National Toxicology Program
U.S. Department of Health and Human Services

REVISED DRAFT NTP APPROACH FOR SYSTEMATIC REVIEW AND EVIDENCE INTEGRATION FOR LITERATURE-BASED HEALTH ASSESSMENTS

Division of the National Toxicology Program

National Institute of Environmental Health Sciences

National Institutes of Health

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES

December 11, 2012

INTRODUCTION

The National Toxicology Program (NTP) conducts literature-based evaluations to assess the evidence that environmental chemicals, physical substances, or mixtures (collectively referred to as "substances") cause adverse health effects and to examine the state of the science. The NTP is adopting a systematic review procedure for these evaluations to enhance transparency for reaching and communicating evidence assessment conclusions. The systematic review format provides a transparent structure to perform a literature search, determine whether studies are relevant for inclusion, extract data from studies, assess study quality, and synthesize data for reaching conclusions. The method for data synthesis includes steps to assess confidence within an evidence stream (i.e., human, animal, and other relevant data[1] separately) and then to integrate across evidence streams to reach hazard identification conclusions (**Figure 1**). These methods are being developed, refined, and implemented according to the procedures established for literature-based evaluations through the NTP's Office of Health Assessment and Translation (OHAT). [2]

Step 1: Prepare Topic

Prior to conducting an evaluation, the NTP scopes and focuses the topic. Once a topic is identified, a draft protocol is developed that outlines the proposed approach to answer the specific question or questions to be addressed in the evaluation. The objective(s) can be to identify a potential health hazard or to summarize data gaps and identify research needs. In either case, each question is formulated based on PECO principles (**Population** of interest, **Exposure** or Intervention, **Control** or comparator group, and **Outcomes** of interest). For example, Do women exposed to chemical X in non-occupational settings have reduced fertility? The detailed protocol documents the strategy to be used in the evaluation and contains project-specific details for key aspects of the methods including: (1) a comprehensive literature search strategy, (2) criteria for selection of studies relevant to address the question, (3) grouping and hierarchy of outcomes pertinent to the question, (4) data extraction elements, (5) risk of bias assessment, and (6) evaluation of confidence in the body of evidence for answering the question. The protocol contains project-specific details as to how human, animal, and other relevant data will be evaluated and utilized.

The topic for the evaluation and the protocol are developed through an iterative process in which information is obtained by outreach to federal partners, use of technical experts as needed, comment from the public, and consultation from the NTP Board of Scientific Counselors. Project-specific decisions for key aspects of the evaluation are made and documented in the protocol before proceeding with the evaluation. However, it is also recognized that valid reasons for modifying a protocol during the course of an evaluation may be encountered (e.g., see FDA 2010). Revisions to the protocol are permitted in these situations; revisions are documented and justified with notation of when in the process the revisions were made.

Step 2: Search for and Select Studies for Inclusion

Searching for Studies: A comprehensive search of the primary scientific literature is performed. The search covers multiple databases (including, but not limited to, PubMed, TOXNET, Scopus, etc.), with sufficient details of the search strategy documented in the protocol such that it could be reproduced. Specifics of the search also list the dates of the search, frequency of updates, and any limits placed on the search (e.g., language or date of publication). The protocol establishes minimum requirements for inclusion of data from meeting abstracts or other unpublished literature. If a study that may be critical to the evaluation has not been peer reviewed, the NTP will have it peer reviewed.

[1] See http://oehha.ca.gov/multimedia/green/pdf/GC_Regtext011912.pdf for definition and discussion of "Other relevant data"; in brief it refers to non-endpoint data, including chemical, physical, biochemical, biological or other data that may be important for consideration in an evaluation.

[2] See schematic of the OHAT evaluation process at http://ntp.niehs.nih.gov/go/38138

Selecting Studies for Inclusion: All references identified in the search are screened to select studies relevant to answering the question of the evaluation. The protocol establishes criteria for including or excluding references based on applicable outcomes, relevant exposures, and types of studies. The criteria contain sufficient detail such that use of scientific judgment during the selection process is limited. If major limitations in a specific study type or design for addressing the question are known in advance (e.g., unreliable methods to assess exposure or health outcome), a basis for excluding those studies may be described *a priori* in the protocol. The protocol also outlines the specific plans for: reviewing studies for inclusion, resolving conflicts between reviewers, and documenting the reasons that studies were excluded. Two reviewers screen all references at the title/abstract level and resolve disagreements by reaching consensus through discussion. Any reference possibly meeting the inclusion criteria is retrieved for full text review. Procedures for full text review are tailored to the scope of the review and follow procedures established in the protocol.

Step 3: Extract Data from Studies

Relevant data are extracted from individual studies selected for inclusion using separate template forms for human, animal, and *in vitro* studies that are customized as needed for specific evaluations. For each study, data extraction is performed by one member of the evaluation team with quality assurance procedures in place and specified in the protocol (e.g., review by another team member). Following completion of an evaluation, data extraction files will be made publicly available.

