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

SYSTEMATIC REVIEW TO EVALUATE THE EVIDENCE FOR AN ASSOCIATION BETWEEN PERFLUOROOCTANOIC ACID (PFOA) OR PERFLUROOCTANE SULFONATE (PFOS) EXPOSURE AND IMMUNOTOXICITY

April 9 2013

Office of Health Assessment and Translation
Division of the National Toxicology Program
National Institute of Environmental Health Sciences

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

TABLE OF CONTENTS

SYSTEMATIC REVIEW TO EVALUATE THE EVIDENCE FOR AN ASSOCIATION BETWEEN PERFLUOROOCTANOIC ACID (PFOA) OR PERFLUROOCTANE SULFONATE (PFOS) EXPOSURE AND IMMUNOTOXICITY

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

Perflurooctanoic acid (PFOA) and perfluorooctane sulfonate (PFOS) are extremely persistent chemicals and widely distributed in the environment. In terms of toxicity and exposure, PFOA and PFOS are the best studied perfluoroalkyl acids, a group of compounds used extensively over the last 40 years in commercial and industrial applications including food packaging, lubricants, water-resistant coatings, and fire-retarding foams. Through voluntary agreements, the primary manufacturer of PFOS phased out production in 2002 and PFOS is no longer manufactured in the United States (US EPA 2006, ATSDR 2009, US EPA 2009). Similar arrangements have been made for PFOA and eight companies that manufacture PFOA have committed to eliminate emissions and product content by 2015 (US EPA 2006, ATSDR 2009, US EPA 2012b, a).

Although emissions have been dramatically reduced, the persistence and bioaccumulation of both PFOA and PFOS result in detectable levels in the U.S. population and is a cause for concern for human health (US EPA 2012a). PFOA and PFOS were present in all of the 1562 serum samples analyzed as part of a study of 11 perfluorinated compounds in the National Health and Nutrition Examination Survey (NHANES 1999-2000) (Calafat *et al.* 2007) and remain the two highest concentrations among perfluorinated compounds measured in blood samples reported from a representative sample of the U.S. population in the most recent National Report on Human Exposure to Environmental Chemicals for 2009-2010 (CDC 2012).

Several recent publications have linked PFOA and PFOS exposure to functional immune changes in humans that are consistent with evidence of PFOA- and PFOS-related immunotoxicity from animal studies. Immune-related health effects including suppression of the antibody response to vaccines have been reported in children with higher current or pre-natal blood levels of PFOA and PFOS in prospective cohort studies in the Faroe Islands (Grandjean *et al.* 2012) and Norway (Granum *et al.* 2013). Similar immunosuppression of the antibody response has been shown in mice at blood concentrations of PFOS occurring in the general US population (e.g., Peden-Adams *et al.* 2008, Fair *et al.* 2011, CDC 2012, DeWitt *et al.* 2012). Experimental studies of PFOA and PFOS in laboratory animals have also reported other immune changes including altered inflammatory response, cytokine signaling, and measures of both innate and adaptive immunity. Wildlife studies in species ranging from loggerhead sea turtles to sea otters have also reported widespread exposure and altered immune measures associated with PFOA and PFOS (e.g., Keller *et al.* 2005, Kannan *et al.* 2006, Hart *et al.* 2009).

Although some health effects of PFOA and PFOS are dependent on peroxisome proliferator-activated receptor alpha (PPARα which shows strong species differences that may affect the relevance of animal data for human health), immune effects reported in laboratory animals appear to be partially or wholly independent of PPARα (DeWitt *et al.* 2009, DeWitt *et al.* 2012)

To help assess the science in this area, OHAT is conducting a systematic review to evaluate the association between exposure to PFOA or PFOS and immunotoxicity or immune-related health effects. To our knowledge a systematic review on this topic has not been conducted.

Use of protocol as a case study to assess OHAT's Draft Approach for Systematic Review and Evidence Integration for Literature-Based Health Assessments

The current protocol is one of two case studies that illustrate the specific details for how OHAT plans to implement the framework described in the "Draft Office of Health Assessment and Translation Approach for Systematic Review and Evidence Integration for Literature-Based Health Assessments – February 2013" ("Draft OHAT Approach – February 2013[1]" (HHS 2013)(available at http://ntp.niehs.nih.gov/go/38673)). These two case studies will be conducted to provide input and experience 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. Individuals interested in receiving updates on this project 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 to PFOA or PFOS is associated with changes in immune-related measures in humans based on integrating the evidence from human and animal data and considering the evidence for biological plausibility provided by other relevant data (e.g., *in vitro* or mechanistic studies).

[1] The approach described in the draft document is based on guidance from the Grading of Recommendations Assessment, Development and Evaluation (GRADE) Working Group (Balshem *et al.* 2011, 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 (defined here as other than whole animal studies, and including both cell systems, computational toxicology 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 – February 2013 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 evaluations are typically to identify potential adverse health effects and randomized clinical trials are not considered ideal for this purpose (Oxman *et al.* 2006, Silbergeld and Scherer 2013). The most relevant data for addressing environmental health questions are human observational epidemiology and experimental animal studies and these data need to be considered with clear appreciation for their strengths and limitations.

Specific aims:

- Provide a summary of the literature and rate our confidence in studies that assess the association between PFOA and PFOS exposure and immune-related health effects in human studies of children and adults.
- Provide a summary of the literature and rate our confidence in studies that assess the association between PFOA and PFOS exposure and immune-related health effects in whole animal models.
- Evaluate evidence for biological plausibility provided by other relevant data (e.g., *in vitro* or mechanistic studies) that assess the effects of PFOA and PFOS on immune-related endpoints.
- 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 other relevant data.

Eligibility criteria for considering studies for this review

Types of studies

There are no restrictions based on study design.

Types of human studies and model systems

Studies of humans, animals (experimental and wildlife [e.g., observational animal studies]), or *in vitro* model systems of immune endpoints are considered relevant. There are no restrictions based on lifestage at exposure or assessment, sex, animal species or strain, or immune model system.

Types of exposures

Exposure to PFOA (CAS# 335-67-1) and PFOS (CAS# 1763-23-1) 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 PFOA or PFOS, differences in the sensitivities of these methods will be considered when assessing the risk of bias ("internal validity") of individual studies.

Types of outcomes

Immunotoxicity considered in this evaluation is defined in the context of immune responses and changes in immune-related measures that reflect the four main categories of immune response: immunosuppression, immunostimulation, sensitization and allergic response, and autoimmunity. Publications must include an indicator of PFOA or PFOS 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 the most direct, or applicable, to the project. Secondary outcomes are relevant, but less direct and can include upstream indicators, risk factors, intermediate outcomes, or related measures to our primary outcomes.

For the evaluation of immunotoxicity, primary outcomes are those with more predictive value for immunotoxicity such as disease resistance assays and functional immune parameters. Secondary outcomes are those with less predictive value for immunotoxicity such as observational parameters including cell counts or cytokine levels. This dichotomy separating the more and less predictive measures

Table 1. Outcomes considered relevant for study eligibility

Humans	Animals*	*In vitro* Assays
Primary outcomes	*Primary outcomes*	*Primary outcomes*
Immune-related diseases and measures of immune function *Immunosuppression* (e.g., otitis, infections, or decreased vaccine antibody response); *Sensitization and allergic response* (e.g., atopic dermatitis or asthma); *Autoimmunity* (e.g., thyroiditis or systemic lupus erythematosus)	Disease resistance assay or measures of immune function *Disease resistance assays* (e.g., host resistance to influenza A or trichinella, changes in incidence or progression in animal models of autoimmune disease) *Immune function assays following in vivo exposure to the test substance* (e.g., antibody response [T-cell dependent IgM antibody response (TDAR)], natural killer cell [NK] activity, delayed-type hypersensitivity [DTH] response, phagocytosis by monocytes, local lymph-node assay [LLNA])	*Immune function assays following in vitro exposure to the test substance* (e.g., natural killer cell [NK] activity, phagocytosis or bacterial killing by monocytes, proliferation following anti-CD3 antibody stimulation of spleen cells or lymphocytes)
Secondary outcomes	*Secondary outcomes*	*Secondary outcomes*
*Immunostimulation*** (e.g., unintended stimulation of humoral immune function) *Observational immune endpoints* (e.g., lymphocyte counts, lymphocyte proliferation, cytokine levels, serum antibody levels, or serum autoantibody levels)	*Observational immune endpoints* (e.g., lymphoid organ weight, lymphocyte counts or subpopulations, lymphocyte proliferation, cytokine production, serum antibody levels, serum or tissue autoantibody levels, or histopathological changes in immune organs)	*Observational immune endpoints following in vitro exposure to the test substance* (e.g., general mitogen-stimulated lymphocyte proliferation, cytokine production)

* Note that the protocol will consider experimental animal studies and observational animal studies (e.g., wildlife studies without a controlled exposure).

** Note that stimulation of the immune response is not adverse per se and most vaccine preparations include adjuvants to aid in stimulation of an immune response to microbes. It is generally agreed that stimulation of the immune system should not be disregarded (WHO 2012). Unintended immunostimulation will be considered for possible hazard in the context of potency and persistence of the elevated immune response. Because evaluation of immunostimulation is less well established for health assessment, outcomes that could be evaluated under autoimmunity or sensitization will be evaluated under these more established categories when possible.

of immunotoxicity is consistent with testing strategies that rely on more sensitive and predictive immune assays (see Luster *et al.* 1992, US EPA 1996a, b, 1998) and the NTP and WHO methods to categorize the evidence of immune system toxicity. Under these systems, measures of immune function or the ability of the immune system to respond to a challenge are weighed more heavily than observational parameters (Germolec 2009, WHO 2012).

For *in vitro* studies, we are interested in immune measures that may support the biological plausibility of observed immune outcomes (see "**Assessment of biological plausibility provided by other relevant studies**" for further discussion). For example, *in vitro* stimulation of immunoglobulin E (IgE) production would support a functional measure of sensitization or allergic response, but it would not support suppression of the natural killer (NK) response. It is generally accepted that *in vitro* systems to evaluate sensitization or immunosuppression would not be able to reproduce the complexity of cellular and soluble interactions that are involved in immune response (this is not unique to the evaluation of immunotoxicity). However, tiered approaches for *in vitro* assays have been proposed and progress has been made in developing assays or groups of assays to assess immunotoxicity with *in vitro* tests (Gennari *et al.* 2005, Carfi *et al.* 2007, Galbiati *et al.* 2010, Lankveld *et al.* 2010). Given the complexity of the immune response, the *in vitro* assessment of immunotoxicity is more likely to have predictive value when the substance evaluated is a direct immunotoxicant, such as a chemical that displays myelotoxicity (killing of immune cells).

Currently within the field of immunotoxicology, *in vitro* data in the absence of *in vivo* human or animal data are considered to provide evidence that is of low predictive value for hazard identification conclusions (see **Step 7:** Integrate evidence to develop hazard identification conclusions for further discussion of the process of integrating the evidence to develop hazard identification conclusions). *In vitro* approaches play a role as a screening tool to identify chemicals that should be subjected to more predictive immunotoxicity testing (Galbiati *et al.* 2010, WHO 2012). In the context of this evaluation, it is envisioned that strong evidence for a relevant immune process from mechanistic or *in vitro* data alone could indicate a greater potential that the substance is an immune hazard to humans and *in vivo* studies are suggested for a more definitive conclusion.

Types of publications

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

There are no date or language restrictions. Review articles and health assessments will be collected for the purposes of reviewing the reference list and will not contribute to the final number of studies considered eligible unless they contain original data.

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:

- Cochrane Library
- EMBASE
- EPA's ACToR (Aggregated Computational Toxicology Resource)
- PubChem
- The Environmental Protection Agency's (EPA) 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)
- Latin American and Caribbean Health Science Information database (LILACS)

- PubMed
- Scopus
- Toxline
- Web of Science

The search terms were identified by (1) reviewing Medical Subject Headings for relevant and appropriate terms, (2) extracting key terminology from reviews and a sample of relevant primary data studies, and (3) review of PFOA search terms from a draft systematic review of developmental PFOA exposure and fetal growth (Johnson *et al.* 2013, Koustas *et al.* 2013) [note that no similar review of PFOS was located so the search for PFOS was developed using search terms from methods #1 and #2 and by analogy to the published PFOA review]. 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 was used in conjunction with text word searches. **Appendix 1** shows the search strategy and specific terminology for PubMed and other databases.

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

Hand searches

Hand searches 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 PFOA and PFOS 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 Program Results and Status Search 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 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). We will also consult subject matter experts and agencies represented on the NTP Executive Committee[2] that may have data that addresses this topic. 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 listserv 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

NTP will only consider publically available information. If a study that may be critical to the evaluation has not been peer-reviewed, the 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 details 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 peer-review would be conducted through the use of experts who have been screened for conflict of interest before confirming 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 Software[3]. 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

[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).

[3] DistillerSR® (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.

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 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 discussion, involving a third member of the review team if necessary. A copy of articles that clearly 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 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 represented in **Figure 1** as recommended by Moher et al. (2009). The study flow schematic in **Figure 1** distinguishes between full-text assessment to determine eligibility and any subsequent exclusion based on full-text review with documentation of reason. The PRISMA Flow Diagram Generator can be used to develop study flow schematics (http://theta.utoronto.ca/tools/prisma).

