



National Toxicology Program

U.S. Department of Health and Human Services

**DRAFT PROTOCOL FOR SYSTEMATIC REVIEW TO EVALUATE THE
EVIDENCE FOR AN ASSOCIATION BETWEEN BISPHENOL A (BPA)
EXPOSURE AND OBESITY**

April 9, 2013

Office of Health Assessment and Translation (OHAT)

Division of the National Toxicology Program

National Institute of Environmental Health Sciences

This draft protocol is being disseminated to obtain public comment. It does not represent and should not be construed to represent final NTP determination or policy.

TABLE OF CONTENTS

STEP 1: Prepare the topic	1
Background.....	1
Objectives.....	2
Eligibility criteria for considering studies for this review	3
Types of studies	3
Types of participants and model systems.....	3
Types of exposures.....	3
Types of outcomes.....	3
Types of publications	4
STEP 2: Search for and select studies for inclusion.....	6
Electronic searches.....	6
Databases to be searched	6
Ongoing Trials databases	6
Searching other resources.....	7
Handsearches.....	7
Grey literature and public request for information.....	7
Duplicate citations	8
Screening studies for eligibility	8
Planned interim analyses.....	11
STEP 3: Extract data from studies	11
Data extraction and management	11
Summarizing study design, experimental model, methodology, and results.....	12
STEP 4: Assess quality of individual studies	17
Human and animal studies.....	17
Determining Tiers of Study Quality.....	22
<i>In vitro</i> studies	25
Data Display	25
Software used for data management, analysis, and display.....	25
STEP 5: Rate confidence in body of evidence.....	35
Planned interim analyses.....	36
Initial confidence based on study design	40
Domains that can reduce confidence.....	40
Risk of bias across studies	40
Summary of risk of bias ratings	40
Consideration of whether to downgrade confidence based on risk of bias.....	42
Unexplained inconsistency	43
Planned interim analyses.....	44
Directness and applicability	47
Consideration of dose or exposure level	47

DRAFT (April 9, 2013)

Planned interim analyses.....47
Tabular summary of guidance for evaluating directness49
Imprecision.....49
Publication bias51
Domains that can increase confidence52
Large magnitude of association or effect52
Dose-response53
Plausible confounding or other residual biases that would increase our confidence in
estimated effect.....56
Consistency across study types, experimental model systems, or populations.....56
Other57
Combine confidence conclusions for all study types and multiple outcomes57
STEP 6: Translate confidence ratings into evidence of health effect conclusions.....58
STEP 7: Integrate evidence to develop hazard identification conclusions59
Assessment of biological plausibility provided by “supportive” evidence.....61
Peer-Review64
Review Team64
Author declarations of interest.....64
Sources of support64
Technical advisors64
Protocol history & revisions65
References.....65
Appendices.....71

DRAFT PROTOCOL FOR SYSTEMATIC REVIEW TO EVALUATE THE EVIDENCE FOR AN ASSOCIATION BETWEEN Bisphenol A (BPA) EXPOSURE AND OBESITY

National Toxicology Program (NTP), Office of Health Assessment and Translation (OHAT), National Institute of Environmental Health Sciences (NIEHS)

STEP 1: PREPARE THE TOPIC

Background

Rationale for topic

The rise in obesity is a major threat to public health in the US and abroad (CDC 2012; IASO 2012; Ogden and Carroll 2010; WHO 2011a). Research addressing the role of environmental chemicals in obesity has rapidly expanded in the past several years (Heindel and vom Saal 2009; Janesick and Blumberg 2011; NTP 2011; Thayer et al. 2012). The May 2010 White House Task Force on Childhood Obesity (2010) and the March 2011 NIH Strategic Plan for Obesity (2011) acknowledge the growing science base in this area and cite the need to understand more about the role of environmental exposures as part of identifying future research and prevention strategies.

To help assess the science in this area OHAT is conducting a systematic review to evaluate the association between exposure to bisphenol A (BPA) and obesity based on the growth in the literature base over the past several years and public interest in this topic (Blue 2013; Kristof 2013; Szabo 2012). To our knowledge a systematic review on this topic has not been conducted and obesity as a health outcome has not been addressed in other safety evaluations of BPA.

Use of protocol as a case study to assess the draft Office of Health Assessment and Translation Approach for Systematic Review and Evidence Integration for Literature-Based Health Assessments (February 2013)

OHAT is conducting two case studies to test implementation of the revised “Draft Office of Health Assessment and Translation Approach for Systematic Review and Evidence Integration for Literature-Based Health Assessments (February 2013)”¹ (HHS 2013) and this is the draft protocol for 1 of the 2 case studies. The other

¹ The approach described in the draft document is based on guidance from the Grading of Recommendations Assessment, Development and Evaluation (GRADE) Working Group (Guyatt et al. 2011a), a framework applied most often to evaluate the quality of evidence and strength of recommendations for health care intervention decisions based on human studies (typically randomized clinical trials). The appeal of the GRADE framework is that it is (1) widely used (Guyatt et al. 2011f), (2) conceptually similar to the approach used by the Agency for Healthcare Research and Quality (AHRQ 2012) for grading the strength of a body of evidence of human studies, and (3) the Cochrane Collaboration has adopted the principles of the GRADE system for evaluating the quality of evidence for outcomes reported in systematic reviews (Higgins and Green 2011). However, none of these existing frameworks (GRADE, AHRQ, and the Cochrane Collaboration) address approaches for considering animal studies or *in vitro* studies. In addition, the guidance provided by GRADE, AHRQ, and the Cochrane

protocol is entitled “Systematic review to evaluate the evidence for an association between perfluorooctanoic acid (PFOA) or perfluorooctane sulfonate (PFOS) exposure and immunotoxicity.” These two case studies will be conducted as guidance for whether changes are needed in the revised framework. Future updates on this project will be posted at <http://ntp.niehs.nih.gov/go/evals> and individuals interested in receiving updates are encouraged to register to the NTP Listserv (<http://ntp.niehs.nih.gov/go/getnews>).

Objectives

Develop hazard identification conclusions (“known”, “presumed”, “suspected”, or “not classifiable”) that exposure BPA is associated with overweight/obesity in humans based on integrating the evidence from human and animal data and considering the evidence for biological plausibility provided by data from *in vitro* studies of adipocytes; *ex vivo*, cellular, genomic, or mechanistic outcomes reported in eligible animal or human studies; and data on interactions with key receptors involved in regulating adipogenesis [e.g., PPAR γ , RXR, LXR, GR, AR, and estrogen receptors (ER α , ER β , and “non-classical”)].

Specific aims:

- Provide a summary of the literature and rate our confidence in studies that assess the association between BPA exposure and overweight/obesity-related outcomes in human studies of children and adults.
- Provide a summary of the literature and rate our confidence in studies that assess the effect of BPA on adiposity-related outcomes in whole animal models.
- Evaluate the evidence for biological plausibility provided by *in vitro* and mechanistic studies that assess the effects of BPA in *in vitro* studies of adipocytes; *ex vivo*, cellular, genomic, or mechanistic outcomes reported in eligible animal or human studies; and data from cell systems, computational toxicology, high throughput screening data, and *in silico* models on interactions with key receptors involved in regulating adipogenesis [e.g., peroxisome proliferator-activated receptors (PPAR), retinoid X receptor (RXR), liver X receptor (LXR), glucocorticoid receptor (GR), androgen receptor (AR), and estrogen receptors (ER α , ER β , and “non-classical”)].
- Develop hazard identification conclusions (“known”, “presumed”, “suspected”, or “not classifiable”) based on integrating the confidence ratings from human and animal data and considering the extent of support for biological plausibility provided by *in vitro* studies (defined here as other than whole animal studies, and including cell systems, computational toxicology, high throughput screening data, and *in silico* methods).

Collaboration is less developed for observational human studies compared to randomized clinical trials. For these reasons the draft OHAT approach includes a number of refinements to GRADE that were considered necessary in order to accommodate our need to integrate data from multiple evidence streams (human, animal, *in vitro*) and focus on observational human studies rather than the randomized clinical trials more commonly encountered in the health care intervention field. This latter point is important because the objectives of OHAT reviews are typically to identify potential adverse effects and randomized clinical trials are not considered ideal for this purpose (Oxman et al. 2006; Silbergeld and Scherer 2013). In environmental health, the most appropriate data is human observational epidemiology and experimental animal studies and these data need to be considered with clear appreciation for their strengths and limitations.

Eligibility criteria for considering studies for this review

Types of studies

Only studies with a control or referent group will be included. Case studies, case reports, and ecological studies in humans will be excluded. Animal and *in vitro* studies without a concurrent control will be excluded..

Types of participants and model systems

Studies of humans, experimental animals, and from “supporting evidence” provided by *in vitro* studies of adipocytes; *ex vivo*, cellular, genomic, or mechanistic outcomes reported in eligible animal or human studies; and data from cell systems, computational toxicology, high throughput screening data, and *in silico* methods on interactions with key receptors involved in regulating adipogenesis [e.g., peroxisome proliferator-activated receptors (PPAR), retinoid X receptor (RXR), liver X receptor (LXR), glucocorticoid receptor (GR), androgen receptor (AR), and estrogen receptors (ER α , ER β , and “non-classical”).

There are no restrictions based on lifestage at exposure or assessment, sex, animal species or strain, or adipocyte model system.

Types of exposures

Exposure to BPA (CAS# 80-05-7) based on administered dose or concentration, biomonitoring data (e.g., urine, blood, or other specimens), environmental measures (e.g., air, water levels), or indirect measures such as job title.

There will be no exclusions based on the analytical method used to measure BPA, differences in the sensitivities of these methods will be considered when assessing the risk of bias (“internal validity”) of individual studies.

Types of outcomes

Publications must include an indicator of BPA exposure analyzed in relation to any one of the following primary or secondary outcomes listed in [Table 1](#) for human and animal studies. Primary outcomes are considered to be most direct, or applicable, to the evaluation. Secondary outcomes are relevant, but less direct and can include upstream indicators, risk factors, intermediate outcomes, or related measures to our primary outcomes.

For human studies, standard diagnostic criteria will be used as measures of overweight and obesity (BMJ Group; NHLBI 2012) ([Table 1](#)). For animals there is no standard or clear definition of obesity and the term has been used in the environmental health literature to describe statistical significant increases in body weight or body weight gains. However, reliance on body weight as a measure of an obesity phenotype is problematic for several reasons and will not be used as the sole health outcome to determine inclusion eligibility for animal studies. First, there is general acceptance that body weight is a relatively crude indicator of internal body fat (“adiposity”) in rodent models. For example, no changes in body weight were observed by Ohlsson et al.(2000) in estrogen receptor- α (ER α) knockout mice compared to wild-type despite having visibly greater amounts of adipose tissue. Preferred measures of adiposity in rodents include fat mass, fat pad weight, and adipose tissue cellularity because there is a metabolic benefit of smaller adipocytes and a metabolic detriment of adipocyte hypertrophy (cell size increase) or hyperplasia (cell number increase) (Jo et al. 2009; Thayer et al. 2012). Second, “high” dose levels of BPA (>100 mg/kg bw) can cause systemic toxicity which is often manifest as body weight loss (Chapin et al. 2008; NTP 2011), so these studies are not likely informative for determining

whether BPA can cause an “obese” phenotype at lower dose levels. Finally, there is a very large literature evaluating the effects of BPA on body weight, and previous surveys of these studies do not indicate that BPA exposure causes “obesity” as defined by a consistent reporting of increased body weight or growth (NTP 2008a). This conclusion remained true in a 2010 analysis restricted to only those studies that tested low doses of BPA, defined as less than 5 mg/kg bw, during development and reported an effect of BPA on a health outcome [(NTP 2011), see Appendix Table A in the draft literature review document for BPA]. This strategy was used to avoid the issue of considering studies in the analysis that have been criticized as being insensitive to detect low dose effects (Myers et al. 2009). Many of the studies did not detect an effect on body weight and the magnitude of the effect in cases where an increase in body weight was observed ranged from 3% to 50%, with most reporting increases of 10% or less (Akingbemi et al. 2004; Alonso-Magdalena et al. 2010; Howdeshell et al. 1999; Kubo et al. 2003; Miyawaki et al. 2007; Nikaido et al. 2004; Okada and Kai 2008; Patisaul and Bateman 2008; Rubin et al. 2001; Ryan et al. 2010; Salian et al. 2009; Somm et al. 2009). It is possible that differences in results were due to differences in experimental design across studies, e.g., diet, sample size, species/strain, sex. Although in some cases divergent results were reported by the same laboratory in separate experiments using the same rodent strain under similar experimental protocols. For these reasons body weight will not be used as the basis for determining eligibility for animal studies and our focus will be primarily on differences in adiposity observed between control and BPA-treated animals where increases in adiposity will be considered consistent with an “obese” phenotype. However, we will conduct a secondary analysis of body weight in studies that also reported a measure of adiposity to assess the relative sensitivity of these measures.

For evidence from *in vitro*, *ex vivo*, or mechanistic outcomes studies (“supportive” evidence” in [Table 1](#)), we are interested in measures of phenotypic or apical response in adipocytes (e.g., lipid accumulation), potential pathways for mediating the phenotypic response (e.g., PPAR γ activation), and other cellular responses in adipocytes.

Types of publications

Publications must be peer-reviewed articles or meet the guidelines for hand-collection or grey literature described below.

There are no language or date restrictions.

*Review articles and health assessments of BPA will be collected for the purposes of reviewing the bibliography list and will not contribute to the final number of studies considered eligible unless they also contain original data.

Table 1. Outcomes considered relevant for study eligibility

Humans	Animals	Supporting evidence
<p><i>Primary outcomes</i></p> <p>Overweight or obesity based on body mass index (BMI), waist circumference, or waist to height ratio (BMJ Group; NHLBI 2012)</p> <p><u>BMI</u></p> <p>Overweight in adults: BMI 25.0-29.9 Obese in adults: BMI 30.0-39.9 Extremely or morbidly obese in adults: BMI ≥40.0</p> <p>Overweight in children and adolescents: BMI 85th to 94th percentile for age and sex Obese in children and adolescents: BMI ≥95th percentile and weight ≥95th percentile for height for age and sex</p> <p><u>Waist circumference (WC)</u></p> <p>Men: WC >102 cm Women: WC >88 cm</p> <p><u>Waist to height ratio (WHR)</u></p> <p>Men: Ideal WHR 0.9; increased risk WHR >1.0 Women: Ideal WHR 0.7; 1ncreased risk WHR >0.85</p>	<p><i>Primary outcomes</i></p> <p>adiposity (e.g., fat mass, percent fat, fat pad weight, adipose tissue cellularity)</p>	<p><i>Phenotypic or “apical” outcomes from in vitro studies of adipocytes</i></p> <p>e.g., adipogenic endpoints such as adipocyte number or size, adipocyte differentiation, or adipocyte lipid accumulation, reprogramming of multipotent stem cell fate toward adipogenic lineage</p>
<p><i>Secondary outcomes</i></p> <p>BMI z-score, measures of adiposity (e.g., fat composition, skin-fold thickness), growth curves, adipokines, ghrelin, leptin, adiponectin, resistin, feeding behavior</p>	<p><i>Secondary outcomes</i></p> <p>adipokines, ghrelin, leptin, adiponectin, resistin, feeding behavior, energy expenditure, and body weight or body weight gain in studies that also included an adiposity measure</p>	<p><i>Pathway and cellular endpoints</i></p> <p>e.g., <i>ex vivo</i>, cellular, genomic, epigenomic, or mechanistic outcomes reported in eligible animal or human studies; cellular, genomic, epigenomic, or mechanistic outcomes reported in <i>in vitro</i> studies of adipocytes; interactions with key receptors involved in regulating adipogenesis, e.g., peroxisome proliferator-activated receptors (PPAR), retinoid X receptor (RXR), liver X receptor (LXR), glucocorticoid receptor (GR), androgen receptor (AR), and estrogen receptors (ERα, ERβ, and “non-classical”)², in any <i>in vitro</i> model or high throughput screening system</p>

²Given the very large literature summarizing BPA’s interactions with estrogen and other nuclear receptors we will rely on secondary citations to summarize this information when possible.

STEP 2: SEARCH FOR AND SELECT STUDIES FOR INCLUSION

Electronic searches

Databases to be searched

The following databases will be searched from inception to the present:

- African Index Medicus
- Cochrane Library
- DART-Europe (E-Theses)
- Embase
- EPA's [ACToR](#) (Aggregated Computational Toxicology Resource)
- EPA's [Chemical Data Access Tool](#) to find health and safety data that has been submitted to the Agency, under authorities in sections 4, 5, and 8 of the Toxic Substances Control Act (TSCA)
- IMSEAR (Index Medicus for South-East Asia Region)
- IndMed
- KoreaMed
- LILACS
- Panteleimon
- Open Access Theses and Dissertations
- PubChem
- PubMed
- Scopus
- Toxline
- Web of Science
- WPRIM (Western Pacific Region)

[Appendix 1](#) shows the search strategy and specific terminology for PubMed and other databases. The search terms were identified by (1) reviewing Medical Subject Headings for relevant and appropriate terms and (2) extracting key terminology from reviews and a sample of relevant primary data studies. A combination of relevant subject headings and keywords were subsequently identified. A test set of relevant studies was used to ensure the search terms retrieve 100% of the test set. The search strategy was tailored for each database. When available, controlled vocabulary is used in conjunction with text word searches.

Ongoing Trials databases

We will search the following ongoing trials registers to identify relevant trials:

- The metaRegister of Controlled Trials on www.controlled-trials.com.
- The US National Institutes of Health Ongoing Trials Register on www.clinicaltrials.gov.
- The World Health Organization International Clinical Trials Registry Platform on www.who.int/trialsearch.

Searching other resources

Handsearches

Handsearches will not be done for any specific journals.

We will scan the bibliographies of the included studies, relevant reviews, government reports and other “grey literature” (see below) for relevant references, a process referred to as “snowballing”.

Grey literature and public request for information

Grey literature refers to reports that are difficult to find via conventional channels such as published journals. Examples of grey literature include technical reports from government agencies or scientific research groups, working papers from research groups or committees, white papers, conference proceedings and abstracts, theses and dissertations, or unpublished research reports.

We will review the contents and reference list of evaluations of BPA that might have been conducted by government or public health entities that routinely produce health assessments, including:

- ATSDR Toxicological Profiles <http://www.atsdr.cdc.gov/toxpro2.html>
- CalEPA Office of Environmental Health Hazard Assessment <http://www.oehha.ca.gov/risk.html>
- European Chemicals Agency <http://echa.europa.eu/en/information-on-chemicals>
- European Food and Safety Authority (EFSA) <http://www.efsa.europa.eu/>
- Health Canada <http://www.hc-sc.gc.ca/index-eng.php>
- US National Toxicology http://ntpserver.niehs.nih.gov/main_pages/NTP_ALL_STDY_PG.html
- WHO assessments – CICADS, EHC <http://www.who.int/ipcs/assessment/en/>

We will also consult subject matter experts and agencies represented on the NTP Executive Committee² to potentially identify data that address this topic. We will attempt to identify grey literature and information on ongoing studies from the research and other stakeholder communities through a public request for information advertised through the NTP listserv (<http://ntp.niehs.nih.gov/go/getnews>) and a query of NIH Research Portfolio Online Reporting Tools (RePORT, <http://report.nih.gov/index.aspx>).

² The NTP Executive Committee provides programmatic and policy oversight to the NTP Director. The Executive Committee meets once or twice a year in closed forum. Members of this committee include the heads (or their designees) from the following federal agencies: Consumer Product Safety Commission (CPSC), Department of Defense (DoD), US Environmental Protection Agency (EPA), Food and Drug Administration (FDA), National Cancer Institute (NCI), National Center for Environmental Health/Agency for Toxic Substances and Disease Registry (NCEH/ATSDR), National Institute of Environmental Health Sciences (NIEHS), National Institute for Occupational Safety and Health (NIOSH), Occupational Safety and Health Administration (OSHA).

In addition, the results of the literature screening will be posted on the OHAT website and we will invite review by the public through the NTP list serve as an additional mechanism to identify relevant studies. The literature search results will also be forwarded to the corresponding authors of the set of relevant studies identified from the literature search to ask for knowledge of other published studies, ongoing research, or grey literature.

Criteria for consideration of relevant unpublished data

The NTP will only consider publically available information. If a study that may be critical to the evaluation has not been peer reviewed, NTP policy is to have it peer reviewed through the use of experts if the owners of the data are willing to have the study made publically accessible. The level of detail provided for methodology and results must be sufficient to permit peer-review, i.e., at least comparable to a journal publication. Any potential peer reviewers would be screened for conflict of interest prior to confirming them for service.

Grey literature such as meeting abstracts for which additional study details are not available will be used to assess potential publication bias but will not be considered an eligible study.

Unpublished data from personal author communication can supplement a peer-reviewed study, so long as it can be made publically available.

Duplicate citations

The results of the literature search will be downloaded into Endnote X5 software. Exact article duplicates will be removed using Endnote X5 software prior to uploading into DistillerSR® Web-Based Systematic Review Software³. The duplicate detection feature in DistillerSR® will also be used to detect and remove duplication citations; this feature looks for similarities in articles based on author and title content. If an article is a duplicate, a member of the review team “quarantines” the article such that it is removed from the main project with an annotation for reason, although the article is not deleted and can be retrieved later if needed. Multiple publications from the same study population identified during full-text review will be evaluated for duplicate data. For studies with multiple publications on the same population, we will select the publication with the longest follow-up as the primary report for data analysis and consider the other as secondary publications. For studies with equivalent follow-up periods, we will select the study with the largest number of cases or the most recent publication as the primary report.

