EPA United States Environmental Protection Agency

Systematic Review for Chemical Assessments: Core Elements and Considerations for Rapid Response

Kris Thayer, National Center for Environmental Assessment (NCEA) Integrated Risk Information System (IRIS) Division Director

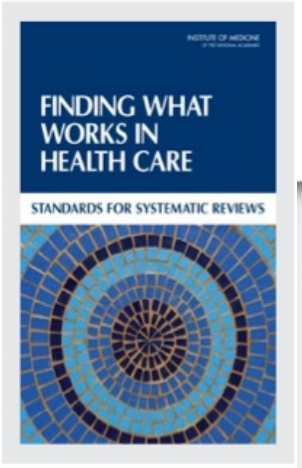
EPA's Computational Toxicology Communities of Practice

November 16, 2017

Outline

- What is a systematic review?
  - Core elements in context of IRIS assessments
  - Cross-walking terminology for study quality and weight of evidence assessment
- Potential areas of overlapping interest with CompTox community
  - Use of structured frameworks for expressing confidence in conclusions
  - Ensuring transparency during rapid response
  - Use of specialized SR software applications/automation for efficiency and data sharing

Systematic Review



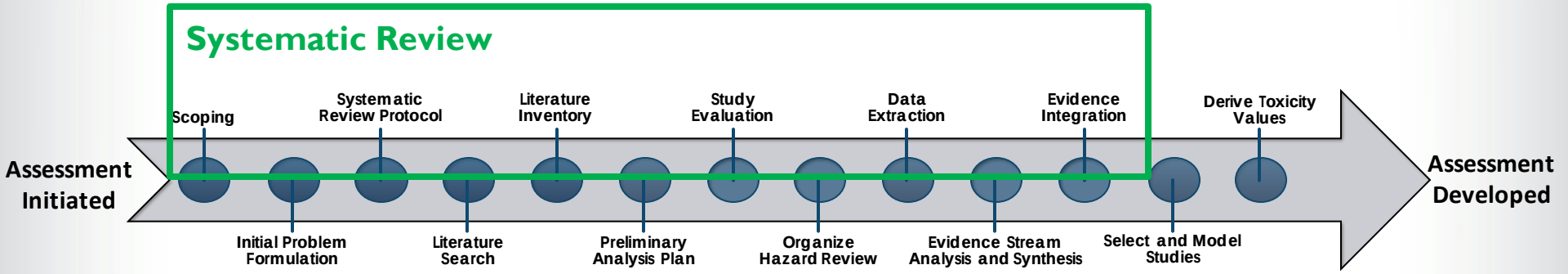
A structured and documented process for transparent literature review[1,2]

"… systematic review is a scientific investigation that focuses on a specific question and uses explicit, pre-specified scientific methods to identify, select, assess, and summarize the findings of similar but separate studies. The goal of systematic review methods is to ensure that the review is complete, unbiased, reproducible, and transparent"

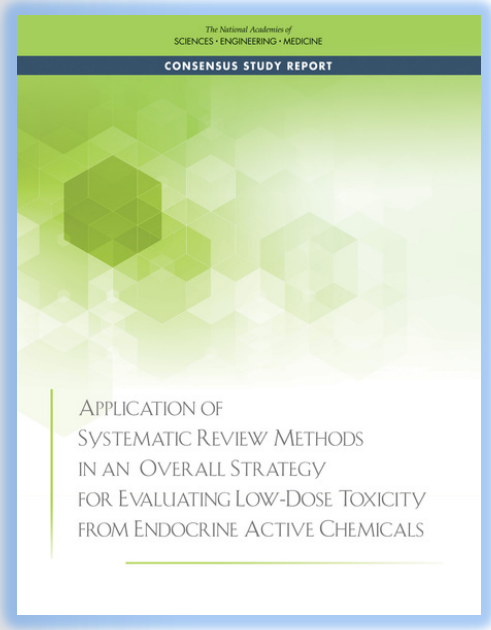
[1] Procedures for Chemical Risk Evaluation Under the Amended Toxic Substances Control Act. EPA-HQ-OPPT-2016-0654. https://www.epa.gov/sites/production/files/2017-06/documents/prepubcopy_tsca_riskeval_final_rule_2017-06-22.pdf

[2] Institute of Medicine. Finding What works in Health Care: Standards for Systematic Reviews. p.13-34. The National Academies Press. Washington, D.C. 2011

Systematic Review Methods in IRIS Assessments



NAS (2017): Reflections and Lessons Learned from the Systematic Review



"….one disadvantage in conducting a systematic review is that it can be time and resource intensive, particularly for individuals that have not previously conducted a systematic review." [p.157]

"The committee discussed at length whether it could provide EPA with advice about when a systematic review should be performed but decided it could not be more specific because that decision will depend on the availability of data and resources, the anticipated actions, the time frame for decision making, and other factors." [p.157]

