COMMENTS ON
THE EPA IRIS PROGRAM'S SUBMISSIONS
TO THE NRC IRIS COMMITTEE

Submitted by
The American Chemistry Council
and
The Center for Advancing Risk Assessment Science and Policy

June 17, 2013

EXECUTIVE SUMMARY

The American Chemistry Council (ACC)[1] and our Center for Advancing Risk Assessment Science and Policy (ARASP)[2] have been actively engaged in reviewing and providing scientific information to EPA's Integrated Risk Information System (IRIS) Program. We are committed to promoting the development and application of up-to-date, science-based methods for conducting chemical assessments. In that regard, we strongly support EPA's activities to significantly improve IRIS to ensure that the Program produces high quality and scientifically sound chemical assessments.

We appreciate the efforts of the IRIS Program to improve its documentation and enhance consistency and transparency in the Agency's approach to develop hazard assessments as evidenced by EPA's January 30, 2013 submission (EPA Submission) to the National Research Council (NRC) IRIS Review Committee (Committee).[3] We have reviewed the EPA Submission and welcome the opportunity to provide comments and share our perspectives on these documents.

1. General Comments

It's now been more than two years since the NRC Committee issued its *Review of the Environmental Protection Agency's Draft IRIS Assessment of Formaldehyde*,[4] and although some important upgrades in IRIS have been made, progress has been slow, and many necessary improvements have yet to be implemented. There are also missing elements in the EPA Submission, which EPA clearly acknowledges, including: "integrating across evidence (epidemiological, toxicological, and mechanistic data) to identify hazards and transition to dose-

---

[1] The American Chemistry Council (ACC) represents the leading companies engaged in the business of chemistry. ACC members apply the science of chemistry to make innovative products and services that make people's lives better, healthier and safer. ACC is committed to improved environmental, health and safety performance through Responsible Care®, common sense advocacy designed to address major public policy issues, and health and environmental research and product testing.

[2] ARASP is a coalition of 19 organizations focused on promoting the development and application of up-to-date, scientifically sound methods for conducting chemical assessments. ARASP supports science based chemical assessments that utilize objective, transparent data acquisition and evaluation criteria. ARASP also advocates for the use of mode of action data in risk assessment. ARASP members include: Acrylonitrile Group, ACC Chlorine Chemistry Division, ACC Ethylene Oxide Panel, ACC Formaldehyde Panel, ACC Hexavalent Chromium Panel ACC High Phthalates Panel, ACC Hydrocarbon Solvents Panel, ACC Oxo Process Panel, ACC Propylene Oxide/Propylene Glycol Panel, ACC Public Health and Science Policy Team, ACC Olefins Panel, ACC Vinyl Chloride Health Committee, American Cleaning Institute, American Petroleum Institute, CropLife America Halogenated Solvents Industry Alliance, Nickel Producers Environmental Research Association Silicones Environmental, Health and Safety Council of North America, Styrene Information and Research Center

[3] EPA Jan 30, 2013, Part 1: Status of Implementation of Recommendations http://www.epa.gov/IRIS/pdfs/IRIS%20Program%20Materials%20to%20NRC_Part%201.pdf and EPA Jan 30, 2013, Part 2: Chemical -Specific Examples. http://www.epa.gov/IRIS/pdfs/IRIS%20Program%20Materials%20to%20NRC_Part%202.pdf .

[4] National Research Council, Review of the Environmental Protection Agency's Draft IRIS Assessment of Formaldehyde (2011). Available at: https://download.nap.edu/catalog.php?record_id=13142

response analysis; conducting dose-response modeling; extrapolating to lower doses and response levels; considering susceptible populations and lifestyles; developing candidate toxicity values; characterizing confidence and uncertainty in toxicity values; and selecting final toxicity values."[5] We recommend that EPA accelerate its efforts to address these elements, and in addition, promptly move ahead to:

- Enhance problem formulation to include early stakeholder and independent expert consultations on proposed procedures and scientific methods for hazard evaluation and dose response assessment.

- Adopt transparent and consistent procedures for evaluating and integrating evidence using uniform evaluation methods to determine quality and reliability for the different types of studies (epidemiology, *in vivo* toxicology, *in vitro* toxicology and mechanistic studies) that are involved, to varying degrees, in every IRIS assessment.

- Apply a scientifically-solid framework for integrating study results based on a weight of evidence approach to establish cause and effect which incorporates modern knowledge of mode of action (MOA) to determine potential risks to humans at environmentally relevant exposures. This is needed to comply with the NRC recommendation that outcomes should be unified around common modes of action rather than considering multiple outcomes separately.

- Improve the representation and communication of toxicity values to more accurately reflect scientific certainties and uncertainties, to include assessment of the sensitivity of derived estimates to model assumptions and end points selected and employ appropriate tabular and graphic displays to illustrate the range of the estimates and the effect of uncertainty on the estimates.

- Upgrade the peer review of IRIS chemical assessments to ensure effective and robust input from EPA's Chemical Assessment Advisory Committee to evaluate cross cutting issues, not solely chemical-specific assessments.

- Strengthen the peer review process to promote opportunities for more meaningful input from stakeholders and independent experts and by improving the transparency and responsiveness to both public comments and peer review recommendations.

Given the important and influential impact that EPA's new IRIS documents[6] will have, when adopted and implemented, on the conduct and quality of IRIS assessments, EPA's current approach for stakeholder input falls short. We believe that these documents will qualify as

[5] See EPA Jan 30, 2013, Part 1: Status of Implementation of Recommendations http://www.epa.gov/IRIS/pdfs/IRIS%20Program%20Materials%20to%20NRC_Part%201.pdf at page F-1.
[6] op. cit. reference 3.

economically significant guidance documents which are subject to the requirements of the Office of Management (OMB) Final Bulletin for Agency Good Guidance Practices.[7] In addition, consistent with OMB Memorandum on Guidance for Regulatory Review,[8] such significant guidance documents should be subject to review by the Office of Information and Regulatory Affairs under Executive Orders 12866 and 13563.

2. Specific Comments

In response to the NRC's recommendations,[9] the IRIS program has made some changes to streamline the assessment development process, improve transparency, and create efficiencies within the program. We conducted a detailed analysis of the EPA Submission and have identified significant shortcomings with EPA's guidance and assessment methods. To address these, we provide specific recommendations. Our key findings and recommendations include:

A. The current preamble does not provide a clear description of specific search strategies, exclusion and inclusion criteria, and weight of evidence approaches as the NRC recommended. Instead, it provides an abbreviated view of EPA policies, guidance documents and standard practices, but fails to include the detail necessary to provide useful information on how the Agency reviews or weighs the scientific information for inclusion in the particular toxicological review. In providing this abbreviated view, critical information has been omitted and the preamble may lead readers to incorrectly interpret EPA guidance.

B. The scoping phase of the IRIS development should include a transparent problem formulation step where each analysis begins with a set of proposed hypotheses that incorporates MOA, the adverse effect(s) of concern, and the exposure level(s) of concern.

C. When conducting and evaluating literature searches, EPA should consider critical toxicology information, including studies on absorption, distribution, metabolism and excretion (ADME) to be of primary relevance, not as additional resource information. This critical toxicological information is necessary to understand MOA and human relevance.

D. EPA must conduct a critical review of the quality and relevance of studies that are included in an evaluation. Clear criteria should define how studies were selected for inclusion. Without a critical assessment of quality and relevance those that read, review and use evidence tables will mistakenly think that each study should be treated equally.

---

[7] See OMB, Final Bulletin for Agency Good Guidance Practices (Jan. 18, 2007). Available at: http://www.whitehouse.gov/sites/default/files/omb/assets/regulatory_matters_pdf/m07-07.pdf
[8] See OMB, Guidance for Regulatory Review (Mar. 4, 2009). Available at: http://www.whitehouse.gov/sites/default/files/omb/assets/memoranda_fy2009/m09-13.pdf.
[9] Op cit. reference 4.

E. Significant improvements are needed to EPA's evaluation of epidemiologic information to make the guidance consistent with current scientific practice. The assessment of causality needs critical improvements and clarifications.

F. To improve the evaluation and display of individual studies, the tables need to provide more details on the statistical results and methodologies used. This information and data should be made publicly available to enable independent analysis and verification of EPA's conclusions.

G. EPA's discussion of dose-response modeling is insufficient in that there is no guidance on extrapolation to lower doses and response levels. The guidance should strengthen the support for conducting physiologically based pharmacokinetic modeling. In addition, the IRIS Program direction to contractors should include guidance for conducting non-linear cancer extrapolation.

H. When evaluating and integrating evidence, a discussion of biological plausibility must be provided and EPA must improve the consideration and incorporation of MOA information. MOA needs to be the central organizing principle in conducting hazard and risk assessments. In accordance with established best practices of systematic evidence-based reviews, EPA should employ a consistent weight of evidence framework, based on specific hypothesized MOAs to permit data from laboratory experiments, epidemiological investigations, and cutting-edge mechanistic research to be integrated in a manner that provides a robust understanding of the MOA and the potential hazards and risks that exposures to a substance could pose to humans.

I. EPA's peer review enhancements are not yet sufficient. EPA should also use the Chemical Assessment Advisory Committee to review cross-cutting IRIS issues. In addition, the IRIS process would be strengthened by use of an independent monitor who can ensure that comments from all reviewers are appropriately and sufficiently addressed.

3. Necessary Next Steps

While the IRIS Program has indicated EPA is accepting comments on the documents submitted to the NRC IRIS Committee, the comment period was not formally announced, nor was a docket created to receive submission of detailed comments and attachments. We recommend EPA create a docket on regulations.gov and announce a formal 60-day comment period via the Federal Register. In addition, the NRC IRIS Committee should hold an open public meeting to discuss the EPA's Draft Handbook for IRIS Assessment Development, to encourage further public input and robust discussion into the EPA revisions to the IRIS Program. Finally, the IRIS handbook and associated documents should be treated as economically significant guidance documents subject to the requirements of the OMB Final Bulletin for Agency Good Guidance

Practices and subject to review by the Office of Information and Regulatory Affairs under Executive Orders 12866 and 13563.

TABLE OF CONTENTS

EXECUTIVE SUMMARY ii

I. INTRODUCTION 1

II. GENERAL COMMENTS 1

A. IMPROVEMENTS NEEDED IN PROCEDURES COMMON TO ALL IRIS ASSESSMENT 1

B. STAKEHOLDER INPUT ON EPA'S SUBMISSION TO THE NRC IRIS COMMITTEE 2

III. SPECIFIC COMMENTS 3

A. IRIS TOXICOLOGICAL REVIEW TEMPLATE 3

B. PREAMBLE TO IRIS TOXICOLOGICAL REVIEWS 4

C. EXAMPLES OF IRIS PROGRAM DIRECTION TO CONTRACTORS 9

D. INFORMATION MANAGEMENT TOOL: COMMENT TRACKER DATABASE 11

E. SCOPING TO INFORM THE DEVELOPMENT OF IRIS ASSESSMENTS 12

F. DRAFT HANDBOOK FOR IRIS ASSESSMENT DEVELOPMENT 13

1. Literature Search Strategies 13
2. Evaluation and Display of Individual Studies 14
3. Evaluating Data Quality 19
4. Evaluating and Integrating Evidence 21
5. Dose-Response Analysis 33

G. EXTERNAL PEER REVIEW ENHANCEMENTS 38

IV. NECESSARY NEXT STEPS 38

V. REFERENCES 39

I. INTRODUCTION

On January 30, 2013 the EPA submitted materials to the NRC IRIS Committee (Committee) which included: (a) Part 1: Status of Implementation of Recommendations and (b) Part 2: Chemical-Specific Examples. The Committee is charged with reviewing the changes being implemented by the IRIS Program, including a discussion of EPA's weight-of-evidence analysis practices, and identifying further opportunities to improve the scientific and technical performance of the IRIS Program. Given the Committee's critically important task of assessing the scientific, technical, and process changes being implemented by EPA to the IRIS Program, it is imperative that a robust review of the materials submitted by EPA be conducted.

According to the EPA Submission to the Committee (hereinafter referred to as "the EPA Submission"), EPA asserts that it plans to fully implement the recommendations from the 2011 NRC *Review of the Environmental Protection Agency's Draft IRIS Assessment of Formaldehyde*. In response to the NRC's recommendations, the IRIS program has made some changes to streamline the assessment development process, improve transparency, and create efficiencies within the Program. While we applaud EPA for making progress, we are concerned that key elements concerning the adoption of transparent procedures for determining study quality, integrating evidence and effectively communicating uncertainty have yet to be addressed in the past two years. We have identified specific recommendations for improving several items in the EPA Submission including:

(a) IRIS Toxicological Review Template;
(b) Preamble to IRIS Toxicological Reviews;
(c) Example of IRIS Program Direction to Contractors;
(d) Information Management Tool: Comment Tracker Database;
(e) Scoping to Inform the Development of IRIS Assessments;
(f) Draft Handbook for IRIS Assessment Development; and
(g) External Peer Review Enhancements.

