August 16, 2018

Comments from Academics, Scientists and Clinicians on: The Application of Systematic Review in TSCA Risk Evaluations.

Submitted online via *Regulations.gov* to docket EPA-HQ-OPPT-2018-0210

These comments are submitted on behalf of the undersigned academic, scientists, and clinicians. We declare collectively that we have no direct or indirect financial or fiduciary interest in any chemical under consideration in these risk evaluations. The co-signers' institutional affiliations are included for identification purposes only and do not necessarily imply any institutional endorsement or support, unless indicated otherwise.

We appreciate the opportunity to provide written comments on the Application of Systematic Review in TSCA Risk Evaluations,[a] pursuant to the Toxic Substances Control Act (TSCA), as amended by the Frank R. Lautenberg Chemical Safety of the 21st Century Act (Lautenberg TSCA). TSCA requires that EPA make decisions about chemical risks based on the "best available science" and the "weight of the scientific evidence"[b] which EPA defined in regulation as "…a systematic review method, applied in a manner suited to the nature of the evidence or decision, that uses a pre-established protocol to comprehensively, objectively, transparently, and consistently identify and evaluate each stream of evidence, including strengths, limitations, and relevance of each study and to integrate evidence as necessary and appropriate based upon strengths, limitations, and relevance.''[c]

Systematic review methods originated more than 40 years ago in psychology. The methodology was soon adapted to evaluating the effectiveness of clinical interventions in medicine and related disciplines in response to empirical evidence demonstrating the need to apply scientific principles not only to primary research, but also to research synthesis methods that inform decision-making in healthcare (1-3). Almost a decade ago, these empirically-proven methods for research synthesis were adapted to environmental health (4, 5). To date, science-based methods for systematic review in environmental health have been demonstrated in case studies in the peer-reviewed literature (6-13), and adopted by the National Toxicology Program (14) and the U.S. EPA's Integrated Risk Information System (IRIS) program (15).

EPA's systematic review framework under TSCA establishes EPA's "rules" for assembling and interpreting the scientific evidence on chemicals in commerce. These "rules" will determine, whether explicitly, implicitly, and/or by default, *what* evidence EPA will consider, and *how* it will evaluate that evidence when it is making decisions about potentially hazardous chemicals in commerce. Exposure to industrial, commercial, and consumer product chemicals is ubiquitous from the time of conception until death. As such, EPA's rules for gathering and interpreting the science that evaluates the relationship between these exposures and adverse health effects are of profound importance to the general public, and will have even greater impact on the potentially exposed or susceptible sub-populations Congress explicitly mandated EPA to protect: pregnant women, children, individuals with underlying health conditions, workers, and those with greater exposure and/or greater vulnerability to chemical toxicity and exposure.

[a] 83 FR 26998, June 11, 2018
[b] 15 USC §2625 (h)-(i)
[c] 40 CFR 704.33

With so much at stake, we are deeply concerned by EPA's ad hoc and incomplete TSCA systematic review framework, which is inconsistent with current, established, best available empirical methods for systematic review. Moreover, as we detail below, the application of EPA's TSCA framework would likely result in the exclusion of quality research from EPA's decision-making. Accordingly, the TSCA systematic review method does not meet the mandate of the law to use the "best available science." [d]

Based on the most current empirically demonstrated principles of systematic review methods, we provide EPA with concrete recommendations and approaches to correct its methodology and inform timely science-based decision-making to achieve the Agency's mission of protecting the public from harmful chemicals.

Our comments address the following six main points:

1. **EPA's TSCA systematic review framework is ad hoc, incomplete, and does not follow established methods for systematic review that are based on the best available science.**

   We recommend: EPA should implement a systematic review method that is compatible with empirically based existing methods and aligns with the Institute of Medicine's[e] definition of a systematic review, including but not limited to, using explicit and pre-specified scientific methods for every step of the review. EPA should consider methods demonstrated for use in environmental health, and which have been endorsed and utilized by the National Academy of Sciences, i.e., the National Toxicology's Office of Heath Assessment and Translation systematic review method, and the Navigation Guide Systematic Review Method. EPA's TSCA systematic review framework should be peer-reviewed by qualified external experts in the field.

2. **EPA's TSCA systematic review framework utilizes a quantitative scoring method that is incompatible with the best available science in fundamental ways:**

   a. **Quantitative scores for assessing the quality of an individual study are arbitrary and not science-based; the Cochrane Collaboration and National Academy of Sciences recommend against such scoring methods.**
   b. **EPA's scoring method wrongly conflates how well a study is reported with how well the underlying research was conducted; and**
   c. **EPA's scoring method excludes research based on one single reporting or methodological limitation.**

   We recommend: EPA should not use a quantitative scoring method to assess quality in individual studies; it should not conflate study reporting with study quality; and it should not exclude otherwise quality research based on a single reporting or methodological limitation. Rather EPA should employ a scientifically valid method to assess risk of bias of individual studies.

3. **EPA's TSCA systematic review framework does not consider financial conflicts of interest as a potential source of bias in research.**

[d] 15 USC §2625 (h)

[e] The Institute of Medicine is now the National Academy of Medicine.

We recommend: EPA should assess study and author funding source as a risk of bias domain for individual studies.

4. **The literature review step of EPA's TSCA systematic review framework incorporates select best practices, but also falls short of, or is unclear about, many other best practices for conducting a systematic and transparent literature review.**

   We recommend: EPA should make its framework for conducting a literature review congruent with all of the Institute of Medicine's best practices and explicitly include rules for when the list of relevant studies will be considered final.

5. **EPA's TSCA systematic review framework correctly recognizes that mechanistic data are not required for a hazard assessment, but EPA is not clear that these data, if available, can only be used to increase, and not to decrease, confidence in a body of evidence.**

   We recommend: EPA should be explicit that mechanistic data can only be used to upgrade a hazard classification, or increase the confidence of a finding made based on evaluation of animal and human data, and that these data will not be used to decrease confidence in a body of evidence.

6. **EPA's TSCA systematic review framework is not independent of the regulatory end user of the review.**

   We recommend: EPA's TSCA systematic reviews should be produced independently of the regulatory end user of the review.

We are appreciative of the opportunity to provide public input. Please do not hesitate to contact us with any questions regarding these comments.

Sincerely,

Veena Singla, PhD
Associate Director, Science and Policy, Program on Reproductive Health and the Environment
University of California, San Francisco

Patrice Sutton, MPH
Research Scientist, Program on Reproductive Health and the Environment
University of California, San Francisco

Tracey Woodruff, PhD, MPH
Director, Program on Reproductive Health and the Environment
University of California, San Francisco

Juleen Lam, PhD, MHS, MS
Assistant Professor, Department of Health Sciences
California State University, East Bay

Patricia D. Koman, PhD, MPP
President and Senior Health Scientist
Green Barn Research Associates*
Ann Arbor, Michigan

Lisa Bero, PhD
Chair of Medicines Use and Health Outcomes, Charles Perkins Centre
The University of Sydney

Liz Borkowski, MPH
Senior Research Scientist, Milken Institute School of Public Health
George Washington University

Sheila Brear, BDS
Associate Dean, Academic Affairs, School of Dentistry
University of California, San Francisco

Adelita G. Cantu, PhD, RN
Associate Professor
Alliance of Nurses for Healthy Environments

Courtney Carignan, PhD
Assistant Professor
Michigan State University

Daniel M. Fox, PhD
President Emeritus
Milbank Memorial Fund

Danielle Fries, MPH
Science Associate, Program on Reproductive Health and the Environment
University of California, San Francisco

Mary Gant, MS
Retired Policy Analyst
National Institute of Environmental Health Sciences

Steven G. Gilbert, PhD, DABT
Affiliate Professor
University of Washington

Robert M. Gould, MD
Associate Adjunct Professor, Department of Obstetrics, Gynecology and Reproductive
University of California, San Francisco
Past-President, Physicians for Social Responsibility

Maeve Howett, PhD, APRN, CPNP, IBCLC, CNE
Clinical Professor and Assistant Dean

University of Massachusetts Amherst

Jyotsna Jagai, MS, MPH, PhD
Research Assistant Professor, School of Public Health
University of Illinois at Chicago

Paula I. Johnson, PhD, MPH
Research Scientist, Safe Cosmetics Program
California Department of Public Health

