Table 1 Overview of Guidance Systems for the Quality Evaluation of Human Studies

Guidance Criteria	IRIS RoB*	OHAT*	STROBE	Money *et al.* (2013)[a]	Navigation Guide[b,*]
Study Objectives			Report		Report
Study Design and Setting (*e.g.*, Date, Location)	Report	Report	Report		
Participant Characteristics (*e.g.*, Age, Race, Sex, Eligibility Criteria)	Report	Report	Report[c]		
Study Size	Report	Report	Report		
Sufficient so that estimates not subject to high imprecision				Y/N	
Consistent Recruiting Methods					Score
Study Power Analysis	Report				
Blinding					
Participants	Score	Score			
Outcome assessors	Score	Score			Score
Participation Rate/Attrition		Report	Report		
Attrition similar across groups	Score	Score			
Loss to follow-up minimized	Score	Score		Y/N	
Potential for selection bias	Discuss				
Inclusion/Exclusion Criteria		Report			
Comparison Groups		Report	Report		
Similar to cases/exposed	Score	Score			
Statistical Methods	Score	Score	Report		
Appropriate techniques				Y/N*	
Data Sources			Report		
Data Measurement Methods	Report	Report	Report		
Sources of Bias and Confounding	Discuss	Discuss	Report		
How Confounding and Bias Addressed	Report, Score	Report, Score	Report		Score
Co-exposures controlled for	Score	Score			
Possibility for bias reduced through design[d]				Y/N*	
Exposure Characterization					
Exposure levels and unit of measurement	Report	Report	Report		
Measurement sensitive and applied consistently	Score	Score			Score
Exposure assessment made independent of outcome[e]				Y/N	
Validated Outcome Assessment Methods	Score	Score	Report	Y/N	
Outcome assessed independent of exposure status				Y/N	
Potential for outcome misclassification	Discuss				
Adherence to Study Protocol	Score	Score			
Study Results	Report	Report	Report	Report	Report
Detailed results, adjusted & unadjusted analyses	Report	Report	Report		Report
Results of sensitivity or other analyses			Report		
All measured outcomes reported	Score	Score			Score
Limitations			Report	Report	
Interpretation			Report		
Unambiguous interpretation[f]				Y/N	
Generalizability	Discuss		Report		
Funding Source/COI Statement		Report	Report	Report	Score
"Other" Bias					Score

Notes:

COI = Conflict of Interest, IRIS RoB = Integrated Risk Information System Risk of Bias; OHAT = The Office of Health Assessment and Translation (Part of the National Toxicology Program); STROBE = Strengthening the Reporting of Observational Studies in Epidemiology.

Guideline Key: Discuss = Address this issue in some way (no specific criteria, and not considered directly as part of the scored 15 questions in the IRIS RoB framework); Report = Reporting Requirement; Score = Scored for category based on the extent that issues were addressed; Y/N = Criteria Fulfilled (*i.e.*, "Yes" or "No").

Sources:

IRIS RoB = US EPA (2013, 2014).

OHAT = NTP (2013a,b).

STROBE = von Elm *et al.* (2007a-e).

Navigation Guide = Koustas *et al.* (2014, 2013); Woodruff and Sutton (2014); Johnson *et al.* (2014); Lam *et al.* (2014).

* Indicates a criteria (or system) that is specifically stated as a risk of bias consideration. All the criteria in the IRIS, OHAT, and Navigation Guide approaches are considered risk of bias issues. The only exception is the "Generalizability" criteria for IRIS, which is discussed in the context of study quality in US EPA's original guidance document (US EPA, 2013).

(a) The quality criteria below are specific to Money *et al.* (2013); the authors also state that all methodology and results should be "comprehensively and transparently" reported according to guidelines such as the STROBE guidelines. If all the criteria detailed in this table are fulfilled, overall, the study is considered "reliable without restriction." Money *et al.* (2013) also provide guidelines for overall ranking of a study if some criteria are missed, which correspond with overall ratings of "reliable with restriction," "not reliable," or "not assignable."

(b) The Navigation Guide was originally developed for systematic reviews of animal studies, but it has also been applied for epidemiology studies in a systematic review of perfluorinated compounds (Johnson *et al.*, 2014).