As the Committee conducts its review, and as EPA engages stakeholders on these implemented and pending IRIS Program enhancements, reviewers and authors should take note of methods and approaches that continue to need upgrades and revisions, including aspects that EPA considers to be implemented, and undertake improvements as necessary.

II. GENERAL COMMENTS

A. IMPROVEMENTS NEEDED IN PROCEDURES COMMON TO ALL IRIS ASSESSMENTS

Many fundamental improvements are needed to the policies and practices of the IRIS Program. These include improvements in the upfront design of IRIS assessments, the adoption of consistent and transparent study evaluation methods to determine quality and reliability, an improved framework for integrating study results based on a weight of evidence approach which

incorporates modern knowledge of mode action to establish cause and effect, and improvements in peer review and Agency accountability in addressing both public comments and peer review recommendations. Progress in implementation has been slow, and we recommend that EPA accelerate action in the following areas:

- Adopting Transparent and Consistent Procedures for Evaluating and Integrating Evidence – While EPA has made progress in adopting modern and more transparent procedures for literature searches and identifying relevant studies, the Agency has yet to adopt consistent and transparent study evaluation methods to determine quality and reliability for the different types of studies (epidemiology, *in vivo* toxicology, *in vitro* toxicology and mechanistic studies) that are involved, to varying degrees, in every IRIS assessment. One of the most critical improvements needed in the IRIS Program is a scientifically-solid framework for integrating study results based on a weight of evidence approach to establish cause and effect which incorporates modern knowledge of MOA to determine potential risks to humans at environmentally relevant exposures.
- Enhancing Problem Formulation – Improvements in the upfront design of IRIS assessments are needed so that problem formulation appropriately covers the specific questions and issues to be addressed in the IRIS assessment. The final design should be informed by early stakeholder and independent expert consultations on proposed procedures and scientific methods for hazard evaluation and dose response assessment.
- Improving Communication and Presentation of Toxicity Values – There is a significant need to improve how hazard values are presented and communicated in order to accurately reflect scientific certainties and uncertainties.
- Upgrading Peer Review of Chemical Assessments – While establishment of the EPA Science Advisory Board Chemical Assessment Advisory Committee (CAAC) is a positive step to improve IRIS peer review, in order to ensure effective and robust input from the CAAC there will need to be opportunities for the CAAC to evaluate cross cutting issues, not solely chemical-specific assessments. The peer review process should also be strengthened to enhance opportunities for more meaningful input from stakeholders and independent experts and by improving the transparency and responsiveness to both public comments and peer review recommendations.

B. STAKEHOLDER INPUT ON EPA'S SUBMISSION TO THE NRC IRIS COMMITTEE

While the IRIS Program has noted on its website that EPA is accepting comments on the documents submitted to the NRC IRIS Committee[10], the comment period was not formally announced and the website does not allow for the submission of attachments or detailed written comments. Given the important and influential impact that these documents will have, when adopted and implemented, on the conduct and quality of IRIS assessments, EPA's current

[10] http://www.epa.gov/IRIS/iris-nrc.htm

approach for stakeholder input falls short. We believe that these documents, when final, will qualify as economically significant guidance documents which are subject to the requirements of the Office of Management (OMB) Final Bulletin for Agency Good Guidance Practices.[11] In addition, consistent with OMB Memorandum on Guidance for Regulatory Review,[12] such significant guidance documents should be subject to review by the Office of Information and Regulatory Affairs under Executive Orders 12866 and 13563.

Therefore, we recommend that EPA:

1) Announce a formal 60-day comment period, through the Federal Register, seeking public comment on all aspects of the January 2013 submission to the NAS.
2) Create a public docket on regulations.gov for documents included in the January 2013 submission and for submission of comments.
3) Request that NRC hold an open public meeting to discuss the EPA's Draft Handbook for IRIS Assessment Development, which will encourage public input into the EPA revisions to the IRIS Program.
4) Treat these documents as economically significant guidance and submit them to OMB for coordinated interagency review as required by Executive Orders 12866 and 13563.

III. SPECIFIC COMMENTS

A. IRIS TOXICOLOGICAL REVIEW TEMPLATE

EPA discusses its new template for IRIS Toxicological Reviews in Appendix A of Part 1 of the EPA Submission. We support a new streamlined document structure to improve the clarity and readability of the documents. However, we encourage EPA to include all appropriate information in the main body of the document, instead of the moving all documentation to the appendices. It is important that assessments contain sufficient information to be reproducible. While EPA states that changes to the template are fully implemented, further improvements are necessary. Below are specific suggestions for improving the new structure:

- The new preface refers to assessments conducted by "Other National and International Health Agencies." We suggest that EPA broaden this to include peer reviewed assessments from all national and international agencies, as well as other independent authoritative reviews. For instance the National Library of Medicine (NLM) provides, through the International Toxicity Estimates for Risk (ITER) and Toxicology Data Network (TOXNET), data sources that include, but are not limited to, state agencies, international non-health agencies, and other peer reviewed sources.[13]

[11] See OMB, Final Bulletin for Agency Good Guidance Practices (Jan. 18, 2007). Available at: http://www.whitehouse.gov/sites/default/files/omb/assets/regulatory_matters_pdf/m07-07.pdf
[12] See OMB, Guidance for Regulatory Review (Mar. 4, 2009). Available at: http://www.whitehouse.gov/sites/default/files/omb/assets/memoranda_fy2009/m09-13.pdf.
[13] See NLM fact sheet for ITER available at: http://www.nlm.nih.gov/pubs/factsheets/iterfs.html.

- When discussing organ/system specific reference doses (RfDs) and reference concentrations (RfCs), EPA should always provide the uncertainty factors that were applied to develop the reference values (RfVs). Similarly, tables like Table 2-5 of Appendix A of Part 1 of the EPA Submission should also include the confidence rating for the study noted.

- While the new preface includes a "confidence statement" in section 2.1.7 for the RfD and 2.2.7 for the RfC, the EPA has not included a similar section for the cancer values (the oral slope factor and the inhalation unit risk). A discussion of agency confidence should accompany all quantitative values, both cancer and non-cancer values. There is no scientific rationale for treating the non-cancer and cancer effects differently when it comes to transparently describing the confidence in the values.

B. PREAMBLE TO IRIS TOXICOLOGICAL REVIEWS

EPA discusses the preamble in Appendix B of Part 1 of the EPA Submission. On August 6, 2012, ACC submitted comments to EPA on the draft ammonia assessment where we provided detailed comments on the draft preamble.[14] EPA has noted that the new preamble was developed by the Agency in response to a recommendation from the NRC. However, in its review of EPA's draft formaldehyde assessment, NRC stated:

> Chapter 1 needs to be expanded to describe more fully the methods of the assessment, including a description of search strategies used to identify studies with the exclusion and inclusion criteria articulated and a better description of the outcomes of the searches and clear descriptions of the weight-of-evidence approaches used for the various non-cancer outcomes. The committee emphasizes that it is not recommending the addition of long descriptions of EPA guidelines to the introduction, but rather clear concise statements of criteria used to exclude, include, and advance studies for derivation of the RfCs and unit risk estimates.

The current preamble does not sufficiently address the NRC's recommendation as it does not provide a clear description of specific search strategies, exclusion and inclusion criteria, and weight of evidence approaches. Specifically, as noted in our August 2012 comments on the preamble, as currently written, the preamble offers an abbreviated view of EPA policies, guidance documents and standard practices but fails to include the detail necessary to provide useful information on how the Agency reviews or weighs the scientific information for inclusion in the particular toxicological review. Unfortunately, in providing this abbreviated view, critical information has been omitted and the preamble may unduly lead readers to incorrectly interpret EPA guidance.

[14] Comment submitted by Center for Advancing Risk Assessment Science and Policy (ARASP), http://www.regulations.gov/#!documentDetail;D=EPA-HQ-ORD-2012-0399-0017

In addition to the general comments noted above we also provide some specific recommendations for improvements to the preamble. While EPA states that changes to the Preamble are fully implemented, further improvements are necessary.

- Section 2, Process for developing and peer-reviewing IRIS assessments: In this section the EPA has provided an overview of the May 2009 revised process for developing IRIS assessments.[15] In step 4 of the development process, EPA estimates at least 3 ½ months for external peer review and comment, but does not specify specific time frames for public input prior to the draft assessment being released or denote a time frame for delivery of public comments to the peer review panel prior to the peer review meeting. Currently, when the draft toxicological reviews are released by the Agency they are near final – decisions about the main conclusions are presented as a *fait accompli*, stifling valuable input. Involving the public and other stakeholders earlier in the process will enable a more meaningful dialogue that can contribute to the development of the draft toxicological review. This engagement with stakeholders should include the identification of useful MOA information, applicable data evaluation frameworks to synthesize the scientific information being reviewed, relevant studies and data, as well as other relevant topics.

- Section 3, Identifying and selecting pertinent studies: This section provides a summary of the basic search strategy the Agency utilizes to gather scientific information for inclusion in the toxicological review and offers the key considerations used to select pertinent epidemiological and experimental studies. However, there are several areas where this section could be greatly improved.

    - Section 3.2 provides some key considerations for selecting epidemiological studies and specifically states that "Cohort studies…provide the strongest epidemiological evidence, as they collect information about individual exposure." However, not all cohort studies collect information based on individual exposure level; one example of this is cohort air pollution studies that are based on group level exposure (e.g., ambient monitoring). This section should provide clear guidance as to what type of information would generally be given more or less weight in the data evaluation framework.

    - Sections 3.2 and 3.3 of the preamble purport to provide the key design considerations for selecting pertinent epidemiological and/or experimental studies from the results of the literature search and note exposure route and duration as key considerations. However, these sections do not provide the criteria used by the Agency for selecting studies. These sections should include all the considerations EPA utilizes in selecting a study for inclusion in the toxicological review and which of the criteria are deemed most necessary. Furthermore, EPA does not provide information that would allow

[15] U.S. EPA. (2009). EPA's Integrated Risk Information System: Assessment development process. Available at http://epa.gov/iris/process.htm.

the public to replicate EPA's literature selection process for the chemical being assessed as recommended by NRC.

- Section 4, Evaluating the quality of individual studies: This section provides basic information on how the assessment evaluates various design and methodological aspects of the data that could increase or decrease the weight given to a study in the overall evaluations. Some examples listed in this section include: documentation of study design, exposure classification, disease classification and sample size. However, it is not clear which elements EPA deems most valuable for a study to possess for use in its data evaluation. The 2011 NRC report explicitly called on EPA to adopt standard data evaluation procedures/protocols for each of the major types of studies that typically need to be reviewed in conducting an IRIS assessment. To date, EPA has provided only very general considerations for study evaluations, and this falls short of what was recommended by the NRC. EPA can improve this section by:

    - Adopting clear and consistent guidance for evaluating studies. ARASP's recent review of the existing methods currently used by environmental health agencies globally to establish study reliability and data quality for *in vivo* and *in vitro* studies shows that there are best practices the IRIS Program can immediately implement for these types of studies.[16]

    - Providing the specific elements or characteristics that would increase or decrease a study's weight (e.g., does a low sample size decrease the weight of a study in the overall evaluation of the available scientific information). This section should include a list of the design or methodological aspects that increase weight and a list of the aspects that decrease weight.

    - Expanding the discussion on the use of historical controls. The draft assessment should clearly note that EPA's Cancer Guidelines[17] discussion on the use of historical controls clearly states: "However, caution should be used in interpreting results."

- Section 5, Evaluating the overall evidence of each effect: This section discusses how the Agency evaluates the scientific evidence as a whole to determine the extent to which any observed association may be causally linked to the chemical of interest. EPA notes that positive, negative and null results are given weight according to the study quality and provides some aspects to consider in making that association to causality (i.e., strength of association, temporal relations, and biological plausibility). However, the section does not indicate how EPA assigns weight to studies or whether, for instance, studies of similar quality are given equal weight regardless of whether the study's results are positive, negative or null. EPA's weighting scheme should be discussed in more detail and clear criteria should be provided for increasing and decreasing weight. Information should be included in this

[16] Available at http://arasp.americanchemistry.com/Data-Quality-Evaluation
[17] U.S EPA (2005a) Guidelines for carcinogen risk assessment (EPA/630/P-03/001F). http://www.epa.gov/cancerguidelines/.

section on how positive, negative and null studies are evaluated and weighted (i.e., are they given equal weight). The preamble also does not clearly identify which weight of evidence approach(es) EPA supports or utilizes. EPA should provide a listing of data evaluation practices that are used in the toxicological review. Additional examples where the section could be improved are provided below:

- Section 5.1 begins to discuss the criteria for causality, but then moves away from causality to focus on determining whether or not an "association" exists. IRIS assessments should retain a focus on whether evidence of causality exists.