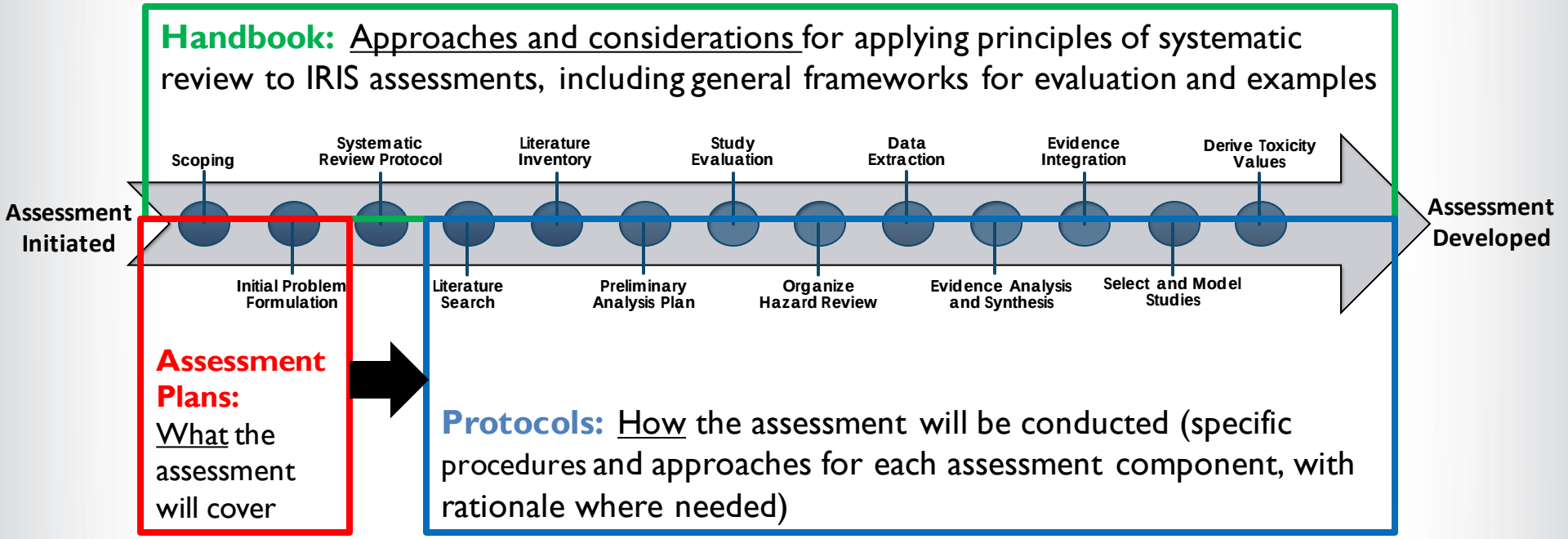
"The committee also recognized that it might be advantageous for EPA to build on existing systematic reviews that are published in the peer-reviewed literature." [p.157]

"The committee recognizes that the methods and role of systematic review and meta-analysis in toxicology are evolving rapidly and EPA will need to stay abreast of these developments, strive for transparency, and use appropriate methods to address its questions." [p.157]

Making Systematic Review (SR) Pragmatic and Feasible For IRIS

- Standard operating procedures (IRIS Handbook), templates (draft assessment plans, chemical-specific protocols), and regular training
- Solicit early feedback during scoping and problem formulation via assessment plans
- Utilize iterative protocols to ensure communication on included studies and focus on best-available and most-informative evidence as the assessment progresses
- Multiple assessment products ("modularity")
- Targeted focus, especially for evidence-rich topics
  - Make better use of existing assessments as starting point
- Use of specialized SR software applications/automation for efficiency

IRIS Systematic Review Documents

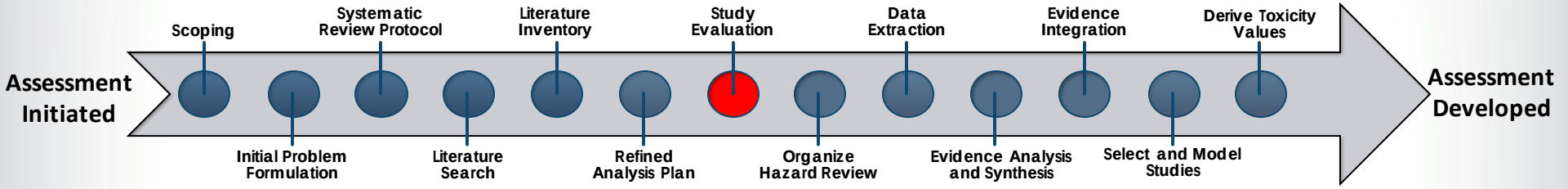


EPA
United States
Environmental Protection
Agency

Study Quality Evaluation

Office of Research and Development
NCEA, IRIS

Individual Study Evaluation



- General approach same for human and animal studies
- Evaluation process focused on:
  - Internal validity/bias
  - Sensitivity
  - Applicability (relevance to the question)
  - Reporting quality

Overview of Study Evaluation in IRIS

Individual study level domains

Animal	Epidemiological
Reporting Quality	Exposure measurement
Selection or Performance Bias	Outcome ascertainment
Confounding/Variable Control	Population Selection
Reporting or Attrition Bias	Confounding
Exposure Methods Sensitivity	Analysis
Outcome Measures and Results Display	Sensitivity
Other	

Domain judgements

- Good
- Adequate
- Poor
- Critically Deficient

	Judgement	Interpretation
	Good	Appropriate study conduct relating to the domain & minor deficiencies not expected to influence results.
	Adequate	A study that may have some limitations, but not likely to be severe or to have a substantive impact on results.
	Poor	Identified biases or deficiencies interpreted as likely to have had a substantial impact on the results or prevent reliable interpretation of study findings.
	Critically Deficient	A flaw that is so serious that the study could not be used.