Step 4: Assess the Quality of Individual Studies

"Study quality" or the risk of bias of individual studies is assessed on an outcome-specific basis by using a set of questions to evaluate study design and performance. The risk of bias tool considers guidance from the Agency for Healthcare Research and Quality (AHRQ) (Viswanathan *et al.* 2012), which uses specific questions under five domains (selection, performance, attrition, detection, and reporting bias). Individual questions are designated as only applicable to certain general types of study designs (randomized controlled trials, cohorts, case-control studies, cross sectional studies, case series, case reports, and experimental animal studies), with a subset of the questions applying to each study design (see Appendix A for the questions used to assess risk of bias in human and experimental animal studies along with applicability by study design). The protocol details project-specific factors of study design and performance that result in specific risk of bias ratings for each question. For each study outcome, all of the applicable questions are answered with one of four options (definitely low, probably low, probably high, or definitely high risk of bias (Guyatt 2012)) following pre-specified criteria detailed in the protocol. All references are assessed on an outcome basis for risk of bias by two reviewers. Discrepancies between the reviewers are resolved by reaching consensus through discussion.

To the extent possible, other relevant data (e.g., exposure data, mechanistic or *in vitro* studies) studies are subject to an assessment of study quality or risk of bias, the details of which are included in the protocol.

Step 5: Rate the Confidence in the Body of Evidence

A confidence rating for the body of evidence for a given outcome is developed by considering the strengths and weaknesses of a collection of studies. These ratings reflect confidence that the study findings accurately reflect the true association between exposure to a substance and an effect. The NTP's method is based on the GRADE[3] and AHRQ approaches (Balshem *et al.* 2011, Lohr 2012), which are conceptually very

[3] Grading of Recommendations Assessment, Development and Evaluation Working Group (http://www.gradeworkinggroup.org/). Note, the GRADE guidelines have been adopted by the Cochrane Collaboration (Schünemann *et al.* 2012).

similar. The method uses 3 descriptors to indicate the level of confidence in the body of evidence (Definitions Box 1). In the context of identifying research needs, a conclusion of "High Confidence" indicates that further research is very unlikely to change our confidence in the apparent relationship between exposure to the substance and the outcome. Conversely, a conclusion of "Low Confidence" suggests that further research is very likely to impact confidence in the apparent relationship. Human and animal data are considered separately, as are other relevant data (e.g., exposure data, mechanistic or *in vitro* studies) to the extent possible and/or necessary. When other relevant data are necessary to address the question of the evaluation, the specific methods used to determine confidence for these studies are explained in the protocol. The methods outlined below apply to the bodies of evidence for human studies and experimental animal studies that address a health outcome.

- **High Confidence (+++)** in the association between exposure to the substance and the outcome. The true effect is highly likely to be reflected by the apparent relationship.
- **Moderate Confidence (++)** in the association between exposure to the substance and the outcome. The true effect may be reflected in the apparent relationship.
- **Low Confidence (+)** in the association between exposure to the substance and the outcome. The true effect is highly likely to be different than the apparent relationship.

Definitions Box 1: Confidence Ratings in the Body of Evidence

Conclusions developed in the subsequent steps of the method are based on the evidence with the highest confidence. The protocol can be used in Step 2 to exclude studies when major problems in study design or conduct are known in advance and the basis for excluding these studies is described *a priori* in the protocol. The risk of bias evaluations that are given to individual studies on an outcome basis in Step 4 are another key means to select the studies for a given outcome with the highest confidence (e.g., see AHRQ 2012a) that will move forward and be used in decision-making at later steps.

For each outcome, collections of studies are given an initial confidence rating by study design (see Figure 1 for Step 5 schematic). The initial rating is downgraded for factors that decrease confidence and upgraded for factors that increase confidence in the results. Then, confidence across all available study designs is assessed. A single, well conducted study may provide evidence of toxicity or a health effect associated with exposure to the substance in question (e.g., see Germolec (2009) and Foster (2009) for explanation of the NTP levels of evidence for determination of "toxicity" for individual studies). If a sufficient body of very similar studies is available, a quantitative meta-analysis may be completed to generate an overall estimate of effect. Finally, confidence conclusions are developed across multiple outcomes for those outcomes that are biologically related. It is recognized that the scientific judgments involved in these confidence ratings are inherently subjective; however, this process provides a transparent framework to document and justify the decisions made to arrive at a final confidence rating.

Initial confidence set by study design for each outcome

An initial confidence rating is given to each study design type based on its ability to address causality as reflected in the confidence that exposure preceded the outcome and was associated with the outcome (see Figure 1, Step 5, column 1). Experimental studies such as animal or human randomized controlled trials (RCT) have an initial rating of "High Confidence". Human observational studies are stratified into different initial confidence levels: cohort and nested case-control studies are designated "Moderate Confidence" and cross-sectional, case-control, case-series, and case-control studies are designated "Low Confidence". Although observational studies in this approach refer to human studies, observational animal studies could be considered using the same initial confidence designations. The initial ratings are the starting points that reflect the general features of each study design, and then studies for a given outcome are evaluated for factors that would downgrade or upgrade confidence in the evidence.

Downgrade confidence rating

Five properties of the body of evidence (risk of bias, unexplained inconsistency, indirectness, imprecision, and publication bias) are considered to determine if the initial confidence rating should be downgraded (see **Figure 1**, Step 5, column 2). For each of the 5 properties, a judgment is made and documented regarding whether or not there are issues that decrease the confidence rating in each aspect of the body of evidence for the outcome. Factors that would downgrade confidence by one versus two levels are specified in the protocol. The reasons for downgrading confidence may not fit neatly into a single property of the body of evidence. If the decision to downgrade is borderline for two properties, the body of evidence is downgraded once to account for both partial concerns. Similarly, the body of evidence is not downgraded twice for what is essentially the same limitation (or upgraded twice for the same asset) that could be considered applicable to more than one property of the body of evidence.

