

PROTOCOL FOR SYSTEMATIC REVIEW OF EFFECTS OF FLUORIDE EXPOSURE ON NEURODEVELOPMENT

Project Leader: Kyla Taylor, PhD, Office of Health Assessment and Translation (OHAT), DNTP

Summary: OHAT is conducting a systematic review to evaluate the evidence for an association between fluoride exposure and neurodevelopmental and cognitive effects. The protocol is detailed in this document.

Date Original Protocol Published: June 2017

Date First Revised Protocol Published: May 29, 2019

Date Second Revised Protocol Published: September 16, 2020

Documentation of Revisions to the Protocol: The principal revisions are detailed in the Protocol History and Revisions Table on Page 36 including the reasons for each revision. In addition, updated language or new text starts with the word "Revision:" Strikethrough text indicates original text that has been modified, and new text is marked in bold text.

TABLE OF CONTENTS

Table of Contents	ii
Tables	iii
Figures	iv
BACKGROUND	1
Exposure	1
Water Fluoridation	1
Concerns for Potential Fluoride Toxicity	1
Nominations to NTP	2
OVERALL OBJECTIVES AND SPECIFIC AIMS	3
Specific Aims	3
PECO Statement	4
METHODS	5
Step 1. Problem Formulation	5
REVISION: Problem formulation activities to develop this protocol	5
REVISION: NASEM Review.....	5
Step 2. Search and Select Studies for Inclusion	6
Searching electronic databases.....	6
REVISION: Supplemental searching in Chinese databases.....	6
Searching other resources	7
Selection criteria for the evidence	7
Step 3. Data Extraction and Content Management	10
Standardizing results from behavioral tests and dose levels in experimental animal studies.....	10
Step 4. Quality Assessment of Individual Studies	11
Critical risk-of-bias domains for epidemiology studies.....	12
Rationale for critical risk-of-bias domains for human studies	12
Critical risk-of-bias domains for animal studies.....	14
Rationale for critical risk-of-bias domains for animal studies.....	15
Missing information for risk-of-bias assessment	16
Step 5. Assessment of Confidence in the Body of Evidence	18
Evidence synthesis.....	20
Step 6. Preparation of Level of Evidence Statement	23
Step 7. Integrate Evidence to Develop Hazard Identification Conclusions	23
Consideration of mechanistic data	26

NTP Monograph	28
Evaluation Process	28
Methodology	28
Results	29
Discussion	29
Conclusion	29
References	30
About the protocol	38
Contributors	38
Evaluation Team	38
Contract support	38
Technical Advisors	39
Sources of Support	39
Protocol History and Revisions.....	39
APPENDICES	41
Appendix 1. Electronic Database Search Strategies	41
REVISION: Main literature databases search	41
REVISION: Supplemental Chinese database literature search.....	50
Appendix 2. Data Extraction Elements for HAWC: Human Studies	51
Appendix 3. Data Extraction Elements for HAWC: Animal Studies	53
Appendix 4. Data Extraction Elements for HAWC: In vitro Studies	55
Appendix 5. Risk-of-Bias Criteria	57
Observational Studies (Human studies).....	57
Experimental Animal Studies	71
REVISION: Appendix 6. Meta-analysis	80
Background	80
Objectives	82
Methods	83
Meta-analysis Protocol History and Revisions	88

TABLES

REVISION: Note a new table 2 for inclusion and exclusion criteria was added and therefore subsequent tables have been renumbered (e.g., “Table 2. Response Options for Each RoB Question” from the original protocol is now “Table 3. Response Options for Each RoB Question”).

Table 1. PECO (Populations, Exposures, Comparators, Outcomes) Statement	4
Table 2. Inclusion and Exclusion Criteria to Determine Study Eligibility	9
Table 3. Response Options for Each RoB Question.....	12
Table 4. OHAT Risk-of-bias Tool	17
Table 5. Key Factors when Considering Whether to Downgrade or Upgrade Across a Body of Evidence.....	18
Table 6. Neurological Outcomes Grouping for Human Studies	20

Table 7. Neurological Outcomes Grouping for Animal Studies..... 21

FIGURES

Figure 1. Assessing Confidence in the Body of Evidence 18
Figure 2. Translate Confidence Ratings into Evidence of Health Effect Conclusions..... 23
Figure 3. Hazard Identification Scheme for Neurodevelopmental or Cognitive Effects..... 25
Figure 4. Factors Considered in Evaluating the Support for Biological Plausibility 27
Figure 5. Evaluation Process for OHAT Monographs 28

BACKGROUND

Exposure

Fluoride, a naturally occurring element, is used for the prevention of dental caries. Humans are exposed to fluoride via dental products (e.g., toothpaste, mouth rinses, supplements) and fluoride-supplemented drinking water. Fluoride also can occur naturally in drinking water. Other sources of human exposure include foods, beverages, industrial emissions, pharmaceuticals, and pesticides (e.g., cryolite, sulfuric fluoride). Soil ingestion is another source of fluoride exposure in young children (EPA 2010).