Overall study rating

- High
- Medium
- Low
- Uninformative

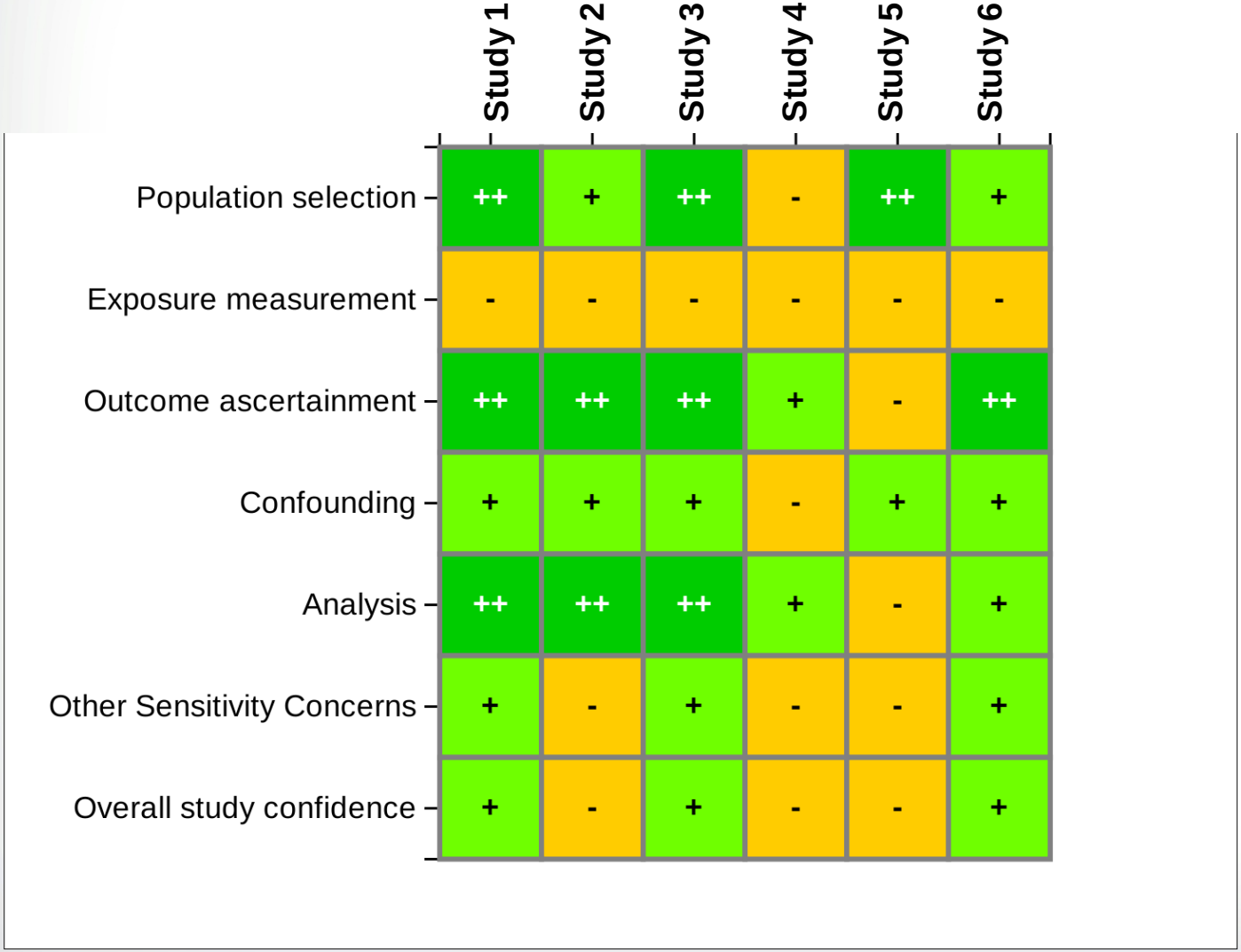
Rating	Interpretation
High	No notable deficiencies or concerns identified; potential for bias unlikely or minimal and sensitive methodology.
Medium	Possible deficiencies or concerns noted, but resulting bias or lack of sensitivity would be unlikely to be of a substantive degree.
Low	Deficiencies or concerns were noted, and the potential for substantive bias or inadequate sensitivity could have a significant impact on the study results or their interpretation.
Uninformative	Serious flaw(s) makes study results unusable for hazard identification

Individual Epidemiological Study Examples

Medium confidence

Uninformative

Across Study Evaluations

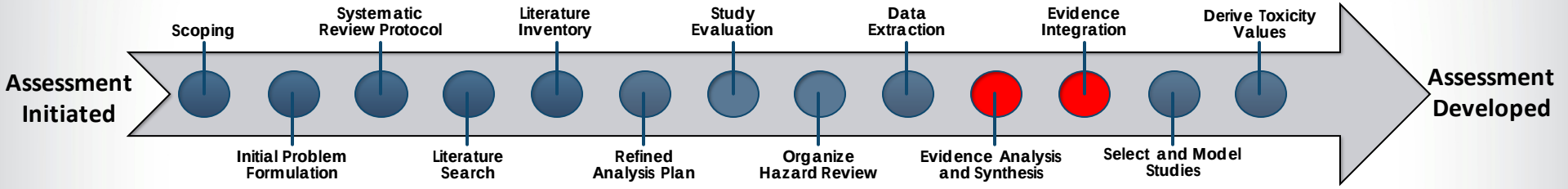


United States
Environmental Protection
Agency

Weight of Evidence

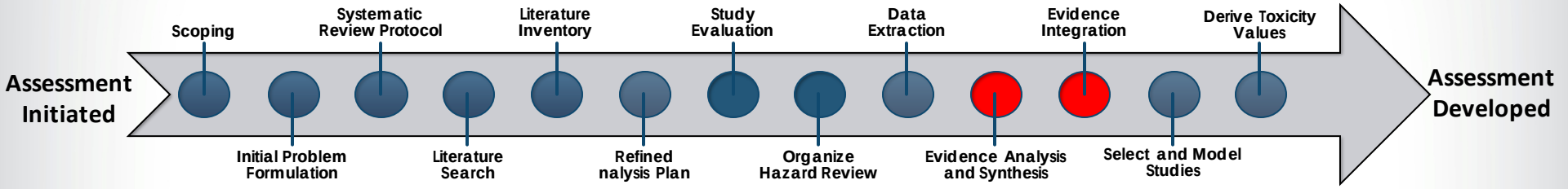
Office of Research and Development
NCEA, IRIS

Evidence Synthesis and Integration



- Synthesis of evidence is more than counting the number of "positive" and "negative" studies
- Consider the influence of bias and sensitivity when describing study results and synthesizing evidence
  - Synthesis should primarily be based on studies of medium and high confidence (when available)
- Use structured framework to aid in transparency

Moving from Synthesis to Integration



Step 1: Within Evidence Stream Judgements

- Results of Human Health Effect Study Synthesis
- Results of Animal Health Effect Study Synthesis
- Results of Synthesis of Mechanistic Evidence Informing the Human and Animal Syntheses