Jean-Marie Kauth, PhD, MPH
Professor
Benedictine University

Carol Kwiatkowski, PhD
Executive Director
The Endocrine Disruption Exchange*

Joseph Laakso, PhD
Director, Science Policy
Endocrine Society*

Gail Lee, RD, REHS Hem
Sustainability Director
University of California, San Francisco

Michael J. Martin, MD, MPH, MBA
Associate Clinical Professor
University of California, San Francisco

Rachel Morello-Frosch, PhD, MPH
Professor, School of Public Health and Department of Environmental Science, Policy and Management
University of California, Berkeley

Katherine Pelch, PhD
Senior Scientist
The Endocrine Disruption Exchange

Janet Pelrman, MD, MPH
Physician
Stanford Children's Hospital

Jeanne Rizzo, RN
President & CEO
Breast Cancer Prevention Partners

Ted Schettler, MD, MPH
Science Director

Science and Environmental Health Network

Rachel M. Shaffer, MPH
PhD Candidate, School of Public Health
University of Washington, Seattle

Laura N. Vandenberg, PhD
Associate Professor
University of Massachusetts, Amherst

Ellen M. Wells, PhD
Assistant Professor of Environmental & Occupational Health
Purdue University School of Health Sciences

Nsedu Obot Witherspoon, MPH
Executive Director
Children's Environmental Health Network

Marya Zlatnik, MD, MMS
Professor, Department of Obstetrics, Gynecology & Reproductive Sciences
University of California, San Francisco

*indicates organizational support

DETAILED COMMENTS

1. EPA's TSCA systematic review framework is ad hoc, incomplete, and does not follow established methods for systematic review that are based on the best available science.

The best available scientific method for a systematic review (SR) specifies that all components of a review be established in a publically available protocol written *prior* to conducting the review to minimize bias and to ensure transparency in decision-making. For example, the Institute of Medicine defines a systematic review as a "scientific investigation that focuses on a specific question and uses explicit, prespecified scientific methods to identify, select, assess, and summarize the findings of similar but separate studies" (emphasis added) (16)(p.1). A fatal flaw in EPA's SR framework is that it lacks essential SR elements, including but not limited to: (1) a protocol for executing a SR developed *prior* to conducting the SR; (2) an explicit method for evaluating the overall body of each evidence stream, i.e., animal, human, etc.; and (3) an explicit method for integrating two or more streams of evidence, including defined criteria for the type and level of evidence needed for a decision by EPA.

Notably, EPA's TSCA SR Framework presents a diagram of a complete SR framework in Figure 3-1 (page 15) and states in footnote 4 on that page that the:

> Diagram depicts systematic review process to guide the first ten TSCA risk evaluations. It is anticipated that the same basic process will be used to guide future risk evaluations with some potential refinements reflecting efficiencies and other adjustments adopted as EPA/OPPT gains experience in implementing systematic review methods and/or approaches to support risk evaluations within statutory deadlines (e.g., aspects of protocol development would be better defined prior to starting scoping/problem formulation).

However, EPA's TSCA SR Framework then proceeds to describe an ad hoc and highly flawed method limited to only the data collection and, to a limited extent, the data evaluation components of a SR. Specifically, Figure S-1 below, excerpted from the National Academy of Sciences 2014 review of the EPA IRIS program's systematic review method (17), presents all of the components of a science-based SR. The red box indicates the parts of a SR method that EPA has included in its proposed framework.

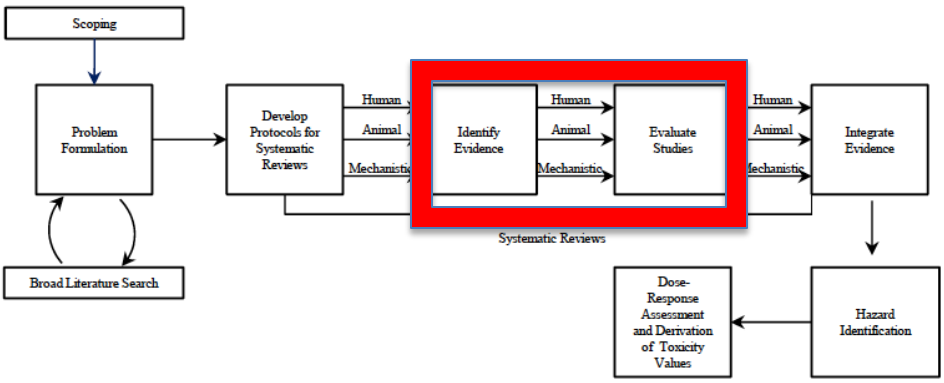


FIGURE S-1 Systematic review in the context of the IRIS process. The committee views public input and peer review as integral parts of the IRIS process, although they are not specifically noted in the figure.

EPA's piecemeal approach is not only in direct contradiction with the best available scientific methods for SR, but also incompatible with the regulatory definition of[f] "weight of evidence" in the risk evaluation rule, which specifies a complete method spelled out in a protocol developed *before* conducting the review. Therefore, the TSCA systematic review method violates both TSCA statute and regulation. [g]

EPA explicitly states that it is proceeding with its first ten risk assessments in the absence of a pre-defined protocol and a complete method for systematic review. Specifically, EPA's SR Framework states:

> (p. 9) … the purpose of the document is internal guidance that … sets out general principles to guide EPA's application of systematic review in the risk evaluation process for the first ten chemicals … **EPA had limited ability to develop a protocol document detailing the systematic review approaches and/or methods prior to the initiation of the risk evaluation process for the first ten chemical substances**. **For these reasons, the protocol development is staged in phases while conducting the assessment work" (emphasis added).** Additional details on the approach for the evidence synthesis and integration will be included with the publication of the draft TSCA risk evaluations.

In effect, EPA is saying it does not have time to comply with its regulatory requirement to conduct a science-based systematic review, and will not actually develop its protocol until it completes the first ten systematic reviews.

First, this approach is in clear violation with scientifically-validated approaches to conducting systematic reviews. In its review of the EPA's Integrated Risk Information System (IRIS) program's proposed SR methods, the National Academy of Sciences specified that, "Completing the literature search as part of

[f] EPA's risk evaluation rule (40 CFR 704.33) states: ''Weight of the scientific evidence means a systematic review method, applied in a manner suited to the nature of the evidence or decision, that uses a pre-established protocol to comprehensively, objectively, transparently, and consistently identify and evaluate each stream of evidence, including strengths, limitations, and relevance of each study and to integrate evidence as necessary and appropriate based upon strengths, limitations, and relevance.''

[g] 15 USC §2625 (h)-(i) and 40 CFR 704.33

protocol development is inconsistent with current best practices for systematic review, and the IRIS program is encouraged to complete the public-comment process and finalize the protocol before initiating the systematic review" (15)(Pg. 8). In the case of TSCA risk assessments, EPA is not only completing the literature search as part of protocol development, it is completing the entire systematic review in the absence of a protocol and complete method. It is blatantly biased to write the rules of evidence assembly and interpretation at the same time one is applying the rules, and as such, this method cannot be validly referred to as a science-based systematic review.

Second, a lack of time is not a credible rationale for EPA's failure to conduct a science-based systematic review for the first ten TSCA chemicals. There are multiple well-developed, science-based, peer-reviewed and validated methods for conducting systematic reviews in environmental health that EPA could readily apply, including the SR method and handbook developed by the Office of Health Assessment and Translation at the National Toxicology Program (14), and the Navigation Guide Systematic Review Method, which has been demonstrated in six case studies (6-13). The National Academy of Sciences cited both of these SR methods as exemplary of the type of methods EPA should use in hazard and risk assessment (17, 18). Further, the National Academy of Sciences utilized both methods in its 2017 assessment of the potential health impacts of endocrine active environmental chemicals (19). Specifically, in its 2017 review the National Academy of Sciences found:

> The two approaches [OHAT and Navigation Guide] are very similar … and they are based on the same established methodology for the conduct of systematic review and evidence assessment (e.g., Cochrane Collaboration, AHRQ Evidence-based Practice Center Program, and GRADE). Both the OHAT and Navigation Guide methods include the key steps recommended by a previous National Academies committee (NRC 2014) for problem formulation, protocol development, specifying a study question, developing PECO statement, identifying and selecting the evidence, evaluating the evidence, and integrating the evidence" (19)(page 119).