Risk of bias of the body of evidence: Risk of bias criteria were described in Step 4 where study quality issues for individual studies are evaluated on an outcome-specific basis. In this step, the previous risk of bias assessments for individual studies now serve as the basis for an overall risk of bias conclusion for the entire body of evidence (Guyatt *et al.* 2011e).

Unexplained inconsistency: Inconsistency, or large variability in the magnitude or direction of estimates of effect, that cannot be explained, reduces confidence in the body of evidence. Large inconsistency across studies should be explored, preferably through *a priori* hypotheses that might explain the heterogeneity. If there is less inconsistency within subgroups of the body of evidence (e.g., men versus women), the protocol can also be amended to consider these *post hoc* groupings.

Indirectness: Indirectness can refer to external validity or indirect measures of the health outcome. Indirectness can lower confidence in the body of evidence when the population, exposure, or outcomes measured differ from those that are of most interest. Concerns about directness could apply to the relationship between a measured outcome and a health effect (i.e., upstream biomarker of a health effect), the route of exposure and the typical human exposure, or the study population and the population of interest (Guyatt *et al.* 2011c, Lohr 2012).

Imprecision: Imprecision is the lack of certainty for an estimate of effect for a specific outcome. A precise estimate enables the evaluator to determine whether or not there is an effect (i.e., it is different from the comparison group). Confidence intervals (CIs) of the estimates of effect provide the primary evidence used in considering the imprecision of the body of evidence (Guyatt *et al.* 2011b).

Publication bias: Publication bias specifically pertains to the body of evidence, as selective reporting within a study is covered in risk of bias criteria addressing these limitations (Guyatt *et al.* 2011d). There is empirical evidence that studies with negative results (no association) are less likely to be in the published literature. Negative studies may also be affected by "lag bias" or longer time to publication. While some publication bias is inevitable, downgrading is reserved for when serious concern for publication bias significantly decreases confidence in the body of evidence.

Upgrade confidence rating

Four properties of the body of evidence (large magnitude of effect, dose-response, all plausible confounding, and cross-species/population/study consistency) are considered to determine if the confidence rating should be upgraded (see **Figure 1**, Step 5, column 3). For each of the 4 properties, a judgment is made and documented regarding whether or not there are factors that increase the confidence rating in each aspect of the body of evidence for the outcome. Factors that would upgrade confidence by one versus two levels are specified in the protocol.

Large magnitude of effect: A large magnitude of effect is defined as an observed effect that is sufficiently large that it is unlikely to have occurred as a result of bias from potential confounding factors.

Dose-response: A plausible dose-response relationship between level of exposure and the outcome increases confidence in the result because it reduces concern that the result could be due to chance. Multiple observational human studies with varied exposure levels can contribute to an overall picture of the dose-response. It is important to recognize that the dose-response relationship may not be monotonic and that biological plausibility should be considered in evaluating the dose-response relationship.

All plausible confounding: This element refers to consideration of confounding, healthy worker effect, or effect modification that would bias the effect estimate towards the null. When a body of evidence is potentially biased by one of these factors in a direction that strengthens the findings (i.e., counter to the observed effect), confidence in the results is increased.

Cross-species/population/study consistency: Three types of consistency in the body of evidence can increase confidence in the results: across animal studies - consistent results reported in multiple experimental animal models or species; across dissimilar populations - consistent results reported across populations that differ in factors such as time, location, and/or exposure; and across study types - consistent results reported from different study designs.

Other: Additional factors specific to the topic being evaluated (for example, particularly rare outcomes) may result in increasing a confidence rating. These other factors would be specified and defined in the protocol.

Combine confidence conclusions for all study types and multiple outcomes

Conclusions are based on the evidence with the highest confidence when considering evidence across study types and multiple outcomes. Confidence ratings are initially set based on available study designs for a given outcome (e.g., for prospective studies separately from cross-sectional studies). The study type with the highest confidence rating forms the basis for the confidence conclusion. As outlined previously, consistent results across study-designs increases confidence in the combined body of evidence and can result in an upgraded confidence rating moving forward to Step 6.

After confidence conclusions are developed for a given outcome, conclusions for multiple outcomes and the entire evaluation are developed. The project-specific definition of an outcome and the grouping of biologically related outcomes used in this step follow the definitions developed *a priori* in the protocol; deviations are taken with care, justified, and documented. When outcomes are sufficiently biologically related that they may inform confidence on the overall health outcome, confidence conclusions may be developed in two steps. Each outcome would first be considered separately. Then, the related outcomes would be considered together and re-evaluated for properties that relate to downgrading and upgrading the body of evidence. The project-specific explanation of the strategy used to combine confidence ratings across multiple outcomes is documented in the protocol.[4]