Step 2: Across Evidence Stream Integration

Within Evidence Stream Conclusions Prior to Across

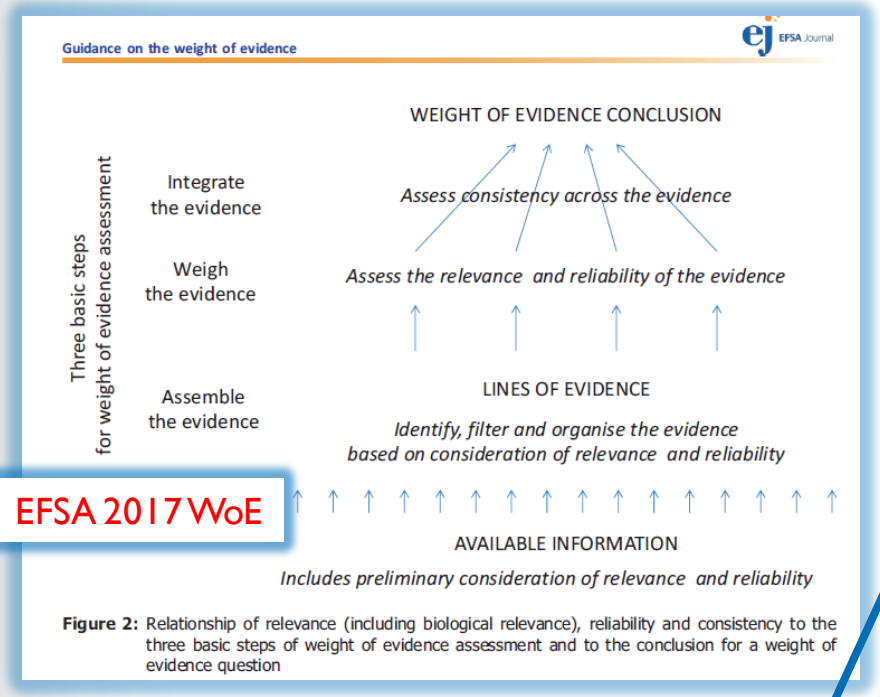
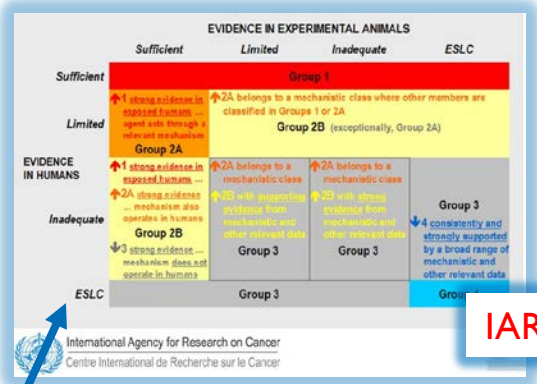
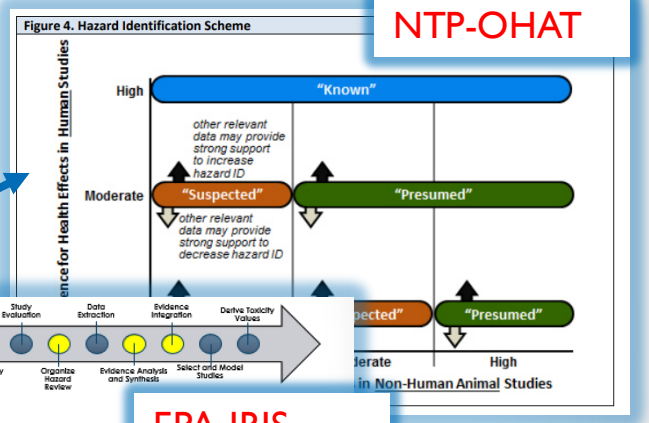


Figure 2: Relationship of relevance (including biological relevance), reliability and consistency to the three basic steps of weight of evidence assessment and to the conclusion for a weight of evidence question

EFSA 2017 WoE

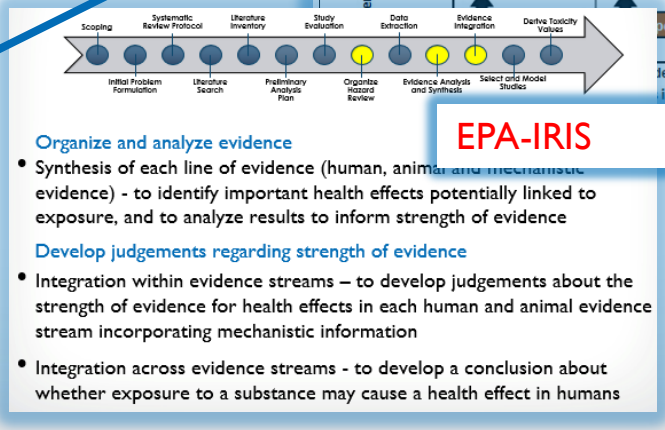


IARC



NTP-OHAT

mechanistic information used to increase/decrease integrated conclusions from human and nonhuman animal evidence



EPA-IRIS

*EFSA 2017 WoE Guidance https://www.efsa.europa.eu/en/efsajournal/pub/4971

EFSA Key Considerations for Weighing Evidence*

- **Reliability** is the extent to which the information comprising a piece or line of evidence is correct, i.e. how closely it represents the quantity, characteristic or event that it refers to. This includes both accuracy (degree of systematic error or bias) and precision (degree of random error).
- **Relevance** is the contribution a piece or line of evidence would make to answer a specified question, if the information comprising the evidence was fully reliable. In other words, how close is the quantity, characteristic or event that the evidence represents to the quantity, characteristic or event that is required in the assessment. This includes biological relevance (EFSA, **2017**) as well as relevance based on other considerations, e.g. temporal, spatial, chemical, etc.
- **Consistency** is the extent to which the contributions of different pieces or lines of evidence to answering the specified question are compatible

*Section 2.5, EFSA 2017 WoE Guidance https://www.efsa.europa.eu/en/efsajournal/pub/4971

Examples of WoE Criteria and Guidance*

Table B.3: Examples of criteria for weighing evidence from the published literature, mapped onto the three basic concepts of reliability, relevance and consistency introduced in Section 2.5