Water Fluoridation

The U.S. Public Health Service (PHS) first recommended communities add fluoride to drinking water in 1962. PHS guidance is advisory, not regulatory, which means that, while PHS recommends community water fluoridation as an effective public health intervention, the decision to fluoridate water systems is made by state and local governments. For community water systems that add fluoride, PHS now recommends a fluoride concentration of 0.7 milligrams/liter (mg/L).¹ Under the Safe Drinking Water Act, the U.S. Environmental Protection Agency (EPA) sets maximum exposure level standards for drinking water quality. For fluoride, the enforceable standard is set at 4.0 mg/L, to protect against skeletal fluorosis. A secondary drinking water standard of 2.0 mg/L protects against moderate to severe dental fluorosis. The secondary standard is not enforceable but requires systems to notify the public if the average levels exceed it. EPA is reviewing the current drinking water standards for fluoride (EPA 2013).

Concerns for Potential Fluoride Toxicity

Controversy over community water fluoridation stems from concerns about the potential harmful effects of fluoride and the ethics of water fluoridation. The most commonly cited health concerns related to fluoride and water fluoridation are bone fractures and skeletal fluorosis, decreased intelligence quotient (IQ) and other neurological effects, cancer, and endocrine disruption. Effects on neurological function, endocrine function (e.g., thyroid, parathyroid, pineal), metabolic function (e.g., glucose metabolism), and carcinogenicity were assessed in the 2006 National Research Council (NRC) report *Fluoride in Drinking Water: A Scientific Review of EPA's Standards* (NRC 2006). The NRC review considered adverse effects of water fluoride, focusing on a range of concentrations (2–4 mg/L) above the current 0.7-mg/L recommendation for community water fluoridation (NRC 2006). At fluoride levels below 4.0 mg/L, NRC did not find sufficient evidence of negative health effects, other than severe dental fluorosis. The conclusions from the NRC review were the primary source of information for the potential hazard summary in a 2015 report by the U.S. Department of Health and Human Services (DHHS), *Federal Panel on Community Water Fluoridation*. The NRC report noted several challenges to evaluating the literature, including: deficiencies in reporting quality, consideration of all sources of fluoride exposure, consideration of potential confounding, selection of appropriate control subject populations in epidemiology studies, demonstrated clinical significance of endocrine effects, and the biological relationship between histological, biochemical, and molecular alterations with behavioral effects.

¹For many years, most fluoridated community water systems used fluoride concentrations ranging from 0.8 to 1.2 mg/L (US DHHS 2015)

Nominations to NTP

In 2015, the National Toxicology Program (NTP) received nominations from the public to conduct analyses of fluoride and developmental neurobehavioral toxicity, endocrine disruption, and cancer.

REVISION: As part of NTP's consideration of the nomination, problem formulation activities were conducted (see Problem Formulation below for additional details) starting with the publication of a request for information about fluoride in a Federal Register notice [October 7, 2015 (<https://www.govinfo.gov/content/pkg/FR-2015-10-07/pdf/2015-25434.pdf>); amended November 23, 2015 (<https://www.govinfo.gov/content/pkg/FR-2015-11-23/pdf/2015-29734.pdf>)].

NTP is moving forward with the consideration of the developmental neurobehavioral evidence. For cancer and endocrine disruption, NTP is analyzing the amount of evidence available and the merit of pursuing systematic reviews, given factors such as the extent of new research published since previous evaluations and whether these new reports address or correct the deficiencies noted in the literature (NRC 2006, OEHA 2011, SCHER 2011).

Regarding neurotoxicity and neurobehavioral effects, the main conclusions in the 2006 NRC report were:

“Animal and human studies of fluoride have been published reporting adverse cognitive and behavioral effects. A few epidemiologic studies of Chinese populations have reported IQ deficits in children exposed to fluoride at 2.5 to 4 mg/L in drinking water. Although the studies lacked sufficient detail for the committee to fully assess their quality and relevance to U.S. populations, the consistency of the results appears significant enough to warrant additional research on the effects of fluoride on intelligence.” [p. 8] (NRC 2006)

“A few animal studies have reported alterations in the behavior of rodents after treatment with fluoride, but the committee did not find the changes to be substantial in magnitude. More compelling were studies on molecular, cellular, and anatomical changes in the nervous system found after fluoride exposure, suggesting that functional changes could occur. These changes might be subtle or seen only under certain physiological or environmental conditions. More research is needed to clarify the effect of fluoride on brain chemistry and function.” [p. 8] (NRC 2006)

REVISION: Since the 2006 NRC report was released, 10+ epidemiological studies and 45+ experimental animal studies have been published addressing the potential **neurodevelopmental and cognitive** neurobehavioral effects of fluoride. Recent reviews of the human literature suggest that high levels of naturally occurring fluoride in water (>1.5 parts per million [ppm]) could be associated with negative health effects, including lower IQ (Choi *et al.* 2012, Sutton *et al.* 2015). Overall, many of these studies were considered low quality, as they did not fully account for known confounding factors regarding IQ (e.g., nutritional status, socioeconomic status) or other potential factors influencing IQ (e.g., iodine deficiency, chemical contaminants in the ground water such as arsenic and lead). Very few studies have assessed the association between fluoride levels relevant to community water fluoridation practices and neurodevelopmental and cognitive effects. Based primarily on an analysis of a prospective cohort study conducted in New Zealand (Broadbent *et al.* 2015), Sutton *et al.* (2015) concluded there was no evidence of an association with lowered IQ in studies of community water fluoridation.