Step 6: Translate Confidence Ratings into Level of Evidence for Health Effect

The level of evidence is assessed separately within the human, experimental animal, and to the extent possible and necessary, other relevant data sets. The level of evidence for health effects conclusions reflect both the overall confidence in the association between exposure to the substance and the outcome (effect or no effect) and the direction of the effect (toxicity or no toxicity; see Figure 1 for Step 6 schematic). The strategy uses 4 terms to describe the level of evidence for health effects. These descriptors reflect both the confidence in the body of evidence for a given outcome and the direction of effect. There are 3 descriptors ("High Level of Evidence," "Moderate Level of Evidence," and "Low Level of Evidence") that directly translate from the confidence ratings that exposure to the substance is associated with a heath effect

[4] The product of an OHAT evaluation may vary (e.g., NTP monograph or peer-reviewed publication). For example, in state of the science evaluations, it may be appropriate to end the process after rating the confidence in the available evidence in Step 5 and developing a summary of data gaps and research needs.

and a fourth designation ("Evidence of No Health Effect") to indicate confidence that the substance is not associated with a health effect (Definitions Box 2). Because of the inherent difficulty in proving a negative, a conclusion of evidence of no health effect is only reached when there is high confidence in the body of evidence. A low or moderate level of evidence results in a conclusion of inadequate evidence to reach a conclusion.

- **High Level of Evidence:** There is high confidence in the body of evidence for an association between exposure to the substance and the health outcome(s).
- **Moderate Level of Evidence:** There is moderate confidence in the body of evidence for an association between exposure to the substance and the health outcome(s).
- **Low Level of Evidence:** There is low confidence in the body of evidence for an association between exposure to the substance and the health outcome(s), or no data are available.
- **Evidence of No Health Effect:** There is high confidence in the body of evidence that exposure to the substance is not associated with the health outcome(s).

Definitions Box 2: Level of Evidence for Health Effects Descriptors

Although the conclusions describe associations, a causal relationship is implied and the ratings describe the level of evidence for health effects in terms of confidence in the association or the estimate of effect determined from the body of evidence (see Table 1 for discussion of the relationship between the Bradford Hill Criteria (Hill 1965) of causality and the approach for upgrading and downgrading confidence in a body of evidence (based on the GRADE approach as described in Schünemann *et al.* 2011)).

Table 1: Relationship of the Hill Criteria to the NTP Approach

Hill Criteria	Consideration in the NTP Approach
Strength	Considered in upgrading the confidence in the body of evidence for ***large magnitude of effect*** and downgrading confidence for ***Imprecision***
Consistency	Considered in downgrading confidence in the body of evidence for ***unexplained inconsistency;*** also considered in upgrading confidence in the body of evidence for ***consistency across study types***, ***across dissimilar populations***, or ***across animal species***; *and in integrating the body of evidence among human, animal, and other relevant data*
Temporality	Considered in ***initial confidence ratings*** by study design, for example experimental studies have an initial rating of "High Confidence" because of the increased confidence that exposure preceded outcome
Biological gradient	Considered in upgrading the confidence in the body of evidence for evidence of a ***dose-response*** relationship
Biological plausibility	Considered in downgrading the confidence in the body of evidence for ***indirectness***; also in examining non monotonic ***dose-response*** relationships. Other relevant data that inform plausibility such as PBPK and mechanistic studies are considered in integrating the evidence. Plausibility is also considered in developing confidence conclusions across biologically related outcomes, particularly for outcomes along a pathway to disease.
Experimental evidence	Considered in downgrading for ***risk of bias*** and ***initial confidence ratings*** by study design

Step 7: Integrate Evidence to Develop Hazard Identification Conclusions

To determine the hazard identification conclusion, the highest level of evidence for a health effect from each of the evidence streams is combined in the final step of the evidence assessment process. Hazard identification conclusions may be reached on individual outcomes (health effects) or groups of biologically related outcomes, as appropriate, based on the evaluation's objectives and the available data. The rationale for such conclusions are documented as the evidence is combined within and across evidence streams and the conclusions are clearly stated as to which outcomes are incorporated into each conclusion. The four hazard identification conclusion categories are:

- Known to be a hazard to humans
- Presumed to be a hazard to humans
- Suspected to be a hazard to humans,
- Not classifiable or not identified to be a hazard to humans

In Step 7, the evidence streams for human studies and non-human animal studies, which have remained separate through the previous steps, are integrated along with other relevant data such as supporting evidence from mechanistic studies and, if necessary, with consideration of special situations related to exposure information that may apply across evidence streams. Hazard identification conclusions are developed by integrating the highest level of evidence for health effects conclusions from the human and the animal evidence streams. First, the level of evidence for health effects conclusion for human data from Step 6 ("High," "Moderate," or "Low") is considered together with the level of evidence for health effects conclusion for non-human animal data to reach one of four hazard identification conclusions as outlined in Step 7 schematic in **Figure 1**.

- If the human level of evidence conclusion is high, the hazard identification conclusion is "known" based on the human data alone.
- If the human level of evidence conclusion is moderate, the hazard identification conclusion depends on the strength of the non-human evidence. The hazard identification conclusion is "presumed" if the non-human evidence conclusion is high or "suspected" if the non-human evidence conclusion is moderate or low.
- If the human level of evidence conclusion is low, the hazard identification conclusion again depends on the strength of the non-human evidence. The hazard identification conclusion is "suspected" if the non-human level of evidence conclusion is high or "not classifiable" if the non-human evidence conclusion is moderate or low.