Publication	Reliability	Relevance	Combination of reliability and relevance	Consistency	Other
Bradford Hill (1965)		Temporality Experimentation Specificity	Strength of association Biological gradient	Consistency of association Biological plausibility Coherence	
Collier et al. (2016)	Uncertainty and variability (treatment of)	Applicability and uti Essentiality of key e			
ECHA (2010)	Reliability	Relevance			
US EPA (1998)	Adequacy and quality of data Degree and type of uncertainty associated with the evidence	Relationship of the to the risk assessm questions			
EPA (2003)	Uncertainty and variability (treatment of)	Applicability and uti			
Hope and Clarkson (2014)	Study quality	Site specificity Spatial representati Temporal represent Specificity to stress			
Hull and Swanson (2006)		Specificity of cause			
		Essentiality of key e			

Guidance on the weight of evidence

Publication	Reliability	Relevance	Combination of reliability and relevance	Consistency	Other
Meek et al. (2014)				Consistency Biological concordance Concordance of empirical observations among key events Analogy (to other chemicals)	
Morgan et al. (2016) (GRADE)	Risk of bias Imprecision Publication bias	Indirectness Confounding Study design (randomised or observational)	Effect size Dose response	Inconsistency	
Lorenz et al. (2013)	Study design Bias/chance Reliability Statistical methods Internal consistency	Confounders Temporality Relevance	Strengths & weaknesses Dose response Predictivity Strength of association	Replicability (if observed) Biological plausibility	Adequacy
Rooney et al. (2014) (OHAT)	Risk of bias (15 subquestions) Imprecision Publication bias Rare outcomes	Indirectness Residual confounding	Effect magnitude Dose response	Consistency	'Other' (unspecified)
SCENIHR (2012)	Quality Reliability	Relevance/potential importance The characterisation of the stressor The relevance of the set of data for a particular endpoint	Utility (combining quality and relevance) Soundness and appropriateness of the methodology used The extent to which the full details of methodology are provided	The reproducibility of findings between experiments Consistency	Validity Uncertainties in the judgement used
Suter and Cormier (2011)	Performance Statistical analysis Potential for bias	Relevance Inherent weights of study types (e.g. randomised vs observational, field vs lab)	Study design Reporting Strength	Number of pieces Coherence Diversity	Case-specific criteria
Vermeire et al. (2013)	Sensitivity Reliability	Relevance Specificity	Predictivity		Adequacy Validity

Section 2.5, Table B.3 EFSA 2017 WoE Guidance https://www.efsa.europa.eu/en/efsajournal/pub/4971

Observations on the WoE Examples from Table B.3

- Many - probably most - are a list of considerations (e.g., Bradford Hill) rather than structured frameworks that provide guidance for how to apply the considerations
- Same essential content, but variation in terminology
  - e.g., relevance ≈ directness ≈ applicability
  - e.g., reliability ≈ study quality ≈ risk of bias
- Use of structured frameworks for WoE becoming more common in chemical assessments for evidence synthesis/integration
  - GRADE is common starting point (Morgan et al. 2016) in Table B.3
  - NTP OHAT (Rooney et al. 2014 in Table B.3), UCSF Navigation Guide, EPA IRIS are derived from GRADE
    - Ongoing collaborations with GRADE Working Group to develop GRADE guidance to avoid derivatives

EPA
United States
Environmental Protection
Agency

Use of Structured Frameworks to Increase Transparency of WoE Judgements

Office of Research and Development
NCEA, IRIS

GRADE Structured Framework

- Widely used (100+ organizations)
- Includes consideration of WoE factors (as characterized in Table B.3)
  - Reliability (risk of bias, imprecision, publication bias)
  - Relevance (directness, confounding, study design)
  - Combination reliability/relevance (effect size, dose-response)
  - Consistency (unexplained inconsistency)
- Compared to other approaches in EFSA WoE Table B.3, GRADE conducts research and develops guidance to operationalize consideration of WoE factors
  - Publications, handbook, software application (GRADEpro/GDT), bi-annual meetings, use of case examples to address methodological challenges
  - GRADE Working Group has open and free membership www.gradeworkingroup.org
- GRADE is dedicated to method development and adaptable, e.g., has GRADE frameworks for interventions, prognostic factors, values and preferences, etc.

Certainty in the Evidence: How Confident in the Research

- Are the research studies well done? Risk of bias
- Are the results consistent across studies ? Inconsistency
- How directly do the results relate to the question? Indirectness
- Is the association precise - due to random error? Imprecision
- Are these all of the studies that have been conducted? Pub. Bias
- Is there anything else that makes us particularly certain? Large associations, worst case scenario predictors still allows strong conclusions, exposure-effect relation

Interpreting the certainty in the evidence

Certainty rating	Definitions
⊕⊕⊕⊕ High	The panel is very confident that the true association lies close to that of the estimate of the association
⊕⊕⊕◯ Moderate	The panel is moderately confident in the association: The true association is likely to be close to the estimate of the association, but there is a possibility that it is substantially different
⊕⊕◯◯ Low	The panel's confidence in the association is limited: The true association may be substantially different from the estimate of the association
⊕◯◯◯ Very low	The panel has very little confidence in the association: The true association is likely to be substantially different from the estimate of association

NAS (2017) Low Dose Toxicity From Endocrine Active Chemicals

TABLE 3-9 Profile of the Confidence in the Body of Evidence on DEHP and AGD in Humans

			Factors Decreasing Confidence "—" If No Concern; "↓" If Serious Concern to Downgrade Confidence					Factors Increasing Confidence "—" If Not Present; "↑" If Sufficient to Upgrade Confidence			
Phthalate	Metabolite(s)	INITIAL CONFIDENCE RATING (# of studies)	Risk of Bias	Unexplained Inconsistency	Indirectness	Imprecision	Publication Bias	Large Magnitude	Dose Response	Residual Confounding	FINAL CONFIDENCE RATING
DEHP	MEHP; 5-oxo-MEHP; 5OH-MEHP; sumDEHP metabolites	Moderate (6 prospective)[a]	—	—	—	—	—	—	—	—	Moderate

[a]Swan et al. (2008); Bustamante-Montes et al. (2013); Bornehag et al. (2015); Swan et al. (2015); Jensen et al. (2016); Martino-Andrade et al. (2016).