- Section 5.2 provides some standard descriptors that may be used. EPA implies that suggestive epidemiologic information will be "consistent with causation" and the Agency does not seem to envision a scenario where there is suggestive epidemiologic information but a causal relation does not exist. The provided descriptors should capture all the realistic scenarios.
    - EPA's standard for suggestive evidence is typified when bias and confounding cannot be ruled out. However, such weak epidemiological evidence may not be consistent with causation. As currently formulated, EPA's criteria does not adequately capture such scenario and is needs to be modified.

- Section 5.4, discusses evaluating MOA data and adverse outcome pathways. However the section does not discuss the concept of "significant biological support." This is an important concept in EPA's Cancer Guidelines. For instance, at page 3-23, the Cancer Guidelines state: "Nonlinear extrapolation having a significant biological support may be presented in addition to a linear approach when the available data and a weight of evidence evaluation support a nonlinear approach, but the data are not strong enough to ascertain the mode of action applying the Agency's mode of action framework." Different modeling approaches can be used even when there is a lack of MOA information.

- Section 5.4 states: "Key data include the ability of the agent or a metabolite to react with or bind to DNA, positive results in multiple test systems, or similar properties and structure-activity relationships to mutagenic carcinogens (U.S. EPA, 2005a)." This statement, which implies that negative data would not be equally considered if it was of equal quality, does not appear to be included in EPA's Cancer Guidelines. EPA should not use the preamble to establish new guidance. This sentence should be removed from the preamble.

- Section 5.5 seems to focus only on characterizing the overall weight of evidence for cancer and provides no guidance for non-cancer evaluations. In discussing the cancer evaluation, EPA notes that a narrative is provided that includes a standard hazard descriptor. EPA then provides the descriptors but provides no guidance for the narrative. This oversight should be corrected as the Cancer Guidelines correctly note that the complete narrative "preserves the complexity that is an essential part of the hazard characterization." Guidance on preparing this narrative should be provided.

- Section 5.5 also provides an example of standard descriptors used for evaluating criteria pollutants. It is unclear what purpose this serves in the preamble. If EPA is suggesting that this approach will be adopted for use in the assessment, this should be clearly stated. Before implementation of a new approach, EPA must seek appropriate peer review and public comment.

- Section 6, Selecting studies for derivation of toxicity values: In this section EPA should be clear about existing guidance for when a toxicity value would not be derived. In particular, the Cancer Guidelines state:

  > When there is suggestive evidence, the Agency generally would not attempt a dose-response assessment, as the nature of the data generally would not support one; however, when the evidence includes a well-conducted study, quantitative analyses may be useful for some purposes, for example, providing a sense of the magnitude and uncertainty of potential risks, ranking potential hazards, or setting research priorities. In each case, the rationale for the quantitative analysis is explained, considering the uncertainty in the data and the suggestive nature of the weight of evidence. These analyses generally would not be considered Agency consensus estimates. Dose-response assessments are generally not done when there is inadequate evidence, although calculating a bounding estimate from an epidemiologic or experimental study that does not show positive results can indicate the study's level of sensitivity and capacity to detect risk levels of concern.

- Section 7, Deriving toxicity values: This section discusses how EPA derives toxicity values and conducts extrapolation to low doses. However, some oversights and inconsistencies should be addressed:
  - In Section 7.3, when discussing extrapolation and selection of a response level, EPA should note that the Benchmark Dose Technical Guidance[18] suggests that an extra risk of 10% is recommended as a standard reporting level for quantal data, for the purpose of making comparisons across chemicals or endpoints. For determination of a point of departure, a lower (or sometimes higher) response is often used based on statistical and biological considerations; nevertheless, for reporting purposes, it is recommended that the benchmark dose (BMD) corresponding to 10% extra risk always be presented.

  - Section 7.4 incorrectly states that "linear extrapolation is also used if there is an absence of sufficient information on modes of action." As note previously, EPA's Cancer Guidelines indicate that if there is "significant biological support", and not a known mode of action, a non-linear extrapolation can be presented. Similarly, in describing when non-linear extrapolation is used, EPA again suggests that the MOA must be ascertained. This is not consistent with the Cancer Guidelines (see page 3-

[18] EPA's Benchmark Dose Technical Guidance (Risk Assessment Forum) June 2012. http://www.epa.gov/osa/raf/publications/benchmark_dose_guidance.pdf

23). In addition, the Cancer Guidelines state that "Where alternative approaches with significant biological support are available for the same tumor response and no scientific consensus favors a single approach, an assessment may present results based on more than one approach."

- The approach described in Section 7.4 inappropriately interjects risk management into an IRIS assessment, under the veil of "scientific analysis." EPA essentially asserts the default as "truth" and then requires that "sufficient" data be developed to refute the default. "Sufficient data" is never defined, and seems to be an ever moving target. This undermines research focused on applying modern techniques to improve the scientific evaluation of specific hypothesis as part of determining relevant modes of action. Instead of trying to ask and answer the question of "how much data and knowledge is enough to overrule a default?" what is needed is a framework that uses all of the relevant and reliable data and knowledge of hypothesized modes of action, in an open, objective and transparent manner, including, if warranted, valuation of the hypothesized MOA underlying the default.

- Section 7.6 does not adequately characterize what an oral reference dose (RfD) or an inhalation reference concentration (RfC) are because the text does not clearly state that RfD and RfC values are estimates, with uncertainty spanning perhaps an order of magnitude. EPA should correct its description in the preamble.

- Section 7.6 provides some discussion regarding uncertainty factors (UFs) however it is unclear what the Agency's policy is on the application of UFs. In this section, EPA appears to create new policy by stating that the UF for human variation is reduced only if the point of departure is derived specifically for susceptible individuals. EPA should provide clear criteria for the application of UFs and discuss how the Agency considers UFs in totality to ensure that any compounding conservatism in the derivation of a toxicity value does not lead to an unrealistic final value.

C. EXAMPLES OF IRIS PROGRAM DIRECTION TO CONTRACTORS

EPA discusses its implementation of dose-response modeling in Part 1, Appendix C and Part 2, Example 6 of the EPA Submission. This section of the EPA Submission provides examples of the IRIS Program direction to contractors and notes that it focuses only on animal data of "standard experimental designs" and does not address more complex study designs. As IRIS substances can often be data rich, they may also have more complex study designs. Thus any specialized methods developed for the analysis of "complex experimental designs" must also be adequately documented and peer reviewed. Some specific suggestions to improve this section are included below:

- Appendix C provides no guidance on the conduct of non-linear cancer dose-response analysis. Contractors should be provided guidance on how to conduct non-linear evaluations and EPA should devote a section specifically to non-linear approaches for cancer modeling.

- Guidance on the analysis of epidemiological studies should be provided. EPA should not treat epidemiological data as requiring "specialized methods" that should be conducted on a case by case basis. Standardization in review and modeling of epidemiological data is equally, if not more important, than animal data, particularly because the IRIS Program has a stated preference for the use of human data.

- As written, it appears that EPA is no longer strongly supporting the use of physiologically based pharmacokinetic (PBPK) models. Page C-6 states, "…this approach [PBPK modeling] is suggested, though is not necessary." EPA also states that PBPK modeling "makes it harder to update an assessment" if the model or pharmacokinetic data are later changed or updated. These comments are not consistent with other EPA documents that state a preference for the use of PBPK models in risk assessments. We support the use of well-designed PBPK models and encourage EPA to strengthen this language in Appendix C.

- The section on modeling cancer endpoints (beginning on page C-8) should be improved. EPA states in this section that when low-dose linearity is expected, the cancer (i.e., Multistage) model should be run, and that other models should only be run when a p-value <0.05 is achieved. Since BMD modeling is essentially a curve fitting exercise, it is unclear why the Multistage model would be preferable to other potentially better fitting models.

- The section on modeling cancer endpoints incorrectly states that if low-dose nonlinearity is expected, then all models can be run on a "relevant precursor effect." Models can and should also be run on the tumor data itself. EPA's Cancer Guidelines indicate that precursor key events can be used as surrogates for the apical endpoint of tumors; in such cases, a reference dose protective of the surrogate endpoint will also be protective of tumors. This is the case with the IRIS assessment of chloroform. In addition, this section appears to exclude the possibility that non-linear modeling could be used in cases where the dose-response for a cancer can empirically appear highly nonlinear, and there may be very plausible factors contributing to such nonlinearity. If non-linearity is expected, it is not clear why, even in the absence of an identified surrogate endpoint, tumor incidence itself could not provide a point of departure for nonlinear low dose extrapolation.

- The direction to contractors should include a discussion concerning the need to consider decoupling of a mode of action (MOA) in cancer dose-response analysis. The EPA Cancer Guidelines state, "If there are multiple modes of action at a single tumor site, one linear and another nonlinear, **then both approaches are used to** decouple and **consider the respective contributions of each mode of action in different dose ranges**" (emphasis added). For example, a tumor may arise in a study at doses that clearly involve non-mutagenic components such as sustained cell proliferation, yet the potential, albeit vanishingly small, for direct chemical mutation cannot be ruled out based on other known aspects of the chemical. In such a case, the risk associated with a 10% benchmark response (BMR) is comprised of both mutagenic and non-mutagenic components, and thus low-dose extrapolation from such a POD will likely over predict risk at lower exposure levels where toxicity-induced proliferation does not occur (i.e., contribute to cancer). To decouple this MOA and derive a slope factor based solely on the mutagenic component, a lower BMR

(e.g., 1%) might be selected (if the dose-response curve is robust enough to develop a $BMDL_{01}$) that results in a BMD_{01} where proliferation did not occur in the study. The resulting [shallower] slope factor, might allow for more reasonable risk predictions at environmental doses where mutation, but not proliferation, might contribute to cancer.[19,20]

- The document should include a discussion on BMR selection. The EPA BMD Technical Guidance contains useful information that should be conveyed to contractors in Appendix C. The BMD Technical Guidance indicates that a BMR less than 10% could be used if the BMR still falls within the observable range. Furthermore, it states that it is "...important to recognize that the BMR need not correspond to a response that the study could detect as statistically significantly different from the control response, provided that the response is considered biologically significant." Together, these two statements lend support for the decoupling approach discussed above.

- Methods for combining risk should be more fully described. The discussion concerning combined cancer risk (beginning on page C-9) describes two methods. The first is using EPA's "multitumor" BMD model, which is fairly well articulated. The second relies on a method described in NRC (1994).[21] This method should be described more fully. While the BMD modeling examples provided in Part 2 of the EPA Submission are relatively straightforward, more detailed guidance is needed for cases where, for example, a composite slope factor is derived from multiple tumor slope factors. Inclusion of a more intricate case example would improve upon the relatively simple example currently provided in Part 2.

D. INFORMATION MANAGEMENT TOOL: COMMENT TRACKER DATABASE

Appendix D of Part 1 of the EPA Submission discusses an initiative to improve documentation and communication of decision by the Chemical Assessment Support Teams (CAST). To address this need, EPA developed an information management tool that could be used to record, review, and analyze the comments and responses received on IRIS assessments. We appreciate EPA's efforts to improve comment tracking. In particular, if staffed by the appropriate experts, CAST should help EPA improve the quality and consistency of assessments. However, we offer some additional suggestions for improving transparency associated with the CAST and elements of the comment tracker.

- While EPA is silent on the CAST membership, we encourage the agency to include senior scientists, agency scientists from regional offices, and experts from outside of NCEA. A breadth and diversity of perspectives, as well as expertise, from throughout EPA program offices will help to enhance the scientific underpinnings of IRIS assessments.

---

[19] Borgert, C.J., Mihaich, E.M., Ortego, L.S., Bentley, K.S., Holmes, C.M., Levine, S.L., Becker, R.A. (2011). Hypothesis-driven weight of evidence framework for evaluating data within the US EPA's Endocrine Disruptor Screening Program. Regul. Toxicol. Pharmacol. 61(2):185-191.

[20] Another example is the case of naphthalene-induced nasal tumors in rats. Here an adjustment factor is used for mode of action uncertainty with dual-mode carcinogens. See Risk Anal 2008; 28(4):1033-1051.

[21] NRC. (1994) Science and judgment in risk assessment. Washington, DC: National Academy Press.

- The comment tracker database is another important improvement. As this will be used to track all comments on different scientific themes, including comments from within EPA as well as from stakeholders and peer reviewers, EPA should make this important information management tool publicly accessible. With a publicly available database, all stakeholders can be equally informed from discussions and decisions that have taken place.

- EPA should clarify whether or not public and peer reviewers will need to use the comment tracker template when providing comments to the agency. In its current format, this could stifle and limit comments from outside experts, which would be an unfortunate unintended consequence. A further understanding of how EPA will use and populate this database is necessary.