Any impact is then considered of other relevant evidence such as mechanistic data, *in vitro* data, and evidence based on upstream indicators of a health effect, on the hazard identification conclusion derived by integrating the human and non-human animal streams as outlined in Step 7 schematic in **Figure 1**. Other relevant data may increase or decrease the initial hazard identification conclusion. A detailed rationale accompanies the conclusions along with an explanation as to how other relevant data contributed to the final hazard identification conclusion.

- Strong supporting evidence may raise the level of the hazard identification conclusion initially derived by considering the human and animal evidence together. Note that mechanistic or supporting evidence is not required to reach a hazard identification conclusion of "known" if the level of evidence conclusion from human data is high.
- If the hazard identification conclusion was "presumed" based on the human and non-human animal data, strong support from other relevant data may result in an upgraded conclusion of "known." If the hazard identification conclusion was "suspected" based on the human and non-human data, strong support from other relevant data may result in an upgraded conclusion of "presumed."
- If the human level of evidence conclusion is low and the non-human level of evidence is moderate, consideration of other relevant data can be used to reach a hazard identification conclusion of "suspected."
- If the human level of evidence conclusion is low and mechanistic or mode of action data are compelling that evidence from non-human studies is not relevant to human health effects, a hazard identification conclusion of "not classifiable" may be appropriate.

In communicating the outcome of the evaluation, the NTP compiles a draft document that presents the hazard identification conclusion. A summary of key scientific judgments made during development of the conclusions is outlined and justified. As appropriate, the NTP also discusses information about outcomes from evidence streams not used in reaching the final hazard identification conclusions placing them into the proper context of whether or not they are supportive. The draft monograph undergoes peer review and public comment as part of the overall process for it preparation and publication.[5]

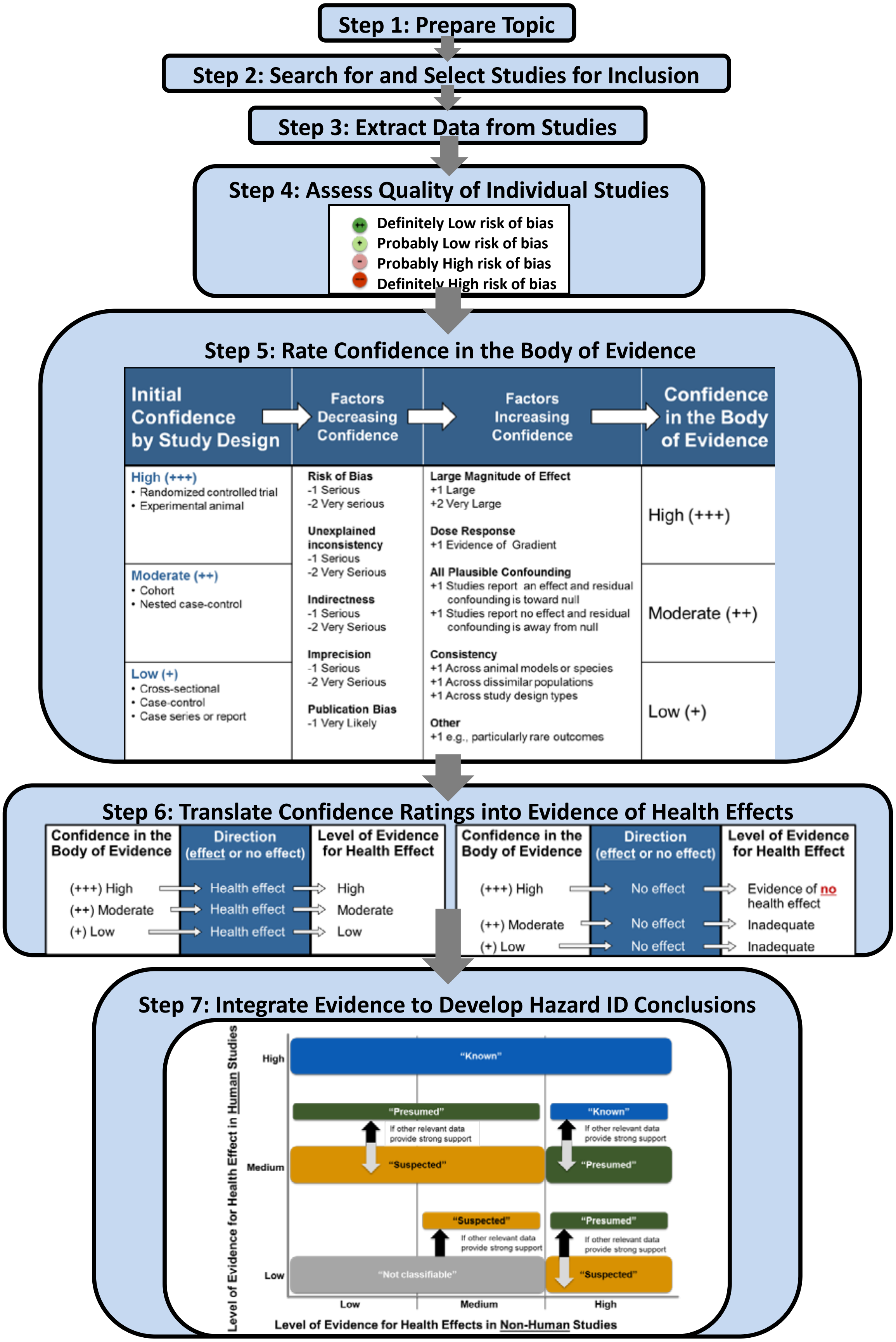
[5] For hazard identification evaluations conducted by the OHAT, the draft monograph undergoes peer review and public comment as part of its overall process for preparation and publication (http://ntp.niehs.nih.gov/go/38138).