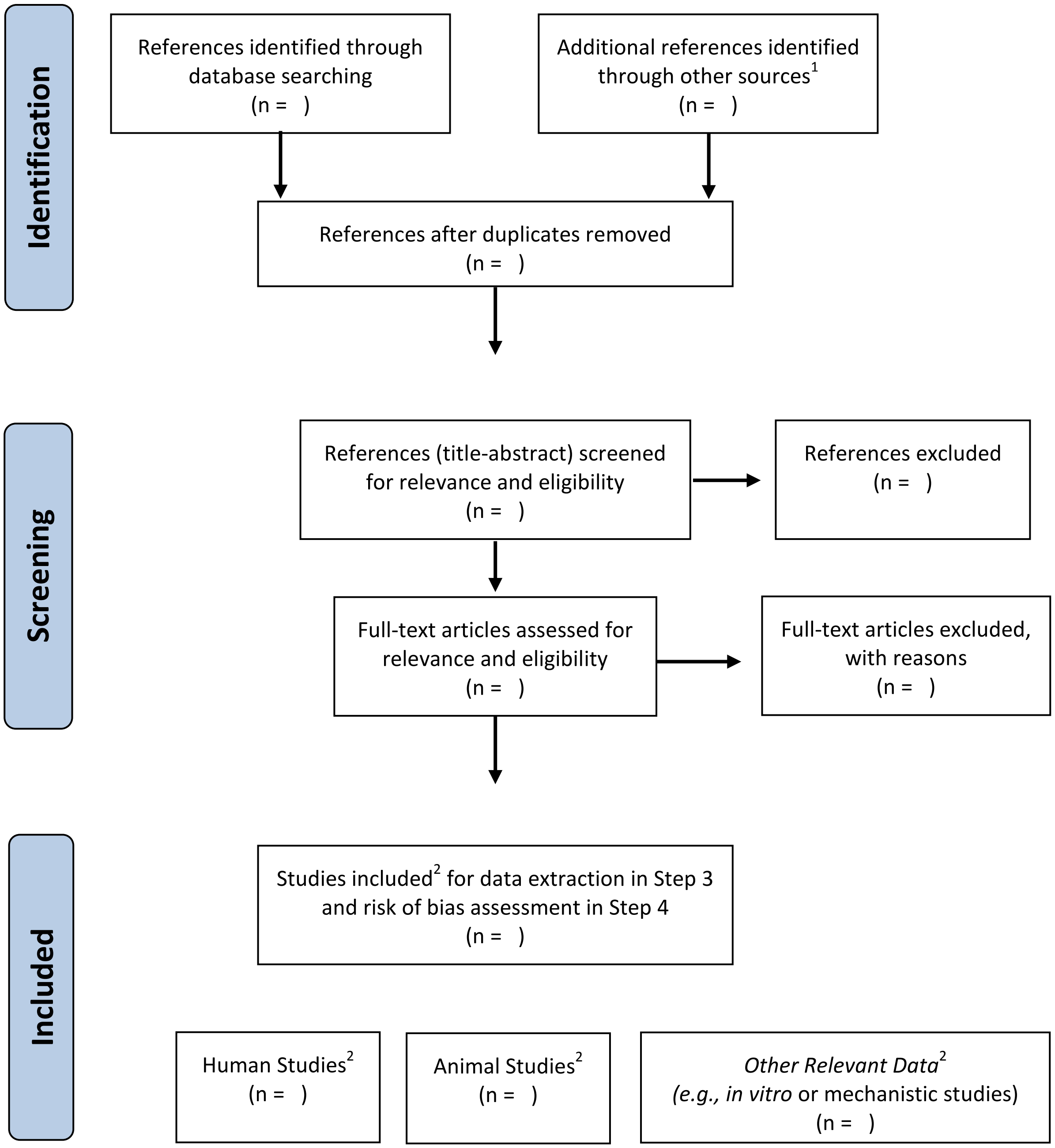
One member of the review team will independently scan the bibliographies of the included studies, relevant reviews, and government reports and other "grey literature" for relevant references that were not identified from database searches. Eligibility will be confirmed by a second screener in order to 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**).

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." If this occurs, we will disseminate these findings as a potential stimulus for further original research.

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.

Figure 1. Flow of information through the phases of the review (adapted from Moher *et al.* 2009)


Identification
References identified through database searching (n =)
Additional references identified through other sources[1] (n =)
References after duplicates removed (n =)
Screening
References (title-abstract) screened for relevance and eligibility (n =)
References excluded (n =)
Full-text articles assessed for relevance and eligibility (n =)
Full-text articles excluded, with reasons (n =)
Included
Studies included[2] for data extraction in Step 3 and risk of bias assessment in Step 4 (n =)
Human Studies[2] (n =)
Animal Studies[2] (n =)
Other Relevant Data[2] *(e.g., in vitro* or mechanistic studies) (n =)

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

[2] The number of studies included in qualitative synthesis will be recorded. If quantitative synthesis or meta-analyses are performed, the number of studies included in the quantitative syntheses will also be recorded.

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 along with the protocol.

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 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 data when a smaller report provides data on an outcome 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 (**Table 3**).

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.* 2011)[4]. In addition, the web-based DistillerSR® 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.	Abee L. Boyles, Ph.D.
Tel (919) 541-5021	Tel (919) 541-7886
thayer@niehs.nih.gov	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].

[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.

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 in the study (e.g., alcohol, smoking in some situations)
	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 (e.g., cardiovascular, immune, nervous, etc.)
	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 a “prospective in spirit” approach to assess ability to detect a 20% change from control or referent group response for continuous data or risk 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 exposure 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)

Table 2. Data extraction and analysis elements to summarize study design, experimental model, methodology and results

	List any other chemicals assessed
***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)
	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)
***methods:** treatment*	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)

Table 2. Data extraction and analysis elements to summarize study design, experimental model, methodology and results	
	Route (e.g., oral, inhalation, dermal, injection)
	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, *ad lib*)
***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 (e.g., cardiovascular, immune, nervous, etc.)
	Endpoint (and unit of measurement)
	Diagnostic or method to measure endpoint.
	Statistical methods
	Statistical power is assed during data extraction using a "prospective in spirit" approach to assess ability to detect a 20% change from control response for continuous data or risk ratio of 1.5 for categorical data using the prevalence of exposure in controls or prevalence of outcome in unexposed 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 unexposed 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)

Table 2. Data extraction and analysis elements to summarize study design, experimental model, methodology and results

other	Documentation of author queries for study details
	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 (e.g., cardiovascular, immune, nervous, etc.)
	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 bias[5], 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** which is included as a separate document "Guidance for using OHAT's risk of bias tool for the evaluation of PFOA or PFOS exposure and immunotoxicity" (Appendix_2_PFOAPFOS_RiskofBias.pdf). 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 2013), staff at other federal agencies, and other risk of bias or study quality tools (Genaidy *et al.* 2007, Dwan *et al.* 2010, Shamliyan *et al.* 2010, Shamliyan *et al.* 2011, Koustas *et al.* 2013, Krauth *et al.* 2013). The tool presents a unified approach to evaluating risk of bias for human and animal studies and therefore allows consideration of risk bias elements across that 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 subjects are 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 **Appendix 2**, the separate document providing detailed risk of bias guidance, 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 (Higgins and Green 2011, Viswanathan *et al.* 2012, CLARITY Group at McMaster University 2013) 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 OHAT 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.

[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, 2012).

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 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 the supplemental document (**Appendix 2**) the separate document, describes individual 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 with the specifics tailored to the features of each study design type. 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.

Table 3. Risk of bias assessment

	Experimental Animal[1]	Human Controlled Trials[2]	Cohort	Case-control[3]	Cross-sectional	Case Series
SELECTION BIAS						
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).	X	X				
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. *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.*	X	X				
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**.**			X	X	X	
CONFOUNDING BIAS						
Did the study design or analysis account for important confounding and modifying variables? *Note: a parallel question under detection bias addresses reliability of the measurement of confounding variables.*	X	X	X	X	X	X
Did researchers adjust or control for other exposures that are anticipated to bias results?	X	X	X	X	X	X
PERFORMANCE BIAS						
Were experimental conditions identical across study groups?	X					
Did deviations from the study protocol impact the results? *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.*	X	X	X	X	X	X
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.	X	X				

[1] Experimental animal studies are controlled exposure studies. Non-human animal observational studies could be evaluated using the design features of observational human studies such as cross-sectional study design.

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

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

Risk of bias assessment table continued

	Experimental Animal	Human Controlled Trials	Cohort	Case-control	Cross-sectional	Case Series
ATTRITION/EXCLUSION BIAS						
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	
INFORMATION/DETECTION BIAS						
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. *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.*	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 BIAS						
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

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 (Viswanathan *et al.* 2012, Krauth *et al.* 2013). 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).

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, 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. 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. The schematics include all of the applicable risk of bias questions with several questions identified as key criteria for evaluating study quality. The clear majority of human studies available for addressing environmental health questions are observational studies (e.g., cohort, case-control, and cross sectional studies), and for observational studies the key questions address exposure characterization, outcome assessment, and confounding. The majority of animal studies are experimental studies, and because these studies have controlled exposure, the only key question for experimental studies is on outcome assessment (see also **Table 4** for observational studies and **Table 5** for experimental studies).

First Tier

For observational studies (most human studies), to be placed in the 1st tier, a study must be rated as "definitely low" or "probably low" for the following key risk of bias elements AND have at least 50 percent of the other applicable items answered "definitely low" or "probably low" 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 experimental studies (most animal studies), to be placed in the 1st tier, a study must be rated as "definitely low" or "probably low" for the following key risk of bias element AND have at least 50 percent of the other applicable items answered "definitely low" or "probably low" risk of bias.

- Can we be confident in the outcome assessment?

Third Tier

For observational studies (most human studies), to be placed in the 3rd tier, a study must be rated as "definitely high" or "probably high" for the following key 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 experimental studies (most animal studies), to be placed in the 3rd tier, a study must be rated as "definitely high" or "probably high" for the following key risk of bias element 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?

Second Tier

For either observational or experimental human or 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 observational studies

		Risk of Bias Criteria and 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?	Were the comparison groups appropriate?	Did researchers adjust or control for other exposures that are anticipated to bias results?	Did deviations from the study protocol impact the results?	Were outcome data incomplete due to attrition or exclusion from analysis?	Were the outcome assessors blinded to study group or exposure level?	Were confounding variable assessed consistently across groups using valid and reliable measures	Were all measured outcomes reported?	Were there any other potential threats to internal validity (e.g., inappropriate statistical methods)?
1st 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	+	++	+	-	+	+	+	+	+	+	-
2nd tier	**example 1**	-	+	++	++	--	+	-	+	+	+	+
	study does not meet criteria for “low” or “high” **example 2**	+	++	+	+	-	-	-	--	+	-	+
	example 3	--	-	--	++	-	+	+	+	+	+	+
3rd 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	--	-	-	--	+	-	-	+	--	+	--

Risk of bias response options for individual items

++	Definitely low risk of bias	--	Definitely high risk of bias
+	Probably low risk of bias	-	Probably high risk of bias

Studies are evaluated on all applicable risk of bias questions based on study design (columns for 11 of the possible 15 risk of bias questions are shown in this example schematic with 3 key risk of bias items applicable for observational studies (i.e., cohort, case-control, cross-sectional, or case series studies). The rating or answer to each risk of bias question is selected from 4 options: definitely low risk of bias (++), probably low risk of bias (+), probably high risk of bias (-), or definitely high risk of bias (--) on an outcome basis prior to determining the tier.

Table 5. Conceptual schematic for determining tiers of study quality for individual experimental studies

		Risk of Bias Criteria and Ratings													
		Key criteria	**Other criteria**												
Category	**Guidance**	Can we be confident in the outcome assessment?	Was administered dose or exposure level adequately randomized?	Was allocation to study groups adequately concealed?	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? (note: applies to animal studies not HCTs)	Did deviations from the study protocol impact the results?	Were 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 variable assessed consistently across groups using valid and reliable measures	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)?
1st 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	+	+	-	+	+	+	+	-	+	+	+	+	+	-
2nd tier	study does not meet criteria for “low” or “high” **example 1**	-	++	-	-	--		+	-	-	+	+	+	+	+
	example 2	+	+	-	-	-		-	--	-	--	+	-	-	+
3rd 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	--	--	-	-	+	-	-	-	-	+	--	+	+	-

Risk of bias response options for individual items

++	Definitely low risk of bias	--	Definitely high risk of bias
+	Probably low risk of bias	-	Probably high risk of bias

Studies are evaluated on all applicable risk of bias questions based on study design (columns for 14 of the possible 15 risk of bias questions are shown in this example schematic with 1 key risk of bias item applicable for experimental studies (i.e., controlled exposure human [human controlled trials or HCT] or animal studies). The rating or answer to each risk of bias question is selected from 4 options: definitely low risk of bias (++), probably low risk of bias (+), probably high risk of bias (-), or definitely high risk of bias (--) on an outcome basis prior to determining the tier.

In vitro studies and other relevant data

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. ToxRTool[6] (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 (Schneider *et al.* 2009, Bevan and Strother 2012). Reporting quality is not considered an appropriate metric to assess risk of bias (Higgins and Green 2011, Viswanathan *et al.* 2012).

Other relevant data

Similar to *in vitro* studies, to our knowledge no risk of bias tool has been developed for other studies that may contribute to other relevant data such as mechanistic or toxicokinetic studies that are not readily evaluated with traditional animal or human risk of bias tools and none is proposed in the current protocol.

Planned interim analyses

We will conduct an interim analysis after assessment of risk of bias of the human and animal studies to collect a list of any studies that provide other relevant data for immune-related effects of PFOA or PFOS that are not amenable to evaluation of study quality under the risk of bias approach for human and animal studies.

As a near-term future research effort 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. Tables 7-9 refer to an example compound, not PFOA or PFOS, and are provided for illustration purposes only (see **Table 6** for animal study example, **Table 7** for human study example, and **Table 8** for *in vitro* study example). Studies will also be presented graphically across collections of studies based on effect size (for human and animal studies) or concentration-specific response for *in vitro* studies.

[6] The ToxRTool worksheets are freely available in Microsoft Excel® format at the ECVAM website (http://ihcp.jrc.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.

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
- *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 useful 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– not for PFOA or PFOS

Reference, Study Design & Population	Health Outcome	Exposure	Statistical Analysis	Results
(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	**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/m^2)	**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	**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	**adjOR (95% CI)** ***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

Risk of bias response options for individual items: should we delete domains from this table?

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 and reliable measures?	++	yes, used standard NHANES methods
	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

Table 6. Example of tabular summary for an human study– not for PFOA or PFOS

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

Risk of bias response options for individual items:	
++	**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– not for PFOA or PFOS

Reference, Animal Model, and Dosing	Health Outcome
(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 (EE_2) used as postive control	**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

Results

group	**mean ± SE**	**% control (95%CI)***	**mean ± SE**	**% control (95%CI)***
leptin	**males**		**females**	
naive	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				
naive	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

Risk of bias response options for individual items:

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
	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
Performance		🟩	
	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

Table 7. Example of tabular summary for an animal study– not for PFOA or PFOS

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

Risk of bias response options for individual items:	
++	**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 *in vitro* study – not for PFOA or PFOS

Reference, Model, and Treatment	Endpoint	Concentration (µM) Specific 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 reported by authors, ↓ = statistically significant decrease reported by authors

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 confidence 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 2013), **Figure 2**] is primarily based on guidance from the Grading of Recommendations Assessment, Development and Evaluation (GRADE) Working Group (Balshem *et al.* 2011, 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 (defined here as other than whole animal studies, and including both cell systems, computational toxicology 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 – February 2013 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 evaluations are typically to identify potential adverse health effects and randomized clinical trials are not considered ideal for this purpose (Oxman *et al.* 2006, Silbergeld and Scherer 2013). The most relevant data for addressing environmental health questions are human observational epidemiology and experimental animal studies and these data need to be considered with clear appreciation for their strengths and limitations. 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). Aspects of this protocol that address Hill considerations on causality are discussed in Step 6.

The framework described below only applies to human and animal studies. To our knowledge there is no analogous model to develop confidence ratings for other relevant data such as outcomes from *in vitro*, mechanistic, cellular, genomic, or mode of action studies. Thus our current approach for considering the level of support provided by other relevant data including *in vitro* studies is described separately in a later section of this the document (see "**Assessment of biological plausibility provided by other relevant studies**"). But, as a future research effort we are interested in developing a framework for other relevant data 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.