TABLE 3-3 Profile of the Confidence in the Body of Evidence on DEHP and AGD in Animals

		Factors Decreasing Confidence "—" If No Concern; "↓" If Serious Concern to Downgrade Confidence					Factors Increasing Confidence "—" If Not Present; "↑" If Sufficient to Upgrade Confidence					
Phthalate	INITIAL CONFIDENCE RATING (# of studies)	Risk of Bias	Unexplained Inconsistency	Indirectness	Imprecision	Publication Bias	Large Magnitude	Dose Response	Residual Confounding	Consistency Across Species/Models	Rare Outcome	FINAL CONFIDENCE RATING
DEHP	High (16 rat,[a] 3 mouse[b])	↓	—	—	—	—	↑	↑	—	—	—	High

[a]Moore et al. (2001); Borch et al. (2004); Jarfelt et al. (2005); Wolfe and Layton (2005); Andrade et al. (2006); Culty et al. (2008); Lin et al. (2008, 2009); Christiansen et al. (2009, 2010); Gray et al. (2009); Martino-Andrade et al. (2009); Vo et al. (2009); Li et al. (2013); Zhang et al. (2013); Jones et al. (2015).
[b]Liu et al. (2008); Do et al. (2012); Pocar et al. (2012).

Mechanistic evidence: "The mechanistic data developed in vitro and in animal models provide evidence that the DEHP effects on AGD in humans identified by the committee's systematic review are biologically plausible….but were not sufficient to result in an upgrade in the committee's final hazard identification."

Final Hazard Conclusion on AGD

On the basis of the committee's evidence integration of the animal and the human evidence on DEHP and effects on AGD and consideration of relevant mechanistic data, the committee concluded that DEHP is presumed to be a reproductive hazard to humans.

Experience in Applying GRADE to Chemical Assessments

- Initial reactions range from "great, we can work with this" to "too simplistic, based on human randomized clinical trials and won't work for environmental health evidence, devalues epidemiological research, inflexible/algorithmic"
- GRADE Environmental Health Project Group established 2015 to address methodological issues (Morgan et al., Environ Int. 2016 Jul-Aug;92-93:611-6)
- Priority areas:
  - Evaluation of observational studies of environmental and occupational exposure
  - Application of GRADE to animal, mechanistic, and modelled evidence
  - How integrate across evidence streams?
  - How assess biological plausibility?
  - How assess coherence and consistency (GRADE downgrades for unexplained inconsistency, but does not "upgrade" for coherence/consistency)
  - Applying GRADE to non-systematic reviews and under rapid timeframes

GRADE Environmental Health Project Group Activities

Issue	Activity	Impact/Status
Epidemiological Evidence	ROBINS-E RoB tool (uses concept of comparison to ideal target experiment) **Workshop:** "Developing ROBINS-I* for studies of exposures (ROBINS-E). January 30-31, 2017"	Remove study design from initial CiE (all studies start at high), remove double downgrading concerns
Animal Evidence	Numerous groups have applied GRADE to animal evidence (pre-clinical and toxicological)	GRADE factors apply; additional examples and discussion needed to develop guidance
Mechanistic Evidence	----	----
Modelled Evidence	**Workshop: "**GRADE for modelled evidence. May 15-16, 2017. McMaster University. Hamilton"	GRADE factors apply; additional examples and discussion needed to develop guidance
Evidence Integration	Numerous groups are using GRADE-derived approaches for within evidence stream input (parallel consideration approach)	additional examples needed to develop guidance
Rapid Response	Applying GRADE to non-systematic reviews under urgent timelines	GRADE factors apply; additional examples needed to develop guidance

method development activities of potential interest to CompTox community

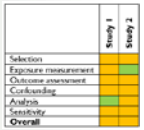
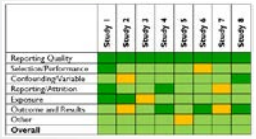
* Stern et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ. 2016 Oct 12;355:i4919. doi: 10.1136/bmj.i4919.

IRIS Structured Framework (Evidence Profile Table)

"No matter what method is used to integrate the different kinds of evidence available for an IRIS assessment, using a template for the evidence-integration narrative could help to make IRIS assessments more transparent." [NAS, 2014]

Studies and interpretation	Factors that increase confidence	Factors that decrease confidence	Summary of findings	Within stream evidence judgements	Inference across evidence streams	Overall confidence conclusion
[Health Effect or Outcome Grouping]						
Evidence from Human Studies (Route)					*Human relevance of findings in animals* • *Cross-stream coherence (i.e. for both health effect-specific and mechanistic data)* • *Other inferences:* o *Information on susceptibility* o *MOA analysis inferences: precursors, cross-species inferences of toxicokinetics, or quantitative implications* o *Relevant information from other sources (e.g., read across; other, potentially related health hazards)*	*Describe conclusion(s) and primary basis for the integration of all available evidence (e.g., across human, animal, and mechanistic):* +++ Strongest conclusion ++◯ +◯◯ Weakest conclusion ◯◯◯ −◯◯ Inadequate −−◯ −−− Conclusion of unlikely to be an effect Summarize the models and range of dose levels upon which the conclusions were primarily reliant
• *References* • *Study confidence (based on evaluation of risk of bias and sensitivity) and explanation* • *Study design description*	• *Consistency* • *Dose-response gradient* • *Coherence of observed effects (apical studies)* • *Effect size (magnitude, severity)* • *Biological plausibility* • *Low risk of bias/ high quality* • *Insensitivity of null/ negative studies* • *Natural experiments* • *Temporality*	• *Unexplained inconsistency* • *Imprecision* • *Indirectness/ applicability* • *Poor study quality/ high risk of bias* • *Other (e.g., Single/Few Studies; small sample size)* • *Evidence demonstrating implausibility*	• *Results information (general endpoints affected/ unaffected) across studies* • *Human evidence informing biological plausibility: discuss how mechanistic data influenced the within stream judgement (e.g., evidence of precursors in exposed humans).* *Could be multiple rows (e.g., grouped by study confidence or population) if this informs results heterogeneity*	*Describe confidence in evidence from human studies, and primary basis:* +++ Strongest evidence ++◯ +◯◯ Weakest evidence ◯◯◯ −◯◯ Inadequate −−◯ −−− Convincing evidence of no effect		
Evidence for an Effect in Animals (Route)						
• *References* • *Study confidence (based on evaluation of risk of bias and sensitivity) and explanation* • *Study design description*	• *Consistency and Replication* • *Dose-response gradient* • *Coherence of observed effects (apical studies)* • *Effect size (magnitude, severity)* • *Biological plausibility* • *Low risk of bias/ high quality* • *Insensitivity of null/ negative studies*	• *Unexplained inconsistency* • *Imprecision* • *Indirectness/ applicability* • *Poor study quality/ high risk of bias* • *Other (e.g., Single/Few Studies; small sample size)* • *Evidence demonstrating implausibility*	• *Results information (general endpoints affected/ unaffected) across studies* • *Evidence informing biological plausibility for effects in animals: discuss how mechanistic data influenced the within stream judgement (e.g., evidence of coherent molecular changes in animal studies)* *Could be multiple rows (e.g., by study confidence, species, or exposure duration) if this informs results heterogeneity*	*Describe confidence in evidence for an effect in animals, and primary basis:* +++ Strongest evidence ++◯ +◯◯ Weakest evidence ◯◯◯ −◯◯ Inadequate −−◯ −−− Convincing evidence of no effect		