E. SCOPING TO INFORM DEVELOPMENT OF IRIS ASSESSMENTS

In Appendix E of Part 1 of the EPA Submission, the Agency discusses an improved scoping phase to put greater attention on the design of the risk assessment. We strongly recommend that this scoping process take place early in the process and be fully transparent. EPA has already begun to pilot implementation of the initiative in the case of inorganic arsenic and we view this as a positive improvement. We offer some additional suggestions below.

- For future scoping sessions, we encourage EPA to ensure that the facilitated discussions are open to all stakeholders and that designated speakers include not only those researchers working collaboratively with EPA, but also other researchers with appropriate expertise. Diverse perspectives, early in the scoping process will help to improve the assessment.

- Scoping sessions should include discussion not only of "what" will be covered, but also of "how" EPA plans to address this issue. This will allow for important early discussions regarding methodologies and approaches that EPA is and should be considering.

- EPA should clarify how problem formulation fits into the planned scoping step and how stakeholders will be engaged for input.

- Problem formulation must include defining the causal question. It is no longer scientifically tenable to simply ask "does X pose a carcinogenic hazard?" Such an approach overemphasizes high dose toxicity studies and perpetuates an overreliance on animal results that may have little to no relevance to humans exposed at environmentally relevant levels. The IRIS Program should begin each analysis with a set of proposed hypotheses that incorporates MOA, the adverse effect(s) of concern, and the exposure level(s) of concern. The available data can then be arrayed to evaluate the extent to which existing data and knowledge does or does not support each hypothesis. In this way, the results of an IRIS analysis can be summarized and presented in a manner which illustrates to risk assessors and risk managers the extent to which each hypothesis is consistent with all of the data – human epidemiology, animal toxicity studies and modern understanding and data on mechanisms of toxicity.

- We recommend that during the scoping process EPA present its "blue print" for the IRIS assessment which should include the Agency's draft plan for conducting data acquisition, identify the objective of the assessment, and present the plan for data evaluation and analysis.

F. DRAFT HANDBOOK FOR IRIS ASSESSMENT DEVELOPMENT

In Appendix F of Part 1 of the EPA Submission the Agency provides a "Draft Handbook for IRIS Development" that includes a discussion of literature search strategies, evaluation and display of individual studies, evaluating data quality, evaluating and integrating evidence, and dose-response analysis. In the comments that follow we will address each of the areas and offer suggestions for improvement.

1. Literature Search Strategies

Appendix F of Part 1 of the EPA Submission and Example 1 of Part 2 contains a description of how EPA will identify and select pertinent studies. Our suggestions on this section are provided below.

- As EPA conducts literature searches, it should engage in active consultation with outside stakeholders known to be actively engaged in research related to a particular chemical. While we applaud the systematic approach EPA is taking, the approach should not lead to a decrease in communications with stakeholders. This can be particularly important as EPA is including a step which helps to augment the database search.

- EPA should not treat critical toxicology information, including studies on absorption, distribution, metabolism and excretion (ADME) as additional resource information; ADME information should be considered to be of primary relevance. Table F-5 in the EPA Submission refers to ADME information as an additional resource. By not treating this information as critical inputs into the toxicological review, critical information to understand MOA and human relevance could be overlooked.

- More detail is needed regarding the approach EPA will use for deciding when to consider information that is kept as an additional resource. It is unclear how EPA will consider and incorporate information such as editorials, previous reviews and meta-analyses. These evaluations can be very helpful, particularly if they provide analyses which the agency can rely upon rather than conducting de novo analyses. A standard procedure for incorporating and considering this information is necessary.

- In Example 1 of Part 2, EPA notes that 667 studies were removed due to multiple reasons, including that they were reviews, commentaries or risk assessments. However in describing the approach, EPA notes that some of these may be used later as additional resources. EPA should describe the systematic approach that will be used to consider this literature to avoid the perception of later "cherry picking" these additional studies for consideration.

2. Evaluation and Display of Individual Studies

In Part 1, Appendix F and Part 2 of the EPA Submission, the Agency demonstrates the evaluation and display of individual studies. While EPA states that changes to the evaluation and display of studies are fully implemented, further improvements are necessary. Our comments and suggestions are below:

a. Comments on Appendix F: Pages F-21 to F-38

- In some epidemiological studies, cut-points between exposure groups are chosen based on statistical convenience rather than biological relevance. The challenges of exposure classification are well recognized and need to be explicitly addressed. These include: (1) individuals within a single quantile are assumed to be homogenous and the choice of cut-points has the potential to produce both false positives and false negatives; and (2) both reduction in numbers within exposure groups and misclassification due to poorly chosen cut-points diminish the power of the study.

- Tables F-11 and F-12 should clearly describe the effect that is being measured and include a space for comments and additional information.

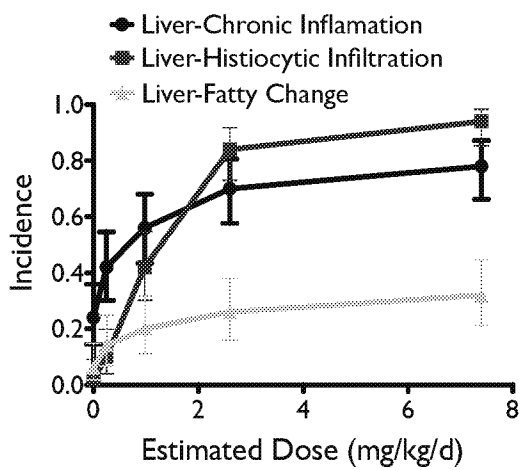
- On page F-38, EPA indicates that results of continuous responses should be presented as percent change only. This is inappropriate because it will prevent readers from knowing whether the observed results were in the normal range. In addition to percent change, results from continuous responses should be provided as means, standard deviations, and numbers (e.g., 6.2+ 2.8 (n=10)). EPA should also include what is considered "normal" range in any table or description, if EPA, in certain summary tables, elects to present results only as a percent change.

- On page F-38, EPA states that these values should be presented as converted doses. The tables assembled by EPA should also include any dose metrics estimated with a PBPK model or measured as an internal biomarker.

- On page F-38, EPA notes that table footnotes should state when EPA performs statistical analyses. EPA should also provide the methods used. For all but the simplest statistical tests (e.g., Student's t-test), details of the test and calculation should be provided, possibly in an appendix.

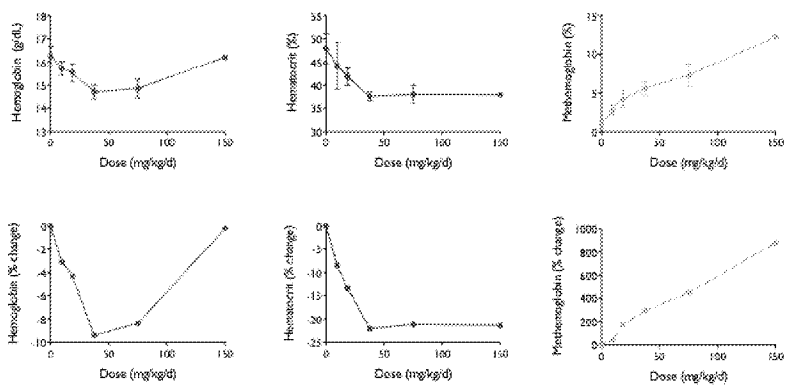
- Two examples of what improved Tables F-11 and F-12 could look like are provided below, as Illustrative Example Table 1 and Illustrative Example Table 2.

Example Table 1: Format for Quantal Effects with Severity Estimates
*Note this table was created for illustrative purposes only.

Table 1—Chronic Studies—Noncancer Effects – Page 1				
Study Design and Reference	**Sex**	**Results**	**Confidence**	**Comment**
Two year corn oil gavage study in F344/N female rats Estimated Doses from Weekly Administration: 0, 0.25, 1, 2.5, 7.5 mg/kg-d Lifetime Average Liver Conc. From interim and terminal sacrifices: 0, 0.01, 0.05, 0.75, 3 µg/kg	F	**Liver** *Chronic Inflammation:* Frequency (avg. severity) 12/50 (1.3), 21/50* (1.2), 28/50** (1.3), 35/50** (1.6), 39/50** (2.1) *Eosinophilic foci:* 1/50 (1.0), 5/50 (1.0), 21/50** (1.3), 42/50** (2.0), 47/50** (2.6) *Fatty change:* 3/50 (3/3), 7/50 (3.6), 10/50* (2.5), 13/50** (2.5), 16/50** (2.8)  	High	Although the biological significance of hepatic eosinophilic foci is unknown), this effect is included here for completeness. Eosinophilic foci were also observed in the abdominal lymph nodes in both mice and rats. The non-cancer effects shown here were measured along with the tumor occurrence. These effects were not used for dose-response evaluation.

		* p<0.05; ** p<0.01; Poly-3 test		

Example Table 2: Format for Continuous Effects
*Note this table was created for illustrative purposes only.

Table 2—Chronic Studies—Noncancer Effects – Page 1				
Study Design and Reference	**Sex**	**Results**	**Confidence**	**Comment**
Two year corn oil gavage study of nitrobenzene in F344/N male rats 10 per group, 1 in highest dose group Doses: 0, 9.38, 18.75, 37.5, 75, 150 mg/kg-d NTP (1983) cited in EPA (2009)	M	**Hematologic Effects** *Hemoglobin (g/dL)* 16.24±0.42, 15.73±0.29*, 15.54±0.37*, 14.72±0.30*, 14.87±0.41, 16.2 0, -3.1%, -4.3%, -9.4%, -8.4%, -0.2% *Hematocrit (%)* 48.82±3.2, 44.19±4.98, 41.84±1.88*, 37.66±0.93*, 38.08±1.96, 38 0, -8.5%, -13.4%, -22.1%, -21.2%, -21.4% *MetHb (%)* 1.13+0.58, 2.75±0.58*, 4.22±1.15*, 5.62±0.85, 7.31±1.44, 12.22 0, 43.3%, 173%, 297%, 447%, 881% *significantly different than control, calculated by study authors 	High	On a percentage basis the effects on hematocrit were considerably greater than the other two hematological parameters evaluated. The change in methemoglobin was clearly adverse and it was chosen as the basis of dose response modeling.

b. Part 2, Example 2 of the EPA Submission

- Regarding Tables 2-1 and 2-2, epidemiologic studies generally do not provide raw data and thus are not amenable to reanalysis by others. In some studies that use generalized linear models or generalized estimating equations, only the means and confidence intervals of beta coefficients are provided. In such cases, EPA should contact the study authors to obtain additional data. These data should be made publicly available to allow IRIS stakeholders to conduct independent analyses and verify EPA's conclusions. The tables are insufficient to allow for independent verification of the results.

- Regarding the use of quantiles, EPA should not accept reported cut points in blind faith; rather, the choice of cut points for quantiles should be an aspect of study quality. The issue of quantiles and cut points has been discussed with regard to Table F-10 above.

- Table 2-3, detailing animal studies, is the same as Table F-9b which provides very little information. Additional narrative in the table boxes would be helpful. Adding two more rows titled "Study Strengths" and "Study Weaknesses" would provide a space to make overall comments on quality.

c. Part 2, Example 3 of the EPA Submission

- Table 3-1 of epidemiological results is incomplete. For example, in the first row, the presentation of the study by Meeker et al. (2009) shows only values for means and confidence intervals (CIs) of beta values. The tables should also provide a discussion of whether or not these hormone changes are within the range of normal.

- The presentation of the studies by Hauser et al. (2006, 2007) shows the need for an analysis of quality and study power. We note that sperm parameters are highly dependent on abstinence time.[22] Whether and how this is potential confounding factor is addressed should be included in the tables.

[22] See, for example, WHO (2010). WHO laboratory manual for the examination and processing of human semen - 5th ed.; Carlsen, E., Petersen, J. H., Andersson, A. M. and Skakkebaek, N. E. (2004). Effects of ejaculatory frequency and season on variations in semen quality. Fertil. Steril. 82, 358-66.; Cooper, T. G., Noonan, E., von Eckardstein, S., Auger, J., Baker, H. W., Behre, H. M., Haugen, T. B., Kruger, T., Wang, C., Mbizvo, M. T. and Vogelsong, K. M. (2010). World Health Organization reference values for human semen characteristics. Hum. Reprod. Update 16, 231-45; Elzanaty, S., Malm, J. and Giwercman, A. (2005). Duration of sexual abstinence: epididymal and accessory sex gland secretions and their relationship to sperm motility. Hum. Reprod. 20, 221-25; Levitas, E., Lunenfeld, E., Weiss, N., Friger, M., Har-Vardi, I., Koifman, A. and Potashnik, G. (2005). Relationship between the duration of sexual abstinence and semen quality: analysis of 9,489 semen samples. Fertil. Steril. 83, 1680-86; and Schwartz, D., Laplanche, A., Jouannet, P. and David, G. (1979). Within-subject variability of human semen in regard to sperm count, volume, total number of spermatozoa and length of abstinence. J. Reprod. Fertil. 57, 391-95.