ACKNOWLEDGMENTS

The NTP's approach for literature-based health assessments was developed by the Office of Health Assessment and Translation (OHAT) and Office of Liaison, Policy and Review (OLPR), within the Division of the National Toxicology Program at the NIEHS. Strong support for undertaking this project was provided by the NTP Board of Scientific Counselors, NTP Executive Committee, public, and other stakeholders (see http://ntp.niehs.nih.gov/go/9741 for meeting minutes). In developing this methodology, the NTP considered authoritative sources on systematic review including, but not limited to, the Agency for Healthcare Research and Quality (AHRQ) (AHRQ 2012b), The Cochrane Collaboration (Higgins and Green 2011), the Grading of Recommendations Assessment, Development and Evaluation (GRADE) Working Group (Guyatt *et al.* 2011a), and the Collaborative Approach to Meta Analysis and Review of Animal Data from Experimental Studies (CAMARADES, see http://www.camarades.info). Additional technical input was provided on portions of this method by experts affiliated with these groups including Lisa Bero, Director, San Francisco Branch, United States Cochrane Center at UC San Francisco; Gordon Guyatt, Co-chair, GRADE working group, McMaster University; Malcolm Macleod, CAMARADES Centre, University of Edinburgh; Karen Robinson, Co-Director, AHRQ Evidence-Based Practice Center, The Johns Hopkins Bloomberg School of Public Health; Holger Schünemann, Co-chair, GRADE working group, McMaster University; and Tracey Woodruff, Director, Program on Reproductive Health and the Environment, UC San Francisco.

The NTP sought consultation on specific topics in an initial draft approach by a working group of the NTP BSC. The NTP plans to present the draft NTP approach for systematic review and evidence integration for literature-based health assessments outlined in this document to the NTP Executive Committee at its meeting in November 2012 and to the NTP BSC in December 2012 where the working group's report will be formally presented. The NTP plans to consider input from these advisory groups as well as any public comments in finalizing the NTP approach for systematic review and evidence integration for literature-based health assessments.

Figure 1: The NTP Approach for Conducting Literature-Based Evidence Assessments



Step 1: Prepare Topic

Step 2: Search for and Select Studies for Inclusion

Step 3: Extract Data from Studies

Step 4: Assess Quality of Individual Studies

- ++ Definitely Low risk of bias
- + Probably Low risk of bias
- - Probably High risk of bias
- -- Definitely High risk of bias

Step 5: Rate Confidence in the Body of Evidence

Initial Confidence by Study Design	Factors Decreasing Confidence	Factors Increasing Confidence	Confidence in the Body of Evidence
High (+++) • Randomized controlled trial • Experimental animal	**Risk of Bias** -1 Serious -2 Very serious **Unexplained inconsistency** -1 Serious -2 Very Serious **Indirectness** -1 Serious -2 Very Serious **Imprecision** -1 Serious -2 Very Serious **Publication Bias** -1 Very Likely	**Large Magnitude of Effect** +1 Large +2 Very Large **Dose Response** +1 Evidence of Gradient **All Plausible Confounding** +1 Studies report an effect and residual confounding is toward null +1 Studies report no effect and residual confounding is away from null **Consistency** +1 Across animal models or species +1 Across dissimilar populations +1 Across study design types **Other** +1 e.g., particularly rare outcomes	High (+++)
Moderate (++) • Cohort • Nested case-control			Moderate (++)
Low (+) • Cross-sectional • Case-control • Case series or report			Low (+)

Step 6: Translate Confidence Ratings into Evidence of Health Effects

Confidence in the Body of Evidence	Direction (effect or no effect)	Level of Evidence for Health Effect
(+++) High	Health effect	High
(++) Moderate	Health effect	Moderate
(+) Low	Health effect	Low

Confidence in the Body of Evidence	Direction (effect or no effect)	Level of Evidence for Health Effect
(+++) High	No effect	Evidence of no health effect
(++) Moderate	No effect	Inadequate
(+) Low	No effect	Inadequate

Step 7: Integrate Evidence to Develop Hazard ID Conclusions

Level of Evidence for Health Effect in Human Studies
High
Medium
Low
"Known"
"Presumed"
"Known"
If other relevant data provide strong support
If other relevant data provide strong support
"Suspected"
"Presumed"
"Suspected"
"Presumed"
If other relevant data provide strong support
If other relevant data provide strong support
"Not classifiable"
"Suspected"
Low
Medium
High
Level of Evidence for Health Effects in Non-Human Studies

REFERENCES

AHRQ. 2012a. Interventions for Adults with Serious Mental Illness Who are Involved with the Criminal Justice System. Available at http://effectivehealthcare.ahrq.gov/ehc/products/406/1259/SMI-in-CJ-System_ResearchProtocol_20120913.pdf [accessed September 26, 2012].

