulnerability to injury across species. *Prog Neurobiol*. 2013; 106-107:1–16. doi: 10.1016/j.pneurobio.2013.04.001. Epub 2013 Apr 11. [PubMed: 23583307]
- Sterne, J.; Higgins, J.; Reeves, B. [March 11, 2016] on behalf of the development group for ROBINS-I: a tool for assessing Risk Of Bias In Non-randomized Studies of Interventions, Version 7 March 2016. Available from <http://www.riskofbias.info>
- United Nations Environment Programme - World Health Organization (UNEP-WHO). Bergman A, Heindel JJ, Jobling S, Kidd KA, Zoeller RT. State of the science of endocrine disrupting chemicals - 2012. 2013
- Viswanathan, M.; Ansari, MT.; Berkman, ND.; Chang, S.; Hartling, L.; McPheeters, M.; Lina Santaguida, P.; Shamliyan, T.; Singh, K.; Tsertsvadze, A.; Treadwell, JR. Methods Guide for Effectiveness and Comparative Effectiveness Reviews. Agency for Healthcare Research and Quality; Rockville, MD: Jan. 2014 Assessing the Risk of Bias of Individual Studies in Systematic Reviews of Health Care Interventions. Chapter 9AHRQ Publication No. 10(14)-EHC063-EFChapters available at: www.effectivehealthcare.ahrq.gov
- Whaley, P.; Halsall, C.; Ågerstrand, M.; Benford, D.; Aiassa, E.; Bilotta, G.; Boyd, I.; Coggon, D.; Collins, C.; Dempsey, C.; Duarte-Davidson, R.; Fitzgerald, R.; Galay Burgos, M.; Gee, D.; Hart, A.; Hoffmann, S.; Lam, J.; Lasseresson, T.; Levy, L.; Lipworth, S.; Mackenzie Ross, S.; Martin, O.; Meads, C.; Meyer- Baron, M.; Miller, J.; Mongelard, P.; Pease, P.; Rooney, A.; Sapiets, A.; Stewart, G.; Taylor, D.; Verloo, D. “Systematic Review Methods for Advancing Chemical Risk Assessment”. Environment International; Implementing systematic review techniques in chemical risk assessments: challenges and opportunities.. in press
- Woodruff TJ, Sutton P. The Navigation Guide systematic review methodology: a rigorous and transparent method for translating environmental health science into better health outcomes. *Environ Health Perspect*. 2014; 122:1007–1014. [PubMed: 24968373]

Highlights

- The concept of sensitivity, the ability to detect a true effect, is discussed.
- Low study sensitivity may result in uninformative results.
- Approaches for systematically evaluating study sensitivity are presented.
- Addressing study sensitivity can improve evaluation of chemical hazard studies.

Table 1

Example considerations relating to sensitivity of study methods (ability to detect an effect) – epidemiology studies

Element	Example Considerations	Example Applications in Risk of Bias Tools
Population and Design	<ul style="list-style-type: none"> • Is the follow-up period adequate to allow for the development of the outcome of interest? • Did entry into the cohort begin with start of the exposure? • Are the numbers of exposed cases adequate to detect an effect in the exposed population and/or subgroups of the exposed population? 	<ul style="list-style-type: none"> • AHRQ EPC: could be considered under “applicability” • NTP OHAT: considered under “applicability” • ROBINS-I: not explicitly considered • AHRQ EPC: not explicitly considered • NTP OHAT: not explicitly considered • ROBINS-I: considered under time-varying confounding and selection of participants • AHRQ EPC, NTP OHAT, and ROBINS-I: may be addressed through assessment of precision (e.g., of summary measure); consider for individual studies when meta-analysis not possible
Outcome	<ul style="list-style-type: none"> • Are the outcome ascertainment methods reliable and valid; is the outcome under study a sensitive indicator of the health effect of interest? 	<ul style="list-style-type: none"> • AHRQ EPC: included in bias related to outcome measures (detection bias) • NTP OHAT: could be incorporated into confidence in outcome characterization • ROBINS-I: not explicitly considered
Exposure	<ul style="list-style-type: none"> • Are the levels, duration, or range of exposure sufficient or adequate to detect an effect of exposure? • Does the exposed group include individuals with a low or unknown probability of exposure? • Does the exposure assessment capture the relevant window of exposure? • Does the exposure assessment capture the relevant exposure metric? 	<ul style="list-style-type: none"> • AHRQ EPC: some (e.g., relevant window of exposure) but not all elements (e.g., level, duration or range of exposure) could be included in bias related to exposure measures (detection bias); • NTP OHAT: some (such as exposure metric and range) could be incorporated into confidence in exposure characterization; other elements (e.g., exposure duration, relevant window of exposure) are considered under applicability or as a inclusion factor • ROBINS-I: could be considered under definition of the intervention (or exposure)
Confounding	<ul style="list-style-type: none"> • Does adjustment for potential confounders include variables on the pathway between exposure and outcome? 	<ul style="list-style-type: none"> • AHRQ EPC: not explicitly considered, but could be included under selection bias criteria • NTP OHAT: could be considered under confounding bias • ROBINS-I: explicitly considered under confounding
Other Aspects of Analysis	<ul style="list-style-type: none"> • Does the analysis appropriately consider subgroups of interest (e.g., based on variability in exposure level, duration or latency)? • Does the analysis use a continuous outcome variable, when appropriate to do so? 	<ul style="list-style-type: none"> • AHRQ EPC: not explicitly considered • NTP OHAT: not explicitly considered (could be included in “other sources of bias”) • ROBINS-I: not explicitly considered