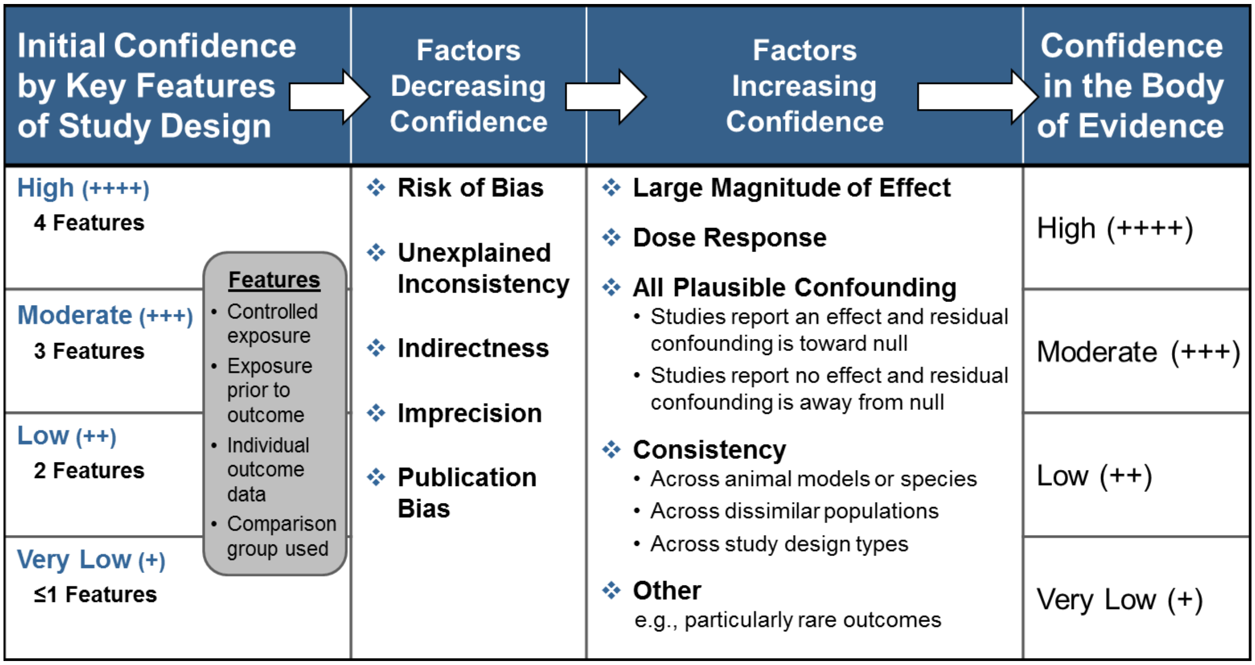
In the context of identifying research needs, a conclusion of "High Confidence" indicates that further research is very unlikely to change our confidence in the apparent relationship between exposure to the substance and the outcome. Conversely, a conclusion of "Very Low Confidence" suggests that further research is very likely to impact confidence in the apparent relationship.

To assess confidence in the body of evidence, available studies on a particular outcome are grouped by key study design features. Then, each collection of studies is given an initial confidence rating by key study design features (**Figure 2**). 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/animal models or species, and consideration of confounding or other biases that increase our confidence in the association or effect)[8]. Consideration of consistency across study designs, human populations, or animal species 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 Approach[9]. Confidence ratings will be summarized in evidence profile tables (see **Table 9** and **Table 10** for examples).

[8] The reasons for downgrading (or upgrading) confidence may not fit neatly into a single domain of the body of evidence. If the decision to downgrade is borderline for two domains, the body of evidence is downgraded once to account for both partial concerns. Similarly, the body of evidence is not downgraded twice for what is essentially the same limitation (or upgraded twice for the same asset) that could be considered applicable to more than one domain of the body of evidence.

[9] (NTP (National Toxicology Program) 2012), see "Meeting Materials and Public Comments", then "NTP BSC Working Group Report on the Draft NTP Approach"

Figure 2. 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	❖ Risk of Bias ❖ Unexplained Inconsistency ❖ Indirectness ❖ Imprecision ❖ Publication Bias	❖ Large Magnitude of Effect ❖ Dose Response ❖ All Plausible Confounding • Studies report an effect and residual confounding is toward null • Studies report no effect and residual confounding is away from null ❖ Consistency • Across animal models or species • Across dissimilar populations • Across study design types ❖ Other e.g., particularly rare outcomes	High (++++)
Moderate (+++) 3 Features Features • Controlled exposure • Exposure prior to outcome • Individual outcome data • Comparison group used			Moderate (+++)
Low (++) 2 Features			Low (++)
Very Low (+) ≤1 Features			Very Low (+)

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).

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 considered to be 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.

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
human prospective cohort studies (n=)								
vaccine antibody response	**↓risk of bias**	**↓inconsistency**	**↓indirectness**	**↓imprecision**	**↓publication bias**			
	• 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)	narrative or results of meta-analysis	• inconsistent (0) • consistent within evidence stream (+1)	
	↑magnitude of effect	**↑dose-response**	**↑plausible confounding**	**↑other**				
	• large (+1) • very large (+2)	• evidence of gradient (+1)	• demonstrated effect or suggest spurious effect when results show no effect (+1)					
human cross-sectional studies (n=)								
vaccine antibody response	**↓risk of bias**	**↓inconsistency**	**↓indirectness**	**↓imprecision**	**↓publication bias**			
	• 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)	narrative or results of meta-analysis		
	↑magnitude of effect	**↑dose-response**	**↑plausible confounding**	**↑other**				
	• large (+1) • very large (+2)	• evidence of gradient (+1)	• demonstrated effect or suggest spurious effect when results show no effect (+1)					

Table 10. Example 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**
experimental animal studies – mice and rats (n=)								
T-cell-dependent antibody response (TDAR)	**↓risk of bias**	**↓inconsistency**	**↓indirectness**	**↓imprecision**	**↓publication bias**			
	• 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)	• narrative or results of meta-analysis	• inconsistent (0) • consistent within evidence stream (+1)	
	↑magnitude of effect	**↑dose-response**	**↑plausible confounding**	**↑other**				
	• large (+1) • very large (+2)	• evidence of gradient (+1)	• demonstrated effect or suggest spurious effect when results show no effect (+1)					
experimental animal studies – non-mammals, zebra fish (n=)								
antibody response	**↓risk of bias**	**↓inconsistency**	**↓indirectness**	**↓imprecision**	**↓publication bias**			
	• 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)	narrative or results of meta-analysis		
	↑magnitude of effect	**↑dose-response**	**↑plausible confounding**	**↑other**				
	• large (+1) • very large (+2)	• evidence of gradient (+1)	• demonstrated effect or suggest spurious effect when results show no effect (+1)					

Initial confidence based on study design

An initial confidence rating is determined by the ability of the study design to address the confidence that exposure preceded and was associated with the outcome (**Figure 2**, column 1). This ability is reflected in the presence or absence of four key study design features that determine initial confidence ratings and studies are differentiated based on whether or not: (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), 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.” See Appendix B of the Draft 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. Observational animal studies could be considered using these same study design features. 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 2**, column 2).

Risk of bias across studies

Risk of bias criteria for individual studies were described earlier in the protocol (see “**Step 4: Assess quality of individual studies**”). In this step, risk of bias for a given health outcome is considered across studies.

Summary of risk of bias ratings for each outcome

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 10 observational human studies). This type of summary 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 risk of bias ratings for each outcome (hypothetical summary for a set of 10 observational human studies)

Questions	20%		40%		60%		80%		100%	
Was administered dose or exposure level adequately randomized?	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Was allocation to study groups adequately concealed?	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/
Were the comparison groups appropriate?	++	++	++	++	+	+	-	-	--	--
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?	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Did deviations from the study protocol impact the results?	++	++	+	+	+	+	+	+	+	-
Were the research personnel and human subjects blinded to the study group during the study?	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
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?	++	+	+	+	+	+	+	+	+	+

++	**definite low risk of bias**
+	**probably low risk of bias**
-	**probably high risk of bias**
--	**definitely high risk of bias**
n/a	**not applicable**

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.

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 (see "**Step 4: Assess quality of individual studies**") 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 1st tier risk of bias	**1st**	**1st**	**2nd**	**1st**	**1st**	**1st**	**1st**	**2nd**	**2nd**	**3rd**
-1 (serious)	most studies are 2nd tier risk of bias	**1st**	**1st**	**1st**	**1st**	**2nd**	**2nd**	**2nd**	**2nd**	**3rd**	**3rd**
-2 (very serious)	most studies are 3rd tier risk of bias	**1st**	**1st**	**2nd**	**3rd**	**3rd**	**3rd**	**3rd**	**3rd**	**3rd**	**3rd**

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 1st tier risk of bias	**1st**	**1st**	**1st**	**1st**	**1st**	**1st**	**1st**	**2nd**	**2nd**	**3rd**
-1 (serious)	most studies are 2nd tier risk of bias	**1st**	**1st**	**1st**	**1st**	**2nd**	**2nd**	**2nd**	**2nd**	**3rd**	**3rd**

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 in the 1st or 2nd tier for risk of bias. This will be done by comparing the consistency of findings from studies in the 3rd tier risk of bias with findings from studies in the 1st and 2nd tier 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 (Guyatt *et al.* 2011d, AHRQ 2012). 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, 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 determine consistency across studies that are similar in study design, dose or exposure level, and the health outcome assessed:

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 thumbs is if χ^2 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.

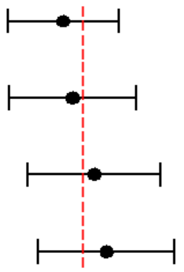
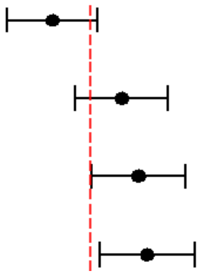
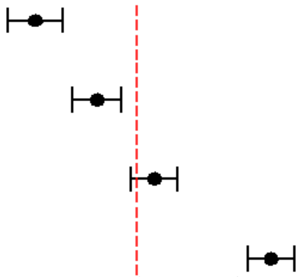
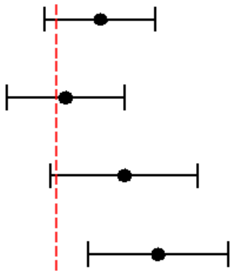
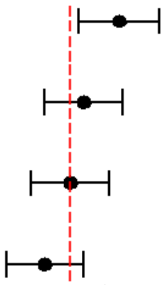
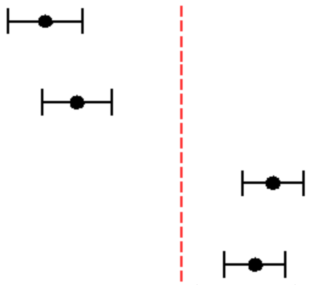
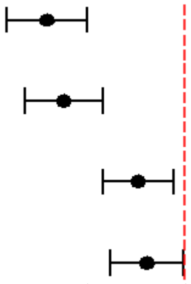
Tau square (T^2, tau^2, τ^2): An estimate of the between-study variance in a random-effects meta-analysis. A τ^2 >1 suggests presence of substantial statistical heterogeneity.

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

Table 14. Factors to consider when considering consistency of results

No downgrade	One level downgrade (serious)	Two level downgrade (very serious)
• Point estimates similar • Confidence intervals overlap • Statistical heterogeneity is non-significant (p≥0.1) • I^2 of ≤50%	• Point estimates vary • Confidence intervals show minimal overlap • Statistical heterogeneity has low p-value (p≤0.1) • I^2 of >50% to 75%	• Point estimates vary widely • Confidence intervals show minimal or no overlap • Statistical heterogeneity has low p-value (p≤0.1) • I^2 of >75%
Example A  χ^2 p-level = 0.767; I^2 = <<1%; τ^2 = <<1	**Example A**  χ^2 p-level = 0.017; I^2 = 71%; τ^2 = 0.044	**Example A**  χ^2 p-level = <0.001; I^2 = 98%; τ^2 = 1.022
Example B  χ^2 p-level = 0.241; I^2 = 29%; τ^2 = 0.046	**Example B**  χ^2 p-level = 0.068; I^2 = 58%; τ^2 = 0.025	**Example B**  χ^2 p-level = <0.001; I^2 = 98%; τ^2 = 0.774
Example C  χ^2 p-level = <0.001; I^2 = 86%; τ^2 = 0.111 *in this case there is less concern for numerical estimates of heterogeneity because point estimates are in the same direction		

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 (e.g., antibody response will not be combined with changes in host resistance to influenza); 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, all eligible studies will be reviewed and included for evaluating and rating the quality/strength of the human evidence.

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

- Concentrations of PFOA or PFOS measured/estimated for each exposure or treatment group
- Estimates of effect for immune-related health 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, 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.

Planned interim analysis

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).

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 (Guyatt *et al.* 2011c, AHRQ 2012).

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[10]. 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 (Twombly 1998, Medlin 2003, Jahnke *et al.* 2005, Shelby 2005). 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.

[10] The appropriateness of the window of exposure given the health outcome measured will generally be considered as part of evaluating directness and applicability (i.e., "Are the results of the study credible?" versus "Did the study design address the topic of the evaluation?"). However, there may be cases where time between exposure and health outcome assessment can be considered a 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. Duration of follow up is also relevant to the indirectness or applicability of a study if the duration of follow up was not sufficient for the development of the outcome of interest (e.g., a 6-week study of cancer endpoints). In this case, an otherwise well-designed and well-conducted study may suffer from indirectness despite having low risk of bias (Viswanathan *et al.* 2012).

Planned interim analyses

The guidance below is meant to be as comprehensive as possible, but it is possible 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.

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 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 the effects of PFOA or PFOS 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.
- *Bird, reptile amphibian, fish, and other non-mammalian vertebrate model systems:* Most immune cell types and function are relatively consistent across vertebrate systems. However, 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 non-mammalian vertebrates 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.
- *Invertebrate model systems:* There is a phylogenetic difference in the ability of the immune cells to confer lifelong protection via adaptive immunity. Invertebrates rely primarily on innate immune responses. For this reason, studies conducted in invertebrates will be downgraded two levels. As with non-mammalian vertebrates, 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.

Health outcomes

- *Primary health outcomes:* The primary outcomes for this evaluation (see "**Types of outcomes**" under criteria for study inclusion) 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.

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, drinking water, or feeding studies are considered relevant because oral exposure through drinking water and food are considered important sources of exposure to PFOS and PFOA in humans (ATSDR 2009).
    - Dermal (no downgrade) – Dermal exposure is considered relevant for contact with surface waters, soil, dusts, soil, and direct contact of skin with consumer products such as treated textiles (e.g., older carpet treatments) (ATSDR 2009).
    - Subcutaneous injection (no downgrade) – A route of administration that bypasses first pass metabolism is relevant for certain exposure scenarios, e.g., the long term goal of several agreements is to eliminate PFOS, however production of PFOS has continued for limited purposes including certain medical devices (ATSDR 2009, OECD 2013).
    - Inhalation (no downgrade) – Inhalation studies are considered relevant, especially to occupational cohorts.
    - Intraperitoneal injection (one level downgrade) – These studies will be downgraded one level because they are not relevant to the nature of human exposure.
    - Water for aquatic species, or culture media for invertebrates (one level downgrade) – These studies will be downgraded one level because they are not relevant to the nature of human exposure.

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

Studies that assess immune-related outcomes following longer periods of exposure are generally expected to be more informative than studies of shorter duration. Many standard testing guidelines for immunotoxicity endpoints specify a minimum of a 28-day exposure period with immune challenge taking place during the end of the exposure period (US EPA 1996a, b, 1998). However, a 14-day exposure period is common in many studies published prior to 2000. The lifetime of cells from the innate and adaptive immune systems range from hours to years and the immune system is in a constant state of renewal. There will be no downgrading for either acute dosing regimens in experimental studies or short window of time between exposure and outcome assessment. However, duration of treatment and window of time between exposure and outcome will be considered in stratification of studies and in examination of possible factors for unexplained inconsistency.