Protocols developed for applying the Navigation Guide and the OHAT method have been published and can serve as a template to further expedite EPA's TSCA reviews.[h]

Furthermore, the language of EPA's systematic review framework is confusing, contradictory, and poorly and incorrectly referenced with little science or policy foundation. This suggests the authors of EPA's TSCA Systematic Review Framework lack sufficient understanding of the scientific process integral to this work. A particularly egregious example is EPA's stated understanding of EPA's TSCA statutory science standards:

> (Pg. 26) EPA/OPPT is required by TSCA to use the weight of the scientific evidence in TSCA risk evaluations. Application of weight of evidence analysis is an integrative and interpretive process that considers both data/information in favor (e.g., positive study) or against (e.g., negative study) a given hypothesis within the context of the assessment question(s) being evaluated in the risk evaluation.

This directly contradicts EPA's own published rule which defines what a systematic review is (see

[h] All Navigation Guide systematic review protocols can be found at: https://prhe.ucsf.edu/navigation-guide The National Toxicology Program's protocol for its systematic review to evaluate the evidence for an association between exposure to PFOA or PFOS and immunotoxicity or immune-related health effects is at: https://ntp.niehs.nih.gov/ntp/ohat/pfoa_pfos/protocol_201506_508.pdf

footnote “e”, above) and such an understanding completely subverts the purpose of a systematic review which is to explicitly avoid a simplistic analysis that would led to erroneous conclusions along the lines of stating that, for instance, “five studies are in favor (positive) and ten are against (negative) and therefore the weight is … ”

Another bewildering statement by EPA concerns its highly quantitative scoring method, which is the main topic of its systematic review framework (see comment #2, below). EPA adds a caveat to the scoring method that says quantitative scoring is actually a qualitative method, and further: “The [scoring] system is not intended to imply precision and/or accuracy of the scoring results” (Pg. 35).

The ad hoc and incomplete nature of EPA’s systematic review framework is incompatible in many additional fundamental ways, described further in detail below, with science based methods of systematic review developed, endorsed, and/or advanced by the: National Academy of Sciences (17-19); the Institute of Medicine (16); the National Toxicology Program (14); the Cochrane Collaboration (20); the Grading of Recommendations Assessment, Development and Evaluation (GRADE) method (21, 22); the international scientific collaboration that developed a framework for the “systematic review and integrated assessment” (SYRINA) of endocrine disrupting chemicals (23); the SYRCLE systematic review method for animal studies (24); the Campbell Collaboration’s methods (25); and the Navigation Guide systematic review method developed by a collaboration of scientists led by the University of California San Francisco (4). Most of these organizations also pre-publish their protocols either online (i.e., the National Toxicology Program) or in PROSPERO[i] (i.e., UCSF).

We recommend: EPA should implement a systematic review method that is compatible with empirically based existing methods and aligns with the Institute of Medicine’s definition of a systematic review, including, but not limited to, using explicit and pre-specified scientific methods for every step of the review. EPA should consider methods demonstrated for use in environmental health, and which have been endorsed and utilized by the National Academy of Sciences, i.e., the National Toxicology’s Office of Heath Assessment and Translation systematic review method, and the Navigation Guide Systematic Review Method. EPA’s TSCA systematic review framework should be peer-reviewed by qualified external experts in the field.

[i] PROSPERO International prospective register of systematic reviews https://www.crd.york.ac.uk/prospero/

2. EPA's TSCA systematic review framework utilizes a quantitative scoring method that is incompatible with the best available science in fundamental ways:

- **a. Quantitative scores for assessing the quality of an individual study are arbitrary and not science-based; the Cochrane Collaboration and National Academy of Sciences recommend against such scoring methods.**
- **b. EPA's scoring method wrongly conflates how well a study is reported with how well the underlying research was conducted; and**
- **c. EPA's scoring method excludes research based on one single reporting or methodological limitation.**

A detailed explanation of each of these scientific shortcomings is provided below.

(a) Quantitative scores for assessing the quality of an individual study are arbitrary and not science-based.

EPA's SR framework employs a quantitative scoring method to assess the quality of individual studies, assigning, based on its "professional judgment", various weights for quality domains and then summing up the quantitative scores to decide whether a study is of "high", "medium", or "low" quality as follows:[j]

> (Pg. 33) A numerical scoring method is used to convert the confidence level for each metric into the overall quality level for the data/information source. The overall study score is equated to an overall quality level (*High*, *Medium*, or *Low*) using the level definitions and scoring scale shown in Table A-1. The scoring scale was obtained by calculating the difference between the highest possible score of 3 and the lowest possible score of 1 (i.e., 3-1= 2) and dividing into three equal parts (2 ÷ 3 = 0.67). This results in a range of approximately 0.7 for each overall data quality level, which was used to estimate the transition points (cut-off values) in the scale between *High* and *Medium* scores, and *Medium* and *Low* scores. These transition points between the ranges of 1 and 3 were calculated as follows: Cut-off values between *High* and *Medium*: 1 + 0.67= 1.67, rounded up to 1.7 (scores lower than 1.7 will be assigned an overall quality level of *High*) Cut-off values between *Medium* and *Low*: 1.67 + 0.67= 2.34, rounded up to 2.3 (scores between 1.7 and lower than 2.3 will be assigned an overall quality level of *Medium)*

<u>This overall scoring method is applied to all streams of evidence, and our comments reflect our objection to EPA's applying scoring to any and all streams of evidence</u>.[k]

Illustrative of the scoring method, in Appendix H "Data Quality Criteria for Epidemiologic Studies," (page

[j] See Appendix A for a more detailed description of the scoring method; how the method will be applied specifically to various streams of evidence, i.e., occupational exposure and release data; animal and in vitro data; epidemiologic studies; etc., is described in subsequent Appendices B-H.

[k] EPA's framework applies quantitative scoring to all types of data; EPA/OPPT "is not applying weighting factors to the general population, consumer, and environmental exposure data types. In practice, it is equivalent to assigning a weighting factor of 1, which statistically assumes that each metric carries an equal amount of weight." (Pg. 96).

225) EPA presents how scoring is further applied to human studies, explaining:

> The critical metrics within each domain are those that cover the most important aspects of the domain and are those that more directly evaluate the role of confounding and bias. After pilot testing the evaluation tool, EPA recognized that more attention (or weight) should be given to studies that measure exposure and disease accurately and allow for the consideration of potential confounding factors. Therefore, metrics deemed as critical metrics are those that identify the major biases associated with the domain, evaluate the measurement of exposure and disease, and/or address any potential confounding. … EPA/OPPT assigned a weighting factor that is twice the value of the other metrics within the same domain to each critical metric. Remaining metrics are assigned a weighting factor of 0.5 times the weighting factor assigned to the critical metric(s) in the domain. The sum of the weighting factors for each domain equals one.

There is no scientific evidence to support EPA's selection of these "critical metrics" as being more important that other metrics, i.e., why within the "study participation" domain "selection" and "attrition" are more important than "comparison group"; and there are no data supporting EPA's choice of particular numbers for weighting these 'critical metrics' (i.e., some metrics are "twice" as important as the other metrics).

Overall, there is no scientific justification for EPA to assign these or any other quantitative scoring measures for assessing the quality of an individual study. The implicit assumption in quantitative scoring methods is that we know empirically how much each risk of bias domain contributes to study quality, and that these domains are independent of each other. This is not a scientifically supportable underlying assumption. Research has documented that scoring methods have, at best, unknown validity, may contain invalid items, and that results of a quality score are not scientifically meaningful or predictive of the quality of studies (26-28). An examination of the application of quality scores in meta-analysis found that quality-score weighting produced biased effect estimates, with the authors explaining that quality is not a singular dimension that is additive, but that it is possibly non-additive and non-linear (29).

Aggregating across quality criteria to produce a single score is recognized by preeminent systematic review methodologists as problematic and unreliable because the weights assigned are arbitrary and focus on the quality of reporting rather than the design and conduct of the research (21, 30). Scoring is not utilized by empirically based systematic review methodologies, such as the Cochrane Collaboration or GRADE (21, 31). As stated by the Institute of Medicine, "… systematic review teams have moved away from scoring systems to assess the quality of individual studies toward a focus on the components of quality and risk of bias" (16).

The Cochrane Collaboration, founded in 1993, is an international non-profit and independent organization that produces and disseminates systematic reviews of healthcare interventions and is a key locus of the world's most authoritative expertise on systematic review methods. Cochrane's methodology states: "The current standard in evaluation of clinical research calls for reporting each component of the assessment tool separately **and not calculating an overall numeric score** **(emphasis added**)"(31).