AHRQ. 2012b. Grading the Strength of a Body of Evidence When Assessing Health Care Interventions: An Update (Draft Report). Available at http://effectivehealthcare.ahrq.gov/search-for-guides-reviews-and-reports/?pageaction=displayproduct&productid=1163 [accessed July 30, 2012].

Balshem H, Helfand M, Schunemann HJ, Oxman AD, Kunz R, Brozek J, Vist GE, Falck-Ytter Y, Meerpohl J, Norris S, Guyatt GH. 2011. GRADE guidelines: 3. Rating the quality of evidence. *Journal of clinical epidemiology* 64(4): 401-406. Available at http://www.sciencedirect.com/science/article/pii/S089543561000332X.

FDA. 2010. *Guidance for Industry: Adaptive Design Clinical Trials for Drugs and Biologics: Draft Guidance*. Silver Spring, MD: US Department of Health and Human Services (DHHS), Food and Drug Administration (FDA), Center for Drug Evaluation and Research (CDER) and Ceter for Biologics Evaluation and Research (CBER). Available at http://www.fda.gov/downloads/Drugs/.../Guidances/ucm201790.pdf.

Foster PM. 2009. *Explanation of Levels of Evidence for Reproductive System Toxicity*. National Toxicology Program. Research Triangle Park, NC: US Department of Health and Human Services. Available at http://ntp.niehs.nih.gov/go/18711.

Germolec D. 2009. *Explanation of Levels of Evidence for Immune System Toxicity*. National Toxicology Program. Research Triangle Park, NC: US Department of Health and Human Services. Available at http://ntp.niehs.nih.gov/go/9399.

Guyatt G, Oxman AD, Akl EA, Kunz R, Vist G, Brozek J, Norris S, Falck-Ytter Y, Glasziou P, Debeer H, Jaeschke R, Rind D, Meerpohl J, Dahm P, Schunemann HJ. 2011a. GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. *Journal of clinical epidemiology* 64(4): 383-394. Available at http://www.sciencedirect.com/science/article/pii/S0895435610003306.

Guyatt G. 2012. Tools to Assess Risk of Bias in Cohort Studies, McMaster University: Available at: http://www.evidencepartners.com/resources/ [accessed July 13, 2012].

Guyatt GH, Oxman AD, Kunz R, Brozek J, Alonso-Coello P, Rind D, Devereaux PJ, Montori VM, Freyschuss B, Vist G, Jaeschke R, Williams JW, Jr., Murad MH, Sinclair D, Falck-Ytter Y, Meerpohl J, Whittington C, Thorlund K, Andrews J, Schunemann HJ. 2011b. GRADE guidelines 6. Rating the quality of evidence--imprecision. *Journal of clinical epidemiology* 64(12): 1283-1293. Available at http://www.sciencedirect.com/science/article/pii/S089543561100206X.

Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, Alonso-Coello P, Falck-Ytter Y, Jaeschke R, Vist G, Akl EA, Post PN, Norris S, Meerpohl J, Shukla VK, Nasser M, Schunemann HJ. 2011c. GRADE guidelines: 8. Rating the quality of evidence--indirectness. *Journal of clinical epidemiology* 64(12): 1303-1310. Available at http://www.sciencedirect.com/science/article/pii/S0895435611001831.

Guyatt GH, Oxman AD, Montori V, Vist G, Kunz R, Brozek J, Alonso-Coello P, Djulbegovic B, Atkins D, Falck-Ytter Y, Williams JW, Jr., Meerpohl J, Norris SL, Akl EA, Schunemann HJ. 2011d. GRADE guidelines: 5. Rating the quality of evidence--publication bias. *Journal of clinical epidemiology* 64(12): 1277-1282. Available at http://www.sciencedirect.com/science/article/pii/S0895435611001818.

Guyatt GH, Oxman AD, Vist G, Kunz R, Brozek J, Alonso-Coello P, Montori V, Akl EA, Djulbegovic B, Falck-Ytter Y, Norris SL, Williams JW, Jr., Atkins D, Meerpohl J, Schunemann HJ. 2011e. GRADE guidelines: 4. Rating the quality of evidence-study limitations (risk of bias). *Journal of Clinical Epidemiology* 64(4): 407-415. Available at http://www.sciencedirect.com/science/article/pii/S0895435610004130.

Higgins J, Green S, eds. 2011. *Cochrane Handbook for Systematic Reviews of Interventions*. Version 5.1.0 [updated March 2011]: The Cochrane Collaboration. Available at: www.cochrane-handbook.org [accessed January 18, 2012].

Hill AB. 1965. The Environment and Disease: Association or Causation? *Proc R Soc Med* 58: 295-300. Available at http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1898525/pdf/procrsmed00196-0010.pdf.

Lohr KN. 2012. Grading the Strenth of Evidence. The Agency for Healthcare Research and Quality (AHRQ) Training Modules for Systematic Reivews Methods Guide, Available at: http://www.effectivehealthcare.ahrq.gov/index.cfm/slides/?pageAction=displaySlides&tk=18 [accessed July 13, 2012].