Tabular summary of guidance for evaluating directness

Table 15. Guidance for downgrading human studies for 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

Table 16. Guidance for downgrading animal studies for directness

Animal model		**Health outcomes**		**Route of administration**		**Time between treatment and outcome assessment**	**Overall downgrade**
Mammalian	0	primary	0	oral, sc injection, dermal, inhalation	0	0	0
				intraperitoneal injection	-1	0	-1
		secondary	-1	oral, injection, dermal, inhalation	0	0	-1
				Intraperitoneal (ip) injection	-1	0	-2
Non-mammalian vertebrates	-1	primary	0	oral, sc injection, dermal, inhalation	0	0	-1
				ip, water for aquatic species	-1	0	-2
		secondary	-1	oral, sc injection, dermal, inhalation	0	0	-2
				ip, water for aquatic species	-1	0	-3
Invertebrates	-2	primary	0	oral, dermal, inhalation	0	0	-2
				ip, water for aquatic species	-1	0	-3
		secondary	-1	oral, dermal, inhalation	0	0	-3
				ip, water for aquatic species	-1	0	-4

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

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 significantly underpowered from the confidence rating phase and analysis of imprecision when confidence 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 powered[11] when assessing precision. When a meta-analysis is not feasible or inappropriate, we will consider the extent to which studies in the evidence base have sufficient power to detect a potentially biologically meaningful difference between groups. 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 due to 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, in most cases where the body of evidence is downgraded for inconsistency in direction of effect we will 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 imprecise.

[11] Statistical power is assed during data extraction using a “prospective in spirit” approach to assess ability to detect a 20% change from control response for continuous data or risk ratio of 1.5 for categorical data using the prevalence of exposure in controls or prevalence of outcome in unexposed 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 unexposed differ, the sample size of the exposed group will be used to determine relative power category.

Table 17. Factors to consider when considering imprecision of results [a]	
0 **(no downgrade)**	**No meta-analysis** • 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. AND • Most studies in the evidence base are "adequately" or "somewhat underpowered" **Meta-analysis** • 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. AND • The sample size for the meta-estimate meets the OIS criterion
-1 downgrade **(serious)**	**Does not clearly meet guidance for 0 (no downgrade) or -2 downgrade**
-2 downgrade **(very serious)**	**No meta-analysis** • 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. AND • Most studies in the evidence base are "underpowered" or "severely underpowered" **Meta-analysis** • 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. AND • 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 preforms 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, non-government organizations (NGOs), or authors with conflicts of interest (reviewed in Guyatt *et al.* 2011e). We will 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 2**, column 3). Consideration of large magnitude of effect, dose-response, and plausible confounding are considered in the GRADE and frameworks (Guyatt *et al.* 2011g, AHRQ 2012). We have added an additional factor to address consistency across human study designs and animal species or 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[12]

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 large magnitude of effect may be based on a single study provided the study is of overall low risk of bias, or few studies provided those studies are of overall low risk of bias and there is no serious unexplained inconsistency among other studies of similar dose or exposure levels. 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 could be 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 0.3-0.6 with no plausible confounders.

[12] Also referred to as *strength of association* or *strength of response*

- 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 the immune outcome under question 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 < 0.3.
- 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 the immune outcome under question 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

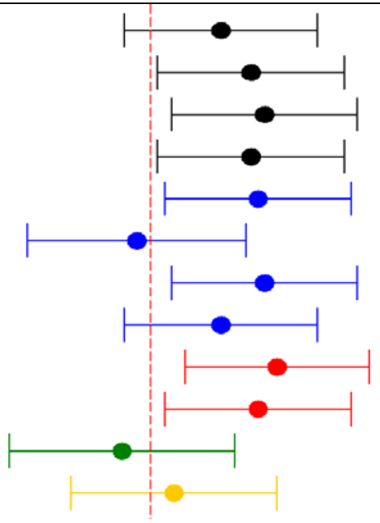
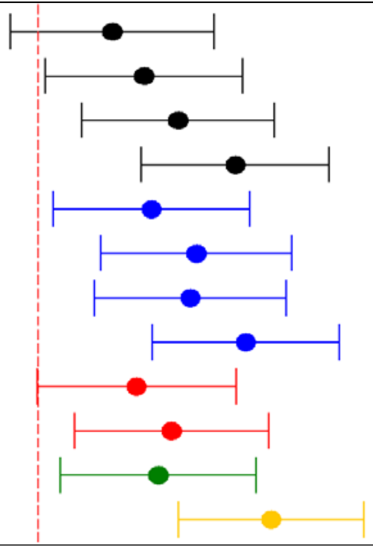
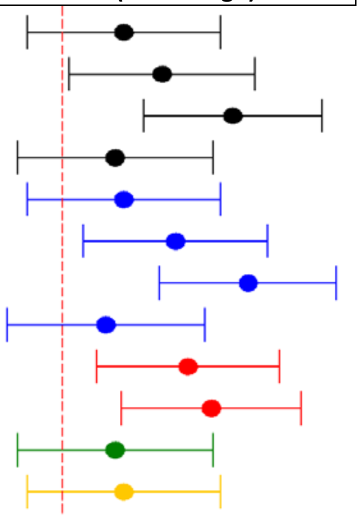
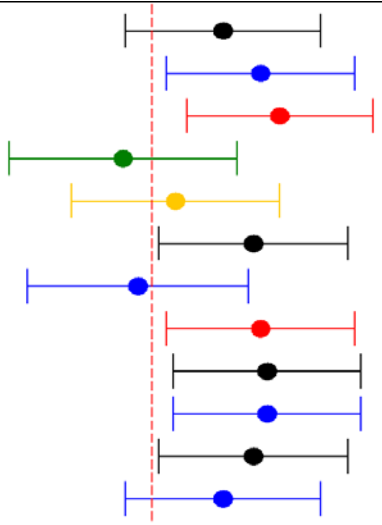
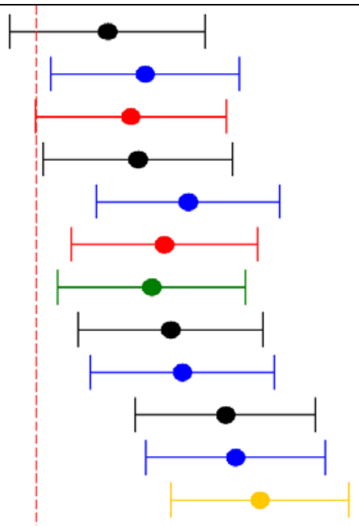
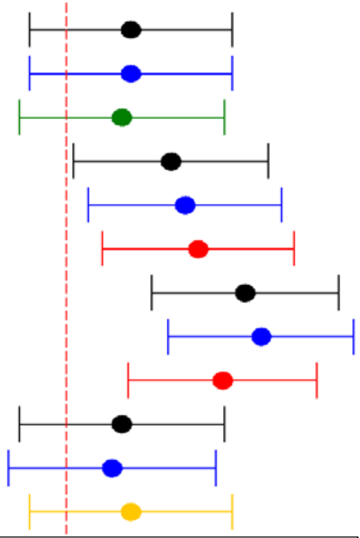
- Non-monotonic dose response is consistently observed in the evidence base

We do not have evidence to suggest that non-monotonicity would be the expected pattern in either human or animal studies of exposure to PFOS or PFOA. However, non-monotonic dose-response relationships have been observed in immunotoxicity (Hastings 2005, Ladics and Loveless 2005). The general concept is for multiple mechanisms of toxicity on the immune system, where each mechanism had a potentially different dose-response curve. Regulatory T cells may be involved in these complex dose-response curves and small changes in the numbers or functions of these cells may affect antigen-driven immune responses, autoimmunity, and other endpoints.

The impact of overt toxicity is another important dose-response consideration. For immune-related endpoints, it is possible that high doses of PFOS or PFOA may cause systemic toxicity and immune effects observed may be indirect effects mediated via stress (e.g., decreased thymus weight is often associated with overt systematic toxicity along with decreased body weight at general high chemical doses).

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 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]
Example A, findings sorted by study and then dose or exposure level (low to high)	**Example B, findings sorted by study and then dose or exposure level (low to high)**	**Example C, findings sorted by study and then dose or exposure level (low to high)**
		
Example A, findings sorted by exposure or dose level (low to high) across studies	**Example B, findings sorted by exposure or dose level (low to high) across studies**	**Example C, findings sorted by exposure or dose level (low to high) across studies**
		

● Study 1 ● Study 2 ● Study 3 ● Study 4 ● Study 5

--------- = null hypothesis reference line

[1] Requires evidence to suggest non-monotonic is expected pattern AND non-monotonic dose response observed in evidence base

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 (Taylor *et al.* 1999, Elliman and Bedford 2001) 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

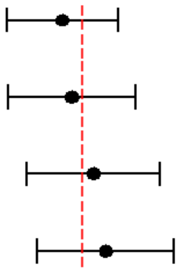
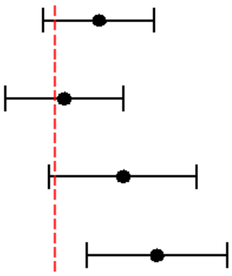
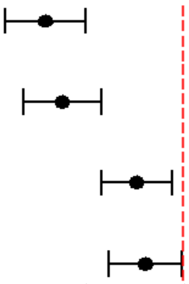
Three types of consistency in the body of evidence can used to support a +1 upgrade:

- <u>across animal models and species</u> - consistent results reported in multiple animal models or species
- <u>across independent studies of different human populations and exposure scenarios</u>
- <u>across study types</u> - 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 ≤50%

Example A	Example B	Example C
		
χ^2 p-level = 0.767; I^2 = <<1%; τ^2 = <<1	χ^2 p-level = 0.241; I^2 = 29%; τ^2 = 0.046	χ^2 p-level = <0.001; I^2 = 86%; τ^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 was 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. We do not anticipate use of the other category for upgrading confidence across the body of studies for this evaluation. If during the course of the evaluation an important additional category for upgrading confidence becomes evident, we will consult experts on the use of an additional factor and changes would be noted as revisions to the protocol.

Combine confidence conclusions for all study types and multiple outcomes

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

After confidence conclusions are developed for a 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.

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

STEP 6: TRANSLATE CONFIDENCE RATINGS INTO LEVEL OF EVIDENCE FOR HEALTH EFFECT

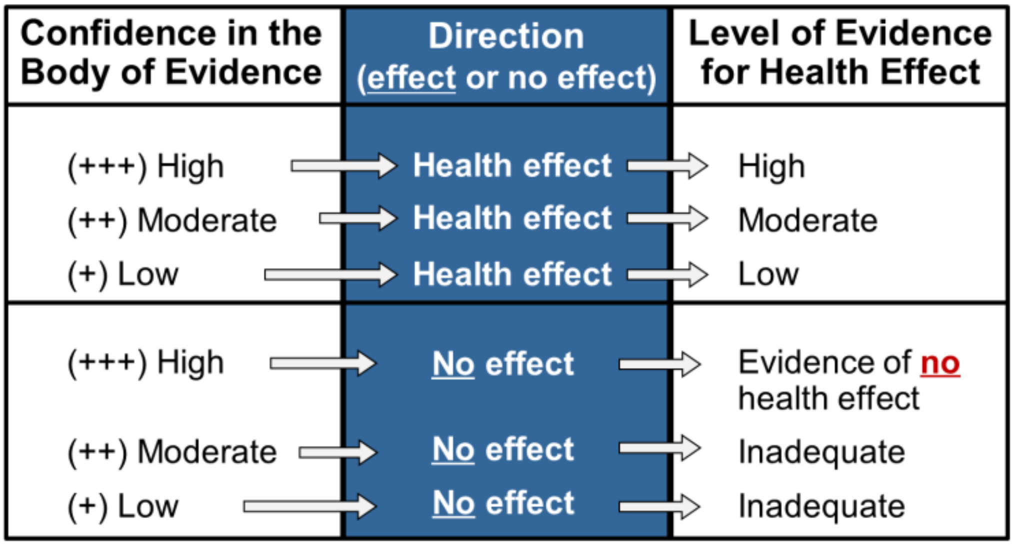
The level of evidence will be assessed separately within the human and non-human animal data sets. The level of evidence for health effect conclusions reflects both the overall confidence in the association between exposure to the substance and the outcome (effect or no effect) and the direction of the 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"[13]. These phrases are defined below and illustrated schematically in **Figure 3.**

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).

[13] 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 3. 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)

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 (Schünemann *et al.* 2011). Some of the causality considerations are also important in step 7 during the process for integrating the evidence to develop conclusions. **Table 20** outlines how these considerations are related to the process of evaluating the confidence in the body of evidence and then integrating the evidence.

Table 20: Aspects of the Hill considerations on causality within the OHAT Approach

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

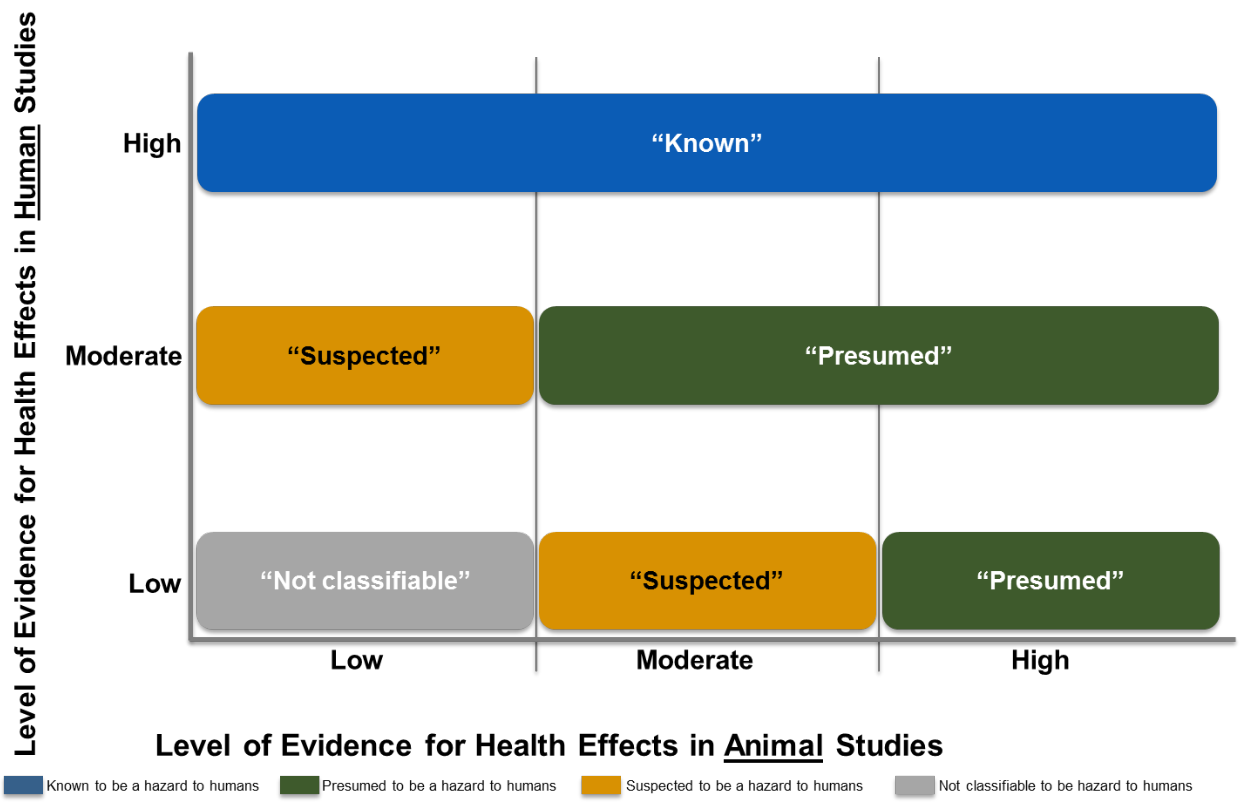
STEP 7: INTEGRATE EVIDENCE TO DEVELOP HAZARD IDENTIFICATION CONCLUSIONS

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.