The National Academy of Sciences in its review of the EPA's IRIS program's method for SR, strongly supported a methodology that did not incorporate quantitative scoring, stating:

> … Cochrane discourages using a numerical scale because calculating a score involves choosing a weighting for the subcomponents, and such scaling generally is nearly impossible to justify (Juni et al. 1999). Furthermore, a study might be well designed to eliminate bias, but because the study failed to report details in the publication under review, it will receive a low score. Most scoring systems mix criteria that assess risk of bias and reporting. However, there is no empirical basis for weighting the different criteria in the scores. Reliability and validity of the scores often are not measured. Furthermore, quality scores have been shown to be invalid for assessing risk of bias in clinical research (Juni et al. 1999). The current standard in evaluation of clinical research calls for reporting each component of the assessment tool separately and not calculating an overall numeric score (Higgins and Green 2008) (17)(Pg. 69).

b) EPA's scoring method wrongly conflates how well a study is reported with how well the underlying research was conducted.

Study reporting addresses how well research findings are written up, i.e., whether there is a complete and transparent description of what was planned, what was done, what was found, and what the results mean. Guidelines and checklists for authors have been developed to help ensure all information pertinent to assessing the quality and meaning of research is included in the report. The "Strengthening of Reporting of Observational Studies in Epidemiology" or "STROBE" Initiative is an example of a checklist of items that should be included in articles reporting such research.[l]

EPA's SR Framework uses reporting measures in its scoring of the quality of human studies, including incorporating reporting guidelines into the reasons for scoring studies "low quality" (Metrics 1 and 15) or "unacceptable for use" (Metrics 2, 3, 4, 6, 7). EPA's SR Framework acknowledges that reporting is not the same as an underlying flaw in study methodology (Pg. 31), but then proceeds to ignore this distinction by using reporting as a measure of the quality of the underlying research. EPA's SR Framework not only does not "untangle" reporting from quality, it specifically conflates the two by using metrics in the STROBE reporting guidelines to score individual studies. The authors of the STROBE guidelines specifically note the guidelines are not a measure of the quality of the underlying research, stating:

> The STROBE Statement is a checklist of items that should be addressed in articles reporting on the 3 main study designs of analytical epidemiology: cohort, case control, and cross-sectional studies. The intention is solely to provide guidance on how to report observational research well; these recommendations are not prescriptions for designing or conducting studies. Also, while clarity of reporting is a prerequisite to evaluation, the checklist is not an instrument to evaluate the quality of observational research (emphasis added). … Our intention is to explain how to report research well, not how research should be done. We offer a detailed explanation for each checklist item. Each explanation is preceded by an example of what we consider transparent reporting. This does not mean that the study from which the example was taken was uniformly well reported or well done; nor does it mean that its findings were reliable, in the sense that they were later confirmed by others: it only means that this particular item was well reported in that study."(32)

How completely and clearly a study is reported is not a scientifically valid measure of the quality of the

[l] See Strobe statement at: https://www.strobe-statement.org/index.php?id=strobe-aims

underlying research (20, 21, 33, 34). As GRADE methodologists have succinctly stated, "… just because a safeguard against bias is not reported does not mean it was neglected"(21). Moreover, including many reporting items that are irrelevant to bias in a quality scoring rule (e.g., an indicator of whether power calculations were reported), will disproportionately reduce some of the resulting scores (29).

The Cochrane Collaboration Handbook for conducting a SR clearly distinguishes reporting and bias, the latter which is defined as "a systematic error, or deviation from the truth, in results or inferences" (20). The Cochrane Manual for conducting systematic reviews is explicit about not conflating reporting with bias, stating:

> Bias may be distinguished from **quality**. The phrase 'assessment of methodological quality' has been used extensively in the context of systematic review methods to refer to the critical appraisal of included studies. The term suggests an investigation of the extent to which study authors conducted their research to the highest possible standards. This *Handbook* draws a distinction between assessment of methodological quality and assessment of risk of bias, and recommends a focus on the latter. The reasons for this distinction include:
>
> 1. The key consideration in a Cochrane review is the extent to which results of included studies should be *believed*. Assessing risk of bias targets this question squarely.
> 2. A study may be performed to the highest possible standards yet still have an important risk of bias. For example, in many situations it is impractical or impossible to blind participants or study personnel to intervention group. It is inappropriately judgemental to describe all such studies as of 'low quality', but that does not mean they are free of bias resulting from knowledge of intervention status.
> 3. Some markers of quality in medical research, such as obtaining ethical approval, performing a sample size calculation and reporting a study in line with the CONSORT Statement (Moher 2001d), are unlikely to have direct implications for risk of bias.
> 4. An emphasis on risk of bias overcomes ambiguity between the quality of reporting and the quality of the underlying research (although does not overcome the problem of having to rely on reports to assess the underlying research).

Importantly, in the application of EPA's SR Framework, studies can be scored as "low quality," and even excluded from EPA's review, based solely on a deficiency in reporting, irrespective of the quality of the underlying research. Research documents that important information is often missing or unclear in published research (35), as word limits, styles, and other specifications are highly variable, and non-standardized among peer-reviewed journals. As such, efforts to improve reporting are focused on uptake of reporting guidelines by journal editors and researchers (32, 36, 37). Improving reporting is needed in academic research, but as stated by the developers of the STROBE guidelines, "We want to provide guidance on how to report observational research well. … the checklist is not an instrument to evaluate the quality of observational research."

Given the historical and present-day deficiencies in how studies are reported in the peer-reviewed literature, and because EPA's scoring system rates as 'unacceptable for use' any human study that does not report even one of five reporting metrics, EPA's proposal could reasonably be expected to lead to the exclusion from EPA's consideration much of the existing body of knowledge on the impact of

environmental chemicals on human health, and is inconsistent with TSCA mandates to use the "best available science" and "reasonably available information." [m] Applying flawed exclusion criteria that directly contradicts widely accepted empirically based SR methodological approaches will almost certainly result in flawed conclusions and threaten the protection of the public's health.

(c) EPA's scoring method excludes research based on one single reporting or methodological limitation.

In the "fatal flaw" component of EPA's SR Framework's scoring system, for each type of evidence stream, i.e., epidemiologic, animal, *in vitro*, etc., EPA created an arbitrary list of metrics that make studies "unacceptable for use in the hazard assessment," stating:

> EPA/OPPT plans to use data with an overall quality level of *High, Medium*, or *Low* confidence to quantitatively or qualitatively support the risk evaluations, but does not plan to use data rated as *Unacceptable*. Studies with any single metric scored as 4 will be automatically assigned an overall quality score of *Unacceptable* and further evaluation of the remaining metrics is not necessary (emphasis added). An *Unacceptable* score means that serious flaws are noted in the domain metric that consequently make the data unusable (or invalid) (Pg. 227).

There is no empirical basis for EPA's selected list of fatal flaws.

Illustrative of this "fatal flaw" aspect of EPA' scoring system, for human epidemiologic studies (See Section H.5, Table H-8 (page 231), EPA lists six domains of study quality, i.e., study participation; exposure characterization; outcome assessment; potential confounding/variable control; analysis**;** and other considerations for biomarker selection and measurement, and 19 metrics to assess the six domains. A study that has even one of the 19 "serious flaws" metrics is considered to be "unacceptable for use."

The underlying assumptions of EPA's "serious flaws" metrics are not science-based because:

- **EPA's list of "serious flaws" are not all equal indicators of study quality**:
  For example, among human observational studies, any one of the list of 19 metrics can eliminate a study from consideration as EPA considers all of these "flaws" to be of equal import; as described in detail above, such weighting is arbitrary and not a science-based method.

