Wikoff et al., Role of Risk of Bias in Systematic Review for Chemical Risk Assessment: A Case Study in Understanding the Relationship Between Congenital Heart Defects (CHDs) and Exposures to Trichloroethylene (TCE)

Supplemental Materials:

Refinements* to OHAT RoB tool for TCE-CHD: Experimental Animal Studies

	RoB Question	*Interpretation (animal studies only)*
Q1a	Was administered dose or exposure level adequately randomized?	**Definitely Low** - explicitly report randomization method. **Probably Low** - explicitly report randomization, but not method. **Probably High** - no information on randomization reported, or suggestion of non-randomization. **Definitely High** - explicitly report non-randomization.
Q1b	Were control and dose groups run concurrently?	**Definitely Low** - explicitly report concurrent administration of control/treatment groups. **Probably Low** - indirect evidence that the study used a concurrent control; OR concurrent administration assumed if authors did not report non-concurrent administration for control and treatment groups. **Probably High** - Indirect evidence that there was a lack of concurrent control group. **Definitely High** - there is direct evidence that there was a lack of a concurrent control group.
Q5a	Were experimental conditions identical across study groups – same vehicles?	**Definitely Low** - direct evidence that same vehicle used in control and experimental animals. **Probably Low** - indirect evidence that same vehicle used in control and experimental animals. **Probably High** - indirect evidence that vehicle differed between control and experimental animals; OR authors did not report the vehicle used. **Definitely High** - direct evidence that vehicle differed between control and experimental animals.
Q5b	Were experimental conditions identical across study groups – non-treatment-related experimental conditions?	**Definitely Low** - direct evidence that non-treatment-related experimental conditions were identical across study groups. **Probably Low** - identical non-treatment-related experimental conditions are assumed if author did not report differences in animal housing or husbandry. **Probably High** - indirect evidence that non-treatment-related experimental conditions were not comparable between study groups. **Definitely High** - direct evidence that non-treatment-related experimental conditions were not comparable between study groups.
Q7	Were outcome data complete without attrition or exclusion from analysis?	**Definitely Low** - direct evidence that loss of animals was adequately addressed and reasons were documented when animals were removed from a study; OR there is direct evidence that no animals died or were removed from the study due to toxicity. **Probably Low** - indirect evidence that loss of animals was adequately addressed and reasons were documented when animals were removed from a study; OR there is indirect evidence that no animals died or were removed from the study due to toxicity. **Probably High** - indirect evidence that loss of animals was unacceptably large and/or not adequately addressed; OR insufficient evidence provided about loss of animals. **Definitely High** - direct evidence that loss of animals was unacceptably large and not adequately addressed.
Q8a	Can we be confident in the exposure characterization? – Test article purity	**Definitely Low** - direct evidence purity confirmed generally ≥ 99% for single or mixed substance. **Probably Low** - indirect evidence purity confirmed generally ≥ 99% for single or mixed substance (chemical supplier documents purity of chemical), or ≥98% for single substance with expectation that 2% impurities would not bias results. **Probably High** - authors did not report chemical purity (NR - insufficient information) **Definitely High** - there is direct evidence that purity was <98% for single substance, and/or impurities would be expected to bias results
Q8b	Can we be confident in the exposure characterization? – test agent solution concentration and stability	**Definitely Low** - direct evidence that exposures were quantitatively characterized prior to and/or during administration (i.e., authors report test agent solution concentrations and/or stability assessed, and method used). **Probably Low** - indirect evidence that exposures were quantitatively characterized prior to and/or during administration (i.e., authors report test agent solution concentrations and/or stability assessed, and method used). **Probably High** - indirect evidence that exposures were quantitatively assessed using poorly validated methods; OR there is insufficient information on quantitative assessment (NR - insufficient information). **Definitely High** - direct evidence that exposures were quantitatively assessed using poorly validated methods.
Q8c	Can we be confident in the exposure characterization? – consistent administration	**Definitely Low** - direct evidence that exposure was consistently administered (i.e., with the same method and time-frame) across treatment groups [e.g., guideline study, daily dose administration times reported (oral gavage), exposure monitoring (inhalation and drinking water), or single animal housing (drinking water studies)]. **Probably Low** - indirect evidence that exposure was consistently administered (i.e., with the same method and time-frame) across treatment groups.

		Probably High - indirect evidence that doses were not administered on a consistent basis between treatment groups (e.g., group housing in drinking water studies); OR there is insufficient information on administration consistency (NR - insufficient information). **Definitely High** - direct evidence that doses were not administered on a consistent basis between treatment groups.
Q9a	Can we be confident in the outcome assessment? – outcome assessment method	**Definitely Low** - direct evidence that the outcome was assessed using well-established methods (e.g., the Staples method (1974) and the close variant published by Stuckhardt and Poppe (1984) are considered the gold standard method for identifying developmental defects). **Probably Low** - indirect evidence that the outcome was assessed using acceptable methods (i.e., deemed valid and reliable but not "gold standard"; e.g., the Wilson (1965) method – an acceptable method commonly used to detect many cardiac defects – is cited in OECD (414) guideline). **Probably High** - indirect evidence that the outcome assessment method is insensitive instrument; OR the method has not been in common use or validated; OR there is insufficient information on the outcome assessment method used related to relevant endpoint (NR). **Definitely High** - direct evidence that the outcome assessment method is an insensitive instrument.
Q9b	Can we be confident in the outcome assessment? – outcome assessors adequately blinded	**Definitely Low** - direct evidence that the outcome assessors were adequately blinded to the study group. **Probably Low** - indirect evidence that the outcome assessors were adequately blinded to the study group. **Probably High** - indirect evidence that it was possible for the 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. **Definitely High** - direct evidence for lack of adequate blinding of outcome assessors.
Q10	Were all measured outcomes reported?	**Definitely Low** - direct evidence that all measured study outcomes (primary and secondary) outlined in the protocol, methods, abstract, and/or introduction (that are relevant for the evaluation) have been reported. **Probably Low** - indirect evidence that all of the measured study outcomes (primary and secondary) outlined in the protocol, methods, abstract, and/or introduction (that are relevant for the evaluation) have been reported. **Probably High** - indirect evidence that all of the measured study outcomes (primary and secondary) outlined in the protocol, methods, abstract, and/or introduction (that are relevant for the evaluation) have not been reported; OR there is insufficient information provided about selective outcome reporting. **Definitely High** - direct evidence that all of the measured study outcomes (primary and secondary) outlined in the protocol, methods, abstract, and/or introduction (that are relevant for the evaluation) have not been reported.
Q11	Were appropriate statistical units evaluated and reported?	**Definitely Low** - developmental effects were evaluated on a per-litter basis, and were reported using a justified statistical significance of the outcome; OR developmental effects were evaluated on a per-litter basis, and there was direct evidence that no outcome effect was observed/reported (e.g., method includes heart defect evaluation, and the paper explicitly states no heart defects were observed). **Probably Low** - developmental effects were evaluated on a per-litter basis, but were not reported using a justified statistical significance of the outcome; OR developmental effects were evaluated on a per-litter basis, and there was indirect evidence that no outcome effect was observed/reported (e.g., method includes heart defect evaluation, but heart defects not mentioned in paper b/c authors only reporting observed effects). **Probably High** - There is insufficient information provided about statistical analysis of developmental effects. **Definitely High** - The investigators evaluated developmental effects on a per-fetus basis only.

*Per OHAT recommendations, investigators should tailor the domains to the specific research question. The table provides description of domains tailored to TCE-CHD evaluation; no refinements were made to domains not listed.