IRIS Structured Framework (Evidence Profile Table)

Studies	Factors that increase confidence	Factors that decrease confidence	Summary of findings	Within stream confidence judgement	Inferences across streams	Hazard assessment conclusion
Chemical X (Health Outcome Y)						
Human (oral)					Findings in animals presumed relevant to humans (no evidence to the contrary); coherent evidence from mechanistic studies mammalian and non-mammalian models. .	⊕⊕○
Case Series Study 1 **Cross-sectional** Study 2 **Risk of bias and sensitivity** 		Few studies Low number of exposed cases (insensitivity) Lack of dose-response High risk of bias	Studies found no significant correlations with chemical X exposure and health outcome y	○○○ INDETERMINATE		
Animal (oral)						
Short-term Study 1 (rat) Study 2 (rat) **Subchronic** Study 3 (rat) Study 4 (mouse) **Developmental/Reproductive** Study 5 (rat) Study 6 (rat) Study 7 (rat) Study 8 (mouse) **Risk of bias and sensitivity** 	Coherence among related endpoints Low risk of bias Dose-response gradient Biological plausibility	Small sample sizes in some studies Some unexplained inconsistency	Similar pattern of changes in hormone A and hormone B were observed in study 1 and study 2. Effects on serum hormone levels are supported by histopathological changes in tissue A (study 1, study 3, study 4, study 5, study 6) and increased tissue A weight (study 1, study, 5, study 6, study 8). Evidence of dose-response gradient in most studies reporting effects. Biological plausibility of the observed effects is supported by mechanistic studies in mammalian and non-mammalian models (see Section 1.2.1 Mechanistic Evidence).	⊕⊕○ MODERATE		

EPA
United States
Environmental Protection
Agency

Systematic Review and Rapid Response

Office of Research and Development
NCEA, IRIS

Approaches for Rapid Response

- Increase staff to conduct a full systematic review
  - Probably not viable for emergency or urgent response (i.e., less than a month)
  - Not viable when resources are constrained
- Methodology should still be described when a systematic review is not practical
  - e.g., use of expert opinion can be considered a method
- Use a structure framework to describe confidence in conclusions - can be done even when analysis is NOT based on a systematic review

Use of Structured Frameworks for Rapid Response

Contents lists available at ScienceDirect

Environment International

journal homepage: www.elsevier.com/locate/en

Preface

Using GRADE to respond to health questions with different leve

Kristina A. Thayer [a], Holger J. Schünemann [b,*]

[a] Division of the National Toxicology Program, National Institute of Environmental Health Sciences, National Institutes of Health, Depc
Mail Drop K2-02, Research Triangle Park, NC 27709, USA
[b] Department of Clinical Epidemiology & Biostatistics, Department of Medicine, McMaster University, Health Sciences Centre, Room 2C

ARTICLE INFO

Article history:
Received 15 March 2016
Received in revised form 21 March 2016
Accepted 21 March 2016
Available online 26 April 2016

ABSTRACT

Increasing interest exists in applying the Grading of Recommer
(GRADE) approach to environmental health evidence. While i
atic reviews and corresponding summary tables, such as evide
that "the evidence that was assessed and the methods that
should be clearly described." In this article, we suggest that
from narrative reviews, modelled (indirect) evidence, or evid
underlying judgments about the certainty in this evidence are
ed transparently. Health questions that require assessing the c
thy answers may range from hours, to days or weeks, to a few
without short-term time pressures. Time frames of emergent,
quire relying on existing summaries or rapidly compiling the
without available full systematic reviews, expressing the cert
for users of the evidence and those who evaluate certainty in ef
between organizations tackling similar questions about the ev
rative or other summaries of the evidence can be presented t

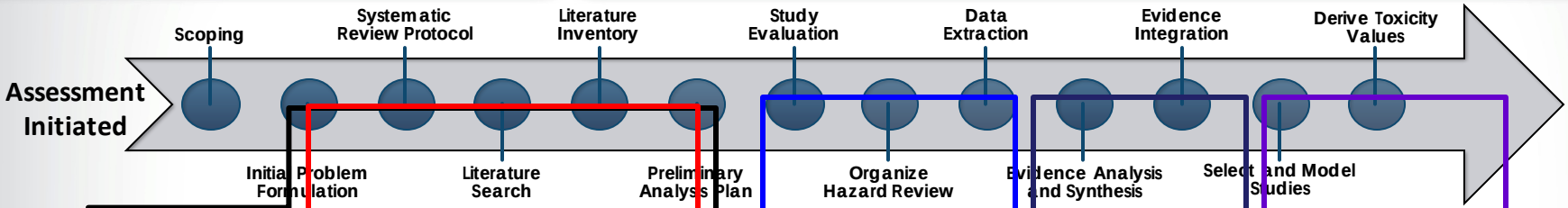
Table 1
Examples of GRADE applied across different time scenarios.