- **EPA's list of "serious flaws" are not all related to real flaws in the underlying research**:

  - **Reporting** guidelines are wrongly equated with "serious flaws" in study quality.
    For example, in scoring the quality of human studies, 5 of 19 "serious flaw" metrics (Table H-8) are STROBE reporting guidelines (STROBE checklist items # 6,7,8,13,15). A study would be scored as "unacceptable for use" by EPA based on any one of these STROBE reporting guidelines. As described above in comment #2a, the STROBE guideline developers explicitly state this is neither the intended nor a scientifically valid use of these guidelines. (32)

[m] 15 USC §2625(h) and (k)

- **Analysis**[n] is equated with a "serious flaw" in study quality, but statistical power[o] alone is not a valid measure of study quality. For example, EPA's framework excludes human studies that do not meet EPA's criteria for "high" in the analysis domain. EPA does not state how it will calculate whether a study is "adequately" powered. According to EPA's framework, to be included in an EPA review, a study must meet the "high" criteria in EPA's "Metric 13, Statistical power (sensitivity, reporting bias)" as presented in the box below. Studies that are not "high" quality for this metric would be designated as "unacceptable for use" by EPA:

EPA Metric 13. Excerpted from Table H-9 (page 243)

Metric 13. Statistical power (sensitivity, reporting bias) **Instructions: To meet criteria for confidence ratings for metrics where 'AND' is included, studies must address both of the conditions where "AND" is stipulated. To meet criteria for confidence ratings for metrics where 'OR' is included studies must address at least one of the conditions stipulated.**	
High (score = 1)	***For cohort and cross-sectional studies:*** The number of participants are adequate to detect an effect in the exposed population and/or subgroups of the total population. **OR** The paper reported statistical power high enough (≥ 80%) to detect an effect in the exposure population and/or subgroups of the total population. ***For case-control studies:*** The number of cases and controls are adequate to detect an effect in the exposed population and/or subgroups of the total population. **OR** The paper reported statistical power was high (≥ 80%) to detect an effect in the exposure population and/or subgroups of the total population.
Medium (score = 2)	• Do not select for this metric.
Low (score = 3)	• Do not select for this metric.
Unacceptable (score = 4)	• ***For cohort and cross-sectional studies:*** The number of participants are inadequate to detect an effect in the exposed population and/or subgroups of the total population. • ***For case-control studies:*** The number of cases and controls are inadequate to detect an effect in the exposed population and/or subgroups of the total population.

[n] See Table H-8 "Serious Flaws that Would Make Epidemiological Studies Unacceptable for Use in the Hazard Assessment" under the "analysis domain" "statistical power/sensitivity" metric (page 233) " in conjunction with Table H-9 "Evaluation Criteria for Epidemiologic Studies, Metric 13 "statistical power (sensitivity, reporting bias) (page 243).

[o] A power calculation is an estimate of the size of the study population needed to detect an effect of a given size.

First and foremost, EPA provides no method for how it will determine the "adequacy" of the statistical power of a study on which to base its score, and provides no rationale for excluding studies with less than 80% statistical power. According to STROBE guideline developers, … "before a study is conducted power calculations are made with many assumptions that once a study is underway may be upended; further, power calculations are most often not reported" (32).

EPA's Metric 13 statistical power/sensitivity also appears to confuse bias with imprecision. Individual studies that are "underpowered" (for example, because in the real world the exposed population may not be large enough for statistical purposes even if they are health impacted) can still be potentially valuable to science-based decision-making. For example a small study may be imprecise but that should not be confused with whether it is biased (20); a small study can be imprecise but at the same time less biased than a larger study (17). Small "underpowered" studies can also be combined in a meta-analysis that increases the statistical power of the body of evidence to reflect the relationship between an exposure and a health impact. Additionally, "underpowered" studies that find a health effect to be present may be indicative of a larger effect size than anticipated. Thus, omitting such studies would severely bias the conclusions of the review.

Illustrative of how EPA's "analysis" metric could result in excluding high quality research that can inform science-based decision-making by EPA, in a 2017 systematic review by Lam et al. *"Developmental PBDE Exposure and IQ/ADHD in Childhood: A Systematic Review and Meta-analysis,"*(12) none of the 4 high-quality[p] studies included in the meta-analysis reported a power calculation, and yet together, these studies found "a 10-fold increase (in other words, times 10) in PBDE exposure associated with a decrement of 3.70 IQ points (95% confidence interval:0.83,6.56)." It is also notable that one of the studies in the meta-analysis, Herbstman et al. 2010, (38) was assessed by the review authors to be "probably high risk of bias" for "Incomplete Outcome Data."[q] As such, this otherwise high quality study, i.e., all of the other domains were "definitely" or "probably" low risk of bias, would meet EPA's criteria for "unacceptable for use" based on STROBE reporting guideline #15, "Report numbers of outcome events or summary measures over time".[r]

In short, the *Lam et al* systematic review, using the best available scientific methods, found that a ubiquitous environmental contaminant is impacting human intelligence, a finding that was subsequently reviewed and endorsed by the National Academy of Sciences (19). Yet EPA's SR review framework would exclude crucial pieces of this body of evidence based on the Agency's inaccurate, non-science-based criteria for deeming studies 'unacceptable.' This is contrary to TSCA's mandate to use the best available science. [s]

- **"Level of exposure"** is equated with a "serious flaw".

[p] "High quality" defined as "definitely" or "probably" low or very low risk of bias (Figure 2a in the *Lam et al* paper) based on specific and detailed definitions of risk of bias established before the review was conducted.

[q] The authors of the systematic review rated the Herbstman 2010 study "probably high risk of bias" for "incomplete outcome data" based on the following rationale: "Concerns regarding missing outcome data at each follow-up time on almost half the cohort of 210 with cord blood PBDE measurements; no argument is presented that would invalidate the possibility of a selection bias (i.e., likelihood that outcome data is missing is related both to outcome status and exposure)."

[r] See Table H-8 "Serious Flaws that Would Make Epidemiological Studies Unacceptable for Use in the Hazard Assessment" under the "outcome assessment domain" "Outcome measurement or characterization" metric (page 232) which specified STROBE guideline #15 to assess this metric.

[s] 15 USC §2625 (h)

EPA's "exposure characterization" domain for human studies includes the level of exposure as a fatal flaw, stating: "For all study types: The **levels** of exposure are not sufficient or adequate (as defined above)[t] to detect an effect of exposure (Cooper et al., 2016)." Unlike human experimental studies, which are largely precluded for ethical reasons, human observational studies can only be based on what exposures actually occur in the real world. EPA offers no explanation of how one could know whether the levels would be "sufficient or adequate" enough to detect an effect. Given the vagaries of this metric, it could be reasonably anticipated that it would permit EPA to arbitrarily exclude quality research from its decision-making.

We recommend: EPA should not use a quantitative scoring method to assess quality in individual studies; it should not conflate study reporting with study quality; and it should not exclude otherwise quality research based on a single reporting or methodological limitation. Rather EPA should employ a scientifically valid method to assess risk of bias of individual studies.

[t] EPA "as defined above" is unclear, presumably "as defined above" refers to the definition of the domain in Table H-2 page 223, "Evaluation of exposure assessment methodology that includes consideration of methodological quality, sensitivity, and validation of the methods used, degree of variation in participants, and an established time order between exposure and outcome."

3. EPA's TSCA systematic review framework does not consider financial conflicts of interest as a potential source of bias in research.

As observed by the Deputy Editor (West) of JAMA in 2010, "the biggest threat to [scientific] integrity [is] financial conflicts of interest" (39). Yet EPA's systematic review framework is silent on how it will take into account this empirically documented influence on the results of scientific research. Underscoring this EPA SR framework deficiency is the fact that recent studies empirically document that industry sponsorship produces research that is favorable to the sponsor (40, 41). The influence of financial ties on research can be traced to a variety of types of biases, and this conflict of interest needs to be distinguished from non-financial interests in the research, which can also affect research (42).

The fact that funding source needs to be accounted for in some manner is empirically supported and not a subject of scientific debate; what scientists differ on is *how* to best address funding as a potential source of bias (43, 44); for example, whether funding source is assessed as a specific risk of bias domain (43) or considered at multiple points in the evaluation (20, 44). For example, funding source is recommended as a factor to consider when evaluating risk of bias of individual studies for selective reporting, and then again for evaluating the body of evidence for publication bias, (45) and/or to be considered as a potential factor to explain apparent inconsistency within a body of evidence (14).

A 2017 Cochrane systematic review of industry sponsorship and research outcome concluded … "industry sponsorship should be treated as bias-inducing and industry bias should be treated as a separate domain" (40). The National Academy of Sciences in its review of the EPA IRIS program's SR method found that "Funding sources should be considered in the risk-of-bias assessment conducted for systematic reviews that are part of an IRIS assessment (17)(p 79).

Notably, EPA's exclusion of consideration of funding source and other potential conflicts of interests is also internally inconsistent with EPA's own improper reliance on STROBE guidelines as quality measures: STROBE guidelines item #22 specified that "the source of funding and the role of funders, could be addressed in an appendix or in the methods section of the article" (32).