NTP recently published a systematic review of the animal evidence on the effects of fluoride on learning and memory (NTP 2016). The systematic review found a low-to-moderate level of evidence that learning and memory deficits occur in experimental animals at fluoride concentrations greater than 0.7 ppm. The evidence was strongest (moderate) in animals exposed as adults and evidence was weaker (low) in animals exposed during development. NTP is conducting additional studies to assess the effect of fluoride exposure on learning and memory. The results from the ongoing experimental animal work will be incorporated into the current review, which will consider the epidemiological, animal, and mechanistic evidence in its conclusions. The NTP review will also identify key research and data gaps for additional study.

OVERALL OBJECTIVES AND SPECIFIC AIMS

The overall objective of this evaluation is to undertake a systematic review of the existing human, experimental animal (non-human mammals), and mechanistic studies to develop hazard identification conclusions about whether fluoride exposure is associated with neurodevelopmental and cognitive effects. The systematic review will be based on guidance outlined in the Office of Health Assessment and Translation (OHAT) Handbook for Conducting a Literature-Based Health Assessment (NTP 2015a, 2019a) **REVISION: new footnote added.**²

Specific Aims

- Identify epidemiological and experimental animal literature (extending the 2016 evaluation) that assesses neurodevelopmental or cognitive health effects, especially outcomes related to learning, memory, and intelligence following exposure during development. Effects on thyroid function will also be assessed to help evaluate potential mechanisms of impaired neurological function. Studies reporting in vitro and other types of mechanistic evidence relating to neurodevelopmental or cognitive outcomes or thyroid function also will be identified.
- Extract data on relevant health outcomes from included epidemiological and experimental animal studies. An iterative approach will be used to determine which in vitro studies are most important to extract or summarize, based on factors such as concentrations tested, directness, and relevance of the in vitro outcomes to the human and animal outcomes of interest.
- Assess risk-of-bias for individual epidemiological and experimental animal studies.
- Synthesize the evidence across studies that assessed learning and memory using a narrative approach or meta-analysis (if appropriate) and evaluate sources of heterogeneity.

REVISION: ²NTP conducts systematic reviews following pre-specified protocols that describe the review procedures selected and applied from the general methods outlined in the OHAT Handbook. The protocol describes project-specific procedures tailored to each systematic review.

Protocol for Systematic Review of Effects of Fluoride Exposure on Neurodevelopment

- Use the GRADE (Grading of Recommendations Assessment, Development, and Evaluation) framework to rate confidence in the body of evidence for effects on learning and memory according to one of four statements: 1. High, 2. Moderate, 3. Low, or 4. Very Low/No Evidence Available.
- Translate confidence ratings into level of evidence for effects on learning and memory for human and animal bodies of evidence separately according to one of four statements: 1. High, 2. Moderate, 3. Low, or 4. Inadequate.
- Combine the level-of-evidence ratings for human and animal bodies of evidence and consider the degree of support from mechanistic data to reach one of five possible hazard identification conclusions: Known, Presumed, Suspected, Not classifiable, or Not identified to be a hazard to humans.
- Characterize uncertainty based on describing limitations of the evidence base, limitations of the systematic review, and consideration of dose relevance and pharmacokinetic differences when extrapolating findings from animal studies to human exposure levels, and identify key data gaps and research needs.

PECO Statement

A PECO statement (Participants/Population, Intervention/Exposure, Comparator, Outcome) was developed to address the overall research question (effects on neurodevelopmental or cognitive function and thyroid associated with fluoride exposure in humans, experimental animals, and in vitro model systems [Table 1]). The PECO statement is used to help develop the specific research questions, search terms, and inclusion/exclusion criteria for the systematic review (Higgins and Green 2011b).

Table 1. PECO (Populations, Exposures, Comparators, Outcomes) Statement	
PECO Element	Evidence
Population	Human: Epidemiological studies, with the exception of case studies and case reports.
	Animal: Non-human mammalian animal species (whole organism).
	In vitro: Human or animal cells, tissues, or biochemical reactions (e.g., ligand binding assays) with in vitro exposure regimens.
Exposure	Exposure to fluoride based on administered dose or concentration, biomonitoring data (e.g., urine, blood, other specimens), environmental measures (e.g., air, water levels), or job title or residence. Relevant forms are those used as additives for water fluoridation: <ul style="list-style-type: none"> • Fluorosilicic acid (also called hydrofluorosilicate; CASRN 16961-83-4) • Sodium hexafluorosilicate (also called disodium hexafluorosilicate or sodium fluorosilicate; CASRN 16893-85-9) • Sodium fluoride (CASRN 7681-49-4) • Other forms of fluoride that readily dissociate into free fluoride ions (e.g., potassium fluoride, calcium fluoride, ammonium fluoride) • REVISION: Aluminum fluoride or aluminum fluoride complexes
Comparator	Human: A comparison population exposed to lower levels of fluoride (e.g., exposure below detection levels) or no fluoride.