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 4**):

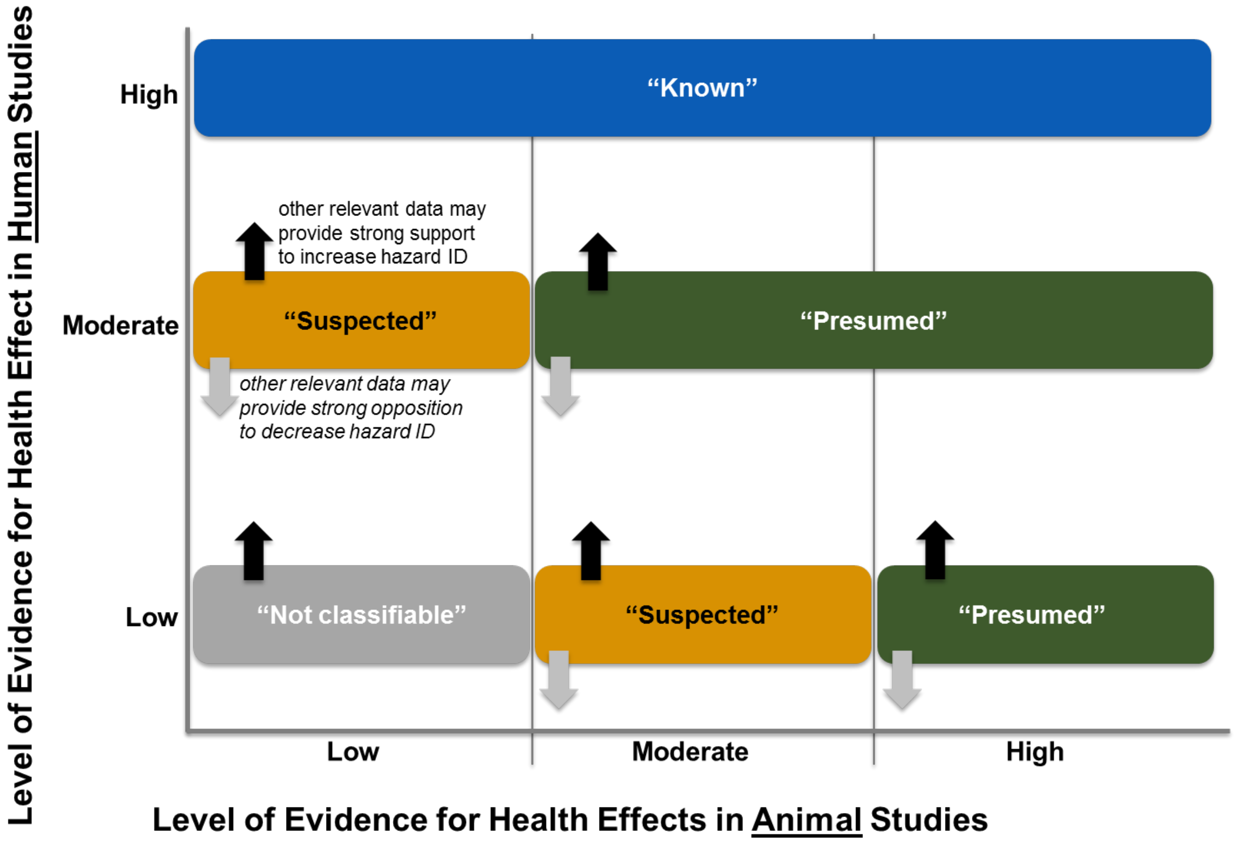
- Known to be a hazard to humans
- Presumed to be a hazard to humans
- Suspected to be a hazard to humans,
- Not classifiable or not identified to be a hazard to humans

Figure 4. Hazard Identification Scheme


Level of Evidence for Health Effects in Human Studies
High
Moderate
Low
"Known"
"Suspected"
"Presumed"
"Not classifiable"
"Suspected"
"Presumed"
Low
Moderate
High
Level of Evidence for Health Effects in Animal Studies
Known to be a hazard to humans
Presumed to be a hazard to humans
Suspected to be a hazard to humans
Not classifiable to be hazard to humans

The NTP does not require mechanistic or mode of action data in order to reach hazard identification conclusions, although when available this and other relevant supporting types of evidence may be used to raise (or lower) the level of the hazard identification conclusion (**Figure 5**). 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 for a relevant biological process from mechanistic or *in vitro* data could result in a conclusion of "suspected" in the absence of human epidemiology or experimental animal data. Alternatively, If the human level of evidence conclusion is low and mechanistic or mode of action data are compelling that evidence from non-human studies is not relevant to human health effects, a hazard identification conclusion of "not classifiable" may be appropriate.

Figure 5. Hazard Identification Scheme with Consideration of Other Relevant Data


Level of Evidence for Health Effect in Human Studies
High
Moderate
Low
"Known"
other relevant data may provide strong support to increase hazard ID
"Suspected"
"Presumed"
other relevant data may provide strong opposition to decrease hazard ID
"Not classifiable"
"Suspected"
"Presumed"
Low
Moderate
High
Level of Evidence for Health Effects in Animal Studies

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 other relevant studies

Any impact of other relevant data on the hazard identification conclusion derived by integrating the human and non-human animal streams is considered in Step 7 (**Figure 5**). Other relevant data could include, but are not limited to *in vitro* or mechanistic data.

- If other relevant data provide strong support for biological plausibility of the relationship between exposure and the health effect, the hazard identification conclusion may be upgraded (indicated by black "up" arrows in **Figure 5**) from that initially derived by considering the human and non-human animal evidence together. 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.
    - To provide support, the mechanistic or *in vitro* data must support biological plausibility of observed immune outcomes from human epidemiology or *in vivo* animal studies.
    - It is also envisioned that strong evidence for a relevant biological process from mechanistic or *in vitro* data could result in a conclusion of "suspected" in the absence of human epidemiology or *in vivo* animal data.
- 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 (indicated by gray "down" arrows in **Figure 5**) from that initially derived by considering the human and non-human animal evidence together.
    - To provide opposition, the mechanistic or *in vitro* data must oppose the biological plausibility of observed immune outcomes from human epidemiology or *in vivo* animal studies.

The strength of the support or opposition presented by the other relevant data is evaluated using the guidance presented in **Figure 6**. The factors outlined in **Figure 6** are conceptually consistent with the factors considered in Step 5 for rating confidence in the body of evidence from human and animal studies. The biological plausibility will be considered for two cases: to inform the biological plausibility of observed outcomes from *in vivo* data, and in the absence of human or animal *in vivo* data.

Data that inform the biological plausibility of observed outcomes from in vivo data

Other relevant studies must be biologically related or along a relevant biological pathway to inform the biological plausibility of observed immune outcomes from *in vivo* human or animal studies. For example, *in vitro* stimulation of immunoglobulin E (IgE) production are relevant to a functional measure of sensitization or allergic response, but it IgE is not relevant to the natural killer (NK) cell response.

Consistency is evaluated within the context of the observed immune outcomes from the *in vivo* human or animal studies. Mechanistic or *in vitro* data that provide information on multiple steps along the relevant biological pathway are more useful in evaluating the biological plausibility. In cases where the mechanism or mode of action is well understood and more widely accepted, consistency between the observed human or animal data and the mechanistic or *in vitro* data are likely to be more informative. Consistency also applies to repeatability or consistent results within the same assay across multiple studies.

Other relevant data must satisfy all of the factors for "strong" evidence in **Figure 6** to provide strong opposition to the biological plausibility of observed outcomes from *in vivo* human or animal studies. The basis for the strong opposition will be described as well as the confidence in the human and animal data from Step 5.

Figure 6. Factors considered when evaluating the support for biological plausibility provided by *in vitro*, cellular, genomic, or mechanistic data

Strong Support[1]	**Weak Support**
Relevance of biological process or pathway to human health	
generally accepted as relevant (e.g., myelotoxicity or bone marrow toxicity)	limited relevance or uncharacterized
Consistency	
Consistency across multiple studies (preferably in more than 2 in different model systems for the same biological pathway)	no studies or unexplained inconsistency
Relevance of concentration	
physiologically relevant or "low" concentration effects (e.g., mean of 3-5ng/ml PFOA and 9–30 ng/ml PFOS in the US population 1999-2010 (CDC 2012) range of 17-5100 ng/ml PFOA and 37-3490 ng/ml PFOS in occupationally exposed adults)	"high" concentration effects (e.g., range above 5100 ng/ml PFOA and 3490 ng/ml PFOS)
Potency	
magnitude of response similar to positive control	weak response relative to positive control
Dose response	
displays expected dose response gradient	no dose response gradient or single concentration tested
Publication bias	
undetected	strongly suspected

[1] A conclusion of "strong" support requires that most elements are met

[2] Physiologically relevant dose range based (Olsen *et al.* 2003a, Olsen *et al.* 2003b, Costa *et al.* 2009, CDC 2012), an effect occurring within an order of magnitude of this range 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

[3] MIAME (Minimum Information About a Microarray Experiment) standards for recording and reporting are recommended by many journals to enable the interpretation of the results of the experiment and potentially to reproduce the experiment (Galbiati *et al.* 2010, http://www.mged.org/Workgroups/MIAME/miame.html).

Strength of in vitro studies in the absence of human or animal in vivo data

We are also interested in whether or not strong evidence for a relevant immunological process from mechanistic or *in vitro* data could result in a conclusion of "suspected" for some immune outcomes in the absence of human epidemiology or *in vivo* animal data. Within the field of immunotoxicology, *in vitro* data in the absence of *in vivo* data are currently considered to provide evidence that is of low predictive value for immunotoxicity hazard identification. If there are in vitro or mechanistic data for immune-related endpoints without relevant human or animal *in vivo* data, the guidance presented in **Figure 6** will also be used to evaluate the strength of support provided by these *in vitro* studies based on whether plausible immunological processes and/or pathways have been identified that are considered relevant to humans.

Mechanistic or *in vitro* data that provide information on multiple steps along a biological pathway are more useful in evaluating the biological plausibility. The relevance of this biological pathway for humans is also applicable here and is related to the consistency of the pathway with the human system. Consistency also applies to repeatability or consistent results within the same assay across multiple studies.

It is generally accepted that *in vitro* systems to evaluate sensitization or immunosuppression would not be able to reproduce the complexity of cellular and soluble interactions that are involved in immune response. This is not unique to the evaluation of immunotoxicity. However, tiered approaches for *in vitro* assays have been proposed to evaluate multiple aspects of the immune response and progress has been made in developing assays or groups of assays to assess immunotoxicity with *in vitro* tests (Gennari *et al.* 2005, Carfi *et al.* 2007, Galbiati *et al.* 2010, Lankveld *et al.* 2010). Given the complexity of the immune response, the *in vitro* assessment of immunotoxicity is more likely to have predictive value when the substance evaluated is a direct immunotoxicant (i.e., kills immune cells), such as a chemical that displays myelotoxicity.

Currently, *in vitro* approaches play a role as a screening tool to identify chemicals that should be subjected to more predictive immunotoxicity testing (Galbiati *et al.* 2010, WHO 2012). In the context of this evaluation, the approach is structured such that strong evidence for a relevant immune process from mechanistic or *in vitro* data could result in a conclusion of "suspected" to be a hazard to humans in the absence of human epidemiology or *in vivo* animal data. A conclusion of strong immunological evidence from mechanistic or *in vitro* data alone is only possible given a body of evidence from *in vitro* studies that satisfies all of the aspects for strong support in **Figure 6**. Given the current low predictive value of *in vitro* immunotoxicological assays, a conclusion based on *in vitro* data alone should be followed up with *in vivo* studies to strengthen the hazard identification conclusion.

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 interest[14]. A more detailed description of the OHAT evaluation process can be found at http://ntp.niehs.nih.gov/go/38138. Confidence ratings and the conclusions derived from them will be finalized after considering this input. When conclusions are oriented towards identifying research

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

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.

REVIEW TEAM

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

AUTHOR DECLARATIONS OF INTEREST

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 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.

Jamie Dewitt, PhD	East Carolina University, Department of Pharmacology and Toxicology
Christopher Lau, PhD	US EPA, ORD/NHEERL
Tony Fletcher, PhD	London School of Hygiene and Tropical Medicine, Department of Social and Environmental Health Research
Dori Germolec, PhD	NIEHS/NTP
Roberta Scherer, PhD	Johns Hopkins University Bloomberg School of Public Health

SOURCES OF SUPPORT

Internal sources

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

External sources

None

PROTOCOL HISTORY AND REVISIONS

Date

March 26, 2013: Draft protocol distributed for comment to technical advisors

April 9, 2013: Draft protocol released publically

REFERENCES

AHRQ. 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]*. Agency for Healthcare Research and Quality.

Andrews N, Miller E, Taylor B, Lingam R, Simmons A, Stowe J, Waight P. 2002. Recall bias, MMR, and autism. *Archives of disease in childhood* 87(6): 493-494.

ATSDR. 2009. Draft Toxcological Profile for Perfluoroalkyls. Registry AfTSaD. Atlanta, GA, U.S. Department of Health and Human Services, Public Health Service, Agency for Toxic Substances and Disease Registry: 404.

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

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*.

Boyles AL, Harris SF, Rooney AA, Thayer KA. 2011. Forest Plot Viewer: a fast, flexible graphing tool. *Epidemiol* 22(5): 746-747.

Calafat AM, Kuklenyik Z, Reidy JA, Caudill SP, Tully JS, Needham LL. 2007. Serum concentrations of 11 polyfluoroalkyl compounds in the u.s. population: data from the national health and nutrition examination survey (NHANES). *Environmental science & technology* 41(7): 2237-2242.

Carfi M, Gennari A, Malerba I, Corsini E, Pallardy M, Pieters R, Van Loveren H, Vohr HW, Hartung T, Gribaldo L. 2007. *In vitro* tests to evaluate immunotoxicity: A preliminary study. *Toxicology* 229(1-2): 11-22.

Carwile JL, Michels KB. 2011. Urinary bisphenol A and obesity: NHANES 2003-2006. *Environmental research* 111(6): 825-830.