Screening studies for eligibility

We will use DistillerSR® for screening studies. Screeners will be trained using written documentation on study eligibility with an initial pilot phase undertaken to improve clarity of the inclusion and exclusion language and to improve accuracy and consistency among screeners. Articles will first be independently reviewed at the title and abstract level by two members of the review team. Disagreements between the 2 screeners will be resolved by each screener independently reviewing the conflicts noted in DistillerSR®, modifying and discussing responses as appropriate to resolve, and arbitration by a third member of the review team if necessary. A copy of articles that appear to meet the inclusion criteria based on the title and abstract screen will be obtained for full-text review unless the article is not available after an attempt has been made to obtain it. Copies of articles that cannot be

³DistillerSR® (<http://systematic-review.net/>) is a proprietary project management tool for tracking studies through the screening process and storing data extracted from these studies. The technical content (i.e., screening results, data extraction) generated by OHAT during an evaluation is not proprietary and will be made publically available.

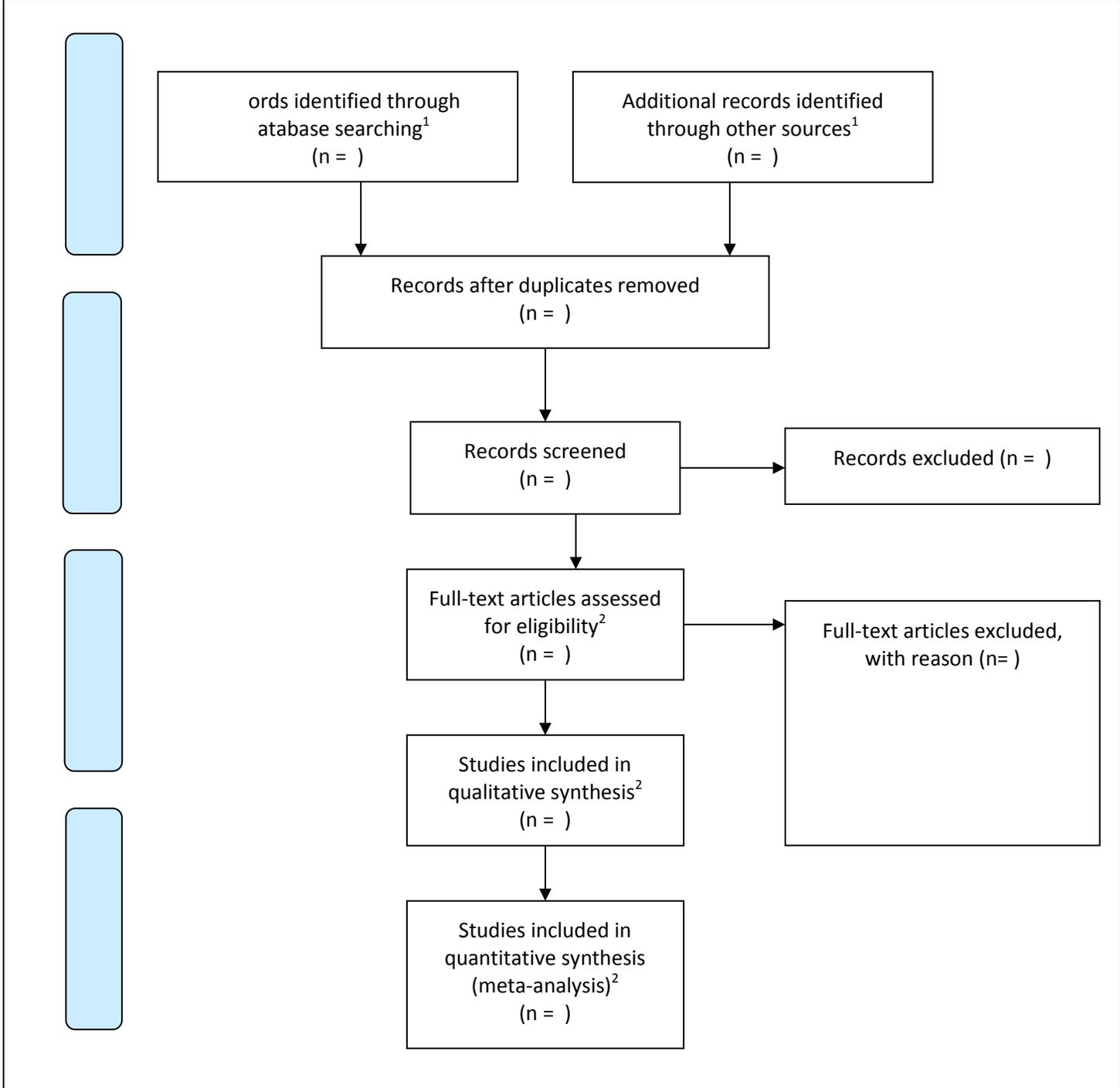
DRAFT (April 9, 2013)

assessed for relevance based on the title and abstract screen will also be obtained to determine eligibility based on full-text review. Studies will not be considered further when the title and abstract clearly indicate that the study does not meet the inclusion criteria described above.

Full-text eligibility review will also be independently conducted by two members of the review team with reasons for exclusion annotated and tracked (e.g., “review paper with no original data”). The primary reason for excluding studies will be if the article does not contain original data relevant to our eligibility criteria. If the full text of an article is not in English, then translation services or consultation with a fluent scientist will be utilized to determine relevance for inclusion. Flow of information through the different phases of the review will be documented in a schematic similar to that represented in [Figure 1](#) as recommended in the PRISMA statement on preferred reporting items for systematic reviews and meta-analyses. (Moher et al. 2009). The PRISMA Flow Diagram Generator can be used to develop study flow schematics (<http://theta.utoronto.ca/tools/prisma>).

One member of the review team will independently scan the bibliographies of the included studies, relevant reviews, and government reports and other “grey literature” for relevant references that were not identified from the database searches. Eligibility will be confirmed by a second screener in order to be included and the source of the citation tracked. Studies considered relevant from this hand searching will be noted separately in the flow of information schematic ([Figure 1](#)).

Figure 1. Flow of information through the different phases of the review



From Moher et al. (2009)

¹The database or other source of article is recorded and presented in evaluation.

²The number of studies are separately identified for human, animal, and supportive study (*in vitro*, *ex vivo*, mechanistic) evidence streams

Planned interim analyses

If no or few (<3) studies are identified that meet our inclusion criteria then we will characterize the evidence base as “insufficient material to conduct a systematic review.”

Although unlikely, it is also possible that we will identify a recent systematic review on this topic during the process of screening studies. In this case we would revisit whether there was still a need to proceed with the proposed evaluation or a portion of the objectives that are not duplicative with the published systematic review.

STEP 3: EXTRACT DATA FROM STUDIES

Data extraction and management

We will use customized data extraction forms in DistillerSR® to collect information on study design, experimental model, methodology and results (see [Table 2](#) for specific items) and internal validity or “risk of bias” for human and animal data. The results of the data extraction will be made publically available in Microsoft Excel® format when the evaluation has been completed. The data extraction files can also be disseminated upon request in CSV or RIS format.

Each team member’s data extraction will be reviewed by one other team member to assure accuracy. The risk of bias questions will be judged independently in duplicate because of the possibility of subjective interpretation. All discrepancies will be resolved with discussion, involving a third member of the review team if necessary.

Multiple publications of the same study (e.g., publications reporting subgroups, other outcomes, or longer follow-up) will be identified by examining author affiliations, study designs, cohort name, enrollment criteria, and enrollment dates. If necessary, study authors will be contacted to clarify any uncertainty about the independence of two or more articles. We will include all reports but select a study to use as the primary report for data analysis and consider the others as secondary publications. The primary report will generally be the publication with the longest follow-up, or for studies with equivalent follow-up periods, we will select the study with the largest number of cases or the most recent publication as the primary report. We will include relevant data from all reports, but if the same outcome is reported in more than one report we will use data from the primary report. To avoid double-counting of subjects when several reports of overlapping subjects are available, only outcome data from the report with the largest number of subjects will be included. We will include the data when a smaller report provides data on an outcome that was not provided by the largest report.

Missing data

We will attempt to contact authors of included studies to obtain missing data considered important to summarize study findings ([Table 2](#)) or evaluate risk of bias.

Note on sharing of data extraction files: The data extraction files are available upon request in an Excel (or similar) format specifically designed to facilitate data display using Meta Data Viewer software (Boyles et al.

DRAFT (April 9, 2013)

2011)⁴. In addition, the web-based DistillerSR[®] screening and data extraction forms can be shared upon request with individuals or organizations that have active licenses to access the software.

For questions on data extraction files, forms, or the Meta Data Viewer graphing program contact:

Kris Thayer, Ph.D.
Tel (919) 541-5021
thayer@niehs.nih.gov

Abee L. Boyles, Ph.D.
Tel (919) 541-7886
boylesa@niehs.nih.gov

Summarizing study design, experimental model, methodology, and results

The elements in [Table 2](#) will be summarized for each study that meets our inclusion criteria. Information that is inferred, converted, or estimated will be marked by brackets, e.g., [n=10], [0.438 µM].

⁴ Meta Data Viewer (http://ntp.niehs.nih.gov/go/tools_metadataviewer) is a graphing program designed to help assess patterns of findings in complex data sets. It can display up to 15 text columns and graph 1-10 numerical values. Users can sort, group, and filter data to look at patterns of findings across studies.

DRAFT (April 9, 2013)

Table 2. Data extraction and analysis elements to summarize study design, experimental model, methodology and results	
HUMAN	
funding	Funding source(s)
	Reporting of COI by authors
subjects	Cohort name (if applicable)
	Number of subjects (total, per group, and participation/follow-up rates by group with calculations)
	Sex
	Geography (country, region, state, etc.)
	Race and ethnicity, socioeconomic background, other variables as reported
	Age at exposure and outcome assessment (e.g., mean, median, measures of variance as presented in paper such as SD, SEM, 75th/90th/95th percentile, minimum/maximum)
	Lifestage at exposure and outcome assessment (e.g., fetus, infancy, adult, older adulthood, etc.)
	Inclusion and exclusion criteria
	Dates of study or sampling time frame (inclusive or recruitment period)
methods: study design	Study design (e.g., prospective cohort, cross-sectional, case-control, case report, etc.)
	Length of follow-up, latency/lag period(s) considered in analysis
methods: health outcome assessment	Endpoint health category (i.e., metabolic)
	Endpoint (and unit of measurement)
	Diagnostic or method to measure health outcome.
	Confounders, modifying factors, or other potential sources of bias considered in analysis and how considered, e.g., included in final model, considered for inclusion but determined not needed.
	Statistical power is assessed during data extraction using an approach to assess ability to detect a 20% change from control or referent group response for continuous data or relative risk or odds ratio of 1.5 for categorical data using the prevalence of exposure or prevalence of outcome in the control or referent group to determine sample size using OpenEpi software, a free open source statistical resource (http://www.openepi.com/OE2.3/menu/openEpiMenu.htm). Recommended sample sizes will be compared to sample sizes used in the study to categorize statistical power as “appears to be adequately powered” (sample size met), somewhat underpowered (sample size is 75% to <100% of recommended), “underpowered” (sample size is 50 to <75% required), or “severely underpowered (sample size is <50% required). For categorical data where the sample sizes in exposed and control or referent groups differ, the sample size of the exposed group will be used to determine relative power category.
methods: exposure assessment	Substance name and CAS number
	Exposure ascertainment (e.g., blood, urine, hair, air, drinking water, job classification, residence, administered treatment in controlled study, etc.)
	Analytical method for exposure assessment (when applicable, e.g., HPLC-MS/MS)
	Time of daily exposure (for occupational exposure e.g., 8 hours/day, 10 hours/day)
	Frequency of exposure when applicable (e.g., in occupational settings exposure might occur 5 days per week)
	List any other chemicals assessed

DRAFT (April 9, 2013)

Table 2. Data extraction and analysis elements to summarize study design, experimental model, methodology and results	
results: exposure assessment	Exposure levels (e.g., mean, median, measures of variance as presented in paper such as SD, SEM, 75th/90th/95th percentile, minimum/maximum) and unit of measurement
	Relative exposure category [general population (e.g., NHANES), occupational, environmental but higher than general population (e.g., living near Superfund or industrial site)]
	Documentation of details for conversion to common exposure unit (when conducted)
results: health outcome	Measures of effect at each exposure level contrast as reported in the paper (e.g., adjusted β , standardized mean difference, adjusted odds ratio, standardized mortality ratio, relative risk, etc.). When possible we will convert measures of effect to a common metric. Most often measures of effect for continuous data will be expressed as mean difference, standardized mean difference, and percent control response. Categorical data will be expressed as odds ratio, relative risk (RR, also called risk ratio), or β values depending on what metric is most commonly reported in the evidence base and our ability to obtain information for effect conversions from the study or through author query. We will calculate 95% confidence intervals (CI) for each type of converted effect size to describe the uncertainty inherent in the point estimates.
	Documentation of details for conversion to common statistic when conducted (e.g., odds ratio)
	Endpoint prevalence (when applicable)
	Statistical significance (author's interpretation)
	Shape of dose response (e.g., description of whether shape appears to be monotonic, non-monotonic, NA for single exposure or treatment group studies)
other	Documentation of author queries for study details
	Documentation of use of digital ruler to obtain data values
ANIMAL	
funding	Funding source(s)
	Reporting of COI by authors
animal model	Sex
	Species
	Strain
	Source
	Age at start of dosing (specific and lifestage)
	Age at start of assessment (specific and lifestage)
methods: treatment	Guideline compliance (i.e., use of EPA, OECD, NTP or other guideline for study design, conducted under GLP guideline conditions, non-GLP but consistent with guideline study, non-guideline peer-reviewed publication)
	Substance name and CAS number
	Source
	Purity
	Dose levels or concentration (as presented and converted to mg/kg bw/d when possible)
	Vehicle used (or untreated control)
	Route (e.g., oral, inhalation, dermal, injection)

DRAFT (April 9, 2013)

Table 2. Data extraction and analysis elements to summarize study design, experimental model, methodology and results	
	Method (e.g., if oral: via feed, gavage, drink from pipette, etc.; if subcutaneous: injection, pump, etc.)
	Documentation of details for dose conversion when conducted
	Any other relevant information, e.g., use of radiolabelled compound
	Duration (e.g., hours, days, weeks when administration was ended)
	Frequency of exposure (e.g., 5 days per week or 7 days per week)
	Time of daily exposure (e.g., 8:00 AM, 8 hours/day, 12 hours/day, <i>ad lib</i>)
methods: diet & husbandry	Diet name
	Diet source
	Diet phytoestrogen content
methods: study design	Study design (e.g., single treatment, acute, subchronic, chronic, multigenerational, developmental, other)
	Number of animals per group (and dams per group in developmental studies)
	Randomization procedure
	Method to control for litter effects in developmental studies
methods: endpoint assessment	Use of positive or negative controls and whether expected response was observed
	Endpoint health category (i.e., metabolic)
	Endpoint (and unit of measurement)
	Diagnostic or method to measure endpoint.
	Statistical methods
	Statistical power is assessed during data extraction using an approach to assess ability to detect a 20% change from control response for continuous data or odds ratio of 1.5 for categorical data using prevalence of outcome in the control group to determine sample size using OpenEpi software, a free open source statistical resource (http://www.openepi.com/OE2.3/menu/openEpiMenu.htm). Recommended sample sizes will be compared to sample sizes used in the study to categorize statistical power as “appears to be adequately powered” (sample size met), somewhat underpowered (sample size is 75% to <100% of recommended), “underpowered” (sample size is 50 to <75% required), or “severely underpowered (sample size is <50% required). For categorical data where the sample sizes in exposed and control groups differ, the sample size of the exposed group will be used to determine relative power category.
results	Endpoint values at each dose or concentration level (e.g., mean, median, frequency, measures of precision or variance)
	Measures of effect at each dose or concentration level. When possible we will convert measures of effect to a common metric. Most often measures of effect for continuous data will be expressed as mean difference, standardized mean difference, and percent control response. Categorical data will be expressed as relative risk (RR, also called risk ratio). We will calculate 95% confidence intervals (CI) for each type of effect size to describe the uncertainty inherent in the point estimates.
	NOEL, LOEL, and statistical significance of other dose levels (author's interpretation)
	Data on internal concentration, toxicokinetics, or toxicodynamics (when reported)
	Shape of dose response (e.g., description of whether shape appears to be monotonic, non-monotonic, NA for single exposure or treatment group studies)
other	Documentation of author queries for study details

DRAFT (April 9, 2013)

Table 2. Data extraction and analysis elements to summarize study design, experimental model, methodology and results	
	Documentation of use of digital ruler to obtain data values
IN VITRO	
funding	Funding source(s) Reporting of COI by authors
cell or tissue model	Cell line, cell type, or tissue Source Species Strain Lifestage Sex
methods: treatment	Dose concentration [as presented and converted to μM and expressed using scientific notation (e.g., 10^{-6}) when possible] Substance name and CAS number Source Purity Vehicle used (or untreated control) Documentation of details for dose conversion when conducted Any other relevant information, e.g., use of radiolabelled compound Duration (e.g., hours, days, weeks when administration was ended)
methods: study design	Number of replicates per group
	Guideline compliance (i.e., use of EPA, OECD, NTP or other guideline for study design, conducted under GLP guideline conditions, non-GLP but consistent with guideline study, non-guideline peer-reviewed publication) Percent serum in medium
methods: endpoint assessment	Use of positive or negative controls and whether expected response was observed Endpoint health category (i.e., metabolic) Endpoint (and unit of measurement) Diagnostic or method to measure endpoint. Statistical methods
results	NOEC, LOEC, statistical significance of other concentration levels, and AC50 (author's interpretation) Shape of dose response (e.g., description of whether shape appears to be monotonic, non-monotonic, NA for single exposure or treatment group studies)
other	Documentation of author queries for study details Documentation of use of digital ruler to obtain data values

STEP 4: ASSESS QUALITY OF INDIVIDUAL STUDIES

Human and animal studies

We will evaluate study quality by assessing risk of bias⁵, also referred to as internal validity (Higgins and Green 2011; IOM 2011; Viswanathan et al. 2012). Risk of bias is an assessment of whether the design and conduct of the study compromised the credibility of the link between exposure and outcome.

Risk of bias for individual studies will be assessed using the elements presented in [Table 3](#) (guidance on how to answer each item is provided in Appendix 2). OHAT's risk of bias rating tool was developed based on guidance from the Agency for Healthcare Research and Quality (Viswanathan et al. 2012), Cochrane Handbook (Higgins and Green 2011), CLARITY Group at McMaster University (2013), consultation with technical advisors (NTP 2013a), staff at other federal agencies, and other risk of bias or study quality tools (Dwan et al. 2010; Genaidy et al. 2007; Johnson et al. 2013; Koustas et al. 2013; Krauth et al. 2013; Shamliyan et al. 2010; Shamliyan et al. 2011). The tool presents a unified approach to evaluating risk of bias for human and animal studies to allow consideration of risk bias elements across a range of study types with common terms and categories. Not every question is applicable to all study designs, and within the tool guidance for assessing risk of bias is further tailored to whether the study is animal or human and the features of each study design type (i.e., controlled exposure, cohort, case-control, cross-sectional, or case series/report). For each study risk of bias is assessed at the outcome level because risk of bias may differ across different outcomes reported within the same study.

Within the risk of bias guidance document (Appendix 2) we note whether there is empirical evidence to support the inclusion of the question as a risk of bias element (and the direction of the bias, if known). However, in certain cases there is currently no or very limited empirical evidence to support consideration as a risk of bias item, but the question is included because it is recommended by groups that develop systematic review guidance (CLARITY Group at McMaster University 2013; Higgins and Green 2011; IOM 2011; Viswanathan et al. 2012) or captures a key epidemiological or toxicological principle in environmental health studies. Over time, we plan to use the risk of bias data collected across OHAT evaluations, as well as related work conducted by others, to develop the empirical support needed to refine the risk of bias tool.

We recognize that given reporting practices it is unlikely that some of the risk of bias items will be informative for the purposes of discriminating between studies of higher risk of bias and studies of lower risk of bias, at least in the short term. However, in the long-term, especially if reporting standards improve, collecting this information will generate data that will allow us to empirically assess evidence of bias or to remove a risk of bias question from consideration if it continues to be uninformative.

Risk of bias will be assessed independently by two data extractors for each study and discrepancies resolved by consensus, arbitration by a third member of the review team, and consultation with technical advisors as

⁵ Risk of bias, defined as the risk of a non-random error or deviation from the truth, in results or inferences, is interchangeable with internal validity, defined as "the extent to which the design and conduct of a study are likely to have prevented bias" or "the extent to which the results of a study are correct for the circumstances being studied" (Viswanathan et al. 2012).

needed. We will pilot test the risk of bias rating tool on a small subset of studies in the evidence base to identify issues and revise the guidance or training as needed.

Each of the risk of bias questions is answered on a 4 point scale:

-  **definitely low risk of bias**
-  **probably low risk of bias**
-  **probably high risk of bias**
-  **definitely high risk of bias**

In general, if information to answer the question is explicitly stated from the study report or through contacting the authors (referred to as “direct” evidence) then “definitely low risk of bias” or “definitely high risk of bias” will be used as responses. If the information is not explicitly reported but can be inferred (referred to as “indirect” evidence) then “probably low risk of bias” or “probably high risk of bias” are typically used as the risk of bias response. The guidance provided in Appendix 2 describes separate instructions for each question to identify what comprises “definitely low risk of bias”, “probably low risk of bias”, “probably high risk of bias”, and “definitely high risk of bias”. An element can be rated as “probably low risk of bias” if it is deemed that deviations from low risk of bias practices during the study would not appreciably bias results, including consideration of direction and magnitude of bias.

Rules for non-reporting: When additional information is required to address an item that is not reported we will attempt to contact the corresponding author of the original reports to provide further details. If we are unable to obtain sufficient information to evaluate the risk of bias question, “probably high risk of bias” will be used as the response except where indicated otherwise based on the guidance.