- Table 3-2 is a good organization for a broad overview of the study design and results. However, the presentation of results is too brief to draw any conclusions. It may be useful to include an additional column in the table that notes "general conclusions" as presented from the study authors. Additionally, for the study by Fujii et al. (2005), providing a legend indicating that the first number is the F0 dose in females and the second number is the F1 dose in females would be helpful (e.g.,"Females (F0/F1 doses in mg/kg/d)"). Although EPA expressed a preference for percent change on page F-38 of Part 1, the presentation of means and standard deviations (or SEM), and sample numbers for continuous endpoints would greatly enhance these tables. Presenting only percent change values without means and standard deviations of the results makes it difficult for external stakeholders to evaluate the studies without having to search for and obtain the publication.

3. Evaluating Data Quality

EPA discusses its implementation of improvements to evaluating and documenting study quality in multiple places. Our suggestions on this section are provided below.

- Sufficient computer resources must be devoted to external stakeholder use. In the section on documenting study quality evaluations, EPA mentions the LitCiter Lite and DistillerSR software. We attempted to search EPA's HERO database to find and evaluate these software packages. One of the choices was to export the results of the search to an EndNote compatible file. For greatest ease of use by external IRIS stakeholders, the ability to export an EndNote compatible file is very helpful; however, at times it appears that the HERO database may not have sufficient bandwidth to keep pace with demands by users. As EPA improves IRIS, concomitant improvements in computer resources should also be made. Databases EPA relies upon should be publicly accessible. A search of EPA's website for "LitCiter Lite" came up with three items. The most informative was an NCEA presentation on carbon monoxide at: http://yosemite.epa.gov/sab/sabproduct.nsf/0c26f31550b283468525766d0046e9d8/$file/ncea+presentation+to+casac+co+panel+11-16-09.pdf. The other two items were an earlier draft of this presentation and minutes from a meeting of the Board of Scientific Counselors (BOSC) held in October of 2010. The BOSC minutes and the presentation indicated the LitCiter was a tool associated with HERO and for internal EPA use only. A search of EPA's website for "DistillerSR" came up with nothing. DistillerSR is proprietary, a web-based systematic review software currently sold by Evidence Partners, Inc. in Ottawa, CA.

- EPA's examples of documentation of study evaluation are not sufficiently helpful. Table F-9b provides very little information. Additional narrative in the table boxes would be helpful. Adding two more rows titled "Study Strengths" and "Study Weaknesses" would provide a space for overall comments on quality.

- In assessing study quality, EPA may wish to consider adopting methods for study quality evaluation based on either the Newcastle-Ottawa scale [23] or that from Health Canada.[24] These references may provide additional ideas for study quality evaluation.

- Table F-7 reflects the need to develop and apply consistent data evaluation procedures to assure transparent and objective evaluation of animal toxicity studies. For the most part, a number of the elements used in existing established animal study data evaluation procedures have been included in Table F-7. ARASP's 2012 white paper on data evaluation procedures provides a comprehensive review of published approaches.[25] However, because Table F-7 lacks a consistent framework for determining the importance of the various elements within each feature and across features, it is difficult to envision how application of the table would be useful. Instead, the table would likely foster subjective and qualitative decisions that are not reproducible across reviewers and across substances over time. Therefore, as discussed in ARASP's 2012 white paper, a systematic and fully transparent approach is needed. We recommend that EPA review the ECETOC-developed enhancements to the Klimisch approach[26] and the ToxRTool and explain why one of these should not be adopted by the IRIS Program for evaluating animal toxicity studies and data.

- Although Table F-7 includes consideration of the validity of test methods within the endpoint evaluation feature, more detail should be provided on analysis of novel or new methods. Methods relied on for regulatory decision making should be scientifically valid, meaning that there needs to be the requisite degree of confidence in the sensitivity, specificity, relevance and reliability of a method for its intended purpose. When new or novel methods are used, until such time as these key attributes have been established, care should be taken in relying on results from such methods for hazard and risk assessment purposes. Furthermore, the manner in which the terms sensitivity and specificity are used in Table F-7

---

[25] Wells GA, Shea B, O'Connell D, Peterson J, Welch V, Losos M, et al. The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses. 2005. www.ohri.ca/programs/clinical_epidemiology/oxford.htm.

[24] See Health Canada. (2009). Guidance Document for Preparing a Submission for Food Health Claims available at http://www.hc-sc.gc.ca/fn-an/legislation/guide-ld/health-claims_guidance-orientation_allegations-sante-eng.php.

[25] See ARASP white paper available at http://arasp.americanchemistry.com/Data-Quality-Evaluation, which cites Klimisch, H.J., Andreae, M., Tillmann, U. (1997). A systematic approach for evaluating the quality of experimental and ecotoxicological data. Regul.Toxicol. Pharmacol. 25(1):1-5 and the ToxRTool described in Schneider et al. (2009) Toxicol. Lett. 189(2):138-44.

[26] Klimisch, H.J., Andreae, M., Tillmann, U. (1997). A systematic approach for evaluating the quality of experimental toxicological and ecotoxicological data. Regul. Toxicol. Pharmacol. 25(1):1-5; see also: ECETOC, JACC report #55 on Linear Polydimethylsiloxanes which incorporates the modified and expanded the justification phrases for each Klimisch reliability category. Available at: http://www.ecetoc.org/jacc-reports.

does not align with the way such terms are standardly used in test method development and validation (see for example, http://arasp.americanchemistry.com/Data-Quality-Evaluation). The table should be revised to be consistent with such standards.

4. Evaluating and Integrating Evidence

Appendix F of Part 1 of the EPA Submission describes how EPA will evaluate and integrate evidence. Suggestions for improving this section are provided below.

a. Utilization of MOA Information

- EPA mentions that determinations of causality involve consideration of information from all available sources, including human, animal and MOA data. However, MOA is only further mentioned in the context of helping to identify an adverse outcome pathway. MOA should be the central organizing principle in conducting hazard and risk assessments. Consistent with established best practices of systematic evidence-based reviews, EPA should employ a consistent weight of evidence framework, based on specific hypothesized MOAs to permit data from laboratory experiments, epidemiological investigations, and cutting-edge mechanistic research to be integrated in a manner that provides a robust understanding of the MOA and the potential hazards and risks that exposures to a substance could pose to humans.

- The MOA/Human Relevance framework was originally developed over a decade ago by the World Health Organization (WHO) International Program for Chemical Safety (IPCS) and specifically focused on chemical carcinogenesis. The original framework was expanded to include a human relevance component and information about the susceptibility of various lifestages and, more recently, the framework has again been enhanced to examine the key events and their dose response relationships in a systematic and quantitative fashion over the full range of responses from early events to the adverse effect of concern.[27] An understanding of the MOA is a fundamental component of risk assessment. Consideration of MOA also allows for an understanding of potentially susceptible human subgroups and different life stages so that the most appropriate adjustments can be factored into quantitative risk assessments. The human relevance of a hypothesized MOA may depend on both qualitative and quantitative factors and can be addressed by examination of the human relevance of each key event in the proposed MOA. EPA's Office of Pesticide Programs uses this approach and assesses both qualitative and quantitative concordance of

[27] See Julien E., Boobis A.R., and Olin S.S. (2009). The Key Events Dose-Response Framework: a cross-disciplinary mode-of-action based approach to examining dose-response and thresholds. Crit. Rev. Food Sci. Nutr. 49, 682-89.

key events between animals and humans.[28] For example, in the early 1990s, a technical panel from EPA concluded that male rat renal tubule tumors from chemicals that induced accumulation of α2u-Globulin were likely not relevant to humans based on qualitative considerations.[29]

- The development of a proposed or hypothesized MOA will necessitate identification of key events and understanding the dose-response and temporal relationships between the various key events and the adverse outcome as well as between the key events themselves. A dose-time concordance table can help to address the temporal aspects of the MOA. We have provided an example below. Please see Illustrative Example Table 3.

- The development of a dose-response concordance table will also be helpful. This table can provide information about both qualitative and quantitative concordance of key events between animals and humans and quantitative dose-response information in both animals and humans. We have provided an example below. Please see Illustrative Example Table 4.

- Quantitative examination of both the dose-response and timing of key events is also necessary to determine human relevance. For example, an MOA may be operative in both animals and humans, but extremely unlikely in humans because of quantitative toxicokinetic or toxicodynamic differences. If the key event has the potential to occur in humans, then this quantitative examination can be used to inform animal-to-human extrapolation. Hence, the quantitative concordance should provide information about the EC50 and/or point-of-departure values for as many key events as possible in both humans and the animal test species.

- Human relevance of the apical endpoint can be determined using a hypothesis based weight-of-evidence approach.[30] To address human relevance of the MOA, qualitative concordance between humans and animals for each key event is considered. If available, *in vitro* data from human or animal cells or tissues and/or *in silico* data should be considered as well. Ideally, the data will be sufficient to determine which of the key events is relevant to humans, and these data may thus be used to support statements about the relevance to humans of the hypothesized MOA in animals.

---

[28] See Dellarco, V., and Fenner-Crisp, P.A. (2012). Mode of Action: Moving toward a More Relevant and Efficient Assessment Paradigm. J. Nutr. 142, 2192S-8S.
[29] See Rodgers, I.S. and Baetcke, K.P. (1993). Interpretation of male rat renal tubule tumors. Environ. Health Perspect. 101 Suppl 645-52.
[30] See Rhomberg, L.R., Bailey, L.A., and Goodman, J.E. (2010). Hypothesis-based weight of evidence: a tool for evaluating and communicating uncertainties and inconsistencies in the large body of evidence in proposing a carcinogenic mode of action--naphthalene as an example. Crit. Rev. Toxicol. 40, 671-96.

- The discussion regarding considerations of consistency would benefit from noting that MOA information can be particularly helpful in this situation.

- EPA's use of MOA data should be improved. EPA's current approach is to assert the default and then require that "sufficient" data be developed to refute the default. "Sufficient data" is never defined, and seems to be an ever moving target. This has led to an impasse that is not sustainable and, moreover, undermines research focused on applying modern techniques, such as knock-out models, to improve the scientific evaluation of specific hypothesis as part of determining relevant MOAs. Instead of trying to ask and answer the question of "how much data and knowledge is enough to overrule a default?" what is needed is a framework that uses all of the relevant and reliable data and knowledge of hypothesized MOAs in an open, objective and transparent manner, including evaluation of the hypothesized MOA underlying the default.

- Additionally, it is no longer scientifically tenable to simply ask "does X pose a carcinogenic hazard?" Such an approach overemphasizes high dose toxicity studies and perpetuates an overreliance on animal results that may have little to no relevance to humans exposed at environmentally relevant levels. The IRIS Program should begin an analysis with a set of proposed hypotheses that incorporates MOA, the adverse effect(s) of concern, and the exposure level(s) of concern. For example, in evaluating carcinogenicity, the two hypotheses below would be examples of a starting point for IRIS:

  - Mutagenic MOA Hypothesis: Chemical X causes cancer by a non-threshold MOA by causing somatic mutations in target cells of the organ Y at doses below Z.

  - Threshold MOA Hypothesis: Chemical X causes cancer by a threshold MOA by causing cytotoxicity at certain doses in target cells of organ Y, leading to compensatory cell proliferation at doses below Z.

  The available data can then be arrayed to evaluate the extent to which existing data and knowledge does or does not support each hypothesis. In this way the results of an IRIS analysis can be summarized and presented in a manner which illustrates to risk assessors and risk managers the extent to which each hypothesis is consistent with all of the data – human epidemiology, animal toxicity studies and modern understanding and data on mechanisms of toxicity. See Rhomberg et al for suggested approaches.[31]

[31] Rhomberg, L.R., Bailey, L.A., Goodman, J.E. (2010). Hypothesis-based weight of evidence: A tool for evaluating and communicating uncertainties and inconsistencies in the large body of evidence in proposing a carcinogenic mode of action – Naphthalene as an example. Crit. Rev. Toxicol. 40:671-696;

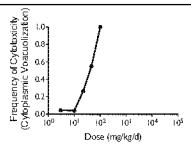
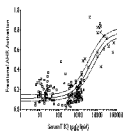
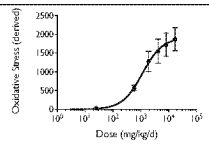
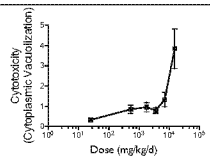
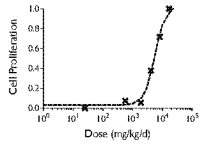
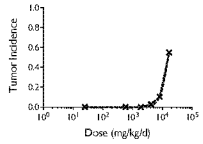
- EPA's guidance regarding mechanistic considerations should be further clarified. The discussion on pages F-51 through F-52, lack details describing how proposed key events will be analyzed for dose response, timing and species concordance. Without such considerations, the ability to provide proper consideration of the potential adverse outcome pathways (AOPs) or proposed MOAs would be limited.