Type of response	Ultra-short emergency response: within one or more hours	Urgent response: one to two weeks	Rapid response: one to three months	Routine response: more than 3 months
Example	West Virginia Elk River spill Population: community exposed to the chemical spill. Intervention/exposure: chemicals in the spill that contaminated water supply. Comparison: no chemicals in the spill. Outcomes: genotoxicity, developmental or reproductive toxicity, liver toxicity and others.	Melamine in composite food products Population: healthy people Intervention/exposure: melamine from composition food products below 0.5 mg/kg body weight per day. Comparison: higher than 0.5 mg/kg body weight of melamine from composition food. Outcomes: renal insufficiency (assessed with renal clearance), urinary tract calculi, urinary tumors (used for this example of the certainty in the evidence).	Avian influenza Population: people with suspected avian influenza infection. Intervention/exposure: oseltamivir. Comparison: no oseltamivir. Outcomes: mortality, duration of hospitalization, incidence of lower respiratory tract complications (used for this example of the certainty assessment below), antiviral drug resistance existing before treatment, and serious adverse events.	PFOA and birth weight Population: women of reproductive age and fetuses (before and/or during pregnancy or development). Intervention/exposure: perfluorooctanoic acid (PFOA; CAS# 335-67-1) or its salts. Comparison: lower levels of PFOA. Outcomes: fetal growth, birth weight, other measures of fetal or newborn size.
Type of evidence	Available evidence: animal toxicology studies in rodents for two chemicals in the spill (a 28-day study and a teratology study) and SAR analyses for other chemicals in the spill with no toxicology data.	Available evidence: animal toxicology studies in rat and mice with exposures to various levels of melamine via feeding, including a control group. The utilized evidence should be supported by a literature search with transparent inclusion and exclusion criteria and a (narrative) summary of that evidence.	Available evidence: five randomized trials in patients with seasonal flu (summarized in systematic reviews), case studies of patients with avian influenza, *in vitro* and *in vivo* animal data.	Available evidence: a systematic review of 18 non-randomized (observational) studies (10 were included in a meta-analysis).
GRADE domains to assess certainty in the evidence: suggested approaches to making judgments or proposed judgments (note these are not necessarily reflecting judgments in the original scenarios).				
Risk of bias	Animal studies: would be assessed by risk of bias (RoB) considerations for animal studies (e.g. randomization, blinding at outcome assessment, sufficient characterization of test compound, or whether all animals were accounted for). Ideally, RoB assessments would be available for individual studies and summarized across studies. In the Elk River example, the number of animal studies was small and could be assessed at the individual level within a short-time frame. A de novo risk of bias evaluation may not be feasible in cases where evidence is drawn from existing narrative risk assessments that summarize a large body of literature. Nevertheless, it may still be possible to assess risk of bias based on the uncertainties and evidence limitations described in the risk assessment. SAR: could be assessed using OECD model validation or similar guidance that recommends presentation of a defined domain of applicability for a defined endpoint supported by appropriate measures of goodness-of-fit (OECD, 2007).	Animal studies: would be assessed by risk of bias (RoB) considerations for animal studies (e.g. randomization, pathologists blinded in their assessments or all animals accounted for). In this case it appears that the animal studies did not report that it was randomized and, thus, may be at risk of bias.	Not serious	Serious based on some concern of risk of bias in the included studies (in the original report, the authors used an approach to rating certainty that accounted for risk of bias by lowering the certainty from high to moderate).
Imprecision	Could be assessed for both animal data and SAR (e.g., considering statistical or numerical uncertainty in model parameters).	While no summary estimates are available, an assessment could be guided by the availability of data from only 100 animals in different exposure groups which would result in wide confidence intervals.	Serious	Not serious
Inconsistency	Could be assessed for both animal data and SAR (e.g., assessing similarity of results based on applying different models).	Only one study was included and therefore no inconsistency is present (Guyatt et al., 2011d).	Not serious	Not serious
Publication bias	Could be assessed for both animal studies and SAR. A judgment of undetected might be reasonable if	Could be assessed using guidance for animal studies but a judgment of undetected might be reasonable if	Undetected	Undetected

EPA
United States
Environmental Protection
Agency

Specialized Software Applications

Systematic Review Tools



HERO
Literature searching, storage and documentation (tagging)

SWIFT Review
Problem formulation

SWIFT Review
DoCTOR
Machine learning for screening prioritization

Distiller
SWIFT Active
Multiple reviewer reference screening and tracking (HERO-tagging)

HAWC
DRAGON
Modules to track multiple reviewer study evaluations

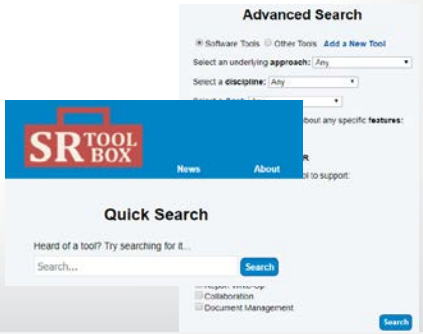
Extracted data storage

HAWC
Interactive graphical display, evidence profile tables

GRADEPro
Evidence profile table, evidence to decision making

METAXL, Metafor
Evaluation of heterogeneity or combined study results analysis

Database of SR software tools:
http://systematicreviewtools.com/



Opportunities for Engagement

- Training on approaches and software tools (either web-based or hands on)
- Engagement with chemical assessment teams
- Additional discussion/case studies to illustrate tenants of transparency applied to non-SR assessments and rapid response
- Academic MOUs to help train next generation

Parting Thoughts

- **Systematic review offers a structured methodology to synthesize and integrate evidence**
  - **Used to characterize what is known and help identify key knowledge gaps**
  - **Similar conceptual methodological challenges for animal bioassay and mechanistic evidence, e.g., characterizing relevance to humans, coherence/consistency in findings = MAJOR opportunity to shape how systematic review is applied beyond human evidence.**
- **Commitment to methodological transparency is critical to ensure credibility and acceptance of "fit for purpose" assessment approach (e.g., rapid response, alternative methods)**

EPA
United States
Environmental Protection
Agency

Questions/Comments?