Table 1. PECO (Populations, Exposures, Comparators, Outcomes) Statement	
PECO Element	Evidence
	Animal and in vitro: Exposed to vehicle-only treatment.
Outcomes	Human and Animal: Neurodevelopmental outcomes including learning, memory, intelligence, other forms of cognitive behavior, other neurological outcomes (e.g., anxiety, aggression, motor activity), or biochemical changes in the brain, nervous system tissue. Also, measures of thyroid function, biochemical changes, or thyroid tissue.
	In vitro: Endpoints related to neurological and thyroid function, including neuronal electrophysiology; mRNA, gene, protein expression; cell proliferation or death in brain or thyroid tissue/cells; neuronal signaling; synaptogenesis, etc.

METHODS

Step 1. Problem Formulation

REVISION: *Problem formulation activities to develop this protocol*

NTP received a nomination from the public in June 2015 to conduct an analysis of fluoride developmental neurobehavioral toxicity. **REVISION: NTP published a request for information about fluoride in a Federal Register notice [October 7, 2015 (<https://www.govinfo.gov/content/pkg/FR-2015-10-07/pdf/2015-25434.pdf>); amended November 23, 2015 (<https://www.govinfo.gov/content/pkg/FR-2015-11-23/pdf/2015-29734.pdf>)].** In November 2015, OHAT released a draft concept for the *Proposed NTP Evaluation on Fluoride Exposure and Potential for Developmental Neurobehavioral Effects* and presented the draft concept at the NTP Board of Scientific Counselors (BSC) meeting on December 1–2, 2015 (see <https://ntp.niehs.nih.gov/go/meeting> for meeting materials). During the meeting there were opportunities for public comment and input from the NTP BSC on draft concept. NTP also received 20 public comments between 2015 and 2016 in response to the request for information (<https://ntp.niehs.nih.gov/go/785076>). All comments received were considered in developing this protocol to guide the systematic review.

REVISION: *NASEM Review*

The July 2019 draft of this protocol and September 6, 2019 draft of this monograph were reviewed by a committee convened by the National Academy of Sciences, Engineering, and Medicine (NASEM). The current draft reflects clarifications and changes in response to that review (NASEM 2020), including the addition of a separate meta-analysis protocol in Appendix 6.

The PECO statement was developed to address this nomination and was presented to the NTP Board of Scientific Counselors during its December 1–2, 2015 meeting.

Step 2. Search and Select Studies for Inclusion

Searching electronic databases

Database search strategies were developed using index terms and text words based on key elements of the PECO Statement. The following databases will be searched (full details of the search strategies are presented in Appendix 1):

REVISION: Main literature databases search

- BIOSIS (Thomson Reuters)
- EMBASE (Elsevier)
- PsycINFO (APA PsycNet)
- PubMed (NLM)
- Scopus (Elsevier)
- Web of Science (Thomson Reuters; Web of Science indexes the journal Fluoride)

REVISION: Supplemental Chinese database literature search

- CNKI
- Wanfang

Searches will not be restricted by publication date. No language restrictions will be applied.

REVISION: The literature search strategy for this protocol was based on the search terms used for the NTP 2016 systematic review of animal studies and refined for the current evaluation, including the addition of search terms to identify human studies (NTP 2016). Note that the assessment of animal data for this systematic review will be limited to studies published since 2015, as the animal data will be considered an extension of the NTP 2016 systematic review of animal studies (NTP 2016). This evaluation will rely on the NTP 2016 systematic review of animal studies as an assessment of the animal literature published prior to 2015. In addition to non-human animal studies published since 2015, the search strategy for this protocol will also identify human and mechanistic studies. The selection of human and mechanistic studies will not be limited by publication year.

REVISION: *Supplemental searching in Chinese databases*

Following NASEM committee peer review in November 2019 (NASEM 2020), additional searches were developed for non-English-language databases to systematically search for studies that were previously identified from other resources (e.g., Chinese-language studies from the Fluoride Action Network website). Multiple non-English language databases were explored before finding two databases (CNKI and Wanfang) that covered studies previously identified from other sources. Search terms from the main literature search were refined to focus on human epidemiology studies. The CNKI and Wanfang databases have character limits in the search strings; therefore, key terms were prioritized using text analytics to identify the most prevalent terms from neurodevelopmental or cognitive human epidemiology studies previously identified as relevant. Search strings were designed to capture known relevant studies that were previously identified from searching other resources without identifying large numbers of non-relevant studies (the search strategy for both databases are available in Appendix 1). These two Chinese electronic databases will be searched with no language restrictions or publication year limits.