Example Table 3: Dose-Time Concordance Table
*Note, this table was created for illustrative purposes only.

Table 3 —Dose-Time Concordance				
Time		**1 month**	**6 months**	**2 years**
Increasing Dose	**Increasing Time** →			
	1		Absorption	No data
	3	Absorption	Absorption Oxidative Stress	Absorption Oxidative Stress Cytotoxicity Proliferation
	20	Absorption Oxidative Stress	Absorption Oxidative Stress Cytotoxicity	Absorption Oxidative Stress Cytotoxicity Proliferation **Tumors**
	50	Absorption Oxidative Stress Cytotoxicity	Absorption Oxidative Stress Cytotoxicity Proliferation	Absorption Oxidative Stress Cytotoxicity Proliferation **Tumors**
	100	Absorption Oxidative Stress Cytotoxicity Proliferation	Absorption Oxidative Stress Cytotoxicity Proliferation	Absorption Oxidative Stress Cytotoxicity Proliferation **Tumors**

Rhomberg, L.R., Bailey, L.A., Goodman, J.E., Hamade, A., Mayfield, D. (2011). Is exposure to formaldehyde in air causally associated with leukemia? – A hypothesis-based weight-of-evidence analysis. Crit. Rev. Toxicol. 41(7):555-621.

Example Table 4: Format for Dose-Response Species Concordance Table for Assessing MOA/Human Relevance
*Note, this table was created for illustrative purposes only.

EVENT OR FACTOR	QUALITATIVE CONCORDANCE			QUANTITATIVE CONCORDANCE AND QUANTITATIVE DOSE-RESPONSE		
	Animals	Humans	Concordance	Strength	Animals	Humans
KEY EVENTS						
Key Event #1 Absorption	Occurs in animals in vivo	Also occurs in humans in vivo	Humans are less sensitive than animals	+++		
Key Event #2 Oxidative Stress	Occurs in animals in vivo	Not measured in humans	Unknown			NA
Key Event #3 Cytotoxicity	Observed in pathology examinations of animals	Not measured in humans	Unknown			NA
Key Event #4 Cell Proliferation	Measured in animals with BrdU labeling for DNA synthesis	Not measured in humans	Unknown			NA
Apical Event Liver Tumors	Known to occur in animals in vivo	Has not been observed in humans	No concordance			NA

b. Synthesis of epidemiology evidence

- This subsection of Appendix F is titled; "Evaluating the Overall Evidence of Each Effect." This title presumes that an "effect" exists. A better title would be: "Evaluating the Overall Evidence" or "Evaluating the Overall Evidence for Hazard Characterization."

- EPA relegates meta-analyses to an "other" category along with reviews, editorials, risk assessments, etc. (see Table F5, p. F-11). This is not consistent with current scientific practice.[32]

- EPA incorrectly elevates case reports to the status of "studies." See, for example, p. F-39, line 19. Case reports and case series generate (rather than test) causal hypotheses about the relationship between exposures and diseases (Holland 1986).[33] Case reports, therefore, only provide clues to etiology (Vandenbroucke, 1999)[34] but should not be used to inform causal inferences. At best, case reports are merely careful observations of events that provide information suggesting that scientific studies, such as analytical epidemiological studies, should be undertaken. There are some very limited exceptions where case reports might inform causal assessments (i.e., very rare, specific outcomes such as mesothelioma, angiosarcoma, or acute conditions where symptom onset is almost immediately after exposure), but these are the exception rather than the rule. There is extensive literature on the methodological problems of case reports that should be considered. Generally, making claims about causality is not recommended when case reports and case series are used as the primary evidentiary source (Jick, 1977 Venning, 1983).[35,36]

- EPA states on page F-40 (line 18) that each study is considered part of the weight of evidence evaluation. This implies that even studies of extremely poor quality will remain within the evaluation of the evidence. Unless there is some process for stratifying the overall evaluation by study quality, this seems inappropriate.

- In assessing aspects of causality, EPA provides a highly selective group of references which does not reflect the full body of historical or current thinking on the Hill "criteria. For instance, EPA cites Hill (1965) but not Hill (1971); Rothman and Greenland (1998) but not MacMahon and Pugh (1970), Mausner and Bahn (1974), Kleinbaum et al. (1982),

---

[32] Aschengrau, A. and Seage, G.R. (2003). The Epidemiologic Approach to Causation. Essentials of Epidemiology in Public Health, 375-401; Bhopal, R. (2005). Cause and effect: the epidemiological approach. Concepts of Epidemiology: An integrated introduction to the ideas, theories, principles and methods of epidemiology, 98-132; Goodman, S.N. and Samet, J.M. (2006). Cause and Cancer Epidemiology. Cancer Epidemiology and Prevention, 3-9; Gordis, L. (2000). From Association to Causation: Deriving Inferences from Epidemiologic Studies. Epidemiology, 184-203.
[33] Holland, P.W. (1986). Statistics and Causal Inference. J. Am. Stat. Assoc., 81(396):945-60.
[34] Vandenbroucke, J.P. (1999). Case reports in an evidence-based world. J. R. Soc. Med., 92(4) 159-62.:
[35] Jick, H. (1977). The Discovery of Drug-Induced Illness. N. Engl. J. Med., 296(9):481-85.
[36] Venning, G.R. (1983). Identification of adverse reactions to new drugs. III: Altering processes and early warning systems. B.M.J., 286:458-60.

Rothman (1986), Beaglehole et al. (1993), Weed (1995), Weed (2000), Vetter and Matthews (1999), Gordis (2000), Rothman (2002), Aschengrau and Seage (2003), or Goodman and Samet (2006). Hill (1971) makes it clear that before these criteria can be "applied" or "considered" there must be evidence of a statistically significant association observed in epidemiology studies. Observing a statistically significant association in an epidemiological study is the only way "chance can be excluded," the first requirement of Hill's approach (1965).

- EPA's discussion of "strength of association" is unclear with regards to the magnitude of the relative risk estimate and the role of chance and bias. To improve clarity, we suggest the language below. EPA also states that "an association of small magnitude (due to factors such as low potency or a low level of exposure) … could lead to a significant public health burden…" Bringing public health burden into a discussion of causality is inappropriate because it infuses values, bias and risk management considerations into what needs to be an objective determination of the overall scientific evidence. The goal is to determine causality, not public health burden. At a minimum, EPA should also have noted that a large relative risk could lead to an insignificant public health burden. For lines 31-37 on page F-40, we suggest the following language:

    - "Strength of Association: refers to the magnitude of the relative risk estimates observed in the epidemiology studies. Typically, the larger the relative risk (RR), the more likely the observed association is causal. Relative risk estimates can be obtained from well-designed cohort and case-control studies comparing the incidence of a condition in those exposed to the putative cause to the incidence of the same condition in those unexposed. Small magnitudes of association (sometimes called "weak" or "modest" associations), e.g., RRs of 2.0 or less, are less likely to represent causal associations in that bias (due especially to uncontrolled confounding) can explain the presence of weak associations."

- In describing consistency of an association EPA incorrectly states "Observing an association in different study types, study populations and exposure scenarios makes it less likely that the association is due to confounding or other factors…." The assessment of confounding can be assessed by examining the extent to which the studies at issue are controlled for known confounders. The assessment of uncontrolled confounding was discussed above in connection with the assessment of strength. Many discussions of the consistency criterion emphasize that the same bias can affect all published studies. EPA should include a discussion of this issue. In addition, EPA's further discussion of consistency seems to ignore the relationship between consistency and meta-analysis (see Weed, 2000).[37] For the paragraph discussing consistency, we suggest replacement with the following language:

[37] Weed, D.L. (2000). Interpreting epidemiological evidence: how meta-analysis and causal inference methods are related. Int.l J. Epidemiol., 29:387-90.

- "Consistency of Association: refers to the extent to which scientific results are similar (e.g., in direction and statistical significance) across the entire body of epidemiological evidence. Typically, the more consistent the results the more likely the observed association is to be considered causal. One of the additional values of meta-analysis is that it provides a quantitative assessment of consistency (Weed, 2000) through tests of heterogeneity; studies homogeneous enough to be combined are, by definition, consistent. Even apparently consistent studies can have the same bias affecting all known studies."

- In the discussion of specificity, it may be helpful to mention that specificity can be used to better characterize the exposure and the outcome. For example, studies of toxaphene exposure and non-Hodgkin lymphoma have more specific measures of exposures and outcomes than, say, studies of pesticides (unspecified) and lymphoma (type unspecified).

- In the discussion of biological gradient or exposure-response relationship, EPA discusses "piecing together evidence" and notes that "a lack of response in any one study does not imply a lack of an association" (page F-41, at lines 28-35). This discussion raises two concerns. First, there is no precedent in the methodological literature for "piecing together" a dose response relationship from more than one study outside the context of meta-analysis. Second, EPA's claim that the lack of a dose-response relationship in one study does not imply a lack of association makes little sense. EPA seems to be comparing apples to oranges. After all, the premise of the application of the Hill considerations is that an association exists.

- When discussing biologic plausibility, coherence, and analogy, EPA seems to link these three items and does not describe the complexities associated with each one. On the topic of complexity, "biologic plausibility" refers to the body of scientific evidence from toxicology and other biological sciences to determine (or not) the existence of a mechanism (or MOA). Coherence, on the other hand, refers to the overall "fit" of the evidence (both epidemiologic and biologic). Finally, analogy refers to the extent to which the evidence for the observed association is similar (or not) to the evidence for a known causal association. These are three distinct and complex considerations. We suggest the following replacement language for this paragraph:

  - "Biologic Plausibility: refers to the extent to which a mechanism of action has been proposed, studied, and demonstrated, typically in toxicological and other types of laboratory-based studies. It is generally accepted that as the evidence explaining the mechanism of action for a disease increases, the more likely the association is causal. A disease mechanism has many features, including but not limited to the many intracellular and extracellular changes that occur from the initiating causal event (e.g., an exposure or some unknown "idiopathic" event) to the subsequent disease event. Indeed, latency (discussed briefly above) can be considered one of many features of a disease mechanism. Assessing biological plausibility also involves distinguishing between what happens in humans and what happens in animals. Although animal testing (also called animal bioassay

testing) has been used for many years as a component of assessing biological plausibility, its relevance to human health is under some scrutiny in the scientific community. The primary concern has always been the extent to which the results of animal testing can be extrapolated to humans. Animal testing typically involves exposing rodents (rats, mice, and hamsters) to excessive doses of the chemical of interest to observe whether these same animals subsequently develop disease (e.g., cancer). The evidentiary concerns about animal testing, however, include the following considerations: (1) that animals are exposed to doses (and durations) that far exceed human exposure conditions, (2) the mechanisms of action in animals are not those found in humans, and (3) the physiology of rodents (and their metabolic pathways) may be different than those in humans.

- Coherence: refers to the extent to which the evidence and hypotheses for the results fit together into a reasonable and well-tested explanation. In the classic description of this so-called criterion, coherence was defined as the extent to which the causal hypothesis does not conflict with the available evidence. Coherence can be assessed in terms of the extent to which other causal criteria (or "guidelines") have been met. The more criteria that are satisfied, the more coherent the causal explanation.

- Analogy: the extent to which the purported exposure-disease relationship under consideration is similar (in types and characteristics of evidence) to other relationships, known to be causal or not.

- In discussing natural experiments, EPA states that natural experiments can "mimic" a controlled experiment or randomized trial. "Natural experiments" cannot control for unknown confounders. "Natural experiments" do not typically have protocols like those used in randomized controlled clinical trials.

- In discussing alternative explanations for observed epidemiologic associations, EPA asserts that confounding and bias can be discounted if the epidemiologic evidence is consistent or if a dose-response relationship is observed. There is no precedent in the scientific literature for using Hill's considerations as a way to determine if bias and confounding exist in a study or set of studies. EPA should remove this assertion.

- While EPA does not state that it will use specific descriptors to describe epidemiology evidence, it does provide some descriptors that are likely under consideration. We have significant concerns with these descriptors and strongly recommend that any proposed descriptors undergo public comment and peer review before being utilized in an assessment. We highlight some additional concerns below:

  - For sufficient evidence of causation, EPA requires: 1) Alternative explanations (confounding, information bias and selection bias) are judged to be unlikely; 2) Evidence of consistency and evidence of a dose-response relationship; 3) Evidence of a "relatively strong association" with the caveat that a weak

association (one small in magnitude) "may not diminish" the judgment; and 4) Evidence of a coherent temporal relationship allowing for latency but "absence of such information does not necessarily detract from the conclusion." These requirements are inconsistent with good methodological practice. Weak associations must "diminish" the judgment given that weak associations increase the likelihood of unknown confounding. The requirements for this category should be revised. In addition, EPA introduces latency without any discussion of how this consideration is evaluated.