(c) The authors stipulate that this information should be provided separately for cases and controls in case-control studies or exposed and unexposed groups in cohort/cross-sectional studies.

(d) Through statistical methods or sensitivity analyses.

(e) Authors emphasize the importance of well-established, validated, quantitative exposure assessment methods at the individual level, with as little measurement error as possible.

(f) Methods (and, thus, results) are without appreciable limitations, such that the reader is able to draw causal inference with respect to the exposure and outcome under consideration.

Table 2 Overview of Guidance Systems for the Quality Evaluation of Animal Studies

Guidance Criteria	ARRIVE	Klimisch	OECD GD 34[a]	ToxRTool[b]	IRIS RoB*	OHAT*	Navigation Guide*
Study Objectives	Report		Report	Optional			Report
Study Design and Setting (*e.g.*, dates of dosing and evaluation periods)	Report		Report		Report[c]	Report[c]	Report[d]
Followed OECD procedure? GLP conditions?				Optional		Report	
Animal Characteristics (Species, Age, Stage, Sex, Weight)	Report	Score	Report	Report	Report[c]	Report[c]	Report
Substance (Composition, CAS #, Purity)	Report	Score	Report	Report	Report	Report	Report
Total Study Size (Number of Control and Experimental Groups)	Report	Score	Report	Y/N	Report	Report	Report
Number of animals per dose group	Report	Score	Report	Report	Report[c]	Report[d]	Report
Source of animals	Report		Report			Report[d]	
Additional relevant information (genetic modification, genotype, health status)	Report		Report		Report		Report
Attrition minimized					Score	Score	Report
Blinding & Subject Randomization	Report*		Report*		Score	Score	Score
Experimental Unit (Single Animal, Cage of Animals)	Report		Report				Report
Husbandry Details (Breeding Program, Access to Food and Water, Light and Dark Cycle)	Report	Score	Report	Y/N	Score	Score[c]	Report
Housing conditions	Report	Score	Report	Y/N	Score	Score	Report
Experimental Procedure	Report		Report		Score	Report[c]	Report
Dose groups, substance preparation, administration route	Report	Score	Report	Report	Score	Report[c]	Report
Time and location of dose administration	Report		Report	Report	Score	Report[c]	Report
Rationale for method used	Report		Report				
Impact of Protocol Deviations					Score	Score	
Outcome Assessment Methods	Report	Score	Report	Y/N	Score	Score	
Statistical Methods Used	Report		Report	Y/N	Score	Score	
Results, Adjusted & Unadjusted	Report		Report	Y/N	Score	Score	Score
Report non-significant results			Report		Score	Score	Score
Baseline Data for Each Experimental Group	Report		Report		Score		
Number of Subjects Included in Statistical Analysis and Rationale for Exclusion of Subjects	Report		Report		Score	Score	
Reliability and Appropriateness of Test for Endpoint Analyzed			Report	Y/N	Score		
Consideration of Confounding or Modifying Variables					Score	Score	
Precision of Results (Standard Deviation, Confidence Interval)	Report		Report				Report
Description of Adverse Events Observed	Report	Score	Report			Score	
Dose/Concentration Relationship		Score	Report				
Limitations	Report		Report				
Interpretation & Implications	Report		Report				
Generalizability	Report		Report	Report			
Funding Source	Report				Report	Report	Score
"Other" Study Design Bias							Discuss

Notes:
ARRIVE = Animal Research: Reporting of *In Vivo* Experiments; CAS # = Chemical Abstracts Service Number; IRIS RoB = Integrated Risk Information System Risk of Bias; OECD GD = Organisation for Economic Co-operation and Development Guidance Document; OHAT = The Office of Health Assessment and Translation (Part of the National Toxicology Program).

Guideline Key: Discuss = Address this issue in some way (no specific criteria, and not considered directly as part of the scored 15 questions in the IRIS RoB framework); Report = Reporting Requirement; Score = Scored for category based on the extent that issues were addressed; Y/N = Criteria Fulfilled (*i.e.*, "Yes" or "No").

Sources:

ARRIVE = Kilkenny *et al.* (2010).
Klimisch = Klimisch *et al.* (1997).
OECD GD 34 = OECD (2005).
ToxRTool = European Commission (Undated).
IRIS RoB = US EPA (2013, 2014).
OHAT = NTP (2013a,b).
Navigation Guide = Koustas *et al.* (2014, 2013); Woodruff and Sutton (2014); Johnson *et al.* (2014); Lam *et al.* (2014).