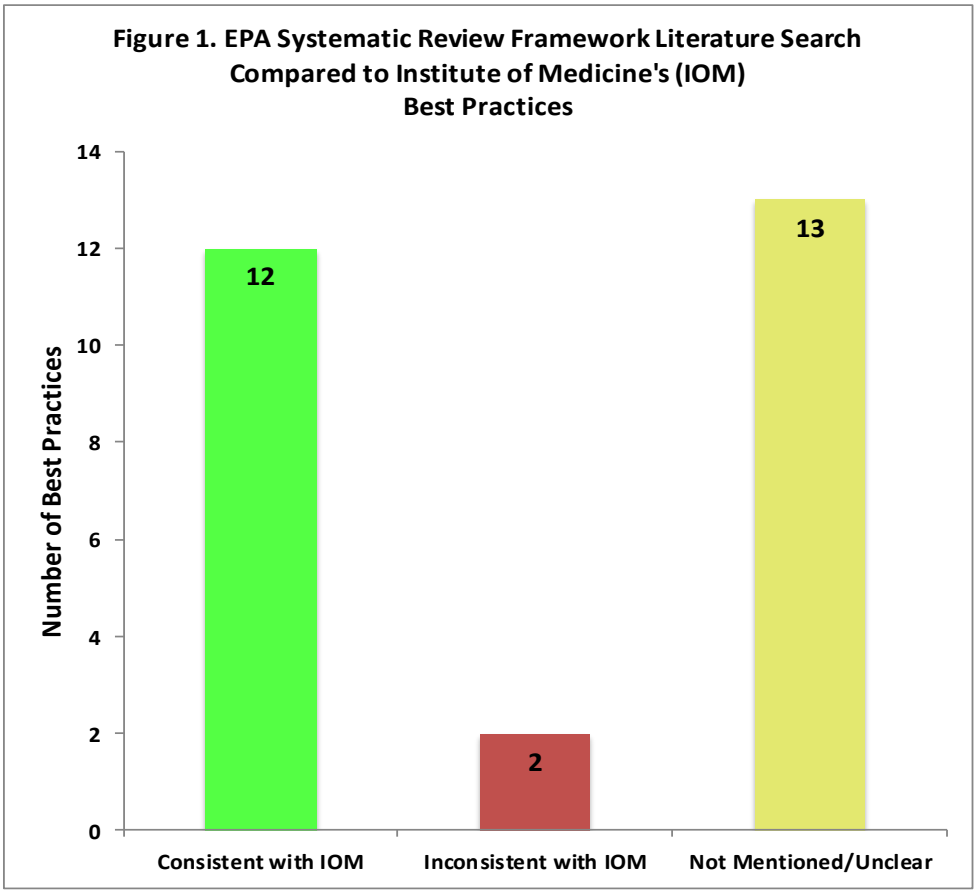
Importantly, including funding as a risk of bias as a domain does not mean excluding industry sponsored studies from EPA's hazard and risk assessment; it only means documenting funding as one of many domains of potential bias and evaluating its impact on the overall quality of the body of evidence.

We recommend: EPA should assess study and author funding source as a risk of bias domain for individual studies.

4. The literature review step of EPA's TSCA systematic review framework incorporates select best practices, but also falls short of, or is unclear about, many other best practices for conducting a systematic and transparent literature review.

Overall, we commend the EPA for its efforts to incorporate many best practices for a comprehensive literature search in its systematic review framework. We compared EPA's framework for systematic review to the Institute of Medicine's (IOM's) best practices for the literature review step of a systematic review (16)(See IOM 2011 Chapter 3. and TABLE E-1), which was applied by the National Academy of Sciences in its review of EPA's IRIS Program methods for systematic review (17)(See Table 4-1 Pp. 43-55).

We found EPA's framework to be consistent with 12 of IOM's 27 best practices for conducting a literature search (Figure 1 and Appendix 1). There are two key features of EPA's framework that are clearly inconsistent with IOM's best practices. EPA fails: (1) to include or exclude studies based on the protocol's pre-specified criteria, a practice that is critical to avoiding results-based decisions;[u] and (2) to use two or more members of the review team, working independently, to screen and select studies, which is an essential quality-assurance measure.[v]



[u] See our Comment #1 regarding the EPA framework's lack of a pre-defined protocol.

[v] EPA's framework, "Summary of the Title/Abstract Screening Conducted for the First Ten TSCA Risk Evaluations" (page 24) states that only one screener conducted the screening and categorization of titles and abstracts.

For the remaining 13 IOM best practices, EPA's framework is either unclearly stated (N=7) or the practice is not mentioned at all (N=6). However, based on the literature review methods presented in the First Ten TSCA Risk Evaluations, EPA's framework appears to have incorporated six additional best practices that are either unclear or not mentioned in EPA's SR framework: (1) work with a librarian or other information specialist trained in performing systematic reviews to plan the search strategy (IOM 3.1.1); (2) Design the search strategy to address each key research question (IOM 3.1.2); (3) Search regional bibliographic databases if other databases are unlikely to provide all relevant evidence (IOM 3.1.9); (4) Conduct a web search (IOM 3.2.5); and (5) Provide a line-by-line description of the search strategy, including the date of search for each database, web browser, etc. (IOM 3.4.1).

EPA should make its framework for conducting a literature review transparently congruent with all of IOM's best practices. This includes addressing two critical inconsistencies: (1) include or exclude studies based on the protocol's pre-specified criteria to prevent results-based decisions; and (2) Use two or more members of the review team, working independently, to screen and select studies, to ensure quality assurance. The transparency of the framework would be improved by specifying how EPA is addressing each best practice; at this juncture, how EPA intends to specifically handle many components of its literature searches could not readily be identified.

For example, the framework is unclear about whether EPA will include papers published in languages other than English. The exclusive reliance on English-language studies may lead to under-representation of the entire body of available evidence, and studies have also suggested that language bias might lead to erroneous conclusions (46). Furthermore, when considering the inclusion or update of an existing systematic review, studies have found that language-inclusive systematic reviews (including studies in languages other than English) were of the highest quality, compared with other types of reviews (47). Online translation tools are readily available to allow screeners to quickly evaluate study abstracts for relevance, and therefore we recommend EPA to incorporate non-English language studies in their screening and not simply exclude in advance these potentially relevant papers.

Additionally, EPA's framework should explicitly include rules for determining when the list of relevant studies will be considered final i.e., "stopping rules." Newer scientific studies will inevitably continue to appear in scientific journals and it will be impossible to continually attempt to include all these studies in a chemical assessment. To meet the deadlines as mandated by the Lautenberg Amendments, EPA should state clear stopping rules in the form of deadlines or criteria for when the body of included relevant studies will be finalized for the purposes of the chemicals assessment. We also strongly encourage EPA in its stated exploration of automation and machine learning tools,[w] which can help speed the production of EPA's systematic reviews.

We recommend: EPA should make its framework for conducting a literature review congruent with all of the Institute of Medicine's best practices, and explicitly include rules for when the list of relevant studies will be considered final.

[w] Footnote 9 page 23 states "In addition to using DistillerSR, EPA/OPPT is exploring automation and machine learning tools for data screening and prioritization activities (e.g., SWIFT-Review, SWIFT-Active Screener, Dragon, DocTER). SWIFT is an acronym for "Sciome Workbench for Interactive Computer-Facilitated Text-mining".

5. EPA's TSCA systematic review framework correctly recognizes that mechanistic data are not required for a hazard assessment, but EPA is not clear that these data, if available, can only be used to increase, and not to decrease, confidence in a body of evidence.

EPA's TSCA framework (page 172) states that EPA will use the evaluation strategies for animal and *in vitro* toxicity data to assess the quality of mechanistic and pharmacokinetic data supporting the model, and may tailor its criteria further to evaluate new approach methodologies (NAMs). We agree with EPA that mechanistic data need to be evaluated in a manner comparable to how other streams of evidence are evaluated. Data generated by alternative test methods (such as high-throughput screening methods) are not different than any other type of *in vitro* or cell-based assay data that would be considered in a systematic review. These kinds of assays provide mechanistic data. However, in this case, as described in comment # 2 above, EPA's use of its evaluation strategies for animal and *in vitro* toxicity data would entail using a quantitative scoring method that is incompatible with the best available science in fundamental ways. EPA should employ a scientifically valid method to assess risk of bias of individual studies in *all* streams of evidence, including mechanistic data.