Consideration of timing and duration of exposure in relation to health outcome assessment: Risk of bias evaluates internal validity: “Are the results of the study credible?” The issue of timing and duration of exposure in relation to health outcome assessment in most cases is an issue of applicability: “Did the study design address the topic of the evaluation?” However, there may be instances where it is best considered as part of risk of bias. For example, if there are differences in the duration of follow-up across study groups, this would be a source of bias considered under detection bias “Can we be confident in the outcome assessment?” If the duration of follow-up was not optimal for the development of the outcome of interest (e.g., short duration of time between exposure and health outcome assessment for chronic disease), then it would be considered under applicability. Ideally, windows of exposure and health outcome assessment that not considered relevant to an evaluation would be considered in determining study eligibility criteria in Step 1.

Consideration of source of funding and disclosed conflict of interest: There is debate on whether financial conflict of interest should be considered a source of bias (Krauth et al. 2013; Viswanathan et al. 2012) and this issue has been raised in the BPA literature (vom Saal and Hughes 2005). Funding source or other conflicts of interest may raise the risk of bias in design, analysis, and reporting (Viswanathan et al. 2012). We will not consider financial conflict of interest as a risk of bias domain or exclude studies where a conflict is reported. However, this information is collected on included studies and is recommended as a factor to consider when evaluating risk of bias for selective reporting (Viswanathan et al. 2012). We may also conduct stratified analyses to assess the impact of disclosed conflict of interest on findings across the body of evidence although it should

be recognized that newer studies may appear to be biased when compared to older studies because of changes in journal reporting standards (Viswanathan et al. 2012).

Table 3. Risk of bias assessment						
	Experimental Animal	Human Controlled Trials ¹	Cohort	Case-control ²	Cross-sectional	Case Series
Selection						
<p>Was administered dose or exposure level adequately randomized? Randomization requires that each human subject or animal had an equal chance of being assigned to any study group including controls (e.g., use of random number table or computer generated randomization).</p>	X	X				
<p>Was allocation to study groups adequately concealed? Allocation concealment requires that research personnel do not know which administered dose or exposure level is assigned at the start of a study. Human studies also require that allocation be concealed from human subjects prior to entering the study. <i>Note: 1) a question under performance bias addresses blinding of personnel and human subjects to treatment during the study; 2) a question under detection bias addresses blinding of outcome assessors.</i></p>	X	X				
<p>Were the comparison groups appropriate? Comparison group appropriateness refers to having similar baseline characteristics and recruited with the same method and inclusion/exclusion criteria between the groups aside from the exposures and outcomes under study.</p>			X	X	X	
Confounding						
<p>Did the study design or analysis account for important confounding and modifying variables? <i>Note: a parallel question under detection bias addresses reliability of the measurement of confounding variables.</i></p>	X	X	X	X	X	X
<p>Did researchers adjust or control for other exposures that are anticipated to bias results?</p>	X	X	X	X	X	X
Performance						
<p>Were experimental conditions identical across study groups?</p>	X					
<p>Did deviations from the study protocol impact the results? <i>Note: it is recognized that protocol deviations are unlikely to be reported given reporting practices. However, in the long-term collecting this information may generate data that will allow us to empirically assess evidence of this bias.</i></p>	X	X	X	X	X	X
<p>Were the research personnel and human subjects blinded to the study group during the study? Blinding requires that study scientists do not know which administered dose or exposure level the human subject or animal is being given (i.e., study group). Human studies also require blinding of the human subjects when possible.</p>	X	X				

¹Human Controlled Trials (HCTs): studies in humans with a controlled exposure, including Randomized Controlled Trials (RCTs) and non-randomized experimental studies

²Cross-sectional studies include population surveys with individual data (e.g., NHANES) and population surveys with aggregate data (i.e., ecological studies).

DRAFT (April 9, 2013)

	Experimental Animal	Human Controlled Trials	Cohort	Case-control	Cross-sectional	Case Series
Attrition/Exclusion						
Were outcome data incomplete due to attrition or exclusion from analysis? Attrition rates are required to be similar and uniformly low across groups with respect to withdrawal or exclusion from analysis.	X	X	X	X	X	X
Information/Detection						
Were the outcome assessors blinded to study group or exposure level? Blinding requires that outcome assessors do not know the study group or exposure level of the human subject or animal when the outcome was assessed.	X	X	X	X	X	X
Were confounding variables assessed consistently across groups using valid and reliable measures? Consistent application of valid, reliable, and sensitive methods of assessing important confounding or modifying variables is required across study groups. <i>Note: a parallel question under confounding bias addresses whether design or analysis account for confounding. Although consistent measurement of variables can be addressed here under detection bias, we are considering whether to move this question to the confounding domain above. Alternately, we may eliminate this as a separate question and cover it under the question on whether design and analysis account for confounding.</i>	X	X	X	X	X	X
Can we be confident in the exposure characterization? Confidence requires valid, reliable, and sensitive methods to measure exposure applied consistently across groups.	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 and the methods should be applied consistently across groups.	X	X	X	X	X	X
Selective Reporting						
Were all measured outcomes reported?	X	X	X	X	X	X
Other						
Were there any other potential threats to internal validity (e.g., inappropriate statistical methods)? On a project specific basis, additional questions for other potential threats to internal validity can be added and applied to study designs as appropriate.	X	X	X	X	X	X

Determining Tiers of Study Quality

Use of summary or composite scores is not recommended to assess the methodological quality of studies (Guyatt et al. 2011h; Higgins and Green 2011; NTP 2013b; Viswanathan et al. 2012). However, we will utilize a tier system to identify studies that are of high risk of bias on many elements for the purposes of potentially omitting studies from additional consideration in Step 5 and for informing overall judgments on quality of the data across the evidence base. The tiers are not intended to be a strict scoring system. Each study will be described as “1st tier,” “2nd tier,” or “3rd tier,” for risk of bias using the method described below (see also [Table 4](#) for human and [Table 5](#) for animal studies).

For human studies, to be placed in the 1st tier a study must be rated as “definitely low” or “probably low” for the following risk of bias elements AND have at least 50 percent of the other applicable items answered “definitely low” or “probably low” risk of bias ([Table 4](#)).

- Can we be confident in the exposure characterization?
- Can we be confident in the outcome assessment?
- Does the study design or analysis account for important confounding and modifying variables?

For animal studies, to be placed in the 1st tier a study must be rated as “definitely low” or “probably low” for the following risk of bias element AND have at least 50 percent of the other applicable items answered “definitely low” or “probably low” risk of bias ([Table 5](#)).

- Can we be confident in the outcome assessment?

For human studies, to be placed in the 3rd tier a study must be rated as “definitely high” or “probably high” for the following risk of bias elements AND have at least 50 percent of the other applicable items answered “definitely high” or “probably high” risk of bias.

- Can we be confident in the exposure characterization?
- Can we be confident in the outcome assessment?
- Does the study design or analysis account for important confounding and modifying variables?

For animal studies, to be placed in the 3rd tier a study must be rated as “definitely high” or “probably high” for the following risk of bias elements AND have at least 50 percent of the other applicable items answered “definitely high” or “probably high” risk of bias.

- Can we be confident in the outcome assessment?

For both human and animal studies, to be placed in the 2nd tier the study meets neither the criteria for 1st or 3rd tiers.

Table 4. Conceptual schematic for determining tiers of study quality for individual human studies																											
		Risk of Bias Criteria & Ratings																									
		key criteria			other criteria																						
Category	Guidance	Can we be confident in the exposure characterization?	Can we be confident in the outcome assessment?	Did the study design or analysis account for important confounding and modifying variables?	Was administered dose or exposure level adequately randomized?	Was allocation to study groups adequately concealed?	Were the comparison groups appropriate?	Did researchers adjust or control for other exposures that are anticipated to bias results?	Were experimental conditions identical across study groups?	Did deviations from the study protocol impact the results?	Were the research personnel and human subjects blinded to the study group during the study?	Were outcome data incomplete due to attrition or exclusion from analysis?	Were the outcome assessors blinded to study group or exposure level?	Were all measured outcomes reported?	Were there any other potential threats to internal validity (e.g., inappropriate statistical methods)?												
1 st tier	– “definitely low” or “probably low” risk of bias for key items AND “definitely low” or “probably low” risk of bias for ≥50% of other applicable criteria	++	+	++	n/a	n/a	+	+	n/a	+	n/a	+	-	+	++												
2 nd tier	study does not meet criteria for “low” or “high”	example 1	+	-	++	n/a	n/a	+	+	n/a	+	n/a	+	-	+	++											
		example 2	++	+	++	n/a	n/a	-	-	n/a	+	n/a	-	-	-	++											
		example 3	-	--	-	n/a	n/a	+	+	n/a	+	n/a	++	-	++	++											
3 rd tier	– “definitely high” or “probably high” risk of bias for key items AND “definitely high” or “probably high” risk of bias for ≥50% of other applicable criteria	--	--	-	n/a	n/a	--	-	n/a	+	n/a	+	-	--	-												
<p>Risk of bias response options for individual items:</p> <table border="0"> <tr> <td>++</td> <td>definitely low risk of bias</td> <td>-</td> <td>probably high risk of bias</td> </tr> <tr> <td>+</td> <td>probably low risk of bias</td> <td>--</td> <td>definitely high risk of bias</td> </tr> <tr> <td>n/a</td> <td>not applicable based on study design</td> <td></td> <td></td> </tr> </table>																++	definitely low risk of bias	-	probably high risk of bias	+	probably low risk of bias	--	definitely high risk of bias	n/a	not applicable based on study design		
++	definitely low risk of bias	-	probably high risk of bias																								
+	probably low risk of bias	--	definitely high risk of bias																								
n/a	not applicable based on study design																										

Table 5. Conceptual schematic for determining tiers of study quality for individual animal studies		Risk of Bias Criteria & Ratings													
Category	Guidance	key criteria			other criteria										
		Can we be confident in the outcome assessment?	Was administered dose or exposure level adequately randomized?	Was allocation to study groups adequately concealed?	Were the comparison groups appropriate?	Did the study design or analysis account for important confounding and modifying variables?	Did researchers adjust or control for other exposures that are anticipated to bias results?	Were experimental conditions identical across study groups?	Did deviations from the study protocol impact the results?	Were the research personnel and human subjects blinded to the study group during the study?	Were outcome data incomplete due to attrition or exclusion from analysis?	Were the outcome assessors blinded to study group or exposure level?	Can we be confident in the exposure characterization?	Were all measured outcomes reported?	Were there any other potential threats to internal validity (e.g., inappropriate statistical methods)?
1 st tier	– “definitely low” or “probably low” risk of bias for key items AND “definitely low” or “probably low” risk of bias for ≥50% of other applicable criteria	++	+	++	n/a	-	+	++	+	-	++	-	+	++	++
2 nd tier	study does not meet criteria for “low” or “high”	-	+	++	n/a	-	+	++	+	-	++	-	+	++	++
3 rd tier	– “definitely high” or “probably high” risk of bias for key items AND “definitely high” or “probably high” risk of bias for ≥50% of other applicable criteria	-	--	-	n/a	-	-	++	+	-	+	-	-	-	++
Risk of bias response options for individual items:															
++	definitely low risk of bias	-	probably high risk of bias												
+	probably low risk of bias	--	definitely high risk of bias												
n/a	not applicable based on study design														

***In vitro* studies**

To our knowledge no risk of bias tool has been developed for *in vitro* studies and none is proposed in the current protocol. ToxRTool⁶ (Toxicological data Reliability Assessment Tool) appears to be the most recommended tool to assess the “reliability”⁷ of *in vitro* studies, although the tool mainly assesses reporting quality (Bevan and Strother 2012; Schneider et al. 2009). Reporting quality is not considered an appropriate metric to assess risk of bias (Higgins and Green 2011; Viswanathan et al. 2012). For this reason we will not use ToxRTool to assess the internal validity of *in vitro* studies. As a near-term future project it may be possible to develop a risk of bias tool for *in vitro* studies that considers items in the risk of bias tool developed for experimental animal studies and items from ToxRTool that do address internal validity.

DATA DISPLAY

Individual study findings and risk of bias ratings (for animal and human studies) will be summarized in tabular format (see Table 7 for human study example, Table 6 for animal study example, and Table 8 for *in vitro* study example). *Ex vivo*, cellular, genomic, or mechanistic outcomes reported in eligible animal or human studies will be noted in animal and human tables but will primarily be summarized and interpreted along with other “supporting evidence” such as results from *in vitro* studies of adipocytes and data on interactions with key receptors involved in regulating adipogenesis.

Data will typically be presented graphically across collections of studies based on effect size (for human and animal studies) or concentration-specific response for *in vitro* studies (see Figure 2 for human study example, Figure 3 for animal study example, and Figure 4 for *in vitro* study example).

The information summarized in tables and graphs represents the basic information typically used to summarize a study’s findings in literature-based evaluations. Additional study details listed in Table 2 are available in the complete data extraction files.

Software used for data management, analysis, and display

- *Comprehensive Meta-Analysis* (www.meta-analysis.com): Used to conduct meta-analysis and to generate statistics for evaluating consistency of data in Step 5.
- *DistillerSR*[®] (<http://systematic-review.net/>): Industry standard systematic review software

⁶ The ToxRTool worksheets are freely available in Microsoft Excel[®] format at the ECVAM website (http://ihcp.irc.ec.europa.eu/our_labs/eurl-ecvam/archive-publications/toxrtool/toxrtool-toxicological-data-reliability-assessment-tool/?searchterm=toxrtool).

⁷ The reliability categories utilized in the ToxRTool are the same as the Klimisch codes of reliability (Klimisch et al. 1997). It should also be noted that Klimisch’s definition of reliability (1997) differs from the more traditional definition. Reliability was defined by Klimisch as “evaluating the inherent quality of a test report or publication relating to preferably standardized methodology and the way the experimental procedure and results are described to give evidence of the clarity and plausibility of the findings” (Klimisch et al. 1997). More traditional definitions of reliability refer to having stable and/or repeatable measures, for example between different raters using the same tool or consistency in test results from one administration to the next.

DRAFT (April 9, 2013)

- *GraphPad Prism*[®] (www.graphpad.com/scientific-software/prism/): Used to prepare additional graphs, such as x versus y plots.
- *MetaData Viewer* (ntp.niehs.nih.gov/go/tools_metadataviewer)(Boyles et al. 2011): Used to visually display data, mostly based on effect size, and allows for sorting and filtering to help assess patterns of findings in complex data sets.
- *OpenEpi* (<http://www.openepi.com/OE2.3/Menu/OpenEpiMenu.htm>): A free and open source software for epidemiologic statistics that provides statistics for counts and measurements in descriptive and analytic studies, stratified analysis with exact confidence limits, matched pair and person-time analysis, sample size and power calculations, random numbers, sensitivity, specificity and other evaluation statistics, R x C tables, chi-square for dose-response, and links to other sites.
- *Quosa Information Manager* (<http://www.quosa.com>): Used to manage personal biomedical literature collections, including batch retrieval of pdf copies of studies.
- *Universal Desktop Ruler* (www.AVPSoft.com): Used to digitally estimate numerical data from graphs presented in included studies.

Table 6. Example of tabular summary for an human study				
Reference, Study Design & Population	Health Outcome	Exposure	Statistical Analysis	Results
<p>(Carwile and Michels 2011) Study Design: cross-sectional Adults who participated in the 2003/04 and 2005/06 National Health and Nutrition Examination Survey (NHANES) and had a spot urine sample analysed for BPA. N: 2747 Location: US, NHANES national survey Sex (% male): ♂♀(49.6%) Sampling time frame: 2003-2006 Age: 18-74 years Exclusions: pregnant women, participants with missing urinary BPA, creatine, BMI, or covariate data Funding Source: NIH National Research Service Award (NRSA) Author conflict of interest: not reported</p>	<p>Diagnostic and prevalence in total cohort: obesity: BMI ≥ 30 (n=932, 34.3%) overweight: 25 ≤ BMI < 30 (n=864, 31.8%) elevated waist circumference (WC): >102 cm in ♂ or ≥ 88 cm in ♀ (n=1330, 50%) *BMI = body mass index (kg/m²)</p>	<p>Exposure assessment: urine (µg/g creatinine or ng/ml and creatinine as adjustment variable) measured by online SPE-HPLC-MS/MS (Ye 2005) Exposure levels: 2.05 µg/g creatinine (geometric mean), 1.18-3.33 (25-75th percentile) Q1: ≤1.1 ng/ml Q2: 1.2-2.3 ng/ml Q3: 2.4-4.6 ng/ml Q4: >4.7 ng/ml</p>	<p>obesity & overweight: polytomous regression elevated WC: logistic regression Adjustment factors: sex, age, race, urinary creatinine, education, smoking Statistical power: “appears to be adequately powered” based on ability to detect an OR of 1.5 with 80% power using Q1 prevalence of 40.4% obesity, 44.4% overweight, and 46% elevated WC</p>	<p>adjOR (95% CI)</p>
				obesity
				Q2 vs Q1: 1.85 (1.22,2.79)
				Q3 vs Q1: 1.60 (1.05,2.44)
				Q4 vs Q1: 1.76 (1.06,2.94)
				overweight
				Q2 vs Q1: 1.66 (1.21,2.27)
				Q3 vs Q1: 1.26 (0.85,1.87)
				Q4 vs Q1: 1.31 (0.80,2.14)
				elevated WC
Q2 vs Q1: 1.62 (1.11,2.36)				
Q3 vs Q1: 1.39 (1.02,1.90)				
Q4 vs Q1: 1.58 (1.03,2.42)				
statistical power as “appears to be adequately powered” (sample size met), somewhat underpowered (sample size is 75% to <100% of recommended), “underpowered” (sample size is 50 to <75% required), or “severely underpowered (sample size is <50% required)				
RISK OF BIAS ASSESSMENT				
<i>Risk of bias response options for individual items:</i>				
Bias Domain	Criterion			Response
Selection	Was administered dose or exposure level adequately randomized?	n/a		not applicable
	Was allocation to study groups adequately concealed?	n/a		not applicable
	Were the comparison groups appropriate?	++		yes, based on quartiles of exposure
Confounding	Does the study design or analysis account for important confounding and modifying variables?	+		yes (sex, age, race, urinary creatinine, education, smoking), but no adjustment for nutritional quality, e.g., soda consumption
	Did researchers adjust or control for other exposures that are anticipated to bias results?	+		no, but not considered to present risk of bias in general population studies
Performance	Were experimental conditions identical across study groups?	n/a		not applicable
	Did deviations from the study protocol impact the results?	+		no deviations reported
	Were the research personnel and human subjects blinded to the study group during the study?	n/a		not applicable
Attrition	Were outcome data incomplete due to attrition or exclusion from analysis?	+		not considered a risk of bias, excluded observations (≤ 87 for any analysis) based on missing BMI or covariate data
Detection	Were the outcome assessors blinded to study group or exposure level?	++		yes, BPA levels not known at time of outcome assessment
	Were confounding variables assessed consistently across groups using valid	++		yes, used standard NHANES methods

Table 6. Example of tabular summary for an human study			
	and reliable measures?		
	Can we be confident in the exposure characterization?	++	yes, NHANES methods are considered “gold standard” for urinary BPA
	Can we be confident in the outcome assessment?	++	yes, used standard diagnostic criteria
Selective Reporting	Were all measured outcomes reported?	++	yes, primary outcomes discussed in methods were presented results section with adequate level of detail for data extraction
Other	Were there any other potential threats to internal validity (e.g., inappropriate statistical methods)?	++	none identified
			1st Tier for risk of bias

RISK OF BIAS

<i>Risk of bias response options for individual items:</i>	
++	definitely low risk of bias
+	probably low risk of bias
-	probably high risk of bias
--	definitely high risk of bias
n/a	not applicable

Table 7. Example of tabular summary for an animal study								
Reference, Animal Model, and Dosing		Health Outcome	Results					
<p>(Ferguson et al. 2011) Species: rat Strain (Source): Sprague-Dawley (NCTR Breeding colony derived from Charles River Crl: COBS CD (SD) BR Rat, Outbred) Sex: ♂♀ Doses: 0.0025 or 0.025 mg/kg/day BPA Purity (Source): >99% (TCI America) Dosing Period: GD6-21 (via dam) and PND 1-21 to pup Route: oral gavage Diet: low-phytoestrogen chow (TestDiet 5K96 [irradiated pellets], Verified Casein Diet 10 IF; TestDiet), low levels of daidzein (< 0.34 ppm) and genistein (< 0.58 ppm) measured in three separate samples Controls: naïve and vehicle control of 0.3% (by weight) aqueous solution of carboxymethylcellulose (CMC) sodium salt Funding Source: National Center for Toxicological Research/Food and Drug Administration Author conflict of interest: not reported Comments: 0.005 or 0.010 mg/kg/day ethinyl estradiol (EE₂) used as postive control</p>		<p>endpoints: leptin & ghrelin measured by ELISA Age at assessment: PND 21 n = 10-17 for males; 13-15 for females Statistical analysis: two-way ANOVAs with treatment and sex as factors Control for litter effects: one offspring/sex/litter Statistical power: “severely underpowered” to detect a change of 10 - 25% control</p>		group	mean ± SE	% control (95%CI)*	mean ± SE	% control (95%CI)*
				leptin	males		females	
				naïve	5.0 ± 1.0		5.8 ± 1.1	
				vehicle	4.7 ± 0.6		5.5 ± 0.8	
				0.0025 BPA	4.2 ± 0.5	-10.6 (-44.6,23.6)	4.1 ± 0.7	-25.5 (-69.4,18.5)
				0.025 BPA	4.7 ± 1.7	0 (-75.2,75.2)	3.3 ± 0.4	-40 (-77.1, -2.9)
				0.005 EE ₂	3.8 ± 0.8	-19.2 (-67.4,29.1)	4.5 ± 1.2	-18.2 (-77.7,41.4)
				0.010 EE ₂	3.1 ± 0.4	-34.0 (-69.6,1.5)	3.2 ± 0.5	-41.8 (-83.7, 0.02)
				ghrelin				
				naïve	1.913 ± 0.179		2.085 ± 0.357	
				vehicle	1.688 ± 0.139		1.953 ± 0.250	
				0.0025 BPA	1.567 ± 0.227	-7.2 (-39.8, 25.5)	1.693 ± 0.170	-13.3 (-45.2,18.6)
				0.025 BPA	1.760 ± 0.193	4.3 (-22.6, 31.2)	1.508 ± 0.140	-22.7 (-53.8,8.2)
				0.005 EE ₂	1.755 ± 0.210	4.0 (-24.5,32.4)	1.823 ± 0.183	-6.6 (-38.5,25.2)
				0.010 EE ₂	1.667 ± 0.201	-1.2 (-29.9,27.4)	1.623 ± 0.184	-16.9 (-50.4,16.6)
* average group size (rounded up when needed) was used to estimate percent control response (14 for males; 14 for females)								
statistical power as “appears to be adequately powered” (sample size met), somewhat underpowered (sample size is 75% to <100% of recommended), “underpowered” (sample size is 50 to <75% required), or “severely underpowered (sample size is <50% required)								
RISK OF BIAS ASSESSMENT								
<i>Risk of bias response options for individual items:</i>								
Bias Domain	Criterion		Response					
Selection	Was administered dose or exposure level adequately randomized?	++	yes, “randomly assigned to treatment within their body weight stratum”					
	Was allocation to study groups adequately concealed?	+	not reported, but lack of adequate allocation concealment at study start not expected to appreciably bias results					
	Were the comparison groups appropriate?	n/a	not applicable					
Confounding	Does the study design or analysis account for important confounding and modifying variables?	+	no, neither litter size or body weight considered as covariates in analysis, but not clear these need to be considered for endpoints reported in study					
Performance	Did researchers adjust or control for other exposures that are anticipated to bias results?	++	yes, low phytoestrogen diet and polysulfone cages with only trace BPA used; levels of BPA in other housing equipment measured					
	Were experimental conditions identical across study groups?	+	assumed yes					
	Did deviations from the study protocol impact the results?	+	no deviations reported					
	Were the research personnel and human subjects blinded to the study group during the study?	+	not reported, but lack of adequate allocation concealment during conduct of study not feasible and not expected to appreciably bias results for this study					