Searching other resources

Reference lists of included studies from the full-text literature screen, reference lists of studies that do not contain original data (i.e., reviews, editorials, commentaries), and the Fluoride Action Network website will be searched for additional relevant publications.

Selection criteria for the evidence

REVISION: Screening and selection of studies from the main literature search

Studies will be screened for inclusion using a structured form in SWIFT-Active Screener, a machine-learning software program used to priority-rank studies for screening. SWIFT-Active Screener employs active learning to incorporate user feedback during the screening process to refine a statistical model that continually ranks the remaining studies according to their likelihood for inclusion. In addition, the software includes a statistical algorithm to estimate predicted recall (percent of truly relevant studies identified) while users work, thus providing a statistical basis for a decision about when to stop screening (Miller *et al.* 2016). The title and abstract screen will be stopped once the statistical algorithm in SWIFT-Active Screener estimates $\geq 98\%$ predicted recall. **REVISION: Howard *et al.* (2020) evaluated the performance of the SWIFT-Active Screener methods for estimating total number of relevant studies using 26 diverse systematic review data sets that were previously screened manually by reviewers. The authors found that, on average, 95% of the relevant articles were identified after screening 40% of the total reference list when using SWIFT-Active Screener.**

REVISION: NTP staff will oversee all levels of screening, and adjustments or clarifications to screening criteria will occur as directed by NTP. A pilot title and abstract screening will be conducted to train the evaluation design team (consisting of ICF and NTP staff) and identify any potential questions related to screening criteria, during which NTP staff will be one of the reviewers for each pilot study. When results between reviewers disagree, the two reviewers will discuss discrepancies and consult with ICF and NTP technical advisor(s) as necessary to resolve each disagreement for inclusion or exclusion. Following the pilot screening, regular title and abstract screening will be conducted. Two members of the evaluation design team will independently conduct a title and abstract screen of the search results to identify studies that meet the eligibility criteria. For citations with no abstract or non-English abstracts, articles will be screened based on title relevance (title should indicate clear relevance), page numbers (articles less than ≤ 2 pages long will be assumed to be conference reports, editorials, or letters), and/or PubMed MeSH headings.

Studies that are not excluded based on the title and abstract will advance to the full-text review. Full-text copies of potentially relevant articles will be screened for inclusion using a structured form in DistillerSR (Evidence Partners; <https://www.evidencepartners.com/products/distillersr-systematic-review-software/>) by two independent reviewers. When results between reviewers disagree, the two reviewers will discuss discrepancies and consult with technical advisor(s) if necessary to decide on the status (include/exclude) of each discrepancy. Translation assistance will be sought to assess the relevance of non-English studies. In addition, full-text copies of potentially relevant review articles also will be screened by two reviewers to identify studies from the review reference lists that satisfy the inclusion criteria.

To be eligible for inclusion, studies must comply with the PECO criteria (Table 1).

REVISION: Inclusion and exclusion criteria based on the PECO statement are detailed in Table 2.

Studies that do not meet the PECO criteria will be excluded. In addition to the PECO criteria, the following exclusion criteria will apply:

- Records that do not contain original data but are relevant to the PECO statement, such as reviews, editorials, or commentaries. Reference lists from these materials, however, will be reviewed to identify potentially relevant studies not identified from the database searches. These studies will be assessed for eligibility for inclusion based on the process described above.
- Conference abstracts or reports. Attempts will be made, however, to contact authors of recent conference abstracts (~past 2–3 years) to assess publication status when a published version of the full study was not identified via the database search.

NTP includes only publicly accessible information in its evaluations. This information is typically based on studies published in peer-reviewed journals. NTP, however, can consider unpublished data or data presented in the grey literature (e.g., conference reports, theses/dissertations, technical reports, white papers) that have not undergone peer review, provided the owners of the data are willing to have the study details and results made publicly accessible. NTP would organize a peer review of any submitted unpublished data (NTP 2015a, 2019a).

The list of included studies will be posted at the NTP website for this project once screening is completed. The results of the literature search will be presented in a PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) diagram (Moher *et al.* 2009).

REVISION: Screening of literature search update after NASEM review

Following the NASEM committee peer review in November 2019 (NASEM 2020), an additional search or update will be conducted using the search strategy presented in Appendix 1. The study screening and selection process will be focused on the human studies with primary outcomes for the evaluation because they form the basis of the conclusions. Animal *in vivo*, human secondary outcome-only, and human and animal mechanistic references identified as part of the screening process will be reviewed by NTP staff. These studies will be scanned for evidence that might extend the information in the September 6, 2019 draft monograph. Studies from this literature search update will be listed in an appendix; however, data from the studies will not be extracted unless it was believed they would materially advance the findings of the review.