EPA's framework (page 172) states, "the availability of a fully elucidated mode of action (MOA) or adverse outcome pathway (AOP) is not required to conduct the human health hazard assessment for a given chemical (emphasis added)." We strongly agree with EPA that mechanistic data are not needed for a hazard assessment. In addition, EPA's framework should be explicit that mechanistic data are only used to increase confidence in a hazard assessment, and never to decrease confidence.

The National Academy of Sciences explicitly considered how mechanistic data could be utilized in a systematic review for evidence integration (19). The committee came to two conclusions. First, the same protocol for evaluating relevance and study quality must be used with mechanistic data as for any other study. For example, in the report's case study on phthalates, the committee was not able to integrate results from high-throughput assays because the cell lines used were of unknown relevance to the *in vivo* mechanism of phthalate toxicity (19)(pg.78). Second, the foundation of the hazard classification in a systematic review is the animal and human data, with the mechanistic data playing a supporting role. If mechanistic data is relevant, it can be used to upgrade a hazard classification, or increase the confidence of a finding made based on evaluation of animal and human data. A hazard classification is never made based on high-throughput or other kinds of mechanistic data alone (19)(Pp. 158-9).

We recommend: EPA should be explicit that mechanistic data can only be used to upgrade a hazard classification, or increase the confidence of a finding made based on evaluation of animal and human data, and that these data will not be used to decrease confidence in a body of evidence.

6. EPA's TSCA systematic review framework is not independent of the regulatory end user of the review.

EPA's TSCA systematic review/risk assessment process is not independent of the TSCA risk management process, a conflict that is incompatible with best scientific methods. EPA's SR framework was developed and is being implemented by the Office of Chemical Safety and Pollution Prevention (OCSPP), which is also responsible for regulating the environmental exposures under TSCA review. In contrast, other EPA chemical assessment programs such as the IRIS program are intentionally placed in a non-regulatory research arm (the Office of Research and Development), to create separation from the Agency's program office responsible for regulatory decisions. This separation supports IRIS's ability to develop impartial chemical toxicity information independent of its ultimate use by EPA's program and regional office in risk assessment and risk management decisions. The National Academy of Sciences supported this in its 2018 report, stating that: "Current best practices [for systematic reviews in other medical disciplines] recommended by the Institute of Medicine (IOM 2011) suggest that the IRIS teams involved in the systematic-review process **should be independent of those involved in regulatory decision-making** who use the products of the systematic-review teams **(emphasis added)**" (15). This same principle should also be implemented across the Agency and specifically for TSCA assessments.

We recommend: EPA's systematic reviews should be produced independently of the regulatory end user of the review.

REFERENCES

1. Rennie D, Chalmers I. Assessing authority. JAMA. 2009;301(17):1819-21. Epub 2009/05/07. doi: 301/17/1819 [pii]10.1001/jama.2009.559. PubMed PMID: 19417202.
2. Fox DM. The Convergence of Science and Governance: Research, Health Policy, and American States. Berkeley, CA: University of California Press; 2010.
3. Antman EM, Lau J, Kupelnick B, Mosteller F, Chalmers TC. A comparison of results of meta-analyses of randomized control trials and recommendations of clinical experts. Treatments for myocardial infarction. JAMA. 1992;268(2):240-8. Epub 1992/07/08. PubMed PMID: 1535110.
4. Woodruff TJ, Sutton P, The Navigation Guide Work Group. An Evidence-Based Medicine Methodology To Bridge The Gap Between Clinical And Environmental Health Sciences. Health Affairs. 2011;30(5):931-7. doi: 10.1377/hlthaff.2010.1219; PMCID: 21555477.
5. Woodruff TJ, Sutton P. The Navigation Guide sytematic review methodology: a rigorous and transparent method for translating environmental health science into better health outcomes. Environmenal Health Perspectives. 2014;122(10):A283.
6. Johnson PI, Sutton P, Atchley DS, Koustas E, Lam J, Sen S, Robinson KA, Axelrad DA, Woodruff TJ. The Navigation Guide - evidence-based medicine meets environmental health: systematic review of human evidence for PFOA effects on fetal growth. Environ Health Perspect. 2014;122(10):1028-39. Epub 2014/06/27. doi: 10.1289/ehp.1307893. PubMed PMID: 24968388; PMCID: 4181929.
7. Koustas E, Lam J, Sutton P, Johnson PI, Atchley DS, Sen S, Robinson KA, Axelrad DA, Woodruff TJ. The Navigation Guide - evidence-based medicine meets environmental health: systematic review of nonhuman evidence for PFOA effects on fetal growth. Environ Health Perspect. 2014;122(10):1015-27. Epub 2014/06/27. doi: 10.1289/ehp.1307177. PubMed PMID: 24968374; PMCID: 4181920.
8. Lam J, Koustas E, Sutton P, Johnson PI, Atchley DS, Sen S, Robinson KA, Axelrad DA, Woodruff TJ. The Navigation Guide - evidence-based medicine meets environmental health: integration of animal and human evidence for PFOA effects on fetal growth. Environ Health Perspect. 2014;122(10):1040-51. Epub 2014/06/27. doi: 10.1289/ehp.1307923. PubMed PMID: 24968389; PMCID: 4181930.
9. Vesterinen H, Johnson P, Atchley D, Sutton P, Lam J, Zlatnik M, Sen S, Woodruff T. The relationship between fetal growth and maternal glomerular filtration rate: a systematic review. J Maternal Fetal Neonatal Med. 2014:1-6. Epub Ahead of Print; PMCID: 25382561.
10. Johnson PI, Koustas E, Vesterinen HM, Sutton P, Atchley DS, Kim AN, Campbell M, Donald JM, Sen S, Bero L, Zeise L, Woodruff TJ. Application of the Navigation Guide systematic review methodology to the evidence for developmental and reproductive toxicity of triclosan. Environ Int. 2016;92-93:716-28. doi: 10.1016/j.envint.2016.03.009. PubMed PMID: 27156197.
11. Lam J, Sutton P, Halladay A, Davidson LI, Lawler C, Newschaffer CJ, Kalkbrenner A, Joseph J. Zilber School of Public Health, Windham GC, Daniels N, Sen S, Woodruff TJ. Applying the Navigation Guide Systematic Review Methodology Case Study #4: Association between Developmental Exposures to Ambient Air Pollution and Autism. PLoS One. 2016;21(11(9)). doi: 10.1371/journal.pone.0161851.