Attrition	Were outcome data incomplete due to attrition or exclusion from analysis?	+	yes, but dead or missing (assumed cannibalized) offspring documented and were generally evenly distributed across groups
Detection	Were the outcome assessors blinded to study group or exposure level?	+	not reported, but not considered a risk of bias for these endpoints (hormone levels) because measurement is not subjective
	Were confounding variables assessed consistently across groups using valid and reliable measures?	n/a	not applicable given that confounding/modifying variables were not included
	Can we be confident in the exposure characterization?	++	yes, purity >99% and dosing solutions measured and were very close to target doses
	Can we be confident in the outcome assessment?	++	yes, used standard kits and inter assay coefficients of variation <4%
Selective Reporting	Were all measured outcomes reported?	++	yes, primary outcomes discussed in methods were presented in results section with adequate level of detail for data extraction
Other	Were there any other potential threats to internal validity (e.g., inappropriate statistical methods)?	++	none identified, potential litter effects were controlled for experimentally
			1st Tier for risk of bias

RISK OF BIAS

<i>Risk of bias response options for individual items:</i>	
++	definitely low risk of bias
+	probably low risk of bias
-	probably high risk of bias
--	definitely high risk of bias
n/a	not applicable

Table 8. Sample tabular summary for an <i>in vitro</i> study		
Reference, Model, and Treatment	Endpoint	Concentrations Tested (μM) and Findings
(Hugo et al. 2008) Species: human Cell-line/Source: explants from breast (8 women undergoing breast reduction surgery) and abdominal subcutaneous adipose (9 women undergoing abdominoplasty) Sex: ♀ Concentrations: 0.0001, 0.001, 0.01, 0.1 μM BPA Purity (Source): >99% (Sigma-Aldrich) Vehicle: <0.001% EtOH Treatment Period: 6h Replicates: Results based on mean of 6 determinations Funding Source: NIH, Department of Defense, Susan G. Komen Breast Cancer Foundation Author conflict of interest: authors declare no competing interest Comments: non-monotonic dose response; response consistent with estradiol positive control	adiponectin release, breast adipose (ng/100 mg/6h):	0.0001(↓), 0.001(↓), 0.01, 0.1
	adiponectin release, abdominal adipose (ng/100 mg/6h):	0.0001(↓), 0.001(↓), 0.01, 0.1
↑ = statistically significant increase as reported by authors, ↓ = statistically significant decrease as reported by authors		

Figure 2. Sample display of human data by effect size

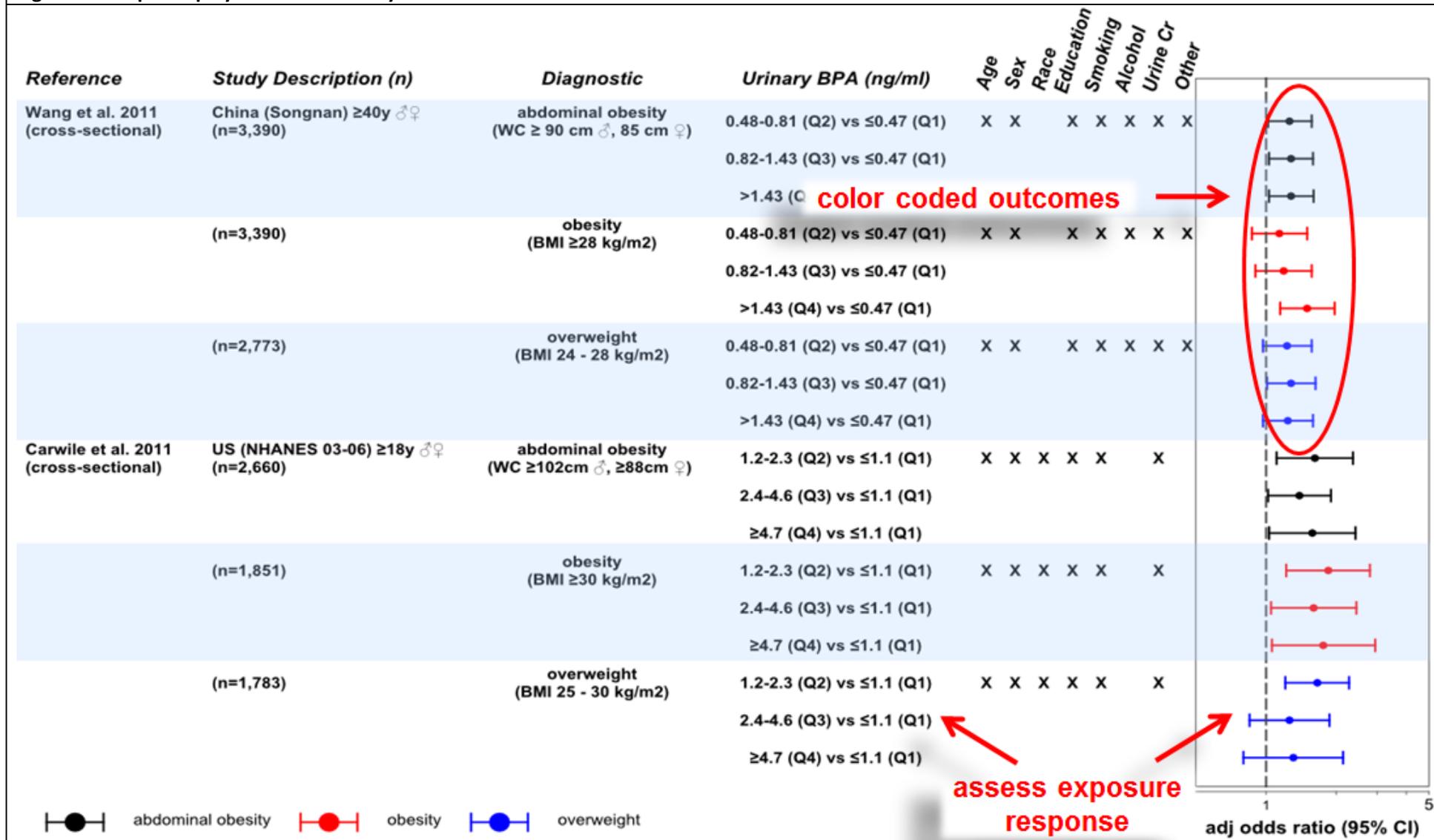


Figure 3. Sample display of animal data by effect size and stratified by species and sex

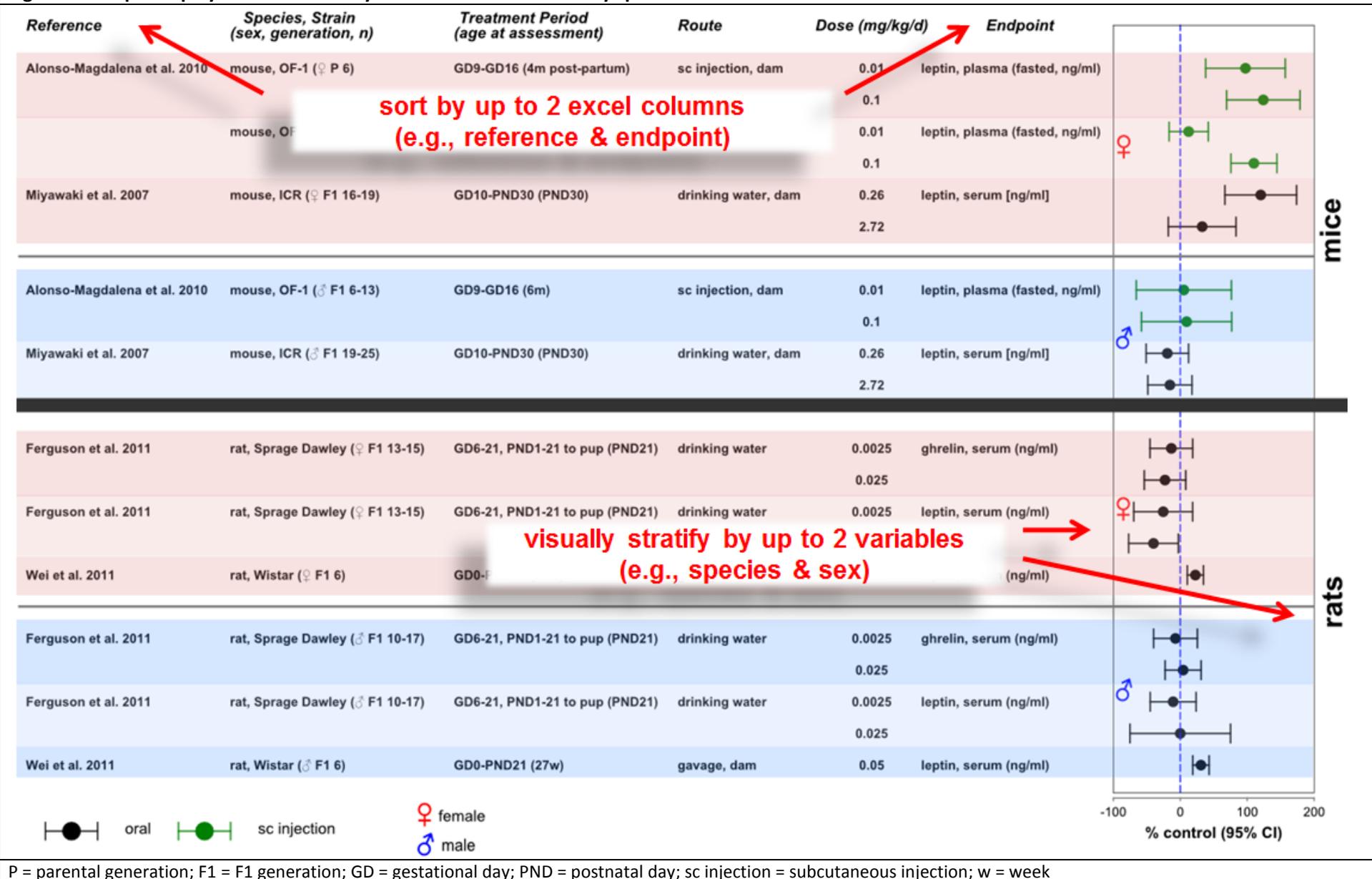
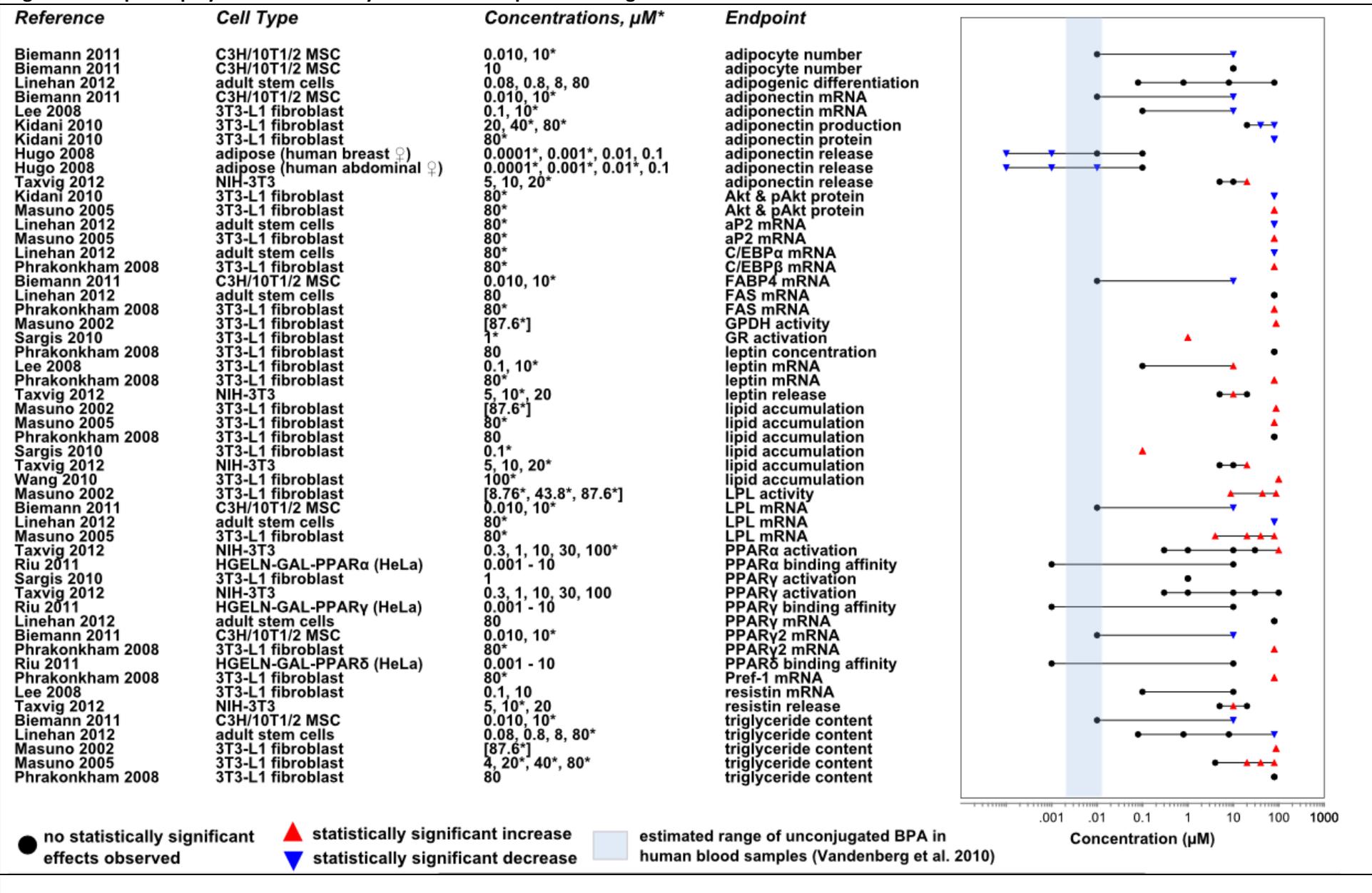


Figure 4. Sample display of *in vitro* data by concentration-specific findings



STEP 5: RATE CONFIDENCE IN BODY OF EVIDENCE

A confidence rating for a given health outcome is developed by considering the strengths and weaknesses in a collection of human and animal studies that constitute the body of evidence. The rating reflects confidence that the study findings accurately reflect the true association between exposure to a substance and an effect. The confidence rating approach described below [(NTP 2013a), Figure 5] is primarily based on guidance from the Grading of Recommendations Assessment, Development and Evaluation (GRADE) Working Group (Balshem et al. 2011), a framework applied most often to evaluate the quality of evidence and strength of recommendations for health care intervention decisions based on human studies (typically randomized clinical trials). The appeal of the GRADE framework is that it is (1) widely used (Guyatt et al. 2011f), (2) conceptually similar to the approach used by the Agency for Healthcare Research and Quality (AHRQ 2012) for grading the strength of a body of evidence of human studies, and (3) the Cochrane Collaboration has adopted the principles of the GRADE system for evaluating the quality of evidence for outcomes reported in systematic reviews (Higgins and Green 2011). However, none of these existing frameworks (GRADE, AHRQ, and the Cochrane Collaboration) address approaches for considering animal studies or *in vitro* studies (defined here as other than whole animal studies, and including cell systems, computational toxicology, high throughput screening data, and *in silico* methods). In addition, the guidance provided by GRADE, AHRQ, and the Cochrane Collaboration is less developed for observational human studies compared to randomized clinical trials. For these reasons the draft OHAT approach includes a number of refinements to GRADE that were considered necessary in order to accommodate our need to integrate data from multiple evidence streams (human, animal, *in vitro*) and focus on observational human studies rather than the randomized clinical trials more commonly encountered in the health care intervention field. Embedded within the GRADE approach is consideration of elements of an association that are consistent with causation as discussed by Bradford Hill (Hill 1965; Schünemann et al. 2011).

The framework described below only applies to human and animal studies. To our knowledge there is no analogous model to develop confidence ratings for “supportive evidence” based on *in vitro*, *ex vivo*, cellular, genomic, or mechanistic outcomes. Our current approach for considering these types of supportive evidence is described in a later section of this protocol. But, as a near-term future research effort we are interested in developing a framework for “supportive evidence” that is conceptually similar to the approach applied to human and animal studies.

Four descriptors are used to indicate the level of confidence in the body of evidence for human and animal studies:

- **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 likely to be different than the apparent relationship.
- **Very 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.

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 “Very Low Confidence” suggests that further research is very likely to impact confidence in the apparent relationship.

For each outcome, collections of studies are given an initial confidence rating by key study design features (Figure 5). This initial rating is downgraded for factors that decrease confidence (risk of bias, unexplained inconsistency, directness or applicability, precision, and publication bias) and upgraded for factors that increase confidence in the results (large magnitude of effect, dose-response, consistency across study designs/populations/experimental animal models, and consideration of confounding or other biases that increase our confidence in the association or effect). Consideration of consistency across study designs, human populations, or experimental animal models is not included in the GRADE guidance (Guyatt et al. 2011a) but was considered appropriate by the NTP BSC Working Group Report on the Draft NTP Approach⁸. Confidence ratings will be summarized in evidence profile tables (see [Table 9](#) and [Table 10](#) for examples).

Each member of the review team will independently develop confidence ratings using the guidance provided below. Members of the review team will then compare their results and reach decisions by consensus discussion. If needed, additional technical input can be obtained. The scientific judgments on whether or not to downgrade or upgrade for each factor will be documented for each outcome in the evidence profile table. The confidence ratings will then be used to develop conclusions related to (1) evidence of health effect and research needs, or (2) evidence of health effect, research needs and hazard identification label, depending on the extent of the available literature.

Planned interim analyses

We will conduct an interim analysis after assessment of risk of bias for individual studies to determine whether confidence ratings will be developed for the primary purpose of developing hazard identification conclusions or to identify research needs. If very few studies are identified that met the eligibility criteria, then a hazard identification analysis will likely not be conducted, especially in cases where those few studies are in the 3rd tier for risk of bias. In this circumstance, confidence ratings will be reached in order to identify key research needs. The outcome of this interim analysis will be noted as a revision to the protocol.