* Indicates a criteria (or system) that is specifically stated as a risk of bias consideration. All the scoring criteria in the IRIS, OHAT, and Navigation Guide approaches are considered risk of bias issues.

(a) OECD Guideline 34, "Guidance Document on the Validation and International Acceptance of New or Updated Test Methods for Hazard Assessment." The OECD guidelines outline criteria for the development of new test methods, rather than assessment criteria for completed studies.

(b) Criteria marked "Report" for ToxRTool must be fulfilled in order to achieve "reliable" score.

(c) Study characteristics that should be reported are not explicitly stated but are provided in example tables for animal studies in the arsenic and perfluorinated compounds assessments for IRIS and OHAT, respectively.

(d) The Navigation system has specific reporting requirements for reproductive and developmental study methodologies.

Table 3 Overview of Guidance Systems for the Quality Evaluation of *In Vitro* Studies

Guidance Criteria	ARRIVE	Klimisch	OECD GD 34[a]	ToxRTool[b]
Study Objectives	Report		Report	Optional
Study Design and Setting (Date, Location, *etc.*)	Report		Report	
Test System and Test Method		Score	Report	Y/N
Followed OECD procedure? GLP conditions?		Score	Report	Optional
Substance (Composition, CAS #, Purity, Source)	Report	Score	Report	Report
Source of Test System and Substance		Score	Report	Y/N
Total Study Size (Number of Control and Experimental Groups)	Report		Report	
Study Size – Number of Replicates			Report	Y/N
Blinding and Subject Randomization	Report*		Report[c]*	
Experimental Procedure	Report	Score	Report	Y/N
Dose group, substance preparation, administration route	Report	Score	Report	Report
Positive or Negative Controls	Report	Score	Report	Report
Outcome Assessment Methods	Report	Score	Report	Y/N
Statistical Methods Used	Report		Report	Y/N
Results, Adjusted and Unadjusted	Report		Report	Y/N
Number of Subjects Included in Statistical Analysis (and Rationale for Exclusion of Subjects)	Report		Report	
Data on Observations That May Influence Interpretation (pH Shift, Impurities, Solubility)		Score	Report	
Precision of Results (Standard Deviation, Confidence Interval)	Report		Report	
Description of Adverse Events Observed	Report		Report	
Dose/Concentration-Response Relationship	Report	Score	Report	
Reliability and Appropriateness of Test for Endpoint Analyzed		Score	Report	Y/N
Limitations	Report		Report	
Interpretation and Implications	Report		Report	
Generalizability	Report		Report	Report
Funding Source	Report			

Notes:

ARRIVE = Animal Research: Reporting of *In Vivo* Experiments; CAS # = Chemical Abstracts Service Number; OECD GD = Organisation for Economic Co-operation and Development Guidance Document.

Guideline Key: Report = Reporting Requirement; Score = Scored for category based on the extent that issues were addressed; Y/N = Criteria Fulfilled (*i.e.*, "Yes" or "No").

Sources:

ARRIVE = Kilkenny *et al.* (2010).
Klimisch = Klimisch *et al.* (1997).
OECD GD 34 = OECD (2005).
ToxRTool = European Commission (Undated).

* Indicates a criteria (or system) that is specifically stated as a risk of bias consideration.

(a) OECD Guideline 34 "Guidance Document on the Validation and International Acceptance of New or Updated Test Methods for Hazard Assessment." The OECD guidelines outline criteria for the development of new test methods, rather than assessment criteria for completed studies.

(b) Criteria marked "Report" for ToxRTool must be fulfilled in order to achieve "reliable" score.