Schünemann H, Hill S, Guyatt G, Akl EA, Ahmed F. 2011. The GRADE approach and Bradford Hill's criteria for causation. *J Epidemiol Community Health* 65(5): 392-395.

Schünemann HJ, Oxman AD, Vist GE, Higgins JPT, Deeks JJ, P. G, Guyatt GH, on behalf of the Cochrane Applicability and Recommendations Methods Group. 2012. Chapter 12: Interpreting results and drawing conclusions. Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [updated March 2011]. Higgins JPT, Green S, eds., The Cochrane Collaboration, 2011. Available at www.cochrane-handbook.org. [accessed July 13, 2012].

Viswanathan M, Ansari MT, Berkman ND, Chang S, Hartling L, McPheeters M, Santaguida PL, Shamliyan T, Singh K, Tsertsvadze A, Treadwell JR. 2012. Assessing the Risk of Bias of Individual Studies in Systematic Reviews of Health Care Interventions. In *Methods Guide for Effectiveness and Comparative Effectiveness Reviews*. Agency for Healthcare Research and Quality (AHRQ) Methods for Effective Health Care, AHRQ Publication No. 12-EHC047-EF. Rockville (MD). Available at: www.effectivehealthcare.ahrq.gov/ [accessed July 13, 2012].

Appendix A: Draft NTP Risk of Bias Questions

	Experimental Animal	RCT	Cohort	Case-control	Cross-sectional	Case Series
SELECTION BIAS						
Was treatment, dose, or exposure adequately randomized? Randomization requires that each human subject or animal had an equal chance of being assigned to any experimental group (e.g., use of random number table or computer generated randomization)?	X	X				
Was treatment, dose, or exposure allocation adequately concealed? Concealment requires that study scientists do not know which treatment, dose, or exposure level the human subject or animal is to be given before it enters the study. Human studies also require that allocation be concealed from human subjects prior to entering the study. *Note: 1) a question under performance bias addresses blinding of scientists and human subjects to treatment during the study; 2) a question under detection bias addresses blinding of outcome assessors.*	X	X				
Were inclusion and exclusion criteria applied consistently across study groups? Consistency refers to criteria used for selection of animals, or during recruitment and selection of human subjects.	X	X	X	X	X	
Is the comparison group appropriate? For human studies: appropriateness includes having similar baseline characteristics between the exposed and comparison groups and having the exposed and non-exposed subjects drawn from the same population. For experimental animal studies: appropriateness includes similar baseline characteristics between the treated and control groups, and use of appropriate vehicle-treatment in the control group.	X	X	X	X	X	X
Does the study design or analysis account for important confounding and modifying variables? *Note: a parallel question under detection bias addresses reliability of the measurement of these variables.*	X	X	X	X	X	X
PERFORMANCE BIAS						
Did researchers adjust or control for other exposures or interventions that are anticipated to bias results?	X	X	X	X	X	X
Were the study scientists and human subjects blinded to treatment, dose, or exposure group? Blinding requires that study scientists do not know which treatment, dose, or exposure level the human subject or animal is being given. Human studies also require blinding of the human subjects.	X	X				
ATTRITION BIAS						
Were attrition rates uniformly low? In RCT, animal, or cohort studies: was the loss of human subjects or animals by treatment, dose, or exposure group consistent across groups? If not, were missing data handled appropriately? And, were reasons documented when human subjects or animals were removed from a study? In case-control studies: is the time period between exposure and outcome the same for cases and controls?	X	X	X	X		

Appendix A: Draft NTP Risk of Bias Questions

	Experimental Animal	RCT	Cohort	Case-control	Cross-sectional	Case Series
DETECTION BIAS						
Were the outcome assessors blinded to treatment, dose, or exposure group? Blinding requires that outcome assessors do not know which treatment, dose, or exposure level the human subject or animal is being given.	X	X	X	X	X	X
Are confounding variables assessed consistently across groups using reliable measures? *Note, a parallel question under selection bias addresses whether design or analysis account for confounding.*			X	X	X	X
Are data analyses appropriate, performed with reliable tests, and implemented consistently?	X	X	X	X	X	X
Can we be confident in the exposure characterization? Confidence requires valid, reliable, and sensitive analytical methods to measure exposure applied consistently across groups, as well as consideration of exposure timing. The time window for treatment or exposure requires it to cover the biologically relevant time period prior to the outcome. For experimental methods confidence in exposure also includes consideration of purity and stability of the test substance.	X	X	X	X	X	X
Can we be confident in the outcome assessment? Confidence requires valid, reliable, and sensitive methods to assess the outcome applied consistently across groups, as well as consideration of the timing of the outcome assessment. The time window for outcome measurement requires a sufficient time to elapse such that the effect could develop before the scheduled assessment of the outcome. For observational studies, this would include consideration of the likely impact of exposure misclassification on the results.	X	X	X	X	X	X
Did the study have sufficient power to detect a biologically meaningful difference between groups? For continuous variables, for example, did the study have sufficient power to detect a 10% change in mean value from the control group with 80% power and alpha of 0.05?	X	X	X	X	X	X
REPORTING BIAS						
Were the potential outcomes pre-specified by the researchers? Are all pre-specified outcomes reported?	X	X	X	X	X	X
OTHER						