⁸(NTP 2012) see “Meeting Materials and Public Comments”, then “NTP BSC Working Group Report on the Draft NTP Approach”

Figure 5. Rating confidence in the body of evidence

Initial Confidence by Key Features of Study Design	Factors Decreasing Confidence	Factors Increasing Confidence	Confidence in the Body of Evidence
High (++++) 4 Features	<ul style="list-style-type: none"> ❖ Risk of Bias ❖ Unexplained Inconsistency ❖ Indirectness ❖ Imprecision ❖ Publication Bias 	<ul style="list-style-type: none"> ❖ Large Magnitude of Effect ❖ Dose Response 	High (++++)
Moderate (+++) 3 Features		<ul style="list-style-type: none"> ❖ All Plausible Confounding <ul style="list-style-type: none"> • Studies report an effect and residual confounding is toward null • Studies report no effect and residual confounding is away from null 	Moderate (+++)
Low (++) 2 Features		<ul style="list-style-type: none"> ❖ Consistency <ul style="list-style-type: none"> • Across animal models or species • Across dissimilar populations • Across study design types 	Low (++)
Very Low (+) ≤1 Features		<ul style="list-style-type: none"> ❖ Other <ul style="list-style-type: none"> e.g., particularly rare outcomes 	Very Low (+)

- Features**

 - Controlled exposure
 - Exposure prior to outcome
 - Individual outcome data
 - Comparison group used

Note: if the only available body of evidence receives a “Very Low Confidence” rating, then conclusions for those outcomes will not move on to Step 6

This figure is reproduced from the Step 5 of the Figure in the Draft OHAT Approach – February 2013 (available at <http://ntp.niehs.nih.gov/go/38673>)

DRAFT (April 9, 2013)

Table 9. Example human evidence profile table								
Outcome	Factors considered in establishing confidence ratings for a body of evidence					Summary of findings	↑Consistency across types of evidence	Confidence in evidence
<i>human prospective cohort studies (n=)</i>								
obesity	risk of bias	inconsistency	indirectness	imprecision	publication bias	narrative or results of meta-analysis	inconsistent (0) consistent within evidence stream (+1)	
	not likely (0) serious (-1) very serious (-2)	no serious (0) serious (-1) very serious (-2)	no serious (0) serious (-1) very serious (-2)	no serious (0) serious (-1) very serious (-2)	undetected (0) strongly suspected (-1)			
	magnitude of effect	dose-response	plausible confounding	other				
	large (+1) very large (+2)	evidence of gradient (+1)	when effect found would decrease magnitude of effect, or if effect not found would lead towards overestimating effect (+1)	(+1)				
<i>human cross-sectional studies (n=)</i>								
obesity	risk of bias	inconsistency	indirectness	imprecision	publication bias	narrative or results of meta-analysis		
	not likely (0) serious (-1) very serious (-2)	no serious (0) serious (-1) very serious (-2)	no serious (0) serious (-1) very serious (-2)	no serious (0) serious (-1) very serious (-2)	undetected (0) strongly suspected (-1)			
	magnitude of effect	dose-response	plausible confounding	other				
	large (+1) very large (+2)	evidence of gradient (+1)	when effect found would decrease magnitude of effect, or if effect not found would lead towards overestimating effect (+1)					

Table 10. Example experimental animal evidence profile table								
Outcome	Factors considered in establishing confidence ratings for a body of evidence					Summary of findings	↑Consistency across types of evidence	Confidence in evidence
<i>mouse studies (n=)</i>								
fat mass	risk of bias	inconsistency	indirectness	imprecision	publication bias	narrative or results of meta-analysis	inconsistent (0) consistent within evidence stream (+1)	
	not likely (0) serious (-1) very serious (-2)	no serious (0) serious (-1) very serious (-2)	no serious (0) serious (-1) very serious (-2)	no serious (0) serious (-1) very serious (-2)	undetected (0) strongly suspected (-1)			
	magnitude of effect	dose-response	plausible confounding	other				
	large (+1) very large (+2)	evidence of gradient (+1)	when effect found would decrease magnitude of effect, or if effect not found would lead towards overestimating effect (+1)					
<i>zebra fish studies (n=)</i>								
fat mass	risk of bias	inconsistency	indirectness	imprecision	publication bias	narrative or results of meta-analysis		
	not likely (0) serious (-1) very serious (-2)	no serious (0) serious (-1) very serious (-2)	no serious (0) serious (-1) very serious (-2)	no serious (0) serious (-1) very serious (-2)	undetected (0) strongly suspected (-1)			
	magnitude of effect	dose-response	plausible confounding	other				
	large (+1) very large (+2)	evidence of gradient (+1)	when effect found would decrease magnitude of effect, or if effect not found would lead towards overestimating effect (+1)					

Initial confidence based on study design

An initial confidence rating is determined by the ability of the study design to address whether exposure preceded and was associated with the outcome (Figure 5, column 1). This ability is reflected in the presence or absence of four key study design features that determine initial confidence ratings: (1) the exposure to the substance is experimentally controlled, (2) the exposure assessment represents exposures occurring prior to the development of the outcome, (3) the outcome is assessed on the individual level (i.e., not population aggregate data like that reported in ecological studies), and (4) a comparison group is used within the study. This first key feature, “controlled exposure” reflects the ability of experimental studies in humans and animals to largely eliminate confounding by randomizing allocation of exposure. Therefore, these studies will usually have all four features and receive an initial rating of “High Confidence.”

Observational studies do not have controlled exposure and are differentiated by presence or absence of the three remaining study design features. For example, prospective cohort studies usually have all three remaining features and receive an initial rating of “Moderate Confidence.” Observational animal studies could be considered using these same study design features. See OHAT Approach – February 2013 for additional examples and discussion (available at <http://ntp.niehs.nih.gov/go/38673>).

These study design features are distinct from the risk of bias assessment. The initial ratings are the starting points that reflect the general strengths of study design features, and then studies are evaluated for factors that would downgrade or upgrade confidence in the evidence for a given outcome.

Domains that can reduce confidence

On an outcome-by-outcome basis, five properties for a body of evidence (risk of bias across studies, unexplained inconsistency, indirectness, imprecision, and publication bias) are used to determine if the initial confidence rating should be downgraded (Figure 5, column 2).

Risk of bias across studies

Risk of bias criteria for individual studies was described earlier in the protocol. In this step, risk of bias for a given health outcome is considered across studies.

Summary of risk of bias ratings

A visual summary of the risk of bias ratings for each outcome will be prepared, one for human studies and one for animal studies (see Table 11 for a hypothetical summary of risk of bias for a set of experimental animal studies). This type of summarization is used to get an appreciation for what the general strengths and weaknesses are for studies included in the analysis. In addition, it highlights particular risk of bias items that could be explored when evaluating inconsistency within the evidence base.

This analysis can also be useful when considering risk of bias in context of direction of bias and magnitude of effect. For example, if most human studies are high risk of bias due to non-differential misclassification of exposure this will generally bias results towards the null, but differential misclassification can bias towards or away from the null so careful consideration of the source, direction, and magnitude of potential biases in the body of evidence is required (Szklo and Nieto 2007).

Table 11. Visual summary of ratings for each risk of bias item for a given health outcome (hypothetical summary for 10 experimental animal studies)										
	20%		40%		60%		80%			100%
Was administered dose or exposure level adequately randomized?	++	+	-	-	-	-	-	-	-	--
Was allocation to study groups adequately concealed?	++	+	+	+	-	-	--	--	--	--
Does the study design or analysis account for important confounding and modifying variables?	++	++	++	+	+	-	-	-	--	--
Did researchers adjust or control for other exposures that are anticipated to bias results?	++	++	++	+	+	+	-	-	--	--
Were experimental conditions identical across study groups?	++	++	++	++	+	+	+	+	+	--
Did deviations from the study protocol impact the results?	++	++	++	+	+	+	+	+	+	-
Were the research personnel and human subjects blinded to the study group during the study?	+	+	-	-	--	--	--	--	--	--
Were outcome data incomplete due to attrition or exclusion from analysis?	++	++	++	+	+	+	+	-	--	--
Were the outcome assessors blinded to study group or exposure level?	++	+	+	-	-	--	--	--	--	--
Were confounding variables assessed consistently across groups using valid and reliable measures?	++	++	++	+	+	+	+	-	-	-
Can we be confident in the exposure characterization?	++	++	++	+	+	+	+	-	-	-
Can we be confident in the outcome assessment?	++	++	++	++	++	+	+	+	-	--
Were all measured outcomes reported?	++	++	++	+	+	+	+	-	-	-

++	definitely low risk of bias
+	probably low risk of bias
-	probably high risk of bias
--	definitely high risk of bias
n/a	not applicable (n/a)

Consideration of whether to downgrade confidence based on risk of bias

The strategy for assessing risk of bias differs depending on whether confidence ratings will be primarily used to identify research needs or to reach conclusions on hazard identification.

a. Confidence ratings to identify research needs

All studies providing data on a given health outcome, regardless of risk of bias tier for the study, will be considered when developing confidence ratings. We will use the approach described earlier for categorizing individual studies as “1st tier,” “2nd tier,” or “3rd tier” risk of bias and the guidance presented in [Table 12](#) when considering the extent to which confidence should be downgraded based on risk of bias across studies.

Table 12. Guidance on when to downgrade for risk of bias across studies when confidence ratings are used to identify research needs												
Downgrade	Guidance	Example										
None	most studies are 1 st tier	1 st	2 nd	2 nd	3 rd							
-1 (serious)	most studies are 2 nd tier	1 st	1 st	1 st	1 st	2 nd	3 rd	3 rd				
-2 (very serious)	most studies are 3 rd tier	1 st	1 st	2 nd	3 rd							

b. Confidence ratings to reach hazard identification conclusions

We will omit the 3rd tier risk of bias studies from consideration when determining confidence ratings. However, such studies will still be considered part of the evidence base and included in the data extraction and summarized in appendix tables. The guidance provided in Table 13 will be used to determine the extent to which confidence for a given health outcome should be downgraded based on risk of bias across studies. Please note the maximum downgrade for risk of bias would be one level after omission of the 3rd tier risk of bias studies.

Table 13. Guidance on when to downgrade for risk of bias across studies when confidence ratings are used to reach hazard identification conclusions												
Downgrade	Guidance	Example										
None	most studies are 1 st tier	1 st	2 nd	2 nd	2 nd							
-1 (serious)	most studies are 2 nd tier	1 st	1 st	1 st	1 st	2 nd						

Although 3rd tier risk of bias studies will be omitted from the confidence rating phase, we will conduct a “sensitivity” analysis to assess the extent to which inclusion of the studies of 3rd tier risk of bias studies might have obscured findings from studies considered in the 1st and 2nd tier for risk of bias. This will be done by comparing the consistency of findings from studies of 3rd tier risk of bias with findings from studies of 1st and 2nd tier for risk of bias. If a meta-analysis is conducted we will conduct a sensitivity analysis to address this issue. When a meta-analysis is not feasible or inappropriate, we will use MetaData Viewer to stratify studies based on risk of bias category to visually compare and assess the impact of omitting studies.

Unexplained inconsistency

Inconsistency, or large variability in the direction or magnitude of individual study effect estimates that cannot be explained, reduces confidence in the body of evidence (AHRQ 2012; Guyatt et al. 2011d). No single measure of consistency is ideal or sufficient, so we will consider the following factors when determining whether to downgrade for inconsistency: (1) similarity of point estimates, (2) extent of overlap between confidence intervals, and (3) results of statistical tests of heterogeneity, e.g., Cochran's Q (chi-square, χ^2), I^2 or τ^2 (tau square). See [Table 14](#) for examples and additional details on guidance.

There will be no downgrade for inconsistency in cases where the evidence base consists of a single study. In this case consistency is unknown and will be documented as such in the summary of findings table.

Sources of inconsistency across studies will be explored, including consideration of population or animal model (e.g., cohort, species, strain, sex, lifestage at exposure and assessment); exposure or treatment duration, level, or timing relative to outcome; study methodology (e.g., route of administration, dietary phytoestrogen content, use of high fat diet, methodology used to measure health outcome); and risk of bias. We will also conduct analyses to evaluate whether source of funding or disclosed conflict of interest may be associated with the studies' results.

The following statistical measures will be used to help assess consistency across studies that are similar in study design, dose or exposure levels, methods to assess exposure, and the health outcome:

Cochran's Q: A statistical test for heterogeneity distributed as a chi-square (χ^2) statistic that tests the null hypothesis that all studies have the same underlying magnitude of effect, thus a low p-value ($P < 0.1$) indicates significant heterogeneity (Higgins and Green 2011). The level of significance for χ^2 is often set at 0.1 due to low power of the test to detect heterogeneity. A rule of thumb is if the χ^2 value is larger than the degrees of freedom (df, number of studies minus 1), then heterogeneity is present. The χ^2 statistic has low power to detect heterogeneity when there are few studies or, conversely, it may detect heterogeneity of minimal biological or clinical importance when the number of studies is large.

Tau square (T^2 , τ^2 , τ^2): An estimate of the between-study variance in a random-effects meta-analysis. A $\tau^2 > 1$ suggests presence of substantial statistical heterogeneity.

I^2 : An index that is not dependent on the number of studies and can be used to quantify the impact of heterogeneity, providing a measure of the degree of inconsistency in the studies' results ($I^2 = [(Q-df)/Q] \times 100\%$). I^2 represents the percentage of the total variation across studies due to heterogeneity rather than sampling error or chance, with values ranging from 0% (no observed heterogeneity) to 100%.

Thresholds for the interpretation of I^2 can be misleading since the importance of the observed value of I^2 depends on the (i) magnitude and direction of effects, and (ii) strength of evidence for heterogeneity (e.g. P value from the chi-squared test, or a confidence interval for I^2). A rough guide to interpretation is as follows (Higgins and Green 2011):

- 0% to 40%: might not be important;
- 30% to 60%: may represent moderate heterogeneity;
- 50% to 90%: may represent substantial heterogeneity;
- 75% to 100%: considerable heterogeneity

Quantitative data synthesis

We will consider performing meta-analyses if we find three or more unique studies with sufficient study level and methodological homogeneity with respect to population or animal model, study design, study duration, dose or exposure level, and health outcome (Fu et al. 2011). Situations in which it may not be appropriate to include a study are: data on exposure or outcome are too different to be combined; there are concerns about high risk of bias, or other circumstances which may indicate that averaging study results would not produce meaningful results. Although certain studies may be excluded from a meta-analysis based on these concerns, these studies will be considered in a qualitative synthesis.

The following fields from the data extraction will be used in the meta-analysis:

- Concentrations of BPA measured/estimated for each exposure or treatment group
- Estimates of effect for obesity and adiposity outcomes for each group
- Upper and lower 95% confidence intervals for outcome measurements for each group

If the type or source of exposure data differs among studies (e.g. biomonitoring data or estimates from dietary intake), then the data will be normalized to the same metric of concentration when possible. If there is a mixture of outcome measurements such that some data are expressed as an empirical or percent change in outcome measurement while other data are expressed as a prevalence of the outcome (such as prevalence of obesity), then the possible combining of these data into one analysis will be explored. For binary outcomes, we would attempt to convert to odds ratio (OR) or relative risk (RR) as the effect measure. For continuous outcomes, we will calculate mean difference and standardized effect sizes, and percent control response. The choice of effect measure is determined primarily by the scale of the available data (Fu et al. 2011). Mean differences can be used if findings are reported with the same or similar scale, standardized mean difference (SMD) are typically used when the outcome is measured using different scales. Percent control response can be helpful to assess dissimilar but related outcomes measured with different scales, e.g., fat mass and percent fat mass; however we would likely not attempt to conduct a meta-analysis on dissimilar health outcomes.

If we are unable to obtain the data for conversion from the study report or authors, then the data will be analyzed separately, as continuous or dichotomous outcomes. Our review team includes a statistician who will be consulted to confirm appropriateness of data conversions and to discuss the feasibility and appropriateness of conducting meta-analysis.

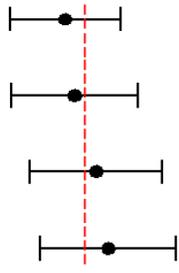
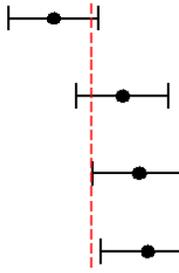
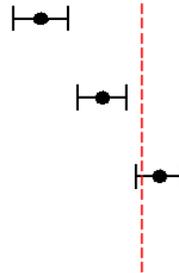
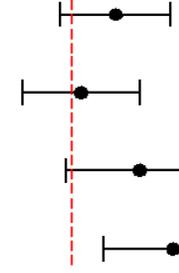
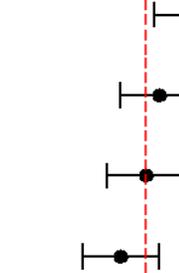
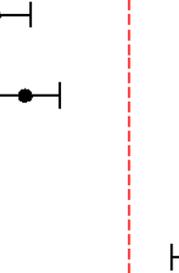
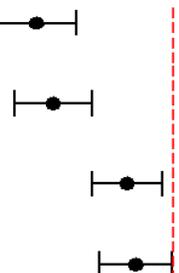
Meta-analysis would be conducted using Comprehensive Meta-Analysis (CMA) software (Biostat, Inc., Englewood, NJ) random-effects model. If there is significant study level heterogeneity or the I^2 statistic is greater than 50%, we will consider conducting subgroup analyses or random effects meta-regression in an attempt to explain the heterogeneity if there are at least 6–10 studies for a continuous variable and at least 4 studies for a categorical variable (Fu et al. 2011). When it is inappropriate or not feasible to conduct a meta-analysis or meta-regression, we will visually display findings using Meta Data Viewer to help present a qualitative synthesis.

Planned interim analyses

DRAFT (April 9, 2013)

The statistical power of studies will also be considered if we detect inconsistency of findings across studies. If we are using confidence ratings for hazard identification purposes and not conducting a meta-analysis, then we will consider omitting studies not reporting an association that are “severely underpowered” from consideration during the confidence rating phase. As described in [Table 2](#), a study will be considered “severely underpowered” if sample size is <50% required to (1) detect a 20% change from control or referent group response for continuous data, or (2) relative risk or odds ratio of 1.5 for categorical data calculated based on the prevalence of exposure or prevalence of outcome in the control or referent group reported in the study. Although no effect/association studies that are significantly underpowered may be omitted from this phase, we will conduct a visual “sensitivity” analysis using MetaData Viewer to assess the extent to which inclusion of the underpowered studies might have obscured ratings based on consideration of studies with better statistical power. **Note:** Consideration of the statistical power of studies remaining in the confidence ratings is formally considered as part of evaluating imprecision (see below).

Table 14. Factors to consider when considering consistency of results

No downgrade	One level downgrade (serious)	Two level downgrade (very serious)
<ul style="list-style-type: none"> • Point estimates similar • Confidence intervals overlap • Statistical heterogeneity is non-significant ($p > 0.1$) • I^2 of $\leq 50\%$ 	<ul style="list-style-type: none"> • Point estimates vary • Confidence intervals show minimal overlap • Statistical heterogeneity has low p-value ($p \leq 0.1$) • I^2 of $> 50\%$ to 75% 	<ul style="list-style-type: none"> • Point estimates vary widely • Confidence intervals show minimal or no overlap • Statistical heterogeneity has low p-value ($p \leq 0.1$) • I^2 of $> 75\%$
<p>Example A</p>  <p>χ^2 p-level = 0.767; $I^2 = << 1\%$; $\tau^2 = << 1$</p>	<p>Example A</p>  <p>χ^2 p-level = 0.017; $I^2 = 71\%$; $\tau^2 = 0.044$</p>	<p>Example A</p>  <p>χ^2 p-level = < 0.001; $I^2 = 98\%$; $\tau^2 = 1.022$</p>
<p>Example B</p>  <p>χ^2 p-level = 0.241; $I^2 = 29\%$; $\tau^2 = 0.046$</p>	<p>Example B</p>  <p>χ^2 p-level = 0.068; $I^2 = 58\%$; $\tau^2 = 0.025$</p>	<p>Example B</p>  <p>χ^2 p-level = < 0.001; $I^2 = 98\%$; $\tau^2 = 0.774$</p>
<p>Example C</p>  <p>χ^2 p-level = < 0.001; $I^2 = 86\%$; $\tau^2 = 0.111$</p> <p>*in this case there is less concern for numerical estimates of heterogeneity because point estimates are in the same direction</p>		

Directness and applicability

Directness refers to the applicability, external validity, generalizability, and relevance of the studies in the evidence base to address the objectives of the evaluation (AHRQ 2012; Guyatt et al. 2011c).

To determine whether to downgrade confidence based on indirectness we will consider factors related to (1) relevance of the animal model to human health; (2) directness of the endpoints to the primary health outcome(s); (3) nature of the exposure in human studies and route of administration in animal studies; and (4) duration of treatment in animal studies and length of time between exposure and outcome assessment in animal and prospective human studies. Studies will be downgraded one level if they are considered indirect based on any one of these factors. Studies will be downgraded two levels if they are considered indirect based on 2 or more factors. A summary of the guidance below is presented in tabular format in [Table 15](#) for human studies and [Table 16](#) for animal studies.

Consideration of dose or exposure level

We recognize that the level of dose or exposure is an important factor when considering the relevance of study findings. However, it is not considered as a factor under directness for the purposes of reaching confidence ratings for evidence of health effects. In OHAT's process this consideration occurs after hazard identification as part of reaching a "level of concern" conclusion, where the health effects are interpreted in the context of what is known regarding the extent and nature of human exposure (Jahnke et al. 2005; Medlin 2003; Shelby 2005; Twombly 1998). We do not currently have updated guidance on how the hazard identification conclusions will be used to reach level of concern conclusions. However, that is OHAT's next phase of work and we expect to have updated draft guidance for reaching level of concern conclusions during FY2014.

While not the case in the current protocol, it is possible that the question being addressed in an evaluation is directed towards a specific range of doses, e.g., "low dose" or "occupationally-relevant". In those cases, dose or exposure levels considered irrelevant to the evaluation topic can be identified in the inclusion and exclusion criteria for study eligibility.

Planned interim analyses

The guidance below is meant to be as comprehensive as possible, but it is possible that during the course of the evaluation we will identify model systems or outcomes in the included studies that are relevant to our question of interest, but have not been *a priori* identified. We will conduct an interim analysis after data extraction to update this guidance to include model systems, primary or secondary health outcomes, or routes of exposure not covered below.

We anticipate that decisions on whether to downgrade for directness in non-traditional or novel model systems and health outcomes will likely be difficult to support based on empirical data. Our strategy in these cases will be to identify any relevant information and engage technical experts, as needed, in order to update the guidance provided below.

a. Relevance of the animal model to human health

- *Rats, mice, and other mammalian model systems:* No limitations of these model systems for our questions of interest have been identified *a priori*. Thus, studies conducted in mammalian model

DRAFT (April 9, 2013)

systems will be assumed to be relevant for humans (i.e., not downgraded) unless compelling data to the contrary is identified during the course of the evaluation. We are not aware of studies that have assessed adipogenic effects of BPA in transgenic animals. However, if encountered, the directness of the transgenic model system will be assessed on a case by case basis and evaluated for directness during the planned interim analysis described above.