REVISION: Screening of supplemental Chinese database search

Publications retrieved from the supplemental search will be compared to publications retrieved from the main literature search and duplicates removed. The remaining relevant publications will be categorized as “references identified through database searches.” New animal and mechanistic references retrieved will be scanned for evidence that might extend the information in the September 6, 2019 draft monograph. Only data that would materially advance the animal and mechanistic findings will be extracted and added to the revised draft. Newly-retrieved human references will be reviewed to identify studies that might impact conclusions with priority given to identifying and translating null studies that may have been missed using previous approaches. Any null study identified will be translated and included.

REVISION: Table 2. Inclusion and Exclusion Criteria to Determine Study Eligibility		
	Inclusion Criteria	Exclusion Criteria (or blank if none)
<i>Population (Human Studies or Experimental Model Systems)</i>		
Human	<ul style="list-style-type: none"> No restrictions on sex, age, or life stage at exposure or outcome assessment 	
Animal	<ul style="list-style-type: none"> No restrictions on sex, age, species, or life stage at exposure or outcome assessment 	<ul style="list-style-type: none"> Animal observational/wildlife studies
Mechanistic	<ul style="list-style-type: none"> Mechanistic studies will be restricted to human or animal cells, tissues, or model systems 	<ul style="list-style-type: none"> Studies in non-animal organisms
<i>Exposure</i>		
Human	<ul style="list-style-type: none"> Known dose or concentration in an experimental protocol Diagnostic biomonitoring data (e.g., fluoride in urine) Environmental detection (e.g., fluoride in drinking water) Corroboration by assessment of direct (in hospital, in clinic) or indirect observation of symptoms of fluoride exposure (e.g., dental fluorosis) 	
Animal	<ul style="list-style-type: none"> Exposure to fluoride based on known administered dose or concentration 	<ul style="list-style-type: none"> Multiple exposures to different chemicals
Mechanistic	<ul style="list-style-type: none"> Exposure to fluoride based on known administered dose or concentration 	<ul style="list-style-type: none"> Multiple exposures to different chemicals
<i>Comparators</i>		
Human	<ul style="list-style-type: none"> Unexposed or lowest-exposure group as the referent group 	
Animals	<ul style="list-style-type: none"> Comparable untreated animal subjects or animals exposed to vehicle-only treatment 	
Mechanistic	<ul style="list-style-type: none"> Study must include vehicle only control group 	
<i>Outcomes</i>		
Human	<ul style="list-style-type: none"> Neurobehavioral/neurodevelopmental outcomes (primarily IQ) Thyroid disease, thyroid gland function 	
Animal	<ul style="list-style-type: none"> Neurobehavioral/neurodevelopmental outcomes Thyroid disease, thyroid gland function 	Goiter
Mechanistic	<ul style="list-style-type: none"> Brain biochemistry, structure, activity 	

REVISION: Table 2. Inclusion and Exclusion Criteria to Determine Study Eligibility		
	Inclusion Criteria	Exclusion Criteria (or blank if none)
<i>Publication Type (e.g., specify any language restrictions, use of conference abstracts, etc.)</i>		
Human, animal, or mechanistic	<ul style="list-style-type: none"> • Report must contain original data in whole or in part relevant to the aims of this evaluation 	<ul style="list-style-type: none"> • Articles with no original data (e.g., editorial or review*) • Studies published in abstract form only (grant awards conference abstracts) • Retracted articles
*Relevant reviews are used as background and for reference scanning		

Step 3. Data Extraction and Content Management

Data will be extracted from individual studies by members of the evaluation team. Data extraction and warehousing will be carried out using Health Assessment Workspace Collaborative (HAWC), a free and open-source, web-based software application.³ Data extraction elements collected from epidemiological studies are listed in Appendix 2, from animal studies in Appendix 3, and in vitro studies in Appendix 4. The content of the data extraction might be revised following the identification of the studies included in the review. The data extraction results for included studies will be presented in the technical report and the data extraction results will be available for download from HAWC in Excel format when the project is completed.

REVISION: NTP staff will oversee the data extraction process. Extractors will be trained with an initial pilot phase of two studies of each study type (e.g., human and animal) undertaken to improve clarity of extraction instructions and to improve consistency among extractors. NTP staff will supervise pilot testing, and adjustments to extraction instructions will occur as directed by NTP. Data extraction will be performed by one member of the evaluation team and checked by one other member. Any discrepancies in data extraction will be resolved by discussion or consultation with a third member of the evaluation team. Missing data from individual animal and in vitro studies generally will not be sought. More attempts, however, will be made for missing data from an otherwise well-reported and well-conducted study, such as missing group size or variance descriptor (standard deviation/standard error from certain animal studies). Routine attempts will be made to obtain missing information from epidemiological studies. Outreach to study authors will be recorded in HAWC as unsuccessful if researchers do not respond to an email or phone request within 1 month of the attempt to contact. Missing information to assess risk-of-bias for animal and epidemiological studies will be sought routinely (see below).