12. Lam J, Lanphear B, Bellinger D, Axelrad D, McPartland J, Sutton P, Davidson LI, Daniels N, Sen S, Woodruff TJ. Developmental PBDE exposure and IQ/ADHD in childhood: A systematic review and meta-analysis. Environmenal Health Perspectives. 2017;125(8). doi: doi: 10.1289/EHP1632.
13. Lam J, Koustas E, Sutton P, Cabana M., Whitaker E., Padula A, Vesterinen H, Daniels N, Woodruff TJ. Applying the Navigation Guide: Case Study #6. Association Between Formaldehyde Exposures and Asthma. In preparation. 2018.
14. National Toxicology Program. Handbook for Conducting a Literature-Based Health Assessment Using OHAT Approach for Systematic Review and Evidence Integration. In: U.S. Department of Health and Human Services, editor.: Office of Health Assessment and Translation, Division of National Toxicology Program, National Institute of Environmental Health Sciences; 2015.
15. National Academies of Sciences, Engineering, and, Medicine. Progress Toward Transforming the Integrated Risk Information System (IRIS) Program: A 2018 Evaluation. Washington, D.C.: The National Academies Press; 2018.
16. Institute of Medicine. Finding What Works in Health Care. Standards for Systematic Review. Washington, D.C.: The National Academies Press.; 2011.
17. National Research Council. Review of EPA's Integrated Risk Information System (IRIS) Process. Washington, DC: National Academies Press; 2014.
18. National Research Council. Review of the Environmental Protection Agency's State-of-the-Science Evaluation of Nonmonotonic Dose–Response Relationships as They Apply to Endocrine Disruptors. Washington, DC: National Academies Press; 2014.
19. National Academies of Sciences, Engineering, and, Medicine. Application of Systematic Review Methods in an Overall Strategy for Evaluating Low-Dose Toxicity from Endocrine Active Chemicals. Washington, DC: 2017 2017. Report No.
20. Higgins JPT, Green S. Cochrane Handbook for Systematic Reviews of Interventions. Version 5.1.0 [Updated March 2011]: The Cochrane Collaboration. Available from http://www.cochrane-handbook.org.; 2011.
21. Guyatt G, Oxman AD, Akl EA, Kunz R, Vist G, Brozek J, Norris S, Falck-Ytter Y, Glasziou P, DeBeer H, Jaeschke R, Rind D, Meerpohl J, Dahm P, Schünemann HJ. GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. Journal of Clinical Epidemiology. 2011;64(4):383-94. doi: 10.1016/j.jclinepi.2010.04.026.
22. Guyatt GH, Oxman AD, Vist G, Kunz R, Brozek J, Alonso-Coello P, Montori V, Akl EA, Djulbegovic B, Falck-Ytter Y, Norris SL, Williams JW, Jr., Atkins D, Meerpohl J, Schünemann HJ. GRADE guidelines: 4. Rating the quality of evidence--study limitations (risk of bias). Journal of Clinical Epidemiology. 2011;64(4):407-15. doi: 10.1016/j.jclinepi.2010.07.017.
23. Vandenberg LN, Ågerstrand M, Beronius A, ., Beausoleild C, Bergman A, Bero LA, Bornehag C, Boyer CS, Cooper GS, Cotgreave I, Gee D, Grandjean P, Guyton KZ, Hass U, Heindel JJ, Jobling S, Kidd KA, Kortenkamp A, Macleod MR, Martin OV, Norinder U, Scheringer M, Thayer KA, Toppari J, Whaley P, Woodrufft TJ, Ruden C. A proposed framework for the systematic review and integrated assessment (SYRINA) of endocrine disrupting chemicals. Environment Health. 2016;In press.

24. Hooijmans CR, Rovers MM, de Vries RB, Leenaars M, Ritskes-Hoitinga M, Langendam MW. SYRCLE's risk of bias tool for animal studies. BMC medical research methodology. 2014;14:43. Epub 2014/03/29. doi: 10.1186/1471-2288-14-43. PubMed PMID: 24667063.
25. Campbell Collaboration. Better evidence for a better world. 2018 [cited 2018 July 29]The Campbell Collaboration promotes positive social and economic change through the production and use of systematic reviews and other evidence synthesis for evidence-based policy and practice.]. Available from: https://campbellcollaboration.org/research-resources/writing-a-campbell-systematic-review.html
26. Herbison P, Hay-Smith J, Gillespie W. Adjustment of meta-analyses on the basis of quality scores should be abandoned. J Clin Epidemiol. 2006;59(12):1249-56. Epub 2006 Sep 11; PMCID: 17098567.
27. Juni P, Witschi A, Bloch R, Egger M. The hazards of scoring the quality of clinical trials for meta-analysis. JAMA. 1999;282(11):1054-60. Epub 1999/09/24. doi: joc81641 [pii]. PubMed PMID: 10493204.
28. Stang A. Critical evaluation of the Newcastle-Ottawa scale for the assessment of the quality of nonrandomized studies in meta-analyses. Eur J Epidemiol. 2010;25(9):603-5. Epub 2010 Jul 22. doi: 10.1007/s10654-010-9491-z.; PMCID: 20652370.
29. Greenland S, O'Rourke K. On the bias produced by quality scores in meta-analysis, and a hierarchical view of proposed solutions. Biostatistics. 2001;2(4):463-71.
30. Moher D, Jadad AR, Nichol G, Penman M, Tugwell P, S. W. Assessing the quality of randomized controlled trials: an annotated bibliography of scales and checklists. Control Clin Trials. 1995;16:62-73.
31. Higgins JPT, Altman DG, Sterne JAC. Chapter 8: Assessing risk of bias in included studies. In: Higgins J, Green S, editors. Cochrane Handbook for Systematic Reviews of Interventions Version 510 [Updated March 2011]: The Cochrane Collaboration. Available from http://www.cochrane-handbook.org.; 2011.
32. Vandenbroucke JP, von Elm E, Altman DG, Gøtzsche PC, Mulrow CD, Pocock SJ, Poole C, Schlesselman JJ, Egger M, Initiative. S. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration. Int J Surg. 2014;12(12):1500-24. doi: 10.1016/j.ijsu.2014.07.014.
33. Devereaux PJ, Choi PT, El-Dika S, Bhandari M, Montori VM, Schünemann HJ, Garg AX, Busse JW, Heels-Ansdell D, Ghali WA, Manns BJ, GH. G. An observational study found that authors of randomized controlled trials frequently use
concealment of randomization and blinding, despite the failure to report these methods. J Clin Epidemiol. 2004;57(12):1232-6; PMCID: 15617948.
34. Soares HP, Daniels S, Kumar A, Clarke M, Scott C, Swann S, B; D, Group. RTO. Bad reporting does not mean bad methods for randomised trials: observational study of randomised controlled trials performed by the Radiation Therapy Oncology Group. BMJ. 2004;328((7430)):22-4.; PMCID: PMC313900.
35. Lee W, Bindman J, Ford T, Glozier N, Moran P, Stewart R, M H. Bias in psychiatric case-control studies: literature survey. Br J Psychiatry. 2007;190:204-9.; PMCID: 17329739.
36. Kilkenny C, Browne W, Cuthill IC, Emerson M, Altman DG. Animal research: reporting in vivo experiments--the ARRIVE guidelines. J Cereb Blood Flow Metab. 2011;31(4):991-3. Epub 2011/01/06. doi: 10.1038/jcbfm.2010.220. PubMed PMID: 21206507; PMCID: 3070981.

37. Shamseer L, Moher D, Clarke M, Ghersi D, Liberati A, Petticrew M, Shekelle P, Stewart LA, Group. P-P. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) :elaboration and explanation. BMJ. 2015;350:(g7647). doi: 10.1136/bmj.g7647.
38. Herbstman JB, Sjödin A, Kurzon M, Lederman SA, Jones RS, Rauh V, Needham LL, Tang D, Niedzwiecki M, Wang RY, Perera F. Prenatal exposure to PBDEs and neurodevelopment. Environ Health Perspect. 2010.
39. Rennie D. Integrity in scientific publishing. Health Serv Res. 2010;45(3):885-96. Epub 2010/03/27. doi: HESR1088 [pii]
10.1111/j.1475-6773.2010.01088.x. PubMed PMID: 20337732.
40. Lundh A, Lexchin J, Mintzes B, Schroll JB, Bero L. Industry sponsorship and research outcome. Cochrane Database Syst Rev. 2017(2:MR000033.). doi: 10.1002/14651858.MR000033.pub3.; PMCID: 28207928.
41. White J, Bero LA. Corporate manipulation of research: Strategies are similar across five industries. . Stanford Law & Policy Review. 2010;21((1)):105-34. .
42. Bero L. Addressing Bias and Conflict of Interest Among Biomedical Researchers. JAMA. 2017;317(17):1723-4. doi: 10.1001/jama.2017.3854; PMCID: 28464166.
43. Bero L. Why the Cochrane risk of bias tool should include funding source as a standard item [editorial]. Cochrane Database Syst Rev. 2013;12:ED000075.
44. Sterne JA. Why the Cochrane risk of bias tool should not include funding source as a standard item. . Cochrane Database Syst Rev. 2013;(12)(:ED000076.). doi: 10.1002/14651858.ED000076.; PMCID: PMID: 24575440.
45. Viswanathan M, Ansari M, Berkman N, Chang S, Hartling L, McPheeters L, Santaguida P, Shamliyan T, Singh K, Tsertsvadze A, Treadwell J. Assessing the Risk of Bias of Individual Studies in Systematic Reviews of Health Care Interventions. 2012 AHRQ Publication No. 12-EHC047-EF.
46. Morrison A, Polisena J, Husereau D, Moulton K, Clark M, Fiander M, Mierzwinski-Urban M, Clifford T, Hutton B, Rabb D. The effect of English-language restriction on systematic review-based meta-analyses: a systematic review of empirical studies. . International journal of technology assessment in health care. 2012;28((2)):138-44.
47. Moher D, Pham B, Lawson ML, Klassen TP. The inclusion of reports of randomised trials published in languages other than English in systematic reviews. . Health Technol Assess. 2003;7((41)):1-90.