Table 4 Criteria for the Quality Evaluation of Systematic Reviews

Guidance Criteria	IRIS	OHAT	AMSTAR	Navigation Guide
Review Objective Identified		Y/N	Y/N	Y/N
A priori Design/Protocol for the Review	Y/N	Y/N	Y/N	Y/N*
Comprehensive Literature Search of More than One Database	Y/N	Y/N	Y/N	Y/N
Details of the Search Strategy (Including: Date of search and any updates, databases used, *a priori* inclusion criteria)	Report	Report	Report	Report
Inclusive Literature Approach Used[a]	Y/N			
Iterative Literature Identification (*i.e.*, contacting subject matter experts for sources for grey literature)		Y/N		
Two Independent Reviewers Of Data	Y/N	Y/N	Y/N	Y/N
Procedure for Disagreements Between Study Reviewers	Y/N	Y/N	Y/N	Y/N
List of Excluded and Included Studies	Y/N*		Y/N	Y/N
Reasons for study exclusion	Y/N*	Y/N*		Y/N
Study Characteristics Reported (*e.g.*, in a table)[b]	Y/N		Y/N	
Study Results Provided without Restriction (based on statistically significant or positive associations)	Y/N*			
Assessment and Documentation of the Scientific Quality of Each Study	Score*	Score[c]*	Y/N[d]*	Score*
If studies assessed for individual quality, considerations/criteria transparently detailed	Y/N	Y/N	Y/N	Y/N
Individual study quality scores provided in tabular format	Y/N	Y/N		Y/N
Classification of individual studies into quality tiers	Optional	Optional		Y/N
Appropriate Methods to Combine Findings Across Studies[e]			Y/N	Y/N
Overall Confidence Rating for Body of Evidence	Score	Score		Score
Consideration of risk of bias, temporality, magnitude of effect, dose-response, unexplained inconsistency, relevance of endpoints, and imprecision	Score*	Score*		Score*
Qualitative Assessment of Publication Bias		Y/N*	Y/N*	Score*
Determination of Level of Evidence for Health Effect		Score		Score
Overall Conclusions for Hazard Identification	Score[f]	Score[g]		Score[h]
Statement of Possible Conflict of Interest in Both Systematic Review and Included Studies		Y/N	Y/N	Score*
Discussion of Deviations from Review Protocol (provided and justified)			Y/N	

Notes:

AMSTAR = Assessment of Multiple Systematic Reviews System; IRIS = Integrated Risk Information System; OHAT = The Office of Health Assessment and Translation.

Guideline Key: Report = Reporting Requirement; Score = Scored for category based on the extent that issues were addressed; Y/N = Criteria Fulfilled (*i.e.*, "Yes" or "No").

Sources:

IRIS = US EPA (2013).

OHAT = NTP (2013a,b).

AMSTAR = Shea *et al.* (2007).

Navigation Guide = Koustas *et al.* (2014, 2013); Woodruff and Sutton (2014); Johnson *et al.* (2014); Lam *et al.* (2014).

* Indicates a criteria (or system) that is specifically stated as a risk of bias consideration.

(a) Reviewers should err on the side of inclusion (*i.e.*, it is better to include a study in the systematic evaluation and examine the impact of potential limitations, rather than exclude a study and lose any information it could have provided).
(b) IRIS criteria require that very specific details be provided (*e.g.*, description of comparison groups and prevalence of important confounders in these groups as well as the preference that reviewers present study sizes by exposure/outcome group).
(c) OHAT's risk of bias system is the same as IRIS but with fewer details provided in the guidance. OHAT states these criteria are based on Guyatt *et al.* (2011) "GRADE" guidelines for risk of bias.
(d) No specific requirements for quality criteria; AMSTAR simply states that the criteria should be developed *a priori* and described.
(e) For pooled results, a test should be done to ensure that studies were combinable, to assess their homogeneity (*i.e.*, Chi-squared test for homogeneity). If heterogeneity exists, a random effects model should be used and/or the clinical appropriateness of combining these results should be considered.
(f) Quality of the individuals are qualitatively evaluated and pooled to form overall conclusions on the body of evidence depending on the likelihood that bias and confounding indicate possible alternative explanations for associations. Categories are "sufficient," "suggestive," or inadequate" epidemiologic evidence of an association consistent with causation, or "epidemiologic evidence consistent with no association."
(g) Based on the evidence, categorize as "known," "presumed," or "suspected" hazard to humans or "not classifiable or not identified to be a hazard to humans."
(h) Based on the evidence, categorize a particular exposure as "known to be toxic," "probably toxic," "possibly toxic," "not classifiable," or "probably not toxic" (in this framework, this is applied specifically to reproductive and developmental health).