- EPA must provide the criteria and/or conditions which are necessary to determine something as suggestive evidence of causation. The purely subjective nature of this category, lacking any conditions, is problematic.

- EPA must provide the criteria and/or conditions to determine something as inadequate evidence to infer causation. The purely subjective nature of this category, lacking any conditions, is problematic.

- EPA must provide the criteria and/or conditions which are necessary to determine that evidence is consistent to illustrate no association. EPA has inappropriately applied a circular definition here and more clarity is necessary.

c. Synthesis of animal toxicology evidence

This section (pages F-45 to F-52) indicates that mechanistic data is useful for informing the plausibility of a causal interpretation in humans and that animal data is generalizable to humans and to the susceptibility of certain populations or lifestages. The section treats mechanistic or MOA as desirable but not quite attainable. Unfortunately, it inappropriately treats MOA data as important for toxicity value calculations but not for hazard assessment.

- EPA has used the term Mode of Action (MOA) as the description of the key events leading to a toxic endpoint. This section however introduces the term Adverse Outcome Pathway (AOP) without a definition. The document implies that this is different from MOA, but the logic of this is unclear. The document implies that ADME data are separate from MOA data, but often ADME processes are part of the key events. If there is a distinction between MOA and AOP, it should be explained.

- MOA data should be part of synthesizing animal toxicity evidence. The section describes how to "Synthesize Animal Toxicology Evidence," then presents an example. Then as a separate section includes "Mechanistic considerations in elucidating AOPs," as if it is not part of animal toxicology evidence. This section discusses ADME and toxicodynamic processes, but gives the impression that such data seldom are useful.

- A major role for MOA data is to understand what is happening in the animal model. It is an integral part of synthesizing the animal toxicology evidence. There is no bright line between toxicity data and mechanistic data. In a necropsy one may observe a toxicity

outcome of "enlarged liver;" histopathologic evaluation may reveal mechanistic data on whether this is hypertrophy of hepatocytes, hyperplasia, or some other finding. The point is that much of the guideline toxicity studies involve mechanistic data. In some cases, non-guideline studies are conducted to provide further mechanistic data. However, mechanistic data is part of the animal toxicity evidence and needs to be included in synthesizing the animal evidence.

- The section says that mechanistic data provide information on how a chemical may disrupt normal biological processes. Data are emerging to indicate that smaller amounts of a chemical generally perturb cells, but they compensate by homeostatic processes. Toxicity occurs when the cell is no longer able to compensate. The mechanistic section incorrectly implies that any difference in a cell is toxicity. Further clarification and guidance is necessary to differentiate cell perturbations from toxicity.

- On page F-47, in describing how to evaluate why a one study is negative and another is positive, EPA provides questions to ask to examine the negative study to determine if it was adequately designed to evaluate the endpoint. Unfortunately, EPA does not mention that the positive study should also be examined to determine if the study design caused the apparent result. The above illustrates a pervasive bias throughout the handbook towards more emphasis and acceptance of positive studies and exclusion and discounting of negative results. A critical examination of both positive and negative studies is equally necessary.

- On page F-48, EPA encourages the use of imprecise terms, such as "demonstrates," "indicates," and "suggests" which are subject to personal interpretation. If EPA is going to recommend such terms, clear guidance on their usage is necessary.

- On page F-51, EPA states that that there may be insufficient data to establish ADME of compound and/or MOA leading to adverse effect due to lack of data. Users of this handbook should evaluate the data sets incorporating all the knowledge available (including all data and information) without any preconceived notions that the data will be inadequate.

d. Part II, Example 4 of EPA Submission

In Part 2 of the EPA Submission to NAS, Example 4 provides an evaluation of the evidence for Hodgkin lymphoma (HL) due to formaldehyde exposure as an example of the integration of evidence from epidemiological studies. Unfortunately, the discussion offers very little in the way of integration of the human data. Instead, it provides a summary of the results of various epidemiology studies, without critical evaluation, and provides little insight into how EPA plans to integrate human and mechanistic data into its assessments. We understand that the discussion is preliminary and that it would be inappropriate to include draft conclusions as to the significance of the human evidence for HL from formaldehyde exposure. Similarly, the discussion provided is only a small part of a larger assessment document and, therefore, does not include relevant information from other sections of the document. Without these

pieces of information, however, it is difficult to assess whether the Agency's proposed approach leads to a full integration of the data and how such integration is incorporated into the conclusion regarding causality. From the information provided, however, it is possible to make a number of observations.

- The discussion provides no evidence of a critical review of the quality of the studies that are included. The evidence tables include basic descriptions of the study, exposures, and results without an assessment of their relevance to EPA's evaluation. It is not clear how the studies were selected for inclusion. For instance, a major study of embalmers[38] is omitted despite the fact that, while no statistical analysis is presented, the researchers provide data and a qualitative assessment of HL incidence. Other studies are included in the analysis where the number of reported cases of HL is too low to provide statistical power. The discussion also fails to address the greater uncertainty in exposure estimates in most of the studies assessed. In many of the studies, exposure is inferred from job title or work area. A critical assessment of quality and relevance must be made otherwise those that read, review and use evidence tables will mistakenly think that each study should be treated equally. This misperception must be corrected.

- In the absence of a critical assessment of study quality and relevance, Example 4 applies a subset of the Bradford-Hill considerations to the 13 studies included in the analysis. Much of the discussion focuses on two studies (Beane Freeman et al., 2009; Coggon, 2003) with a particular focus on the only study reporting an elevated risk of HL death. Many of the concerns discussed above regarding Part 1 of the EPA Submission also can be applied to the Bradford-Hill discussion in this example. Among the key concerns is whether it is appropriate to apply the Bradford-Hill considerations to this particular set of data since only one study reports a statistically significant association. Other key concerns are described below.

- In discussing consistency of the observed association, this section reviews the information for all 13 studies, rather than focusing on consistency within and among the three or four studies of highest quality. Because of the vast differences in the assessment of a quantitative association between exposure and outcome in the four highest quality studies, appropriately addressing the issue of consistency of effect is a critical first step in the assessment of causality. Although it also may be addressed in the discussion of exposure-response, it may be useful to discuss the consistency of the association among the various exposure metrics within the study by Beane Freeman et al. (2009).

- In discussing strength of the observed association, the focus is on only one study. In the case of the study by Beane Freeman et al. (2009), the majority of the magnitudes presented would be considered weak or modest, and less likely to be suggestive of a casual association.

[38] See Hauptmann, M. et al. (2009). Mortality From Lymphohematopoietic Malignancies and Brain Cancer Among Embalmers Exposed to Formaldehyde. JNCI 101(24):1696-1708.

- In discussing the temporal relationship of the observed association, EPA again focuses on just one study. Table 4-1 indicates, however, that at least one other study evaluated duration and time since first exposure (Pinkerton et al., 2004).
- In discussing exposure-response relationships, the discussion makes no attempt to understand the differences in results among the various exposure metrics. For example, it is important to understand whether one should expect to see the highest risk ratios using the peak exposure metric when more traditional exposure metrics (e.g., cumulative exposure) produce weaker associations. The association with peak exposure in the study by Beane Freeman et al. (2009) provides additional rationale for consideration of the study by Hauptmann et al. (2009) of embalmers where peak exposures appear to have been significantly higher. In addition, the study by Hauptmann et al. (2009) analyzed the incidence of lymphohematopoietic (LHP) malignancies based on the duration of employment and, while the number of HL cases is small, reports no association with longer employment. Thus, the limited focus of the EPA evaluation is unclear.

- EPA provides no discussion in the section on biologic plausibility. This is a critical component that is missing from this example. Without understanding how the available mechanistic data are to be incorporated into the assessment in practice, it is impossible to comment on the robustness of the process that EPA proposes to use. The more evidence EPA has to support the MOA for a disease, the greater confidence the Agency can have in the existence of a causal association. In the case of HL, the MOA is unclear. It is critical that EPA discuss whether data exist to support the suggestion of a causal association. In a more general sense, without an example, we cannot provide comments on the critical element of how EPA plans to consider biologic plausibility when integrating evidence.

5. Dose-Response Analysis

Selecting Studies for Derivation of Toxicity Values

EPA discusses its study selection process for dose-response analysis in Part 1, Appendix F as well as in Part 2, Example 5. Generally, EPA's preference for human data should be clarified (or caveated). For selecting studies, EPA states that "human data are preferred to reduce interspecies extrapolation uncertainties." The statement implies that the uncertainties in human data are fewer than the uncertainties associated with interspecies extrapolation. In many instances, there may be equal or larger uncertainties present in human data. As such, the above statement should be further explained to include specifically what kind of human data could potentially reduce interspecies extrapolation uncertainties. As it reads, there is an impression that human data (regardless of quality) is superior to animal data. While EPA states that its approach for selecting studies for dose response analysis is fully implemented, further improvements are necessary. In addition to the comments above, we provide specific comments below.

a. <u>Appendix F, Table F-13 of Part 1 and Example 5 of Part 2 of the EPA Submission</u>

- This table is quite comprehensive, however it does not provide sufficient information regarding whether a study can inform MOA. For human studies, measurement of one or more biomarkers of either exposure or effect might inform MOA; for animal studies, particularly chronic bioassays, interim sacrifices with biochemical and histopathological data are often very helpful in the consideration and elucidation of a proposed MOA.

- The evaluation of data section preceding this table emphasizes biologic significance over statistical significance and the need to understand the biology/toxicology of events. Footnote 2 should be removed. Table F-13 is used to evaluate studies for derivation of toxicity values (NOAELs or BMDLs) and footnote 2 does not make sense in those terms. Results within a study that are not statistically significantly different from control can often figure into BMDL and related calculations for deriving a toxicity value. However, in no case should assessors rely on a study where none of the dose groups is significantly different from the control.

- Table F-13 provides a citation to Hoenig and Helsey (2001). The title of this article is: "The Abuse of Power: The Pervasive Fallacy of Power Calculations for Data Analysis." As this paper addresses clinical trials, it is not clear that its use is applicable to the evaluation of observational epidemiologic studies. In addition, this paper and the similar paper by Bland (2009),[39] advocate for the use of additional information to assess whether the observed difference is a true difference or a random effect due to small samples—an essentially Bayesian approach that uses the heuristic value of knowledge obtained elsewhere.

- Example 5 of Part 2 provides information regarding how the most suitable studies and the studies judged less suitable were identified and included in the toxicity value derivation. While the "Draft assessment text" section provides an overview the reasons why some studies were not considered in derivation, this section should also provide the number of studies that were excluded from consideration based on each of the criteria/reasons noted for exclusion.

b. <u>Part 1, Appendix F (pages F-56 to F-60)</u>

- The advantages and disadvantages of additional BMD modeling software should be considered. The data management section for dose-response modeling includes a discussion of internal database development and quality control measures. Software (e.g., Excel add-in) for housing dose-response data, data manipulations, and dose-response modeling results is also discussed. In addition, it is stated that these add-ins, such as BMDS Wizard and DRAGON, are supposed to help automate model selection. These database management systems were developed, in part, by outside contractors. While they may be helpful, they are not required and, in fact, are separate from the

[39] See Bland, J.M. (2009) The tyranny of power: is there a better way to calculate sample size? Br. Med. J. 339:1133-35.

BMDS program. One concern is that in addition to maintaining the BMDS software, there will become a constellation of BMDS-related programs that will have to work correctly together, as well as with Excel and/or ACCESS. Whether the benefits of these additional software packages outweigh their upkeep should be considered. Moreover, if there is some automated decision/logic to model selection offered by these tools, it would be ideal for them to be incorporated directly into the BMDS program. Our understanding is that WIZARD and DRAGON are works-in progress. Further understanding of these programs by EPA and stakeholders is needed. EPA should consider a public workshop to discuss these data management tools.