CDC. 2012. Fourth National Report on Human Exposure to Environmental Chemicals: Updated Tables, September 2012. U.S. Department of Health and Human Services CfDCaP. Atlanda, GA.

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].

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.

Costa G, Sartori S, Consonni D. 2009. Thirty years of medical surveillance in perfluooctanoic acid production workers. *J Occup Environ Med* 51(3): 364-372.

Detels R, English P, Visscher BR, Jacobson L, Kingsley LA, Chmiel JS, Dudley JP, Eldred LJ, Ginzburg HM. 1989. Seroconversion, sexual activity, and condom use among 2915 HIV seronegative men followed for up to 2 years. *Journal of acquired immune deficiency syndromes* 2(1): 77-83.

DeWitt JC, Shnyra A, Badr MZ, Loveless SE, Hoban D, Frame SR, Cunard R, Anderson SE, Meade BJ, Peden-Adams MM, Luebke RW, Luster MI. 2009. Immunotoxicity of perfluorooctanoic acid and perfluorooctane sulfonate and the role of peroxisome proliferator-activated receptor alpha. *Crit Rev Toxicol* 39(1): 76-94.

DeWitt JC, Peden-Adams MM, Keller JM, Germolec DR. 2012. Immunotoxicity of perfluorinated compounds: recent developments. *Toxicol Pathol* 40(2): 300-311.

Difranceisco W, Ostrow DG, Chmiel JS. 1996. Sexual adventurism, high-risk behavior, and human immunodeficiency virus-1 seroconversion among the Chicago MACS-CCS cohort, 1984 to 1992. A case-control study. *Sexually transmitted diseases* 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. *Archives of disease in childhood* 85(4): 271-274.

Fair PA, Driscoll E, Mollenhauer MA, Bradshaw SG, Yun SH, Kannan K, Bossart GD, Keil DE, Peden-Adams MM. 2011. Effects of environmentally-relevant levels of perfluorooctane sulfonate on clinical parameters and immunological functions in B6C3F1 mice. *J Immunotoxicol* 8(1): 17-29.

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, Griffith L, Oremus M, Raina P, Ismaila A, Santaguida P, Lau J, Trikalinos TA. 2011. Conducting quantitative synthesis when comparing medical interventions: AHRQ and the Effective Health Care Program. *J Clin Epidemiol* 64(11): 1187-1197.

Galbiati V, Mitjans M, Corsini E. 2010. Present and future of in vitro immunotoxicology in drug development. *J Immunotoxicol* 7(4): 255-267.

Genaidy AM, Lemasters GK, Lockey J, Succop P, Deddens J, Sobeih T, Dunning K. 2007. An epidemiological appraisal instrument - a tool for evaluation of epidemiological studies. *Ergonomics* 50(6): 920-960.

Gennari A, Ban M, Braun A, Casati S, Corsini E, Dastych J, Descotes J, Hartung T, Hooghe-Peters R, House R, Pallardy M, Pieters R, Reid L, Tryphonas H, Tschirhart E, Tuschl H, Vandebriel R, Gribaldo L. 2005. The Use of In Vitro Systems for Evaluating Immunotoxicity: The Report and Recommendations of an ECVAM Workshop. *J Immunotoxicol* 2(2): 61-83.

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

Grandjean P, Andersen EW, Budtz-Jorgensen E, Nielsen F, Molbak K, Weihe P, Heilmann C. 2012. Serum vaccine antibody concentrations in children exposed to perfluorinated compounds. *JAMA : the journal of the American Medical Association* 307(4): 391-397.

Granum B, Haug LS, Namork E, Stolevik SB, Thomsen C, Aaberge IS, van Loveren H, Lovik M, Nygaard UC. 2013. Pre-natal exposure to perfluoroalkyl substances may be associated with altered vaccine antibody levels and immune-related health outcomes in early childhood. *J Immunotoxicol*.

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

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

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

Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, Alonso-Coello P, Glasziou P, Jaeschke R, Akl EA, Norris S, Vist G, Dahm P, Shukla VK, Higgins J, Falck-Ytter Y, Schunemann HJ. 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, Alonso-Coello P, Djulbegovic B, Atkins D, Falck-Ytter Y, Williams JW, Jr., Meerpohl J, Norris SL, Akl EA, Schunemann HJ. 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, Atkins D, Kunz R, Brozek J, Montori V, Jaeschke R, Rind D, Dahm P, Meerpohl J, Vist G, Berliner E, Norris S, Falck-Ytter Y, Murad MH, Schunemann HJ. 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, Montori V, Akl EA, Djulbegovic B, Falck-Ytter Y, Norris SL, Williams JW, Jr., Atkins D, Meerpohl J, Schunemann HJ.

2011h. GRADE guidelines: 4. Rating the quality of evidence-study limitations (risk of bias). *Journal of Clinical Epidemiology* 64(4): 407-415.

Hart K, Gill VA, Kannan K. 2009. Temporal trends (1992-2007) of perfluorinated chemicals in Northern Sea Otters (Enhydra lutris kenyoni) from South-Central Alaska. *Arch Environ Contam Toxicol* 56(3): 607-614.

Hastings KL. 2005. Commentary on hormetic dose-response relationships in immunology: occurrence, quantitative features of the dose response, mechanistic foundations, and clinical implications. *Crit Rev Toxicol* 35(2-3): 297-298.

HHS. 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]. Services DoHaH.

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? *Proc Roy Soc Med* 58: 295-300.

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. *Environmental health perspectives* 116(12): 1642-1647.

IOM. 2011. *Finding What Works in Health Care: Standards for Systematic Reviews*. 9780309164252. Washington, DC: Institute of Medicine. 318. http://www.nap.edu/openbook.php?record_id=13059.

Jahnke GD, Iannucci AR, Scialli AR, Shelby MD. 2005. Center for the evaluation of risks to human reproduction--the first five years. *Birth defects research. Part B, Developmental and reproductive toxicology* 74(1): 1-8.

Johnson P, Sutton P, Atchley D, Koustas E, Lam J, Robinson K, Sen S, Axelrad D, Woodruff T. 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.

Kannan K, Perrotta E, Thomas NJ. 2006. Association between perfluorinated compounds and pathological conditions in southern sea otters. *Environmental science & technology* 40(16): 4943-4948.

Keller JM, Kannan K, Taniyasu S, Yamashita N, Day RD, Arendt MD, Segars AL, Kucklick JR. 2005. Perfluorinated compounds in the plasma of loggerhead and Kemp's ridley sea turtles from the southeastern coast of the United States. *Environmental science & technology* 39(23): 9101-9108.

Klimisch HJ, Andreae M, Tillmann U. 1997. A systematic approach for evaluating the quality of experimental toxicological and ecotoxicological data. *Regulatory toxicology and pharmacology : RTP* 25(1): 1-5.

Koustas E, Lam J, Sutton P, Johnson P, Atchley D, Robinson K, Sen S, Axelrad D, Woodruff T. 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).

Ladics GS, Loveless SE. 2005. Commentary on hormetic dose-response relationships in immunology: occurrence, quantitative features of the dose response, mechanistic foundations, and clinical implications. *Crit Rev Toxicol* 35(2-3): 303-304.

Lankveld DPK, Van Loveren H, Baken KA, Vandebriel RJ. 2010. In Vitro Testing for Direct Immunotoxicity: State of the Art. In *Immunotoxicity Testing: Methods and Protocols, Methods in Molecular Biology*. 598. Dietert RR, ed.: Humana Press. 401-423.

Luster MI, Portier C, Pait DG, White KL, Jr., Gennings C, Munson AE, Rosenthal GJ. 1992. Risk assessment in immunotoxicology. I. Sensitivity and predictability of immune tests. *Fundamental and Applied Toxicology* 18(2): 200-210.

Medlin J. 2003. New arrival: CERHR monograph series on reproductive toxicants. *Environmental health perspectives* 111(13): A696-698.

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.

NTP. 2013. Draft OHAT Approach for Systematic Review and Evidence Integration for Literature-based Health Assessments – February 2013. Program NT. RTP, NC, Office of Health Assessment and Translation, Division of the National Toxicology Program.

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].

OECD. 2013. *OECD Portal on Perfluorinated Chemicals*. http://www.oecd.org/ehs/pfc/.

Olsen GW, Burris JM, Burlew MM, Mandel JH. 2003a. Epidemiologic assessment of worker serum perfluorooctanesulfonate (PFOS) and perfluorooctanoate (PFOA) concentrations and medical surveillance examinations. *J Occup Environ Med* 45(3): 260-270.

Olsen GW, Church TR, Miller JP, Burris JM, Hansen KJ, Lundberg JK, Armitage JB, Herron RM, Medhdizadehkashi Z, Nobiletti JB, O'Neill EM, Mandel JH, Zobel LR. 2003b. Perfluorooctanesulfonate and other fluorochemicals in the serum of American Red Cross adult blood donors. *Environmental health perspectives* 111(16): 1892-1901.

Oxman AD, Schunemann HJ, Fretheim A. 2006. Improving the use of research evidence in guideline development: 7. Deciding what evidence to include. *Health research policy and systems / BioMed Central* 4: 19.

Peden-Adams MM, Keller JM, Eudaly JG, Berger J, Gilkeson GS, Keil DE. 2008. Suppression of humoral immunity in mice following exposure to perfluorooctane sulfonate. *Toxicol Sci* 104(1): 144-154.

Schneider K, Schwarz M, Burkholder I, Kopp-Schneider A, Edler L, Kinsner-Ovaskainen A, Hartung T, Hoffmann S. 2009. "ToxRTool", a new tool to assess the reliability of toxicological data. *Toxicol Lett* 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. *J Epidemiol 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, Janes G, Maglione M, Moher D, Nasser M, Robinson KA, Segal JB, Tsouros S. 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 research. Part B, Developmental and reproductive toxicology* 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.

Szklo M, Nieto FJ. 2007. *Epidemiology: Beyond the basics (2nd edition)* 2nd, Sudbury, MA: Jones and Bartlett Publishers.

Taylor B, Miller E, Farrington CP, Petropoulos MC, Favot-Mayaud I, Li J, Waight PA. 1999. Autism and measles, mumps, and rubella vaccine: no epidemiological evidence for a causal association. *Lancet* 353(9169): 2026-2029.

Twombly R. 1998. New NTP centers meet the need to know. *Environmental health perspectives* 106(10): A480-483.

US EPA. 1996a. *Biochemicals Test Guidelines: OPPTS 880.3550 Immunotoxicity*. Report. EPA/712/C-96/280. Washington, DC: Office of Prevention, Pesticides and Toxic Substances. 1-14. http://www.epa.gov/ocspp/pubs/frs/publications/Test_Guidelines/series880.htm.

US EPA. 1996b. *Biochemicals Test Guidelines: OPPTS 880.3800 Immune Response*. Report. EPA/712/C-96/281. Washington, DC: Office of Prevention, Pesticides and Toxic

Substances. 1-8. http://www.epa.gov/ocspp/pubs/frs/publications/Test_Guidelines/series880.htm.

US EPA. 1998. *Health Effects Test Guidelines: OPPTS 870.7800 Immunotoxicity*. Report. EPA/712/C-98/351. Washington, DC: Office of Prevention, Pesticides and Toxic Substances. 1-11. http://www.epa.gov/ocspp/pubs/frs/publications/Test_Guidelines/series870.htm.

US EPA. 2006. 2010/2015 PFOA Stewardship Program.

US EPA. 2009. Long-Chain Perfluorinated Chemicals (PFCs) Action Plan.

US EPA. 2012a. Emerging Contaminants - Perfluorooctane sulfonate (PFOS) and Perfluorooctanoic acid (PFOA).

US EPA. 2012b. Perfluorooctanoic Acid (PFOA) and Fluorinated Telomers.

Viswanathan M, Ansari M, Berkman ND, Chang S, Hartling L, McPheeters LM, Santaguida PL, Shamliyan T, Singh K, Tsertsvadze A, Treadwell JR. 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].

Wakefield AJ, Murch SH, Anthony A, Linnell J, Casson DM, Malik M, Berelowitz M, Dhillon AP, Thomson MA, Harvey P, Valentine A, Davies SE, Walker-Smith JA. 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. 2012. *Guidance for Immunotoxicity Risk Assessment for Chemicals*. IPCS Harmonization Project No. 10. Geneva: International Programme on Chemical Safety, World Health Organization. http://www.inchem.org/documents/harmproj/harmproj/harmproj10.pdf.

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. *Pediatrics international : official journal of the Japan Pediatric Society* 53(6): 826-831.

Ye X, Kuklenyik, 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: Draft Medline search strategy (PubMed)

The strategy for this search is broad for the consideration of immune-related endpoints and comprehensive for PFOA or PFOS as an exposure or treatment in order to ensure inclusion of relevant papers.

COCHRANE LIBRARY **Cochrane Reviews: 2** results – these are associated with ventilators **Trials: 66 results** – vast majority are therapeutic use of perfluorocarbons	Perfluoroalkyl* OR perfluorocaprylic OR perfluorocarbon* OR perfluorocarboxyl* OR perfluorochemical* OR (perfluorinated AND (C8 OR carboxylic OR chemical* OR compound* OR octanoic)) OR PFAA* OR “fluorinated polymer” OR “fluorinated polymers” OR (fluorinated AND (polymer OR polymers)) OR (fluorocarbon AND (polymer OR polymers)) OR Fluoropolymer* OR (fluorinated AND telomer*) OR fluorotelomer* OR fluoro-telomer* OR fluorosurfactant* OR “FC 143” OR FC143 OR Pentadecafluoroctanoate* OR Pentadecafluorooctanoate* OR pentadecafluoroctanoic OR pentadecafluorooctanoic OR “pentadecafluoro-1-octanoic” OR “pentadecafluoro-n-octanoic” OR “perfluoro-1-heptanecarboxylic” OR perfluorocaprylic OR perfluoroheptanecarboxylic OR perfluoroctanoate OR perfluorooctanoate OR “perfluoro octanoate” OR “perfluorooctanoic acid” OR perfluoroctanoic OR perfluorooctanoic OR “perfluoro octanoic” OR “perfluoro-n-octanoic” OR “perfluorooctanoyl chloride” OR PFOA OR APFO OR “1-octanesulfonic acid” OR “1-perfluorooctanesulfonic” OR “1-perfluoroctanesulfonic” OR “heptadecafluoro-1-octanesulfonic” OR “heptadecafluoro-1-octane sulfonic” OR “heptadecafluorooctanesulfonic” OR “heptadecafluorooctane sulfonic” OR “heptadecafluoroctane sulfonic” OR “perfluoroalkyl sulphonate” OR perfluoroctanesulfonate OR perfluorooctanesulfonate OR “perfluoroctane sulfonate” OR “perfluorooctane sulfonate” OR “perfluoro-n-octanesulfonic” OR perfluoroctanesulfonic OR perfluorooctanesulfonic OR “perfluoroctane sulfonic” OR “perfluorooctane sulfonic” OR perfluoroctanesulphonic OR perfluorooctanesulphonic OR “perfluoroctane sulphonic” OR “perfluorooctane sulphonic” OR perfluoroctylsulfonic OR PFOS
EMBASE **1633 results**	Perfluoroalkyl* OR perfluorocaprylic OR perfluorocarbon* OR perfluorocarboxyl* OR perfluorochemical* OR (perfluorinated NEXT/4 (C8 OR carboxylic OR chemical OR chemicals OR compound OR compounds OR octanoic)) OR PFAA* OR “fluorinated polymer” OR “fluorinated polymers” OR (fluorinated NEXT/4 (polymer*)) OR (fluorocarbon NEXT/4 (polymer*)) OR Fluoropolymer* OR (fluorinated NEXT/4 telomer*) OR fluorotelomer* OR fluoro NEXT/0 telomer* OR fluorosurfactant* OR “FC 143” OR FC143 OR Pentadecafluoroctanoate* OR Pentadecafluorooctanoate* OR pentadecafluoroctanoic OR pentadecafluorooctanoic OR “pentadecafluoro-1-octanoic” OR “pentadecafluoro-n-octanoic” OR “perfluoro-1-heptanecarboxylic” OR perfluorocaprylic OR perfluoroheptanecarboxylic OR perfluoroctanoate OR perfluorooctanoate OR “perfluoro octanoate” OR “perfluorooctanoic acid” OR perfluoroctanoic OR perfluorooctanoic OR “perfluoro octanoic” OR “perfluoro-n-octanoic” OR “perfluorooctanoyl