- *Fish, amphibian, C. elegans, and other non-mammalian model systems:* Use of these model systems to address human health is not as well-established as use of the mammalian model systems. For this reason, studies conducted in fish, amphibian, *C. elegans* and other non-mammalian model systems will be downgraded one level. This decision will be re-assessed during the evaluation if information is identified that directly addresses the ability of any of these model systems to predict response in mammalian model systems or humans. If any of the models are considered reasonably predictive, then we will not downgrade based on directness for use of that model system. Our assessment of “predictive” is based on reasonable scientific judgment and does not require formal validation of the nature undertake to gain regulatory acceptance of alternative methods.

b. Health outcomes

- *Primary health outcomes:* The primary outcomes for this evaluation were selected based on their directness for our question of interest. Thus, there will be no downgrades for these outcomes.
- *Secondary health outcomes:* The secondary outcomes for this evaluation were selected because they are relevant to our question of interest; however, they are considered upstream indicators, intermediate outcomes, or related measures to our primary outcomes. Thus, secondary outcomes will be one downgraded one level on their directness for our question of interest. This decision may be re-assessed during the evaluation if information is identified that indicates a secondary outcome is sufficiently predictive or indicative of a primary outcome to serve as a surrogate measure. In this case, the secondary measure would be re-designated as a primary outcome and the change noted in the history of protocol revisions.

c. Exposure

- *Human studies:* All exposure levels and scenarios encountered in the human studies (e.g., general population, occupational settings, etc.) will be considered direct and not downgraded.
- *Dose levels used in animal studies:* There will be no downgrade for dose level used in experimental animal studies. As noted above, we recognize that the level of dose or exposure is an important factor when considering the relevance of animal findings to human health. However, in OHAT’s process the relevance of the dose or exposure level occurs after hazard identification as part of reaching a “level of concern” conclusion.
- *Route of administration in animal studies:* All of the most commonly used routes of administration will be considered direct for the purposes of establishing confidence ratings. We recognize that some of these exposure routes may only be relevant for certain human sub-populations. However, in OHAT’s process this consideration occurs after hazard identification as part of reaching a “level of concern” conclusion.
 - Oral (no downgrade) - Gavage or feeding studies are considered relevant because the primary route of exposure in most humans is oral (NTP 2008b; WHO 2011b)
 - Dermal (no downgrade) – Dermal exposure is considered relevant, e.g., BPA can be detected in thermal paper products, such as cash register receipts (EPA 2012).

DRAFT (April 9, 2013)

- Subcutaneous injection (no downgrade) – A route of administration that bypasses first pass metabolism is relevant for certain exposure scenarios, e.g., BPA can be detected in certain medical devices (Calafat et al. 2009).
- Inhalation (no downgrade) – Inhalation studies are considered relevant, especially to occupational cohorts.
- Intracranial or intraperitoneal injection, water for aquatic species, or culture media for C. elegans (one level downgrade) – These studies will be downgraded one level because they are not relevant to the nature of human exposure.

d. Duration of treatment and window of time between exposure and outcome assessment:

Studies that assess obesity-related outcomes following longer periods of exposure are expected to be more informative than studies of shorter duration. However, there will be no downgrading for either acute dosing regimens in experimental studies or short window of time between exposure and outcome assessment because there is insufficient evidence to identify a minimum amount of time required for many of our primary or secondary outcomes to manifest. We will consider duration of time between exposure and outcome assessment when evaluating consistency of findings across studies of similar experimental design.

Tabular summary of guidance for evaluating directness

Health outcomes		Exposure scenario	Time between exposure and outcome assessment	Overall downgrade
primary	0	0	0	0
secondary	-1	0	0	-1

0 = no downgrade, -1 = one downgrade, -2 two downgrade

Animal model	Health outcomes	Route of administration	Time between treatment and outcome assessment	Overall downgrade
Mammalian	primary	oral, sc injection, dermal, inhalation	0	0
		intracranial or intraperitoneal injection	-1	0
	secondary	oral, injection, dermal, inhalation	0	0
		intracranial or intraperitoneal injection	-1	0
Non-Mammalian	primary	oral, injection, dermal, inhalation	0	0
		water for aquatic species, culture media	-1	0
	secondary	oral, injection, dermal, inhalation	0	0
		water for aquatic species, culture media	-1	0

0 = no downgrade, -1 = one downgrade, -2 two downgrade
sc = subcutaneous

Imprecision

Precision is the degree of certainty surrounding an effect estimate with respect to a given outcome (AHRQ 2012). We will use 95% confidence intervals as the primary method to assess imprecision (Guyatt et al. 2011b). We will also consider whether the studies are adequately powered when assessing

precision, an issue that is especially important when interpreting findings that do not provide support for an association. As noted earlier, if we find that the relative statistical power of a study is a source of inconsistency, then we will consider omitting the no effect/association studies that are severely underpowered from the confidence rating phase when the ratings are used to develop hazard identification conclusions. Although no effect/association studies that are significantly underpowered may be omitted from the confidence rating phase, we will conduct a “sensitivity” analysis to assess the extent to which inclusion of the underpowered studies might be obscuring findings from studies with better statistical power.

When a meta-analysis is not feasible or inappropriate precision will be primarily based on the narrowness of the effect size estimates in the evidence base (AHRQ 2012). Data will be considered imprecise for ratio measures (e.g., OR) when the ratio of the upper to lower 95% CI for most studies is ≥ 10 ; and for absolute measures (e.g., percent control response) when the absolute difference between the upper and lower 95% CI for most studies is ≥ 100 . If a meta-analysis is conducted the same 95% confidence interval assessment will be made based on the meta-estimate of the association.

In addition, we will consider whether the studies are adequately powered⁹ when assessing precision. If a meta-analysis is conducted we will conduct an "optimal information size" (OIS) analysis as an additional indicator of precision for dichotomous and continuous outcomes (Guyatt et al. 2011b). This analysis calculates the sample size required for an adequately powered individual study, referred to as the OIS threshold or criterion (OIS calculator available at <http://www.stat.ubc.ca/~rollin/stats/ssize/>). The threshold for precision is met when the total sample size for the meta-estimate is as great as or greater than the OIS threshold. See Table 17 for a tabular summary of the guidance we will use to assess imprecision.

It is often difficult to distinguish between wide confidence intervals due to inconsistency and imprecision, leading to the question of whether to downgrade once or twice in these circumstances. In most cases a single downgrade for one of these domains is considered sufficient (AHRQ 2012). Thus, when the body of evidence is downgraded for inconsistency in direction of effect we will generally not further downgrade for imprecision. However, it is considered appropriate to downgrade twice if studies are very inconsistent (e.g., Table 14 see downgrade -2 levels, example B) *and* studies are considered very imprecise.

⁹ Statistical power is assessed during data extraction using an approach to assess ability to detect a 20% change from control or referent group response for continuous data or relative risk or odds ratio of 1.5 for categorical data using the prevalence of exposure or prevalence of outcome in the control or referent group to determine sample size using OpenEpi software, a free open source statistical resource (<http://www.openepi.com/OE2.3/menu/openEpiMenu.htm>). Recommended sample sizes will be compared to sample sizes used in the study to categorize statistical power as “appears to be adequately powered” (sample size met), somewhat underpowered (sample size is 75% to <100% of recommended), “underpowered” (sample size is 50 to <75% required), or “severely underpowered (sample size is <50% required). For categorical data where the sample sizes in exposed and control or referent groups differ, the sample size of the exposed group will be used to determine relative power category.

<p>0 (no downgrade)</p>	<p>No meta-analysis</p> <ul style="list-style-type: none"> For ratio measures (e.g., odds ratio, OR) the ratio of the upper to lower 95% CI for most studies is <10; or for absolute measures (e.g., percent control response) the absolute difference between the upper and lower 95% CI for most studies is <100. <p>AND</p> <ul style="list-style-type: none"> Most studies in the evidence base are “adequately” or “somewhat underpowered” <p>Meta-analysis</p> <ul style="list-style-type: none"> For ratio measures (e.g., OR) the ratio of the upper to lower 95% CI for the meta-estimate is <10; or for absolute measures (e.g., percent control response) the absolute difference between the upper and lower 95% CI for the meta-estimate is <100. <p>AND</p> <ul style="list-style-type: none"> The sample size for the meta-estimate meets the OIS criterion
<p>-1 downgrade (serious)</p>	<p>Does not clearly meet guidance for 0 (no downgrade) or -2 downgrade</p>
<p>-2 downgrade (very serious)</p>	<p>No meta-analysis</p> <ul style="list-style-type: none"> For ratio measures (e.g., OR) the ratio of the upper to lower 95% CI for most studies is ≥ 10; or for absolute measures (e.g., percent control response) the absolute difference between the upper and lower 95% CI for most studies is ≥ 100. <p>AND</p> <ul style="list-style-type: none"> Most studies in the evidence base are “underpowered” or “severely underpowered” <p>Meta-analysis</p> <ul style="list-style-type: none"> For ratio measures (e.g., OR) the ratio of the upper to lower 95% CI for the meta-estimate is ≥ 10; or for absolute measures (e.g., percent control response) the absolute difference between the upper and lower 95% CI for the meta-estimate is ≥ 100. <p>AND</p> <ul style="list-style-type: none"> The sample size for the meta-estimate does not meet the OIS criterion

Publication bias

Publication bias will be characterized as “undetected” (no downgrade) or “strongly suspected” (-1 downgrade) as recommended by GRADE (Guyatt et al. 2011e). In general, studies with statistically significant results are more likely to be published than studies without statistically significant results (“negative studies”) (Guyatt et al. 2011e). Thus some degree of publication bias is likely on any topic, but downgrading is reserved for cases where the concern is serious enough to significantly reduce confidence in the body of evidence. Below are some issues we will consider when determining whether to downgrade for publication bias:

- Early positive studies, particularly if small in size, are suspect. Reviews performed early, when only few initial studies are available will tend to overestimate effects [reviewed in (Guyatt et al. 2011e)]. There may be publication lag time for “negative” studies and it will take time for other authors to replicate the early studies. When it is inappropriate or not feasible to conduct a meta-analysis, we will use MetaData Viewer to stratify study findings by publication year and sample size to visually compare and determine if this appears to be an issue. In meta-analyses, a recursive cumulative analysis can be conducted that performs a meta-analysis at the end of each year to note changes in the summary effect.
- Publication bias should be suspected when studies are uniformly small, particularly when sponsored by industries or authors with conflicts of interest (Guyatt et al. 2011e; Viswanathan et al. 2012). We will use MetaData Viewer to stratify findings by funding source or whether the author(s) reported a conflict of interest and visually compare results.
- We will develop funnel plots to visualize asymmetrical or symmetrical patterns of study results to help assess publication bias when adequate data for a specific outcome are available.
- The identification of abstracts or other types of grey literature that do not appear as full-length articles within a reasonable time frame (~3-4 years) can be another indication of publication bias (AHRQ 2012).

Domains that can increase confidence

Four properties for a body of evidence (large magnitude of effect, dose-response, plausible confounding that would impact the observed association, and consistency across study designs and experimental model systems) are used to determine if the initial confidence rating should be upgraded (Figure 5, column 3). Consideration of large magnitude of effect, dose-response, and plausible confounding are considered in the GRADE and frameworks (AHRQ 2012; Guyatt et al. 2011g). We have added an additional factor to address consistency across human study designs and animal model systems to accommodate our focus in environmental health on evaluating observational human of different study designs and experimental animal studies rather than the randomized clinical trials more commonly encountered in the health care intervention field.

Large magnitude of association or effect

The guidance below will be considered when determining whether to upgrade based on magnitude of effect. In general, in order to rate up for large magnitude of effect there should not be any serious problems with risk of bias, precision, and publication bias. Evidence of a large magnitude of effect may be based on a single study provided the study did not have serious risk of bias issues and there was not serious inconsistency in direction of association/effect across studies. The rapidity of the response compared with natural progression of the condition can also be considered when determining large effect size.

For human observational studies of categorical data there is modeling and empirical data to suggest that consideration of associations between causal factors and confounders, and between confounders and outcomes, is unlikely to explain a relative risk (RR) greater than 2 (or less than 0.5), and very unlikely to explain associations with an RR greater than 5 (or less than 0.2) [reviewed in (Guyatt et al. 2011g)]. When the baseline risk is low (<20%), the RR and odds ratio (OR) are similar and the RR guidance can be applied to ORs. When the baseline risk is high (>40%), then the ORs can be much larger in magnitude

than RRs and a higher threshold for ORs might be appropriate. The outcome of obesity has a high baseline risk with more than one-third of U.S. adults (35.7%) and approximately 17% of children and adolescents aged 2–19 years considered obese (CDC 2012). Thus, a higher threshold for ORs is justified, at least for studies of adults. An OR in the range of 3-6 would be similar to ORs that have been reported for well-established risk factors of obesity, such as the association between parental overweight/obesity and childhood obesity (Xu et al. 2011).

Large magnitude of effect (upgrade +1):

- For categorical data: Relative risk (RR) = 2-5 or RR = 0.5-0.2 or odds ratio (OR) = 3-6 or 0.33-0.167 with no plausible confounders.
- For continuous variables: A standardized mean difference with a lower 95% confidence interval of 0.8 to 1.5 or upper 95% confidence interval of -0.8 to -1.5, based on guidance that identifies an effect size based on standardized mean difference of 0.8 as “large” (Cohen 1988).
- If we encounter study findings that cannot be converted to an RR, OR, or standardized mean difference we will attempt to define “large” based on what is known about the relationship between traditional risk factors for obesity or adiposity based on studies that use that measure.

Very large magnitude of effect (upgrade +2):

- RR > 5 or RR < 0.2 or OR > 6 or RR < 0.167.
- For continuous variables: A standardized mean difference with a lower 95% confidence interval of >1.5 or upper 95% confidence interval of >-1.5.
- If we encounter study findings that cannot be converted to an RR, OR, or standardized mean difference we will attempt to define “very large” based on what is known about the relationship between traditional risk factors for obesity or adiposity based on studies that use that measure.

Dose-response

We will upgrade +1 for evidence of a monotonic dose-response gradient (Guyatt et al. 2011g).

We will upgrade +1 for evidence of a non-monotonic dose response when:

- Data fits the expected pattern, i.e., prior knowledge leads to expectation for non-monotonic dose response.

AND

- Non-monotonic dose response is consistently observed in the evidence base

For adiposity-related outcomes in animals, it is possible that high doses of BPA may cause systemic toxicity manifesting as reductions in weight and adiposity..

Patterns of dose response will be considered within and across studies when considering whether to upgrade (Table 18). In order to visualize dose response, effect size data will be sorted in Meta Data Viewer in two ways: (1) by study in order to assess dose response within studies and to assess

DRAFT (April 9, 2013)

consistency of dose-response across studies of similar dose or exposure levels, and (2) by dose or exposure level to assess dose-response across the entire evidence base.

Table 18. Conceptual examples of upgrade decisions for evidence of dose response gradient		
no upgrade	upgrade +1 (monotonic)	upgrade +1 (non-monotonic) ¹
<p>Example A, findings sorted by study and then dose or exposure level (low to high)</p>	<p>Example B, findings sorted by study and then dose or exposure level (low to high)</p>	<p>Example C, findings sorted by study and then dose or exposure level (low to high)</p>
<p>Example A, findings sorted by exposure or dose level (low to high) across studies</p>	<p>Example B, findings sorted by exposure or dose level (low to high) across studies</p>	<p>Example C, findings sorted by exposure or dose level (low to high) across studies</p>
<p> </p>		
<p>¹Requires evidence to suggest non-monotonic is expected pattern AND non-monotonic dose response observed in evidence base</p>		

Plausible confounding or other residual biases that would increase our confidence in estimated effect

This element primarily applies to human studies and refers to consideration of unmeasured determinants of an outcome unaccounted for in an adjusted analysis that are likely to be distributed unequally across groups, referred to “residual confounding” or “residual biases” (Guyatt et al. 2011g).

We will upgrade one level when there are indications that residual confounding or bias would underestimate an apparent association or treatment effect (i.e., bias towards the null), or suggest a spurious effect when results suggest no effect.

Examples of residual bias towards the null: The “healthy worker” effect is one example of residual bias known to bias towards the null. Another example is outlined in the GRADE guidance (Guyatt et al. 2011g) of a systematic review of HIV infection and condom use. The effect estimate from five studies was statistically significant with condom use showing a protective effect compared with no condom use. In two of the studies, number of sexual partners was also considered (Detels et al. 1989; Difranceisco et al. 1996). These studies found that condom users were more likely to have more sexual partners, yet these studies did not adjust for number of partners in their final analyses. Had number of partners been considered in the meta-analysis, it’s likely it would have strengthened the effect estimate in favor of condom use.

Example of residual bias suggesting a spurious effect: This example, also taken from the GRADE guidance (Guyatt et al. 2011g), considers two observational studies (Elliman and Bedford 2001; Taylor et al. 1999) that failed to confirm a well-publicized association between vaccination and autism that was widely discredited and eventually retracted (Wakefield et al. 1998). After the widespread initial publicity, it was empirically confirmed that parents of autistic children were more likely to remember their vaccine experience than parents of children diagnosed before the publicity (Andrews et al. 2002). Parents of non-autistic children are presumed to also be less likely to remember their children’s vaccinations. Thus, the negative findings of the observational studies, despite the demonstrated recall bias, increase the confidence that there is no association and suggest an upgrade to the confidence rating.

Consistency across study types, experimental model systems, or populations

Three types of consistency in the body of evidence can be used to support a +1 upgrade:

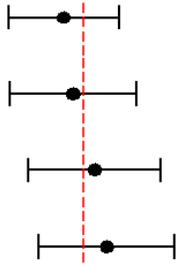
- across animal models and species - consistent results reported in multiple experimental animal models or species
- across independent studies of different human populations and exposure scenarios
- across study types - consistent results reported from study designs with different key features, e.g., between prospective cohort and case-control human studies or between a chronic and multigenerational animal studies.

We will use the guidance described earlier for no downgrade for unexplained inconsistency to determine whether findings are consistent enough within an evidence stream across human studies of different design or populations or across different animal models to warrant a +1 upgrade (Table 19). In general, in order to rate up for consistency there should not be any serious problems with risk of bias.

Table 19. Guidance for upgrading +1 for consistency across study types, experimental model systems, or populations

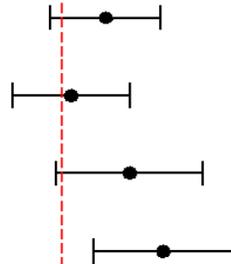
- Point estimates similar
- Confidence intervals overlap
- Statistical heterogeneity is non-significant
- I^2 of $\leq 50\%$

Example A



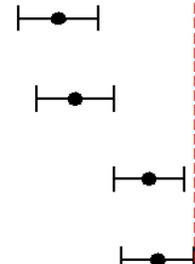
χ^2 p-level = 0.767; $I^2 = \ll 1\%$; $\tau^2 = \ll 1$

Example B



χ^2 p-level = 0.241; $I^2 = 29\%$; $\tau^2 = 0.046$

Example C



χ^2 p-level = < 0.001 ; $I^2 = 86\%$; $\tau^2 = 0.111$
 *considered consistent because point estimates are in the same direction

Other

Additional factors specific to the topic being evaluated. For example specificity of the association in cases where the effect is rare or unlikely to have multiple causes. For example, the observation of cases of clear cell adenocarcinoma, a rare kind of vaginal and cervical cancer, in a group of women in their teens and early twenties was highly unusual, and subsequent investigation determined that this as the result of *in utero* exposure to diethylstilbestrol (DES) (<http://www.cdc.gov/des/consumers/daughters/index.html>). This particularly rare outcome in an unusual population increases confidence in the association despite being based on small observational human studies.

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 specific health outcomes, e.g., hypertension or stroke, confidence ratings can also be developed for an overall health outcome if appropriate, e.g., cardiovascular disease. This is not anticipated in the current protocol.

STEP 6: TRANSLATE CONFIDENCE RATINGS INTO EVIDENCE OF HEALTH EFFECT CONCLUSIONS

The level of evidence will be assessed separately within the human and experimental animal data sets. The level of evidence for health effect conclusions reflects 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). The strategy uses four terms to describe the level of evidence for health effects: “High Level of Evidence,” “Moderate Level of Evidence,” “Low Level of Evidence,” and “Evidence of No Health Effect”¹⁰. These phrases are defined below and illustrated schematically in [Figure 6](#).

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).

Although the conclusions describe associations, Bradford Hill considerations on causality (Hill 1965) are embedded within the process used to evaluate the confidence in the body of evidence in the GRADE approach (NTP 2013a; Schünemann et al. 2011).

¹⁰ If the body of evidence for a health outcome receives a “Very Low Confidence” rating in Step 5, it will not proceed to developing evidence of health effect conclusions in Step 6.

Figure 6. Translation of confidence ratings into evidence of health effect conclusions

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
(+++) High	No effect	Evidence of no health effect
(++) Moderate	No effect	Inadequate
(+) Low	No effect	Inadequate

Note this figure is reproduced from the Step 6 of the Figure in the Draft OHAT Approach – February 2013 (available at <http://ntp.niehs.nih.gov/go/38673>)

STEP 7: INTEGRATE EVIDENCE TO DEVELOP HAZARD IDENTIFICATION CONCLUSIONS

During hazard identification the evidence streams for human studies and animal studies, which have remained separate through the previous steps, are integrated along with other relevant data such as supporting evidence from *in vitro* studies (defined here as other than whole animal studies, and including cell systems, computational toxicology, high throughput screening data, and *in silico* methods).

To determine the initial hazard identification conclusion, the highest level of evidence for a health effect from the human and animal evidence streams are combined. First, the level of evidence for health effects conclusion for human data from (“High,” “Moderate,” or “Low”) is considered together with the level of evidence for health effects conclusion for animal data (“High,” “Moderate,” or “Low”) to reach one of four hazard identification conclusion categories (Figure 7):

- 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

NTP does not require mechanistic or mode of action data in order to reach hazard identification conclusions, although when available this type of evidence may be used to raise (or lower) the level of the hazard identification conclusion when the evidence is “strong” (Figure 8). “Strong” evidence from *in vitro* or mechanistic studies demonstrates that a response is unequivocally associated with a given

health outcome or biological process relevant to a health outcome (e.g., adipogenesis). For example, if the hazard identification conclusion was “presumed” based on the human and 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 animal data, strong support from other relevant data may result in an upgraded conclusion of “presumed.” It is envisioned that strong evidence from *in vitro* or mechanistic studies for an effect on a biological process or pathway considered relevant to humans could result in a conclusion of “suspected” in the absence of human epidemiology or experimental animal data. Alternatively, if other relevant data provide strong opposition for biological plausibility of the relationship between exposure and the health effect, the hazard identification conclusion may be downgraded from that initially derived by considering the human and non-human animal evidence together.