- More guidance is needed regarding the development of composite toxicity values. The section on considerations for selecting organ or system-specific toxicity values contains standard considerations for developing multiple toxicity values prior to selecting a final value. The section then concludes that a tissue or system-specific toxicity value can be based either on a single study, or by deriving a "composite value supported by multiple candidate toxicity values that protects against toxicity *in the given organ or system*" (emphasis added). An example of the 2011 trichloroethylene (TCE) assessment[40] is cited (in the body of the document), where the 2011 TCE assessment "identified multiple candidate RfDs that fell within a narrow dose range, and selected an overall RfD that reflected the midpoint among the similar candidate RfDs." More EPA guidance on this approach would be useful.

c. Part 2, Example 6

In Part 2 of the EPA Submission, Example 6 demonstrates the presentation of dose-response modeling output. We independently verified all modeling results by conducting BMD modeling of the sample datasets. Our comments and suggestions are below:

- The document does not fully reflect the complexity (and latitude) in BMDL selection. The majority of this section is scientifically defensible, with the exception of the position EPA takes with regards to model selection. It is stated that when multiple models provide BMDL values within a three-fold range, the BMDL from the model with the lowest Akaike information criterion (AIC) should be selected; but if the BMDLs differ by more than three-fold, then model with the lowest BMDL (not lowest AIC) should be selected. This ignores that BMD modeling is essentially a curve fitting exercise. Thus, the model with the lowest AIC should usually be selected unless there are other obvious reasons not to select that value. The fact that a BMDL may be greater than three-fold lower than other BMDLs is not sufficient justification for selecting the lower BMDL and is not consistent with guidance in the EPA BMD Technical Guidance, which offers flexibility with respect to selection of the BMDL. In fact, in Section 2.3.9 of the BMD

[40] Toxicological Review of Trichloroethylene(CAS No. 79-01-6) In Support of Summary Information on the Integrated Risk Information System (IRIS), September 2011. http://epa.gov/iris/toxreviews/0199tr/0199tr.pdf

Technical Guidance, EPA states that in situations where the BMDL values are not in a sufficiently narrow range, "the lowest BMDL *may* be selected" (emphasis added). The Technical Guidance also indicates that when multiple models share the lowest AIC, the BMDLs can be averaged. Furthermore, it states that a more complex process of weighted "model averaging" continues to be explored. As such Appendix F should be revised to better reflect the flexibility that the BMD Technical Guidance provides.

- The sample dataset does not reflect the complexity of BMDL selection. In one of the quantal noncancer datasets (Rotorod), the BMDL selected was 93.9 mg/m^3, whereas the four other models (with good fits) ranged from 129-233 mg/m^3. The rationale cited for selecting 93.9 mg/m^3 was that the range of values exceeded three-fold. In fact, fold differences ranged from 1.4-3.6 and, based on the average of the other four models, the difference was less than three-fold (2.7 specifically). A better rationale for selection of 93.9 mg/m^3 is that the model that provided this value actually had the lowest AIC – implying that it indeed had the better fit to the data. However, it is worth considering that the four other good models provided similar results to one another, which might suggest that the higher BMDL is indeed more scientifically justified. For example, it is conceivable that BMDL values for similar central nervous system endpoints are also higher and thus provide insight for evaluating which BMDL to select. This specific dataset highlights the complexity that sometimes occurs in selecting a BMDL value, and that it may be more appropriate to use this example as a case to say that the selection of this particular BMDL might best be determined on an *ad hoc* basis. At the very least, choosing the best fitting model (i.e., lowest AIC) provides a better rationale for selecting the BMDL than simply choosing the lowest BMDL.

- There does not appear to be scientific support for selecting the lowest BMDL when the BMDL values ranged greater than three-fold. The following example from the BMD Technical Guidance implies, at best, that four- or five-fold might be an appropriate cutoff: "Which of the three acceptable models should be used as a basis for a BMD and BMDL? In this case, the BMDLs range about fourfold, from 1.7 to 5.2. Depending on the needs of the application, the BMDLs may not be considered sufficiently close. For risk assessment purposes, for example, the range is large enough that the model with the lowest BMDL would be considered preferable, as a reasonable conservative estimate."

e. Considerations for Selecting Organ/System-Specific or Overall Toxicity Values

There is a limited discussion of this topic in Appendix F (page F-60), as well Part 2 of the EPA Submission in Example 7. Although providing multiple toxicity values can be informative and useful, it is critically important that the values be clearly and transparently presented and communicated to the public. For instance, while there may be a toxicity value developed for multiple endpoints, it is not clear that the development of the value implies that the endpoint is of concern. Thus, there may be cases where a value is developed for developmental effects and it is an order of magnitude above the critical effect, which may be for liver effects. By developing a value for developmental effects, would the chemical then be classified as a developmental toxicant, even though this endpoint is not a driver for the RfV?

- As EPA develops multiple toxicity values, EPA should ensure that the values are appropriately communicated to and understood by all IRIS users.

Comments on Part 2, Example 7

Without seeing the underlying studies and their descriptions and quality reviews, it is difficult to comment on how they are used and whether or not the appropriate uncertainty factors are applied. Thus, we will focus on the presentation of the information.

- Table 7-1 could be improved by more explicitly stating what the endpoint is. For instance, instead of saying "cardiovascular effects in rats," the more specific endpoint that was evaluated should be presented. This type of transparency is important to being able to truly understand the effect of concern. Similar changes should be made in Figure 7-1 as well (e.g., EPA should describe the neurodevelopmental alterations seen in Chen, 2012).

- Table 7-2 clearly presents the reproductive and immunological endpoints, but is vague in describing the specific neurodevelopmental alterations of concern. Unfortunately, the text on page 45 is equally vague so the reader really does not know what the particular endpoint of concern is.

- Page 46 states, at line 16, that fluctuations in exposure could potentially lead to appreciable risk even if average levels over the exposure were less than or equal to the RfD. It is unclear whether EPA is referring to a specific endpoint. This statement implies that one should always be looking at peak exposures. We assume this is not EPA's intent. However more clarity and specific citations are needed to put this statement in a useable context.

- In the text and tables in this example, EPA mentions the confidence in the RfD values. However, how EPA determines confidence is never described. Similarly this section provides no consideration of uncertainty within the individual values and thus the uncertainties associated with deriving organ specific values. More should be stated regarding uncertainties as it is well known that point estimates have a false sense of precision.[41]

[41] See NRC Models in Environmental Regulatory Decision Making (2007) ("[T]here are substantial problems in reducing the results of a large-scale study with many sources of uncertainty to a single number or even a single probability distribution. We contend that such an approach draws the line between the role of analysts and the role of policy makers in decision making at the wrong place."; Id. at 7 "Effective decision making will require providing policy makers with more than a single probability distribution for a model result (and certainly more than just a single number, such as the expected net benefit, with no indication of uncertainty). Such summaries obscure the sensitivities of the outcome to individual sources of uncertainty, thus undermining the ability of policy makers to make informed decisions and constraining the efforts of stakeholders to understand the basis for the decisions.")

G. EXTERNAL PEER REVIEW ENHANCEMENTS

The establishment of the Chemical Assessment Advisory Committee (CAAC) is a positive development. EPA should ensure that members are able to retain their independence from EPA, and provide truly independent and constructive advice to the IRIS Program. While EPA states that it has fully implemented its peer review enhancements, further improvements are necessary. Below are additional suggestions for how the Agency can continue to enhance the peer review process.

- The CAAC should be used to review updates to IRIS guidance, including the handbook, and should also advise the IRIS Program on cross-cutting scientific issues, which can impact multiple assessments.

- EPA should enhance the rigor of the contractor run peer review panels. These panels should be held to the same standards for technical expertise, conflict of interest, and bias as the Scientific Advisory Board (SAB) panels. Similarly, they should follow Federal Advisory Committee Act requirements and be fully transparent to stakeholders. For instance, draft and preliminary peer review reports should be shared with stakeholders as well as EPA.

- As recommended by the EPA's SAB and BOSC, strategies should be developed to more efficiently address peer review comments.[42] In particular, the joint SAB and BOSC report notes the NAS example that uses an independent review monitor to provide critical guidance on addressing comments. Similar to the role of a journal editor, the NAS review monitor helps to ensure that comments from reviewers have been appropriately and sufficiently addressed. Currently, the IRIS process lacks such a step and EPA staff, who are the authors of the draft assessments, have full discretion and oversight in determining which peer review and stakeholder comments are responded to. Similarly, EPA staff have sole discretion in deciding if the responses provided are sufficient. Further improvements are necessary in this area.

IV. NECESSARY NEXT STEPS

ACC and ARASP support EPA's activities to improve the IRIS Program and ensure that the Program produces high quality, scientifically sound chemical assessments. We also commend EPA's efforts to improve its IRIS Program documentation and enhance consistency and transparency in the Agency's approach to develop hazard assessments as presented in the EPA Submission to the NRC IRIS Review Committee. However as noted above there are still opportunities for EPA to improve and refine its processes. Specifically, consistent and transparent study evaluation methods to determine quality and reliability for the different types of studies (epidemiology, *in vivo* toxicology, *in vitro* toxicology and mechanistic studies) should

[42] See Sept 28, 2012 report available at: http://yosemite.epa.gov/sab/sabproduct.nsf/3822EB089FCCB18D85257A8700800679/$File/EPA-SAB-12-012-unsigned.pdf.

be adopted. Additionally, a scientifically sound framework for integrating study results to establish cause and effect which incorporates MOA information to determine potential risks to humans at environmentally relevant exposures should be implemented. ACC and ARASP are firmly committed to promoting the development and application of up-to-date, scientifically sound methods for conducting chemical assessments as well as utilizing MOA information as the organizing principle in chemical assessment. Improving the technical quality and objectivity of EPA IRIS assessments, particularly by ensuring transparency in what science is being considered, how it is being interpreted, and how it is integrated within an assessment, EPA can ensure that potential risks are objectively and consistently evaluated.

While the IRIS Program has indicated EPA is accepting comments on the documents submitted to the NRC IRIS Committee, the comment period was not formally announced, nor was a docket created to receive submission of detailed comments and attachments. We recommend EPA create a docket on regulations.gov and announce a formal 60-day comment period via the Federal Register. In addition, the NRC IRIS Committee should hold an open public meeting to discuss the EPA's Draft Handbook for IRIS Assessment Development, to encourage further public input and robust discussion into the EPA revisions to the IRIS Program. Finally, the IRIS handbook and associated documents should be treated as economically significant guidance documents subject to the requirements of the OMB Final Bulletin for Agency Good Guidance Practices and subject to review by the Office of Information and Regulatory Affairs under Executive Orders 12866 and 13563.

V. REFERENCES

Aschengrau A, Seage GR. (2003). Essentials of Epidemiology in Public Health. Sudbury, Mass: Jones and Bartlett.

Beaglehole R, Bonita R, Kjellstrom, T. (1993). Ch 1 - What is Epidemiology? In Basic Epidemiology. Geneva: WHO: 1 - 11.

Beane Freeman LE, Blair A, Lubin JH, Stewart PA, Hayes RB, Hoover RN, Hauptmann M. Mortality from lymphohematopoietic malignancies among workers in formaldehyde industries: the National Cancer Institute Cohort. (2009) J Natl Cancer Inst. May 20;101(10):751-61

Coggon D, Harris EC, Poole J, Palmer KT. Extended follow-up of a cohort of british chemical workers exposed to formaldehyde. (2003) J Natl Cancer Inst. Nov 5;95(21):1608-15

Goodman SN, Samet JM. (2006). Cause and cancer epidemiology. In: Schottenfeld D, Fraumeni J (eds). Cancer Epidemiology and Prevention, 3rd edition. New York, Oxford University Press; 3-9.

Gordis L. (2000). From association to causation: deriving inferences from epidemiologic studies. In: Epidemiology. Philadelphia: W B Saunders, 184–203.

Hauptmann M, Stewart PA, Lubin JH, Beane Freeman LE, Hornung RW, Herrick RF, Hoover RN, Fraumeni JF Jr, Blair A, Hayes RB. Mortality from lymphohematopoietic malignancies and brain cancer among embalmers exposed to formaldehyde. (2009) J Natl Cancer Inst. 2009 Dec 16;101(24):1696-708.

Pinkerton LE, Hein MJ, Stayner LT. Mortality among a cohort of garment workers exposed to formaldehyde: an update. (2004) Occup Environ Med. Mar;61(3):193-200

Hill AB (1965). The environment and disease: association or causation?. Proceedings of the Royal Society of Medicine, 58(5), 295

Hill AB: Principles of medical statistics. (1971). 9th edition. New York: Oxford University Press
Kleinbaum DG, Kupper LL, Morgenstern H. (1982). Epidemiologic Research: Principles and Quantitative Methods. Lifetime Learning Publications

MacMahon B and Pugh TF. (1970). Epidemiology. Principles and Methods. Little, Brown, Boston.

Mausner JS and Bahn, AK. (1974). Epidemiology: An Introductory Text. Elsievier – Health Science Division

Rothman KJ. (1986). Modern Epidemiology. Boston: Little, Brown, 299-304.

Rotheman KJ. (2002). Epidmiology: An Introduction. Oxford University Press.

Rothman KJ, Greenland S. (2005). Causation and causal inference in epidemiology. Am J Public Health. 95 Suppl 1:S144-50.

Vetter N, Matthews I. (1999). Causation. In: Epidemiology and public health medicine. Edinburgh: Churchill Livingstone, 23–30.

Weed DL. (1995). Causal and preventive inference. Cancer prevention and control, 285, 302.

Weed DL. (2000). Interpreting epidemiological evidence: how meta-analysis and causal inference methods are related. International Journal of Epidemiology, 29(3), 387-390.