	chloride” OR PFOA OR APFO OR “1-octanesulfonic acid” OR “1-perfluorooctanesulfonic” OR “1-perfluoroctanesulfonic” OR “heptadecafluoro-1-octanesulfonic” OR “heptadecafluoro-1-octane sulfonic” OR “heptadecafluorooctanesulfonic” OR “heptadecafluorooctane sulfonic” OR “heptadecafluoroctane sulfonic” OR “perfluoroalkyl sulphonate” OR perfluoroctanesulfonate OR perfluorooctanesulfonate OR “perfluoroctane sulfonate” OR “perfluorooctane sulfonate” OR “perfluoro-n-octanesulfonic” OR perfluoroctanesulfonic OR perfluorooctanesulfonic OR “perfluoroctane sulfonic” OR “perfluorooctane sulfonic” OR perfluoroctanesulphonic OR perfluorooctanesulphonic OR “perfluoroctane sulphonic” OR “perfluorooctane sulphonic” OR perfluoroctylsulfonic OR PFOS OR 307-35-7 OR 1763-23-1 OR 335-67-1 AND immune OR immunocomp* OR immunogen* OR immunolog* OR immunotox OR immunity OR autoimmun* OR "host resistance" OR spleen OR splenic OR splenocyt* OR thymus OR thymic OR thymocyt* OR leukocyt* OR granulocyt* OR basophil* OR eosinophil* OR neutrophil* OR lymph OR lymphoid* OR lymphocyt* OR "b-lymphocyte" OR "b-lymphocytes" OR "t-lymphocyte" OR "t-lymphocytes" OR "killer cell" OR "killer cells" OR "NK cell" OR "NK-cell" OR "NK-cells" OR macrophag* OR "mast cell" OR "mast cells" OR monocyt* OR phagocyt* OR dendrit* OR "t-cell" OR "t cell" OR "t cells" OR "t-cells" OR "T helper" OR "T-helper" OR "b-cell" OR "b cell" OR "b cells" OR "b-cells" OR antibod* OR histamine* OR histocompatib* OR immunoglobulin* OR "immunoglobulin A" OR IgA OR "immunoglobulin D" OR IgD OR "immunoglobulin E" OR IgE OR "immunoglobulin G" OR IgG OR "immunoglobulin M" OR IgM OR antigen OR antigens OR CD3 OR CD4 OR CD8 OR CD25 OR CD27 OR CD28 OR CD29 OR CD45* OR cytokine* OR chemokine* OR inteferon* OR interleukin* OR "IL-6" OR "IL-8" OR lymphokine* OR monokine* OR ("tumor necrosis" NEXT/0 factor*) OR "TNF alpha" OR "TNFalpha" OR autoimmun* OR addison OR rheumatoid OR glomerulonephritis OR diabetes OR graves OR lupus OR thyroiditis OR hypersensitiv* OR sensitization OR hyperresponsiv* OR allerg* OR atopy OR atopic OR dermatitis OR eczema OR otitis OR “ear infection” OR “ear inflammation” OR (respiratory NEXT/2 infection*) OR asthma OR bronchitis OR pneumonia OR bronchiolitis OR rhinitis OR sinusitis OR wheez* OR crackle* OR cough* OR dyspnea OR gastroenteritis OR inflammat* OR pro NEXT/0 inflammat* OR anti NEXT/0 inflamm* OR autacoid* OR eicosanoid* OR prostaglandin* OR immunomodul* OR immunotherap* OR vaccin* OR immuniz* OR immunosuppress* OR desensitiz* OR immunoprotein* OR "c-reactive protein" OR CRP OR "complement component" OR (complement NEXT/2 (C1 OR C2 OR C3 OR C4 OR C5 OR C6 OR C7 OR C8 OR C9))

EPA ACToR	Select “Search on CAS Numbers” Enter each CAS number on a new line: 307-35-7 1763-23-1 335-67-1
EPA Chemical Data Access Tool	307-35-7 OR 1763-23-1 OR 335-67-1
PubChem	307-35-7 OR 1763-23-1 OR 335-67-1
PUBMED Combined strategy = 779 results	perfluoroalkyl*[tiab] OR perfluorocaprylic[tiab] OR perfluorocarbon*[tiab] OR perfluorocarboxyl*[tiab] OR perfluorochemical*[tiab] OR (perfluorinated[tiab] AND (C8[tiab] OR carboxylic[tiab] OR chemical*[tiab] OR compound*[tiab] OR octanoic[tiab])) OR PFAA*[tiab] OR “fluorinated polymer”[tiab] OR “fluorinated polymers”[tiab] OR (fluorinated[tiab] AND (polymer[tiab] OR polymers[tiab])) OR (fluorocarbon[tiab] AND (polymer[tiab] OR polymers[tiab])) OR Fluoropolymer*[tiab] OR (fluorinated[tiab] AND telomer*[tiab]) OR fluorotelomer*[tiab] OR fluoro-telomer*[tiab] OR fluorosurfactant*[tiab] OR “FC 143”[tiab] OR FC143[tiab] OR 335-67-1 [rn] OR Pentadecafluoroctanoate*[tiab] OR Pentadecafluorooctanoate*[tiab] OR pentadecafluoroctanoic[tiab] OR pentadecafluorooctanoic[tiab] OR “pentadecafluoro-1-octanoic”[tiab] OR “pentadecafluoro-n-octanoic”[tiab] OR “perfluoro-1-heptanecarboxylic”[tiab] OR perfluorocaprylic[tiab] OR perfluoroheptanecarboxylic[tiab] OR perfluoroctanoate[tiab] OR perfluorooctanoate[tiab] OR “perfluoro octanoate”[tiab] OR “perfluorooctanoic acid”[nm] OR perfluoroctanoic[tiab] OR perfluorooctanoic[tiab] OR “perfluoro octanoic”[tiab] OR “perfluoro-n-octanoic”[tiab] OR “perfluorooctanoyl chloride”[tiab] OR PFOA[tiab] OR APFO[tiab] OR 1763-23-1[rn] OR 307-35-7[rn] OR “1-octanesulfonic acid”[tiab] OR “1-perfluorooctanesulfonic”[tiab] OR “1-perfluoroctanesulfonic”[tiab] OR “heptadecafluoro-1-octanesulfonic”[tiab] OR “heptadecafluoro-1-octane sulfonic”[tiab] OR “heptadecafluorooctanesulfonic”[tiab] OR “heptadecafluorooctane sulfonic”[tiab] OR “heptadecafluoroctane sulfonic”[tiab] OR “perfluoroalkyl sulphonate”[tiab] OR perfluoroctanesulfonate[tiab] OR perfluorooctanesulfonate[tiab] OR “perfluoroctane sulfonate”[tiab] OR “perfluorooctane sulfonate”[tiab] OR “perfluoro-n-octanesulfonic”[tiab] OR perfluoroctanesulfonic[tiab] OR perfluorooctanesulfonic[tiab] OR “perfluorooctane sulfonic acid”[nm] OR “perfluoroctane sulfonic”[tiab] OR “perfluorooctane sulfonic”[tiab] OR perfluoroctanesulphonic[tiab] OR perfluorooctanesulphonic[tiab] OR “perfluoroctane sulphonic”[tiab] OR “perfluorooctane sulphonic”[tiab] OR

	perfluoroctylsulfonic[tiab] OR PFOS [tiab] AND immunology[sh] OR immune[tiab] OR immunocomp*[tiab] OR immunogen*[tiab] OR immunolog*[tiab] OR immunotox*[tiab] OR immunotoxins[mh] OR immunity[tiab] OR autoimmun*[tiab] OR "host resistance"[tiab] OR immunocompetence[mh] OR "immune system"[mh] OR spleen[tiab] OR splenic[tiab] OR splenocyt*[tiab] OR thymus[tiab] OR thymic[tiab] OR thymocyt*[tiab] OR leukocyt*[tiab] OR granulocyt*[tiab] OR basophil*[tiab] OR eosinophil*[tiab] OR neutrophil*[tiab] OR lymph[tiab] OR lymphoid*[tiab] OR lymphocyt*[tiab] OR "b-lymphocyte"[tiab] OR "b-lymphocytes"[tiab] OR "t-lymphocyte"[tiab] OR "t-lymphocytes"[tiab] OR "killer cell"[tiab] OR "killer cells"[tiab] OR "NK cell"[tiab] OR "NK-cell"[tiab] OR "NK-cells"[tiab] OR macrophag*[tiab] OR "mast cell"[tiab] OR "mast cells"[tiab] OR monocyt*[tiab] OR phagocyt*[tiab] OR dendrit*[tiab] OR "t-cell"[tiab] OR "t cell"[tiab] OR "t cells"[tiab] OR "t-cells"[tiab] OR "T helper"[tiab] OR "T-helper"[tiab] OR "b-cell"[tiab] OR "b cell"[tiab] OR "b cells"[tiab] OR "b-cells"[tiab] OR antibod*[tiab] OR histamine*[tiab] OR histocompatib*[tiab] OR immunoglobulins[mh] OR immunoglobulin*[tiab] OR "immunoglobulin A"[tiab] OR IgA[tiab] OR "immunoglobulin D"[tiab] OR IgD[tiab] OR "immunoglobulin E"[tiab] OR IgE[tiab] OR "immunoglobulin G"[tiab] OR IgG[tiab] OR "immunoglobulin M"[tiab] OR IgM[tiab] OR "antigens, CD"[mh] OR CD3 [tiab] OR CD4 [tiab] OR CD8 [tiab] OR CD25 [tiab] OR CD27 [tiab] OR CD28 [tiab] OR CD29 [tiab] OR CD45*[tiab] OR cytokines[mh] OR cytokine*[tiab] OR chemokine*[tiab] OR inteferon*[tiab] OR interleukin*[tiab] OR "IL-6"[tiab] OR "IL-8"[tiab] OR lymphokine*[tiab] OR monokine*[tiab] OR ("tumor necrosis"[tiab] AND (factor[tiab] OR factors[tiab])) OR "TNF alpha"[tiab] OR "TNFalpha"[tiab] OR "immune system diseases"[mh] OR autoimmun*[tiab] OR addison[tiab] OR rheumatoid[tiab] OR glomerulonephritis[tiab] OR diabetes[tiab] OR graves[tiab] OR lupus[tiab] OR thyroiditis[tiab] OR hypersensitiv*[tiab] OR sensitization OR hyperresponsiv*[tiab] OR allergy[mh] OR allerg*[tiab] OR atopy[tiab] OR atopic[tiab] OR dermatitis[tiab] OR eczema[tiab] OR otitis[tiab] OR “ear infection”[tiab] OR “ear inflammation”[tiab] OR Respiratory tract infections[mh] OR (respiratory[tiab] AND infection*[tiab]) OR asthma[tiab] OR bronchitis[tiab] OR pneumonia[tiab] OR bronchiolitis[tiab] OR rhinitis[tiab] OR sinusitis[tiab] OR wheez*[tiab] OR crackle*[tiab] OR cough[mh] OR cough*[tiab] OR dyspnea[tiab] OR gastroenteritis[tiab] OR inflammation[mh] OR inflammat*[tiab] OR pro-inflammat*[tiab] OR anti-inflamm*[tiab] OR "inflammation mediators"[mh] OR autacoid*[tiab] OR eicosanoid*[tiab] OR prostaglandin*[tiab] OR immunomodulation[mh] OR immunomodul*[tiab] OR immunotherap*[tiab] OR vaccin*[tiab] OR immuniz*[tiab] OR immunosuppress*[tiab] OR desensitiz*[tiab] OR immunoproteins[mh] OR immunoprotein*[tiab] OR "c-reactive protein"[tiab] OR CRP[tiab] OR "complement component"[tiab] OR