Figure 7. Hazard Identification Scheme Based on Human and Animal Data

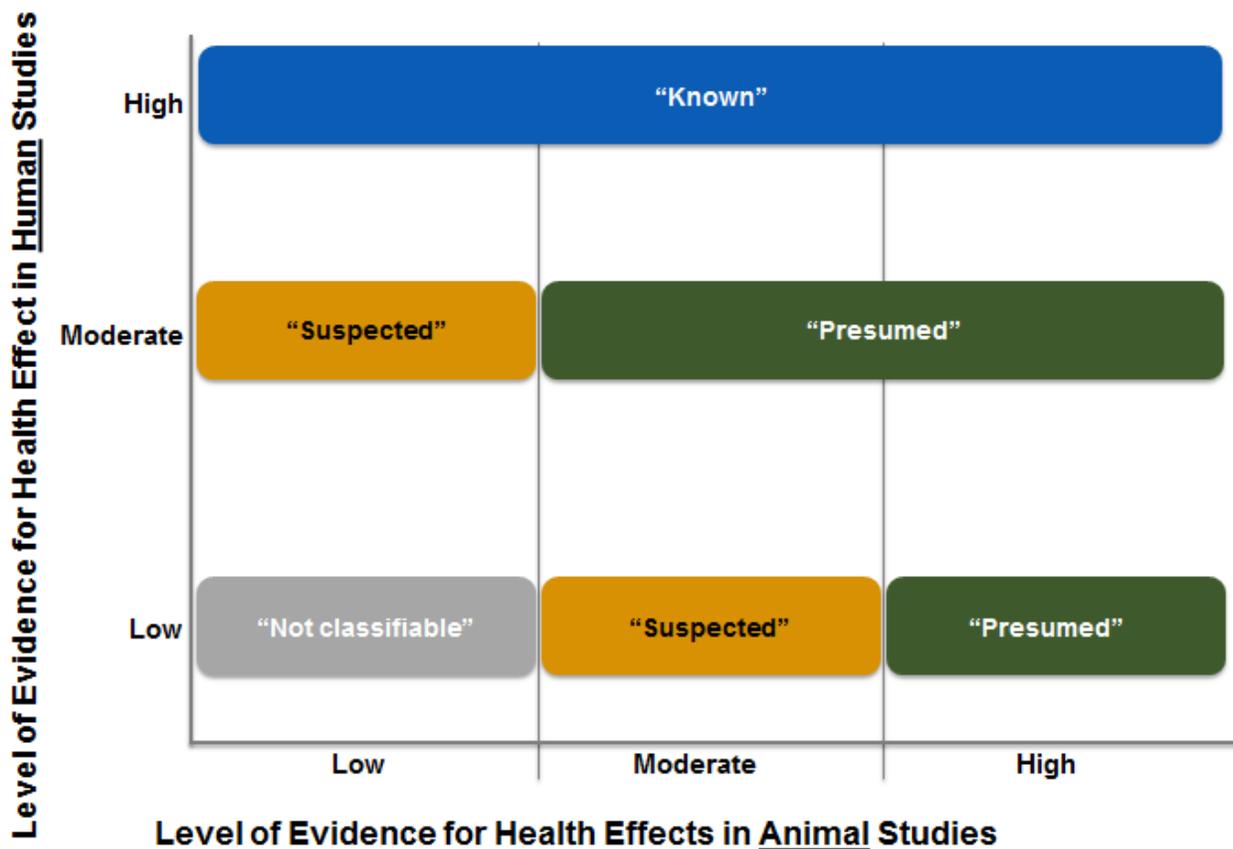
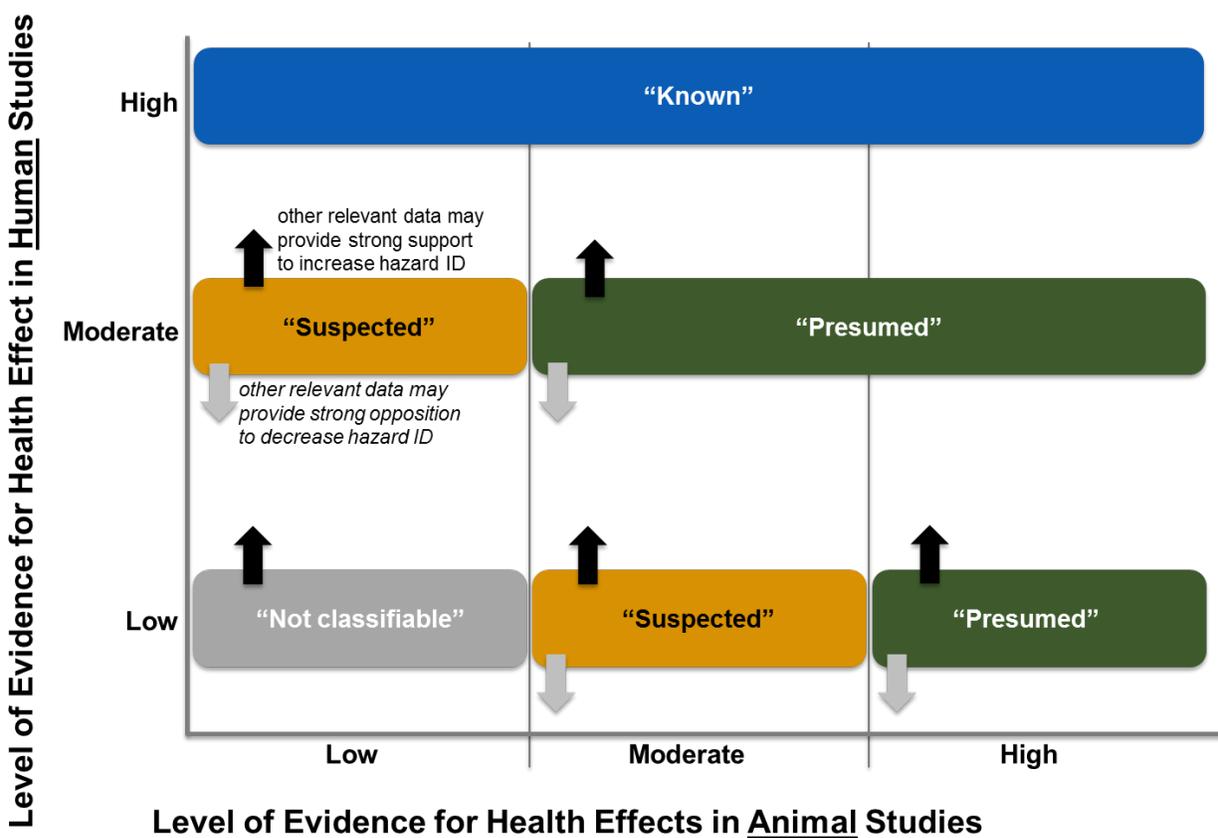


Figure 8. Hazard Identification Scheme Based on Human and Animal Data + Consideration of “Supporting Evidence”



Note this figure is reproduced from the Step 7 of the Figure in the Draft OHAT Approach – February 2013 (available at <http://ntp.niehs.nih.gov/go/38673>)

Assessment of biological plausibility provided by “supportive” evidence

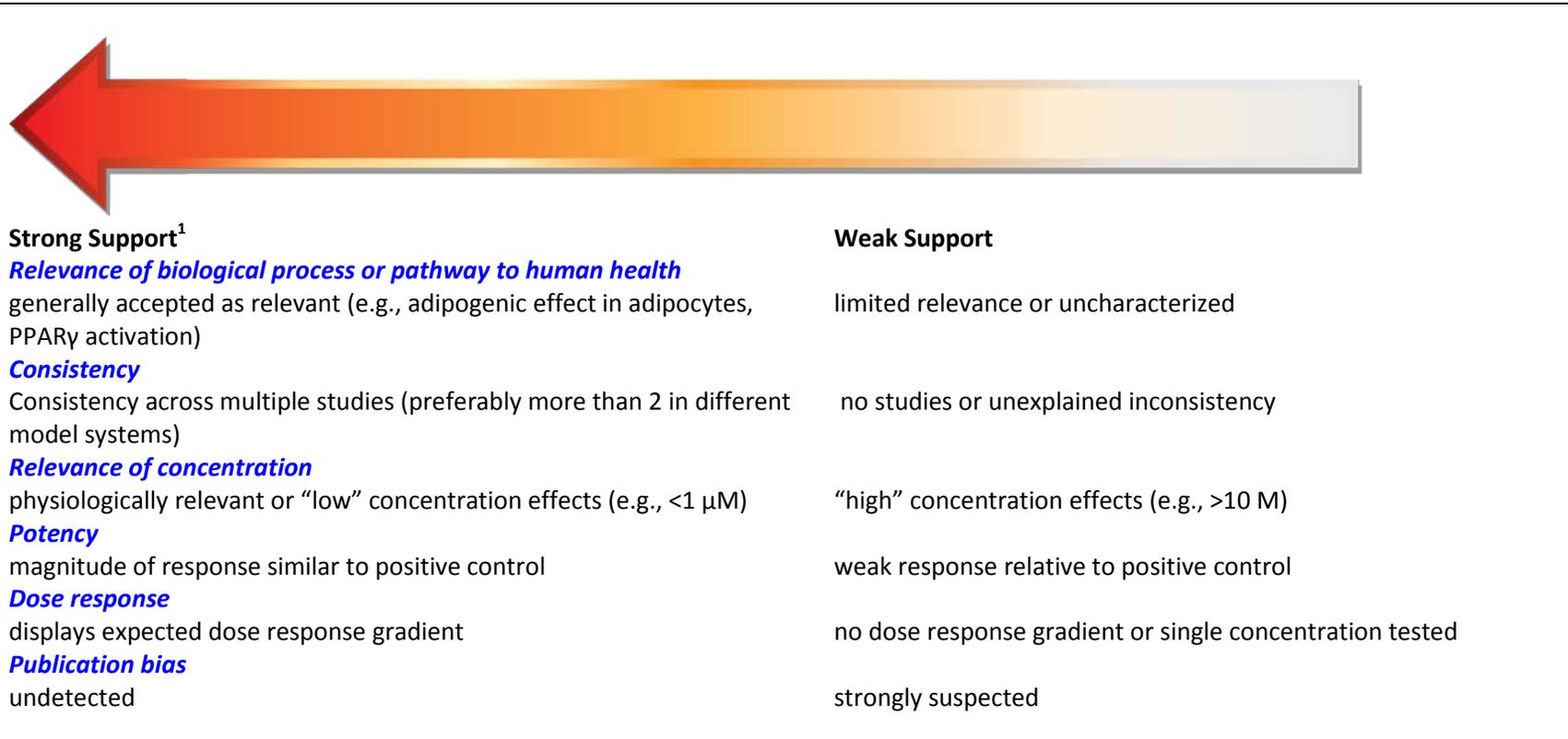
In the current protocol supporting evidence is derived from *in vitro* studies of adipocytes; *ex vivo*, cellular, genomic, or mechanistic outcomes reported in eligible animal or human studies; and data on interactions with key receptors involved in regulating adipogenesis [e.g., PPAR γ , RXR, LXR, GR, AR, and estrogen receptors (ER α , ER β , and “non-classical”)]. We are not aware of any body of evidence that suggests that commonly used adipocyte models, which are often rodent-derived, are not relevant for understanding human adipocyte biology (Hausman et al. 2009; Poulos et al. 2010).

The guidance presented in Figure 9 will be used to evaluate the strength of support provided by “supportive evidence” studies. The factors considered in Figure 9 are conceptually consistent with the factors considered in Step 5 for human and animal studies, although we are interested in further developing this framework as a near-term research activity.

- Consistency
- Directness and applicability \approx relevance of biological process or pathway to human health + relevance of concentration
- Large magnitude of effect \approx potency

- Dose response
- Publication bias

Figure 9. Factors considered when evaluating the support for biological plausibility provided by *in vitro*, cellular, genomic, or mechanistic data



¹A conclusion of “strong” support requires that most elements are met

²Physiologically relevant dose range based on range of unconjugated (“free”) BPA measured in human serum of 0.0022- 0.13 μ M (Vandenberg et al. 2010), an effect occurring within an order of magnitude of this range (~0.00022 – 1.3 μ M) is considered physiological relevant in order to account for unmeasured individual human variability; monotonic concentration-response not necessarily expected, e.g., high concentrations may cause cytotoxicity

PEER-REVIEW

When conclusions include a hazard identification label a draft version of the evaluation will then be disseminated for public comment and peer-reviewed by topic specific experts who are screened for financial conflicts of interest¹¹. Confidence ratings and the conclusions derived from them will be finalized after considering this input. When conclusions are oriented towards identifying research needs (i.e., do not include a hazard identification label), then the evaluation will be peer-reviewed by topic specific experts who are screened for financial conflicts of interest and released as an NTP Monograph or submitted to a peer-reviewed journal for publication. A more detailed description of the OHAT evaluation process can be found at <http://ntp.niehs.nih.gov/go/38138>.

REVIEW TEAM

Kristina Thayer (KAT) (primary author), Andrew Rooney (AAR), Abee Boyles (AB), Stephanie Holmgren (SH), Vickie Walker (VW), Grace Kissling (GK)

AUTHOR DECLARATIONS OF INTEREST

None

SOURCES OF SUPPORT

Internal sources

National Institute of Environmental Health Sciences/Division of the National Toxicology Program

External sources

None

TECHNICAL ADVISORS

Technical advisors are outside experts selected on an “as needed” basis to provide individual advice to the NTP for a specific topic. Potential technical advisors are screened for conflict of interest prior to their service. Depending upon the situation, any potential conflict of interest is acknowledged in the protocol or the person is disqualified from service. Service as a technical advisor does not necessarily indicate that an advisor has read the entire protocol or endorses the final document.

¹¹ Peer-review occurs either by a panel in a public meeting or by ad hoc reviewers by letter review.

DRAFT (April 9, 2013)

Bruce Blumberg, PhD - University of California-Irvine, Developmental & Cell Biology
School of Biological Sciences

Steven G. Hentges, PhD* - Polycarbonate/BPA Global Group of the American Chemistry Council (ACC)

Daniele Mandrioli, MD - Cesare Maltoni Cancer Research Center, Ramazzini Institute

Retha Newbold, MS – NIEHS/NTP (retired)

Ellen Silbergeld, PhD – Johns Hopkins University Bloomberg School of Public Health

Laura Vandenberg, PhD* - Tufts University, Department of Biology and Center for Regenerative & Developmental Biology

Tracey Woodruff, PhD - University of California-San Francisco (UCSF), Program on Reproductive Health and the Environment

*financial conflict of interest disclosed

PROTOCOL HISTORY & REVISIONS

- March 6, 2013: Draft protocol distributed to NTP Executive Committee points of contact
- April 9, 2013: Draft protocol publically released

REFERENCES

- [Anonymous]. 2010. White House Task Force on Childhood Obesity Report to the President: Solving the Problem of Childhood Obesity Within a Generation.
http://www.letsmove.gov/sites/letsmove.gov/files/TaskForce_on_Childhood_Obesity_May2010_FullReport.pdf [accessed 12 December 2011].
- AHRQ (Agency for Healthcare Research and Quality). 2012. 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 30 July 2012].
- Akingbemi BT, Sottas CM, Koulova AI, Klinefelter GR, Hardy MP. 2004. Inhibition of testicular steroidogenesis by the xenoestrogen bisphenol A is associated with reduced pituitary luteinizing hormone secretion and decreased steroidogenic enzyme gene expression in rat Leydig cells. *Endocrinology* 145(2):592-603.
- Alonso-Magdalena P, Vieira E, Soriano S, Menes L, Burks D, Quesada I, et al. 2010. Bisphenol A exposure during pregnancy disrupts glucose homeostasis in mothers and adult male offspring. *Environ Health Perspect* 118(9):1243-1250.
- Andrews N, Miller E, Taylor B, Lingam R, Simmons A, Stowe J, et al. 2002. Recall bias, MMR, and autism. *Arch Dis Child* 87(6):493-494.
- Balshem H, Helfand M, Schunemann HJ, Oxman AD, Kunz R, Brozek J, et al. 2011. GRADE guidelines: 3. Rating the quality of evidence. *Journal of clinical epidemiology* 64(4):401-406.
- Bevan C, Strother D. 2012. Toxicity data evaluation (method validity, data quality, study reliability) for hazard and risk assessments: Best practices (workshop discussion draft). Prepared for American Chemistry Council for December 2012 workshop.
- Blue L. 2013. More health harms for children exposed to BPA. *Time Magazine* 1/9/2013
<http://healthland.time.com/2013/01/09/more-health-harms-for-children-exposed-to-bpa/>
[accessed 19 January 2013].
- BMJ Group. Best Practices Reference Material (<http://bestpractice.bmj.com/best-practice/welcme.html>) [accessed 12 December 2011].

- Boyles AL, Harris SF, Rooney AA, Thayer KA. 2011. Forest Plot Viewer: a fast, flexible graphing tool. *Epidemiol* 22(5):746-747.
- Calafat AM, Weuve J, Ye X, Jia LT, Hu H, Ringer S, et al. 2009. Exposure to bisphenol A and other phenols in neonatal intensive care unit premature infants. *Environ Health Perspect* 117(4):639-644.
- Carwile JL, Michels KB. 2011. Urinary bisphenol A and obesity: NHANES 2003-2006. *Environmental research* 111(6):825-830.
- CDC (Centers for Disease Control and Prevention). 2012. Overweight and Obesity: Data and Statistics. <http://www.cdc.gov/obesity/data/index.html> [accessed 18 December 2012].
- Chapin RE, Adams J, Boekelheide K, Gray LE, Jr., Hayward SW, Lees PS, et al. 2008. NTP-CERHR Expert Panel Report on the Reproductive and Developmental Toxicity of Bisphenol A. *Birth Defects Res B Dev Reprod Toxicol* 83(3):157-395.
- CLARITY Group at McMaster University. 2013. Tools to assess risk of bias in cohort and case control studies; randomized controlled trials; and longitudinal symptom research studies aimed at the general population. <http://www.evidencepartners.com/resources/> [accessed 19 January 2013].
- Cohen J. 1988. *Statistical Power Analysis for the Behavioral Sciences* (2nd edition.). Lawrence Erlbaum Associates.
- Detels R, English P, Visscher BR, Jacobson L, Kingsley LA, Chmiel JS, et al. 1989. Seroconversion, sexual activity, and condom use among 2915 HIV seronegative men followed for up to 2 years. *J Acquir Immune Defic Syndr* 2(1):77-83.
- Difranceisco W, Ostrow DG, Chmiel JS. 1996. Sexual adventurousness, high-risk behavior, and human immunodeficiency virus-1 seroconversion among the Chicago MACS-CCS cohort, 1984 to 1992. A case-control study. *Sex Transm Dis* 23(6):453-460.
- Dwan K, Gamble C, Kolamunnage-Dona R, Mohammed S, Powell C, Williamson PR. 2010. Assessing the potential for outcome reporting bias in a review: a tutorial. *Trials* 11:52.
- Elliman DA, Bedford HE. 2001. MMR vaccine--worries are not justified. *Arch Dis Child* 85(4):271-274.
- EPA (Environmental Protection Agency). 2012. Bisphenol A alternatives in thermal paper (July 2012 DRAFT). Available: <http://www.epa.gov/dfe/pubs/projects/bpa/about.htm> [accessed 1 November 2012].
- Ferguson SA, Law CD, Jr., Abshire JS. 2011. Developmental treatment with bisphenol a or ethinyl estradiol causes few alterations on early preweaning measures. *Toxicological sciences : an official journal of the Society of Toxicology* 124(1):149-160.
- Fu R, Gartlehner G, Grant M, Shamliyan T, Sedrakyan A, Wilt TJ, et al. 2011. Conducting quantitative synthesis when comparing medical interventions: AHRQ and the Effective Health Care Program. *J Clin Epidemiol* 64(11):1187-1197.
- Genaidy AM, Lemasters GK, Lockey J, Succop P, Deddens J, Sobeih T, et al. 2007. An epidemiological appraisal instrument - a tool for evaluation of epidemiological studies. *Ergonomics* 50(6):920-960.
- Guyatt GH, Oxman AD, Akl EA, Kunz R, Vist G, Brozek J, et al. 2011a. GRADE guidelines: 1. Introduction- GRADE evidence profiles and summary of findings tables. *Journal of clinical epidemiology* 64(4):383-394.
- Guyatt GH, Oxman AD, Kunz R, Brozek J, Alonso-Coello P, Rind D, et al. 2011b. GRADE guidelines 6. Rating the quality of evidence--imprecision. *Journal of clinical epidemiology* 64(12):1283-1293.
- Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, et al. 2011c. GRADE guidelines: 8. Rating the quality of evidence--indirectness. *Journal of clinical epidemiology* 64(12):1303-1310.
- Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, et al. 2011d. GRADE guidelines: 7. Rating the quality of evidence--inconsistency. *Journal of clinical epidemiology* 64(12):1294-1302.
- Guyatt GH, Oxman AD, Montori V, Vist G, Kunz R, Brozek J, et al. 2011e. GRADE guidelines: 5. Rating the quality of evidence--publication bias. *Journal of clinical epidemiology* 64(12):1277-1282.