Standardizing results from behavioral tests and dose levels in experimental animal studies

Results from behavioral tests will be transformed, when possible, to a common metric of percent change from control response to help assess dissimilar but related outcomes measured with different scales. In this project, percent control response will be used as the common metric because it is recommended for assessing dissimilar but related outcomes measured with different scales (Vesterinen

³Health Assessment Workspace Collaborative (HAWC): A Modular Web-based Interface to Facilitate Development of Human Health Assessments of Chemicals. <https://hawcproject.org/portal/>.

et al. 2014). Percent control group calculations will be based on sample size, means, and standard deviation or standard error values presented in the studies.

For studies in which experimental animals were dosed with sodium fluoride (NaF) or other forms of dissociable fluoride, dose levels will be converted to fluoride equivalents (F), for example, 100 ppm NaF = 45.3 ppm (mg/L) fluoride. In studies where F was administered directly (often reported simply as “fluoride”), no such conversions are conducted. Fluoride dose levels are standardized to mg/kg-d and ppm (mg/L). Conversions will be made using water consumption rates and body weights for rats and mice reported in the EPA dosimetry (US EPA 1988, 1994). In each case, the “subchronic” values will be used because this period fit the maternal or single-generation dosing periods in most studies. The strain-specific and sex-specific values will be used when available; for strains that are not available, the “other” values will be used. For studies in which dosing is through the feed, the first conversion is from food ppm to mg/kg-d. Then, the “effective water concentration” is estimated by multiplying the converted dietary dose by body weight and dividing by water consumption rate. The uncertainty in these estimates should be considered higher than in water consumption studies.

Unless otherwise reported by study authors, a fluoride background level of 0 ppm (0 mg/kg-d) will be assumed in experimental animal studies. As available, dose levels will be presented as mg/kg-d and ppm. Dose conversions using US EPA (1988, 1994) default food or water consumption rates and body weights will be performed for any studies not reporting dose levels as mg/kg-d. Importantly, dose levels in mg/kg-d can vary for a given ppm across different studies if the studies use different species, strains, or sexes of animals that are assumed to have different food or water consumption rates.

Step 4. Quality Assessment of Individual Studies

Risk-of-bias (RoB) will be assessed for individual studies using a tool developed by OHAT that takes a parallel approach for evaluating RoB from human and animal studies in order to facilitate consideration of RoB across evidence streams with common terms and categories (NTP 2015a, 2019a) (Table 3). The RoB tool consists of a set of 11 questions that are answered based on the specific details of individual studies to develop RoB ratings (using the four options in Table 3) for each question. The subset of questions that will be used to assess RoB for an individual study is based on the study design (Table 3); specific protocols have been developed for this systematic review based on evidence stream and the type of human study design (Appendix 5). For example, the subset of RoB questions applicable to all experimental study designs includes a question on randomization of exposure that would not be applicable to observational study designs. RoB will be assessed at the outcome level because study design or method specifics might increase the RoB for some outcomes and not for others within the same study. Missing information to assess RoB for human and animal studies will be routinely sought. Outreach to study authors will be recorded in HAWC as unsuccessful if researchers do not respond to an email or phone request within 1 month of the attempt to contact. Any information not reported will be assumed as not having been conducted (e.g., randomization, blinding), resulting in an assessment of “probably high” RoB.

REVISION: NTP staff will oversee the risk-of-bias assessment process and supervise the pilot testing; adjustments to the RoB criteria will occur as directed by NTP. If questions arise during piloting or subsequent RoB assessment on the appropriateness of a specific method (e.g., outcome assessment, statistical analyses), experts in the subject area will be contacted. For both epidemiological and experimental animal studies, two reviewers will independently conduct RoB evaluations and reach consensus on disagreements by discussion and consultation with technical expert(s) as needed.

Assessors will be trained using the criteria in Appendix 5 with an initial pilot phase undertaken to improve clarity of criteria that distinguish between adjacent ratings and to improve consistency among assessors. All team members involved in the RoB assessment will be trained on the same set of studies and asked to identify potential ambiguities in the criteria used to assign ratings for each question. Any ambiguities and rating conflicts will be discussed relative to opportunities to refine the criteria to distinguish more clearly between adjacent ratings. If major changes to the RoB criteria are made based on the pilot phase (i.e., those that would likely result in revision of response), they will be documented in a protocol amendment along with the date of modifications and the logic for the changes. Information about confounding, exposure characterization, outcome assessment, and other important issues might be identified during or after data extraction, which can lead to further refinement of the RoB criteria (Sterne *et al.* 2014). After assessors have independently made RoB determinations for a study across all RoB questions, the two assessors will compare their results to identify discrepancies and attempt to resolve them. Any remaining discrepancies will be considered by the project lead and, if needed, other members of the evaluation design team and technical advisors. The final RoB rating for each question will be recorded with a statement of its basis.