AHRQ EPC = Agency for Healthcare Research and Quality Evidence-based Practice Centers; NTP OHAT = National Toxicology Program Office of Health Assessment and Translation; ROBINS-I = Risk of Bias for Non-Randomized Studies – Interventions.

Table 2

Example considerations relating to sensitivity of study methods (ability to detect an effect) – animal toxicology studies

Element	Example Questions	Example Applications in Risk of Bias Tools
Animal model Species, strain, sex	<ul style="list-style-type: none"> Is the selected animal model likely to underrepresent changes in the endpoint(s) under study? [Consider the extent to which the species, strain, or sex is sensitive or suitable for the specific effects or endpoints under evaluation; note: consider current knowledge regarding the relative susceptibility or resistance of the selected model; sensitivity is generally assumed when specific knowledge otherwise is lacking] 	<ul style="list-style-type: none"> NTP OHAT: not explicitly considered; relevance of animal species considered under applicability SciRAP: considered under relevance criteria SYRCLE: not explicitly considered (could be included in “other sources of bias”)
Design Evaluation of relevant endpoints	<ul style="list-style-type: none"> Are sensitive methods used for endpoint evaluation? [or, what is the reliability, specificity, and validity of the method used for endpoint evaluation?] Is the timing of endpoint evaluation appropriate for detecting an effect (e.g., is the latency period sufficient to allow development of the endpoint; does the animal age at evaluation allow for accurate measurement of the endpoint; for some endpoints, evaluation too early or too late may mask an effect) Use of positive and negative assay controls, where applicable (e.g., positive controls may not be needed for established protocols, but could be essential for assays that are under development) Is sampling (e.g., sample size per group; number of observations per endpoint) expected to be sufficient to detect differences across groups? 	<ul style="list-style-type: none"> NTP OHAT: could be incorporated into confidence in outcome characterization; timing of endpoint evaluation considered under applicability SciRAP: negative controls considered under Tier I criteria; Tier II administration of test substance criteria and measurement/data collection criteria includes positive controls and appropriateness of the methods for the endpoints SYRCLE: not explicitly considered (could be included in “other sources of bias”)
Exposure Number of concentrations or dose levels and their range, timing and duration of exposure	<ul style="list-style-type: none"> Do the concentration/dose levels span a range in which effects can be expected to occur? Is the timing and duration of exposure relevant for the endpoints under study? If the sensitive exposure window is unknown, is the exposure period wide enough to cover likely possibilities? Is the route and method (e.g., gavage versus diet) of administration appropriate for the chemical and endpoint(s) being studied? 	<ul style="list-style-type: none"> NTP OHAT: dose, timing and route of exposure considered under applicability SciRAP: some elements considered under relevance criteria; others considered under Tier II test compound and administration of test substance criteria SYRCLE: not explicitly considered (could be included in “other sources of bias”)
Analysis Statistical analysis and presentation (aspects that could obscure or exaggerate an effect)	<ul style="list-style-type: none"> Does the presentation of the data fully and accurately represent the results (e.g., pooling across dose groups or sexes or related endpoints without justification)? Are alternative analyses or displays that are likely to change the interpretation of the results (e.g., presenting continuous data as dichotomized)? 	<ul style="list-style-type: none"> NTP OHAT: not explicitly considered (could be included in “other sources of bias”) SciRAP: considered under Tier II measurement/data collection criteria and statistics criteria SYRCLE: not explicitly considered (could be included in “other sources of bias”)

NTP OHAT = National Toxicology Program Office of Health Assessment and Translation; SciRAP = Science in Risk Assessment and Policy;
SYRCLE = SYstematic Review Centre for Laboratory animal Experimentation