- Guyatt GH, Oxman AD, Schunemann HJ, Tugwell P, Knottnerus A. 2011f. GRADE guidelines: A new series of articles in the Journal of Clinical Epidemiology. *Journal of clinical epidemiology* 64(4):380-382.
- Guyatt GH, Oxman AD, Sultan S, Glasziou P, Akl EA, Alonso-Coello P, et al. 2011g. GRADE guidelines: 9. Rating up the quality of evidence. *Journal of clinical epidemiology* 64(12):1311-1316.
- Guyatt GH, Oxman AD, Vist G, Kunz R, Brozek J, Alonso-Coello P, et al. 2011h. GRADE guidelines: 4. Rating the quality of evidence-study limitations (risk of bias). *Journal of Clinical Epidemiology* 64(4):407-415.
- Hausman GJ, Dodson MV, Ajuwon K, Azain M, Barnes KM, Guan LL, et al. 2009. Board-invited review: the biology and regulation of preadipocytes and adipocytes in meat animals. *J Anim Sci* 87(4):1218-1246.
- Heindel JJ, vom Saal FS. 2009. Role of nutrition and environmental endocrine disrupting chemicals during the perinatal period on the aetiology of obesity. *Mol Cell Endocrinol* 304(1-2):90-96.
- HHS (Department of Health and Human Services). 2013. Draft Office of Health Assessment and Translation Approach for Systematic Review and Evidence Integration for Literature-Based Health Assessments. *Federal Register*, Vol. 78, No. 34, pages 12764-5. February 25, 2013. Available at <http://www.gpo.gov/fdsys/pkg/FR-2013-02-25/pdf/2013-04254.pdf> [accessed 25 February 2013].
- Higgins J, Green S. 2011. *Cochrane Handbook for Systematic Reviews of Interventions*. Version 5.1.0 (updated March 2011). <http://handbook.cochrane.org/> (accessed 3 February 2013).
- Hill AB. 1965. The Environment and Disease: Association or Causation? *Proceedings of the Royal Society of Medicine* 58:295-300.
- Howdeshell KL, Hotchkiss AK, Thayer KA, Vandenberg JG, vom Saal FS. 1999. Exposure to bisphenol A advances puberty. *Nature* 401(6755):763-764.
- Hugo ER, Brandebourg TD, Woo JG, Loftus J, Alexander JW, Ben-Jonathan N. 2008. Bisphenol A at environmentally relevant doses inhibits adiponectin release from human adipose tissue explants and adipocytes. *Environ Health Perspect* 116(12):1642-1647.
- IASO (International Association for the Study of Obesity). 2012. World map of obesity. <http://www.iaso.org/resources/world-map-obesity/> [accessed 18, December 2012].
- IOM (Institute of Medicine). 2011. Finding What Works in Health Care: Standards for Systematic Reviews. http://www.nap.edu/openbook.php?record_id=13059&page=R1 [accessed 13 January 2013].
- Jahnke GD, Iannucci AR, Scialli AR, Shelby MD. 2005. Center for the evaluation of risks to human reproduction--the first five years. *Birth Defects Res B Dev Reprod Toxicol* 74(1):1-8.
- Janesick A, Blumberg B. 2011. Minireview: PPAR β as the target of obesogens. *J Steroid Biochem Mol Biol* 127(1-2):4-8.
- Jo J, Gavrilova O, Pack S, Jou W, Mullen S, Sumner AE, et al. 2009. Hypertrophy and/or Hyperplasia: Dynamics of Adipose Tissue Growth. *PLoS Comput Biol* 5(3):e1000324.
- Johnson PI, Sutton P, Atchley D, Koustas E, Lam J, Robinson K, et al. 2013. Applying the Navigation Guide: Case Study #1. The Impact of Developmental Exposure to Perfluorooctanoic Acid (PFOA) On Fetal Growth. A Systematic Review of the Human Evidence - Protocol.
- Klimisch HJ, 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.
- Koustas E, Lam J, Sutton P, Johnson PI, Atchley D, Robinson K, et al. 2013. Applying the Navigation Guide: Case Study #1. The Impact of Developmental Exposure to Perfluorooctanoic Acid (PFOA) On Fetal Growth. A Systematic Review of the Non-Human Evidence - Protocol.
- Krauth D, Woodrull T, Bero L. 2013. A systematic review of quality assessment instruments for published animal studies (submitted).

DRAFT (April 9, 2013)

- Kristof ND. 2013. Warnings from a flabby mouse. NY Times 1/20/2013.
http://www.nytimes.com/2013/01/20/opinion/sunday/kristof-warnings-from-a-flabby-mouse.html?_r=0 [accessed 20 January 2013].
- Kubo K, Arai O, Omura M, Watanabe R, Ogata R, Aou S. 2003. Low dose effects of bisphenol A on sexual differentiation of the brain and behavior in rats. *Neurosci Res* 45(3):345-356.
- Medlin J. 2003. New arrival: CERHR monograph series on reproductive toxicants. *Environ Health Perspect* 111(13):A696-698.
- Miyawaki J, Sakayama K, Kato H, Yamamoto H, Masuno H. 2007. Perinatal and postnatal exposure to bisphenol A increases adipose tissue mass and serum cholesterol level in mice. *J Atheroscler Thromb* 14(5):245-252.
- Moher D, Liberati A, Tetzlaff J, Altman DG. 2009. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *Journal of Clinical Epidemiology* 62(10):1006-1012.
- Myers JP, vom Saal FS, Akingbemi BT, Arizono K, Belcher S, Colborn T, et al. 2009. Why public health agencies cannot depend on good laboratory practices as a criterion for selecting data: the case of bisphenol A. *Environ Health Perspect* 117(3):309-315.
- NHLBI (National Heart Lung and Blood Institute). 2012. What Are Overweight and Obesity?
<http://www.nhlbi.nih.gov/health/health-topics/topics/obe/> [accessed 4 April 2013].
- NIH Obesity Research Task Force. 2011. Strategic Plan for NIH Obesity Research.
<http://www.obesityresearch.nih.gov/about/strategic-plan.aspx> [accessed 12 December 2011].
- Nikaido Y, Yoshizawa K, Danbara N, Tsujita-Kyutoku M, Yuri T, Uehara N, et al. 2004. Effects of maternal xenoestrogen exposure on development of the reproductive tract and mammary gland in female CD-1 mouse offspring. *Reprod Toxicol* 18(6):803-811.
- NTP (National Toxicology Program). 2008a. NTP-CERHR Monograph on the Potential Human Reproductive and Developmental Effects of Bisphenol A (BPA). September 2008
(<http://cerhr.niehs.nih.gov/chemicals/bisphenol/bisphenol.html>).
- NTP (National Toxicology Program). 2008b. NTP-CERHR Monograph on the Potential Human Reproductive and Developmental Effects of Bisphenol A (BPA). September 2008. Available at
<http://cerhr.niehs.nih.gov/chemicals/bisphenol/bisphenol.html> (accessed 5 March 2013).
- NTP (National Toxicology Program). 2011. "Role of Environmental Chemicals in the Development of Diabetes and Obesity" workshop. <http://ntp.niehs.nih.gov/go/36433> [accessed 8 November 2012].
- NTP (National Toxicology Program). 2012. Board of Scientific Counselors June 21-22, 2012 meeting. Meeting materials available at <http://ntp.niehs.nih.gov/go/9741> [accessed 21 February 2013].
- NTP (National Toxicology Program). 2013a. Draft OHAT Approach for Systematic Review and Evidence Integration for Literature-based Health Assessments – February 2013.
<http://ntp.niehs.nih.gov/go/38138> [accessed 26 January 2013].
- NTP (National Toxicology Program). 2013b. Webinar on the assessment of data quality in animal studies (March 20, 2013). Presentations available at <http://ntp.niehs.nih.gov/go/38752> (accessed 7 April 2013).
- Ogden C, Carroll M. 2010. Prevalence of Obesity Among Children and Adolescents: United States, Trends 1963-1965 Through 2007-2008. CDC-NCHS Health E-Stat.
http://www.cdc.gov/nchs/data/hestat/obesity_child_07_08/obesity_child_07_08.htm
[accessed 12 December 2011].
- Ohlsson C, Hellberg N, Parini P, Vidal O, Bohlooly YM, Rudling M, et al. 2000. Obesity and disturbed lipoprotein profile in estrogen receptor-alpha-deficient male mice. *Biochem Biophys Res Commun* 278(3):640-645.
- Okada A, Kai O. 2008. Effects of estradiol-17beta and bisphenol A administered chronically to mice throughout pregnancy and lactation on the male pups' reproductive system. *Asian J Androl* 10(2):271-276.

DRAFT (April 9, 2013)

- Oxman AD, Schunemann HJ, Fretheim A. 2006. Improving the use of research evidence in guideline development: 7. Deciding what evidence to include. *Health Res Policy Syst* 4:19.
- Patisaul HB, Bateman HL. 2008. Neonatal exposure to endocrine active compounds or an ERbeta agonist increases adult anxiety and aggression in gonadally intact male rats. *Horm Behav* 53(4):580-588.
- Poulos SP, Dodson MV, Hausman GJ. 2010. Cell line models for differentiation: preadipocytes and adipocytes. *Exp Biol Med (Maywood)* 235(10):1185-1193.
- Rubin BS, Murray MK, Damassa DA, King JC, Soto AM. 2001. Perinatal exposure to low doses of bisphenol A affects body weight, patterns of estrous cyclicity, and plasma LH levels. *Environ Health Perspect* 109(7):675-680.
- Ryan KK, Haller AM, Sorrell JE, Woods SC, Jandacek RJ, Seeley RJ. 2010. Perinatal exposure to bisphenol A and the development of metabolic syndrome in CD-1 mice. *Endocrinol* 151(6):2603-2612.
- Salian S, Doshi T, Vanage G. 2009. Perinatal exposure of rats to bisphenol A affects the fertility of male offspring. *Life Sci* 85(21-22):742-752.
- Schneider K, Schwarz M, Burkholder I, Kopp-Schneider A, Edler L, Kinsner-Ovaskainen A, et al. 2009. "ToxRTool", a new tool to assess the reliability of toxicological data. *Toxicology Letters* 189(2):138-144.
- Schünemann H, Hill S, Guyatt G, Akl EA, Ahmed F. 2011. The GRADE approach and Bradford Hill's criteria for causation. *Journal of epidemiology and community health* 65(5):392-395.
- Shamliyan T, Kane RL, Dickinson S. 2010. A systematic review of tools used to assess the quality of observational studies that examine incidence or prevalence and risk factors for diseases. *Journal of Clinical Epidemiology* 63(10):1061-1070.
- Shamliyan TA, Kane RL, Ansari MT, Raman G, Berkman ND, Grant M, et al. 2011. Development of quality criteria to evaluate nontherapeutic studies of incidence, prevalence or risk factors of chronic diseases: Pilot study of new checklists. Available at <http://www.ncbi.nlm.nih.gov/books/NBK53272/> [accessed March 6, 2012]. Rockville (MD): Agency for Healthcare Research and Quality (US); 2011 Jan. Report No.: 11-EHC008-EF. AHRQ Methods for Effective Health Care.
- Shelby MD. 2005. National Toxicology Program Center for the Evaluation of Risks to Human Reproduction: guidelines for CERHR expert panel members. *Birth Defects Res B Dev Reprod Toxicol* 74(1):9-16.
- Silbergeld E, Scherer RW. 2013. Evidence-based toxicology: Strait is the gate, but the road is worth taking. *ALTEX* 30(1):67-73.
- Somm E, Schwitzgebel VM, Toulotte A, Cederroth CR, Combescure C, Nef S, et al. 2009. Perinatal exposure to bisphenol A alters early adipogenesis in the rat. *Environ Health Perspect* 117(10):1549-1555.
- Szabo L. 2012. Study links chemical BPA to obesity in children, teens. *USA Today* 9/18/2012. <http://usatoday30.usatoday.com/news/nation/story/2012/09/18/bpa-link-to-obesity-in-kids/57799902/1> [accessed 19 January 2013].
- Szklo M, Nieto FJ. 2007. *Epidemiology: Beyond the basics* (2nd edition). 2nd ed. Sudbury, MA: Jones and Bartlett Publishers.
- Taylor B, Miller E, Farrington CP, Petropoulos MC, Favot-Mayaud I, Li J, et al. 1999. Autism and measles, mumps, and rubella vaccine: no epidemiological evidence for a causal association. *Lancet* 353(9169):2026-2029.
- Thayer KA, Heindel JJ, Bucher JR, Gallo MA. 2012. Role of environmental chemicals in diabetes and obesity: A National Toxicology Program workshop report. *Environ Health Perspect* 120(6):779-789.
- Twombly R. 1998. New NTP centers meet the need to know. *Environ Health Perspect* 106(10):A480-483.

DRAFT (April 9, 2013)

- Vandenberg LN, Chahoud I, Heindel JJ, Padmanabhan V, Paumgartten FJ, Schoenfelder GC. 2010. Urinary, circulating, and tissue biomonitoring studies indicate widespread exposure to bisphenol A. *Environmental Health Perspectives* 118(8):1055-1070.
- Viswanathan M, Ansari M, Berkman ND, Chang S, Hartling L, McPheeters LM, et al. 2012. Assessing the risk of bias of individual studies when comparing medical interventions (March 8, 2012). Agency for Healthcare Research and Quality Methods Guide for Comparative Effectiveness Reviews. March 2012. AHRQ Publication No. 12-EHC047-EF. Available at: www.effectivehealthcare.ahrq.gov/, or direct link at <http://www.effectivehealthcare.ahrq.gov/index.cfm/search-for-guides-reviews-and-reports/?pageaction=displayproduct&productid=998> [accessed 3 January 2013].
- vom Saal FS, Hughes C. 2005. An extensive new literature concerning low-dose effects of bisphenol A shows the need for a new risk assessment. *Environ Health Perspect* 113(8):926-933.
- Wakefield AJ, Murch SH, Anthony A, Linnell J, Casson DM, Malik M, et al. 1998. Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children. *Lancet* 351(9103):637-641 [RETRACTION: *Lancet*. 2010 Feb 2016;2375(9713):2445].
- WHO (World Health Organization). 2011a. Obesity and Overweight (updated March 2011). <http://www.who.int/mediacentre/factsheets/fs311/en/index.html> [accessed 12 December 2011].
- WHO (World Health Organization). 2011b. Toxicological and Health Aspects of Bisphenol A. Report of Joint FAO/WHO Expert Meeting 2–5 November 2010 and Report of Stakeholder Meeting on Bisphenol A 1 November 2010 in Ottawa, Canada (<http://www.who.int/foodsafety/chem/chemicals/bisphenol/en/index1.html> accessed 27, November 2012).
- Xu L, Dubois L, Burnier D, Girard M, Prud'homme D. 2011. Parental overweight/obesity, social factors, and child overweight/obesity at 7 years of age. *Pediatr Int* 53(6):826-831.
- Ye X, Kuklennyik, Z., Needham, L. L., and Calafat, A. M. 2005. Automated on-line column-switching HPLC-MS/MS method with peak focusing for the determination of nine environmental phenols in urine. *Analytical Chemistry* 77:5407-5413.

APPENDICES

Appendix 1. Database search strategies	
Database and Results	Search Strategy
African Index Medicus – 0 results	"Bisphenol A"
Cochrane Library (from 2005 to the present) – 1 result (economic evaluation)	"Bisphenol A"
DART-Europe (E-Theses) – 51 results	"Bisphenol A"
Embase (from 1947 to the present) – 809 results	Limits: exclude records from Medline (“4,4 isopropylidenediphenol”/exp OR “4,4 isopropylidenediphenol”:ti:ab OR "bisphenol A"/exp OR “bisphenol A”:ti:ab OR 80-05-7:rn) AND ('body weight disorder'/exp OR obes*:ti:ab OR "body mass":ti:ab OR 'body weight'/exp OR "body weight":ti:ab OR "weight gain":ti:ab OR overweight:ti:ab OR "body fat":ti:ab OR adipocyte/exp OR adipocyte*:ti:ab OR 'lipid metabolism'/exp OR lipid*:ti:ab OR adipogen*:ti:ab OR 'adipose tissue'/exp OR 'adipocytokine'/exp OR adipocytokine*:ti:ab OR adipokine*:ti:ab OR 'adiponectin'/exp OR adiponectin*:ti:ab OR adipos*:ti:ab OR ghrelin/exp OR ghrelin:ti:ab OR leptin/exp OR leptin:ti:ab OR resistin/exp OR resistin:ti:ab OR lipogen*:ti:ab OR lipoprotein/exp OR lipoprotein*:ti:ab OR triacylglycerol/exp OR triacylglycerol:ti:ab OR triglyceride*:ti:ab OR 'retinoid x receptor'/exp OR RXR:ti:ab OR "retinoid x":ti:ab OR "9-cis-retinoic":ti:ab OR "peroxisome proliferator-activated receptors"/exp OR PPAR*:ti:ab OR "peroxisome proliferator":ti:ab OR glucocorticoid/exp OR glucocorticoid*:ti:ab OR 'liver x receptor'/exp OR LXR:ti:ab OR "liver x":ti:ab OR Nr1h2:ti:ab)
EPA's ACTor (Aggregated Computational Toxicology Resource)	Search on chemical names: "Bisphenol A" OR Search on CAS Numbers: 80-05-7
EPA's Chemical Data Access Tool to find health and safety data that has been submitted to the Agency, under authorities in sections 4, 5, and 8 of the Toxic Substances Control Act (TSCA)	80-05-7
IMSEAR (Index Medicus for South-East Asia Region) – 25 results	"Bisphenol A"
IndMed – 0 results	"Bisphenol A"
KoreaMed – 30 results	"Bisphenol A"
LILACS – 0 results	"Bisphenol A"
Open Access Theses and Dissertations	"Bisphenol A"

DRAFT (April 9, 2013)

Appendix 1. Database search strategies	
Database and Results	Search Strategy
- 252 results	
Panteleimon - 0 results	"Bisphenol A"
PubChem	Search compound: "4,4' isopropylidenediphenol" OR "bisphenol A" OR 80-05-7
PubMed (from 1948 to the present) - 480 results	("4,4' isopropylidenediphenol" OR "Bisphenol A" OR 80-05-7) AND (Obesity[mh] OR obes*[tiab] OR "body mass index"[mh] OR "body mass"[tiab] OR "body weight"[mh] OR "body weight"[tiab] OR "weight gain"[mh] OR "weight gain"[tiab] OR overweight[tiab] OR "body fat"[tiab] OR adipocyte[mh] OR adipocyte*[tiab] OR adipogenesis[mh] OR adipogen*[tiab] OR "adipose tissue"[mh] OR adipos*[tiab] OR adipokines[mh] OR adipokine*[tiab] OR adipocytokine*[tiab] OR adiponectin[mh] OR adiponectin*[tiab] OR ghrelin[mh] OR ghrelin[tiab] OR leptin[mh] OR leptin*[tiab] OR resistin[mh] OR resistin[tiab] OR Lipid metabolism[mh] OR lipogen*[tiab] OR lipid[tiab] OR lipids[tiab] OR lipoprotein OR triacylglycerol OR triglyceride OR "Retinoid x receptors"[mh] OR RXR[tiab] OR "retinoid x"[tiab] OR "9-cis-retinoic"[tiab] OR "peroxisome proliferator-activated receptors"[mh] OR PPAR*[tiab] OR "peroxisome proliferator"[tiab] OR "receptors, glucocorticoid"[mh] OR glucocorticoid*[tiab] OR "liver x receptor"[supplementary concept] OR LXR*[tiab] OR "liver x"[tiab] OR Nr1h2[tiab])
Scopus (from 1823 to the present) - 734 results	Advanced search: TITLE-ABS-KEY("4,4' isopropylidenediphenol" OR "Bisphenol A" OR 80-05-7) AND TITLE-ABS-KEY(obes* OR "body mass" OR "body weight" OR "weight gain" OR overweight OR "body fat" OR adipocyte* OR adipogen* OR adipos* OR adipokine* OR adipocytokine* OR adiponectin* OR ghrelin OR leptin* OR resistin OR lipogen* OR lipid* OR lipoprotein* OR triglyceride* OR triacylglycerol OR RXR OR "retinoid x" OR "9-cis-retinoic" OR PPAR* OR "peroxisome proliferator" OR glucocorticoid* OR LXR* OR "liver x" OR Nr1h2)
Toxline (from 1965 to the present) - 115 results	Set Limits: For chemicals, add synonyms and CAS number - YES; Include PubMed records - NO ("4,4' isopropylidenediphenol" OR "Bisphenol A" OR 80-05-7) AND (obes* OR "body mass" OR "body weight" OR "weight gain" OR overweight OR "body fat" OR adipocyte* OR adipogen* OR adipos* OR adipokine* OR adipocytokine* OR adiponectin* OR ghrelin OR leptin* OR resistin OR lipogen* OR lipid OR lipids OR lipoprotein* OR triglyceride* OR triacylglycerol OR RXR OR "retinoid x" OR "9-cis-retinoic" OR PPAR* OR "peroxisome proliferator" OR glucocorticoid* OR LXR* OR "liver x" OR Nr1h2)

DRAFT (April 9, 2013)

Appendix 1. Database search strategies	
Database and Results	Search Strategy
Web of Science (from 1900 to the present) – 633 results	Advanced Search: TS=("4,4' isopropylidenediphenol" OR "Bisphenol A" OR 80-05-7) AND TS=(obes* OR "body mass" OR "body weight" OR "weight gain" OR overweight OR "body fat" OR adipocyte* OR adipogen* OR adipos* OR adipokine* OR adipocytokine* OR adiponectin* OR ghrelin OR leptin* OR resistin OR lipogen* OR lipid OR lipids OR lipoprotein* OR triglyceride* OR triacylglycerol OR RXR OR "retinoid x" OR "9-cis-retinoic" OR PPAR* OR "peroxisome proliferator" OR glucocorticoid* OR LXR* OR "liver x" OR Nr1h2)
WPRIM (Western Pacific Region) – 11 results	("4,4' isopropylidenediphenol" OR "Bisphenol A" OR 80-05-7) AND (obes* OR "body mass" OR "body weight" OR "weight gain" OR overweight OR "body fat" OR adipocyte* OR adipogen* OR adipos* OR adipokine* OR adipocytokine* OR adiponectin* OR ghrelin OR leptin* OR resistin OR lipogen* OR lipid OR lipids OR lipoprotein* OR triglyceride* OR triacylglycerol OR RXR OR "retinoid x" OR "9-cis-retinoic" OR PPAR* OR "peroxisome proliferator" OR glucocorticoid* OR LXR* OR "liver x" OR Nr1h2)