	(complement[tiab] AND (C1 OR C2 OR C3 OR C4 OR C5 OR C6 OR C7 OR C8 OR C9))
SCOPUS Combined Strategy = 1369 results Character limit in regular search box, need to copy and paste into Advanced search	TITLE(Perfluoroalkyl* OR perfluorocaprylic OR perfluorocarbon* OR perfluorocarboxyl* OR perfluorochemical* OR (perfluorinated AND (C8 OR carboxylic OR chemical* OR compound* OR octanoic)) OR PFAA* OR “fluorinated polymer” OR “fluorinated polymers” OR (fluorinated AND (polymer OR polymers)) OR (fluorocarbon AND (polymer OR polymers)) OR Fluoropolymer* OR (fluorinated AND telomer*) OR fluorotelomer* OR fluoro-telomer* OR fluorosurfactant* OR “FC 143” OR FC143 OR Pentadecafluoroctanoate* OR Pentadecafluorooctanoate* OR pentadecafluoroctanoic OR pentadecafluorooctanoic OR “pentadecafluoro-1-octanoic” OR “pentadecafluoro-n-octanoic” OR “perfluoro-1-heptanecarboxylic” OR perfluorocaprylic OR perfluoroheptanecarboxylic OR perfluoroctanoate OR perfluorooctanoate OR “perfluoro octanoate” OR “perfluorooctanoic acid” OR perfluoroctanoic OR perfluorooctanoic OR “perfluoro octanoic” OR “perfluoro-n-octanoic” OR “perfluorooctanoyl chloride” OR PFOA OR APFO OR “1-octanesulfonic acid” OR “1-perfluorooctanesulfonic” OR “1-perfluoroctanesulfonic” OR “heptadecafluoro-1-octanesulfonic” OR “heptadecafluoro-1-octane sulfonic” OR “heptadecafluorooctanesulfonic” OR “heptadecafluorooctane sulfonic” OR “heptadecafluoroctane sulfonic” OR “perfluoroalkyl sulphonate” OR perfluoroctanesulfonate OR perfluorooctanesulfonate OR “perfluoroctane sulfonate” OR “perfluorooctane sulfonate” OR “perfluoro-n-octanesulfonic” OR perfluoroctanesulfonic OR perfluorooctanesulfonic OR “perfluoroctane sulfonic” OR “perfluorooctane sulfonic” OR perfluoroctanesulphonic OR perfluorooctanesulphonic OR “perfluoroctane sulphonic” OR “perfluorooctane sulphonic” OR perfluoroctylsulfonic OR PFOS) OR CASREGNUMBER(335-67-1) OR CASREGNUMBER(1763-23-1) OR ABS(Perfluoroalkyl* OR perfluorocaprylic OR perfluorocarbon* OR perfluorocarboxyl* OR perfluorochemical* OR (perfluorinated AND (C8 OR carboxylic OR chemical* OR compound* OR octanoic)) OR PFAA* OR “fluorinated polymer” OR “fluorinated polymers” OR (fluorinated AND (polymer OR polymers)) OR (fluorocarbon AND (polymer OR polymers)) OR Fluoropolymer* OR (fluorinated AND telomer*) OR fluorotelomer* OR fluoro-telomer* OR fluorosurfactant* OR “FC 143” OR FC143 OR Pentadecafluoroctanoate* OR Pentadecafluorooctanoate* OR pentadecafluoroctanoic OR pentadecafluorooctanoic OR “pentadecafluoro-1-octanoic” OR “pentadecafluoro-n-octanoic” OR “perfluoro-1-heptanecarboxylic” OR perfluorocaprylic OR perfluoroheptanecarboxylic OR perfluoroctanoate OR perfluorooctanoate OR “perfluoro octanoate” OR “perfluorooctanoic acid” OR perfluoroctanoic OR perfluorooctanoic OR “perfluoro octanoic” OR “perfluoro-n-octanoic” OR “perfluorooctanoyl chloride” OR PFOA OR APFO OR “1-octanesulfonic acid” OR “1-perfluorooctanesulfonic” OR “1-perfluoroctanesulfonic” OR “heptadecafluoro-1-octanesulfonic” OR “heptadecafluoro-1-octane sulfonic” OR “heptadecafluorooctanesulfonic” OR “heptadecafluorooctane sulfonic” OR “heptadecafluoroctane sulfonic” OR “perfluoroalkyl sulphonate” OR perfluoroctanesulfonate OR perfluorooctanesulfonate OR “perfluoroctane

	sulfonate” OR “perfluorooctane sulfonate” OR “perfluoro-n-octanesulfonic” OR perfluoroctanesulfonic OR perfluorooctanesulfonic OR “perfluoroctane sulfonic” OR “perfluorooctane sulfonic” OR perfluoroctanesulphonic OR perfluorooctanesulphonic OR “perfluoroctane sulphonic” OR “perfluorooctane sulphonic” OR perfluoroctylsulfonic OR PFOS) OR CASREGNUMBER(335-67-1) OR CASREGNUMBER(1763-23-1) AND TITLE(immune OR immunocomp* OR immunogen* OR immunolog* OR immunoto* OR immunotoxins OR immunity OR autoimmun* OR "host resistance" OR spleen OR splenic OR splenocyt* OR thymus OR thymic OR thymocyt* OR leukocyt* OR granulocyt* OR basophil* OR eosinophil* OR neutrophil* OR lymph OR lymphoid* OR lymphocyt* OR "b-lymphocyte" OR "b-lymphocytes" OR "t-lymphocyte" OR "t-lymphocytes" OR "killer cell" OR "killer cells" OR "NK cell" OR "NK-cell" OR "NK-cells" OR macrophag* OR "mast cell" OR "mast cells" OR monocyt* OR phagocyt* OR dendrit* OR "t-cell" OR "t cell" OR "t cells" OR "t-cells" OR "T helper" OR "T-helper" OR "b-cell" OR "b cell" OR "b cells" OR "b-cells" OR antibod* OR histamine* OR histocompatib* OR immunoglobulin* OR "immunoglobulin A" OR IgA OR "immunoglobulin D" OR IgD OR "immunoglobulin E" OR IgE OR "immunoglobulin G" OR IgG OR "immunoglobulin M" OR IgM OR antigen OR antigens OR CD3 OR CD4 OR CD8 OR CD25 OR CD27 OR CD28 OR CD29 OR CD45* OR cytokine* OR chemokine* OR inteferon* OR interleukin* OR "IL-6" OR "IL-8" OR lymphokine* OR monokine* OR ("tumor necrosis" AND factor*) OR "TNF alpha" OR "TNFalpha" OR autoimmun* OR addison OR rheumatoid OR glomerulonephritis OR diabetes OR graves OR lupus OR thyroiditis OR hypersensitiv* OR sensitization OR hyperresponsiv* OR allerg* OR atopy OR atopic OR dermatitis OR eczema OR otitis OR “ear infection” OR “ear inflammation” OR (respiratory AND infection*) OR asthma OR bronchitis OR pneumonia OR bronchiolitis OR rhinitis OR sinusitis OR wheez* OR crackle* OR cough* OR dyspnea OR gastroenteritis OR inflammat* OR pro-inflammat* OR anti-inflamm* OR autacoid* OR eicosanoid* OR prostaglandin* OR immunomodul* OR immunotherap* OR vaccin* OR immuniz* OR immunosuppress* OR desensitiz* OR immunoprotein* OR "c-reactive protein" OR CRP OR "complement component" OR (complement AND (C1 OR C2 OR C3 OR C4 OR C5 OR C6 OR C7 OR C8 OR C9))) OR ABS(immune OR immunocomp* OR immunogen* OR immunolog* OR immunotox OR immunity OR autoimmun* OR "host resistance" OR spleen OR splenic OR splenocyt* OR thymus OR thymic OR thymocyt* OR leukocyt* OR granulocyt* OR basophil* OR eosinophil* OR neutrophil* OR lymph OR lymphoid* OR lymphocyt* OR "b-lymphocyte" OR "b-lymphocytes" OR "t-lymphocyte" OR "t-lymphocytes" OR "killer cell" OR "killer cells" OR "NK cell" OR "NK-cell" OR "NK-cells" OR macrophag* OR "mast cell" OR "mast cells" OR monocyt* OR phagocyt* OR dendrit* OR "t-cell" OR "t cell" OR "t cells" OR "t-cells" OR "T helper" OR "T-

	helper" OR "b-cell" OR "b cell" OR "b cells" OR "b-cells" OR antibod* OR histamine* OR histocompatib* OR immunoglobulin* OR "immunoglobulin A" OR IgA OR "immunoglobulin D" OR IgD OR "immunoglobulin E" OR IgE OR "immunoglobulin G" OR IgG OR "immunoglobulin M" OR IgM OR antigen OR antigens OR CD3 OR CD4 OR CD8 OR CD25 OR CD27 OR CD28 OR CD29 OR CD45* OR cytokine* OR chemokine* OR inteferon* OR interleukin* OR "IL-6" OR "IL-8" OR lymphokine* OR monokine* OR ("tumor necrosis" AND factor*) OR "TNF alpha" OR "TNFalpha" OR autoimmun* OR addison OR rheumatoid OR glomerulonephritis OR diabetes OR graves OR lupus OR thyroiditis OR hypersensitiv* OR sensitization OR hyperresponsiv* OR allerg* OR atopy OR atopic OR dermatitis OR eczema OR otitis OR “ear infection” OR “ear inflammation” OR (respiratory AND infection*) OR asthma OR bronchitis OR pneumonia OR bronchiolitis OR rhinitis OR sinusitis OR wheez* OR crackle* OR cough* OR dyspnea OR gastroenteritis OR inflammat* OR pro-inflammat* OR anti-inflamm* OR autacoid* OR eicosanoid* OR prostaglandin* OR immunomodul* OR immunotherap* OR vaccin* OR immuniz* OR immunosuppress* OR desensitiz* OR immunoprotein* OR "c-reactive protein" OR CRP OR "complement component" OR (complement AND (C1 OR C2 OR C3 OR C4 OR C5 OR C6 OR C7 OR C8 OR C9)))
Toxline **765 results**	*NOTE: Searching on the immune terms will only retrieve the first 50,000 records (Toxline’s display limit). Attempts to break the search up into separate searches is possible, but even a search on the term ‘immunology’ alone will hit the maximum.* Perfluoroalkyl* OR perfluorocaprylic OR perfluorocarbon* OR perfluorocarboxyl* OR perfluorochemical* OR (perfluorinated AND (C8 OR carboxylic OR chemical OR chemicals OR compound OR compounds OR octanoic)) OR PFAA* OR “fluorinated polymer” OR “fluorinated polymers” OR (fluorinated AND (polymer OR polymers)) OR (fluorocarbon AND (polymer OR polymers)) OR Fluoropolymer* OR (fluorinated AND telomer*) OR fluorotelomer* OR fluoro-telomer* OR fluorosurfactant* OR “FC 143” OR FC143 OR Pentadecafluoroctanoate* OR Pentadecafluorooctanoate* OR pentadecafluoroctanoic OR pentadecafluorooctanoic OR “pentadecafluoro-1-octanoic” OR “pentadecafluoro-n-octanoic” OR “perfluoro-1-heptanecarboxylic” OR perfluorocaprylic OR perfluoroheptanecarboxylic OR perfluoroctanoate OR perfluorooctanoate OR “perfluoro octanoate” OR “perfluorooctanoic acid” OR perfluoroctanoic OR perfluorooctanoic OR “perfluoro octanoic” OR “perfluoro-n-octanoic” OR “perfluorooctanoyl chloride” OR PFOA OR APFO OR “1-octanesulfonic acid” OR “1-perfluorooctanesulfonic” OR “1-perfluoroctanesulfonic” OR “heptadecafluoro-1-octanesulfonic” OR “heptadecafluoro-1-octane sulfonic” OR “heptadecafluorooctanesulfonic” OR “heptadecafluorooctane sulfonic” OR “heptadecafluoroctane sulfonic” OR “perfluoroalkyl sulphonate” OR perfluoroctanesulfonate OR perfluorooctanesulfonate OR “perfluoroctane sulfonate” OR “perfluorooctane sulfonate” OR “perfluoro-n-octanesulfonic” OR

	perfluoroctanesulfonic OR perfluorooctanesulfonic OR “perfluoroctane sulfonic” OR “perfluorooctane sulfonic” OR perfluoroctanesulphonic OR perfluorooctanesulphonic OR “perfluoroctane sulphonic” OR “perfluorooctane sulphonic” OR perfluoroctylsulfonic OR PFOS OR 335-67-1 OR 1763-23-1
WEB OF SCIENCE Combined Strategy = 923 results	Perfluoroalkyl OR perfluoroalkyls OR perfluorocaprylic OR perfluorocarbon OR perfluorocarbons OR perfluorocarboxyl* OR perfluorochemical* OR perfluorocarboxyls OR perfluorochemicals OR PFAA* OR “fluorinated polymer” OR “fluorinated polymers” OR Fluoropolymer* OR (fluorinated AND telomer*) OR fluorotelomer* OR fluoro-telomer* OR fluorosurfactant* OR “FC 143” OR FC143 OR Pentadecafluoroctanoate* OR Pentadecafluorooctanoate* OR pentadecafluoroctanoic OR pentadecafluorooctanoic OR “pentadecafluoro-1-octanoic” OR “pentadecafluoro-n-octanoic” OR “perfluoro-1-heptanecarboxylic” OR perfluorocaprylic OR perfluoroheptanecarboxylic OR perfluoroctanoate OR perfluorooctanoate OR “perfluoro octanoate” OR “perfluorooctanoic acid” OR perfluoroctanoic OR perfluorooctanoic OR “perfluoro octanoic” OR “perfluoro-n-octanoic” OR “perfluorooctanoyl chloride” OR PFOA OR APFO OR “1-octanesulfonic acid” OR “1-perfluorooctanesulfonic” OR “1-perfluoroctanesulfonic” OR “heptadecafluoro-1-octanesulfonic” OR “heptadecafluoro-1-octane sulfonic” OR “heptadecafluorooctanesulfonic” OR “heptadecafluorooctane sulfonic” OR “heptadecafluoroctane sulfonic” OR “perfluoroalkyl sulphonate” OR perfluoroctanesulfonate OR perfluorooctanesulfonate OR “perfluoroctane sulfonate” OR “perfluorooctane sulfonate” OR “perfluoro-n-octanesulfonic” OR perfluoroctanesulfonic OR perfluorooctanesulfonic OR “perfluoroctane sulfonic” OR “perfluorooctane sulfonic” OR perfluoroctanesulphonic OR perfluorooctanesulphonic OR “perfluoroctane sulphonic” OR “perfluorooctane sulphonic” OR perfluoroctylsulfonic OR PFOS AND immune OR immunocomp* OR immunogen* OR immunolog* OR immunotox OR immunity OR autoimmun* OR "host resistance" OR spleen OR splenic OR splenocyt* OR thymus OR thymic OR thymocyt* OR leukocyt* OR granulocyt* OR basophil* OR eosinophil* OR neutrophil* OR lymph OR lymphoid* OR lymphocyt* OR "b-lymphocyte" OR "b-lymphocytes" OR "t-lymphocyte" OR "t-lymphocytes" OR "killer cell" OR "killer cells" OR "NK cell" OR "NK-cell" OR "NK-cells" OR macrophag* OR "mast cell" OR "mast cells" OR monocyt* OR phagocyt* OR dendrit* OR "t-cell" OR "t cell" OR "t cells" OR "t-cells" OR "T helper" OR "T-helper" OR "b-cell" OR "b cell" OR "b cells" OR "b-cells" OR antibod* OR histamine* OR histocompatib* OR immunoglobulin* OR "immunoglobulin A" OR IgA OR "immunoglobulin D" OR IgD OR "immunoglobulin E" OR IgE OR "immunoglobulin G" OR IgG OR "immunoglobulin M" OR IgM OR antigen OR antigens OR CD3 OR CD4 OR CD8

	OR CD25 OR CD27 OR CD28 OR CD29 OR CD45* OR cytokine* OR chemokine* OR inteferon* OR interleukin* OR "IL-6" OR "IL-8" OR lymphokine* OR monokine* OR ("tumor necrosis" AND factor*) OR "TNF alpha" OR "TNFalpha" OR autoimmun* OR addison OR rheumatoid OR glomerulonephritis OR diabetes OR graves OR lupus OR thyroiditis OR hypersensitiv* OR sensitization OR hyperresponsiv* OR allerg* OR atopy OR atopic OR dermatitis OR eczema OR otitis OR “ear infection” OR “ear inflammation” OR (respiratory AND infection*) OR asthma OR bronchitis OR pneumonia OR bronchiolitis OR rhinitis OR sinusitis OR wheez* OR crackle* OR cough* OR dyspnea OR gastroenteritis OR inflammat* OR pro-inflammat* OR anti-inflamm* OR autacoid* OR eicosanoid* OR prostaglandin* OR immunomodul* OR immunotherap* OR vaccin* OR immuniz* OR immunosuppress* OR desensitiz* OR immunoprotein* OR "c-reactive protein" OR CRP OR "complement component" OR (complement AND (C1 OR C2 OR C3 OR C4 OR C5 OR C6 OR C7 OR C8 OR C9))

Appendix 2. Instructions to answer risk of bias questions

Appendix 2 is provided in a separate file: Appendix_2_PFOAPFOS_RiskofBias.pdf