Table 3. Response Options for Each RoB Question	
	Definitely Low risk-of-bias: Direct evidence of low risk-of-bias practices (Could include specific examples of relevant low risk-of-bias practices)
	Probably Low risk-of-bias: Indirect evidence of low risk-of-bias practices OR deviations from low risk-of-bias practices for these criteria during the study are deemed not to bias results appreciably, including consideration of direction and magnitude of bias
	Probably High risk-of-bias: Indirect evidence of high risk-of-bias practices OR insufficient information (e.g., not reported or “NR”) is provided about relevant risk-of-bias practices
	Definitely High risk-of-bias: Direct evidence of high risk-of-bias practices (Could include specific examples of relevant high risk-of-bias practices)

Critical risk-of-bias domains for epidemiology studies

Confidence in exposure or exposure assessment, the study design accounting for confounding variables, and the confidence in the outcome assessment (including blinding of outcome assessors to subjects’ exposure levels) are the critical risk-of-bias domains that will be used to evaluate the potential for an overall very serious RoB for individual studies, referred to as a “tier 3” study in the OHAT Handbook (NTP 2015a, 2019a). Studies considered “probably high” or “definitely high” RoB in several of these domains are considered to pose an overall very serious RoB. These studies might be excluded from the analysis when they represent a sizeable portion of the studies considered for evidence integration.

Rationale for critical risk-of-bias domains for human studies

- Differential or non-differential misclassification of the outcome through an improper definition of the outcome status or errors at the data collection stage may lead to an over- or underestimation of the effect size (Szklo *et al.* 2014). Confidence in the outcome assessment

for observational epidemiology studies will be evaluated both in terms of the specific measurement instruments used and with regard to the steps taken towards blinding the assessment of the outcome. Ideally, epidemiologic studies would include independent assessments of outcome measure validity both in the population for which it was originally designed, and with modifications appropriate to the study population of interest (Sabanathan *et al.* 2015). However, studies utilizing well-documented tests with modification for the population being studied would be considered at least "probably low RoB", even without specific validity measures provided. Importantly, a validated outcome assessment instrument may still result in bias if the test assessors are not appropriately blinded to the exposure status. For this reason, failure to provide evidence of blinding at outcome assessment, scoring, and evaluation will be weighed more heavily than the specific outcome assessment measure. In cases where blinding is not possible due to discrete study populations and/or exposure locations, studies should be considered "probably high RoB" unless specific direct or indirect evidence of blinding is provided or steps were taken to minimize the potential bias due to lack of blinding.

- Confirmation of exposure is vital to proper analysis and effect assessment. This should include evidence of consistent assessment methods used throughout the study, detection and quantification limits, and the utilization of well-established methods that directly characterize the exposure or intake. Studies that do not measure individual exposures (that is, that use summary statistics for a given population or group), generally will rate probably or definitely high on exposure assessment risk of bias. Studies where summary statistics are poorly documented with regard to variability or range, numbers of samples from which estimates were derived, and source and timing of measurements, may be assigned definitely high RoB. Studies that measure or estimate individual exposures, biomarker levels (such as urinary fluoride), or fluoride intake will generally be assigned probably or definitely low RoB with regard to exposure assessment. Where non-water sources of fluoride are unlikely to contribute substantially to overall intake, using indirect measures of exposure such as drinking water levels will not, by itself, be sufficient grounds for increasing the risk of bias rating.
- In assessing the effect of an exposure on a given outcome, improper adjustment for confounders can bias the results towards or away from the null (Szklo *et al.* 2014). **REVISION:** ~~Therefore, direct evidence should be provided that adjustments and/or considerations were made for any covariates that are known to effect the relationship between the exposure and outcome of interest in each study. For neurodevelopmental effects of fluoride exposure, key covariates include co-exposure to other chemicals associated with neurotoxicity (e.g., arsenic and lead) and iodine sufficiency. Failure to consider the distribution of the key covariates across the exposure groups will result in a "probably high RoB" or "definitely high RoB", depending on the likelihood of those factors affecting the results of the final analyses. Furthermore, individual and parental demographic, socioeconomic, and health characteristics, nutrition and growth factors, parental IQ, and smoking and smoke exposure, and dental and skeletal fluorosis should all be given either direct or indirect consideration. Dental and skeletal fluorosis are highly correlated with fluoride exposure, so careful consideration should be given to how they are handled in the study, since such physical anomalies may impact performance on neurodevelopmental testing independent of fluoride exposure (von Hilsheimer and Kurko 1979). To receive a "definitely low RoB" rating, it will be necessary that studies both provide quantitative summaries of covariate values across~~

exposure groups or the study population, and adjust for covariates in statistical analyses. Therefore, direct or indirect evidence should be provided that adjustments and/or considerations were made for any covariates that are known to affect the relationship between the exposure and outcome of interest in each study. For neurodevelopmental effects of fluoride exposure, key covariates for all study populations include age, sex, and socioeconomic factors. Co-exposure to other chemicals associated with neurotoxicity (e.g., arsenic and lead) and iodine sufficiency are also of high importance in areas where co-exposures are reasonably expected to occur. Some areas of China, India, and Mexico have known co-exposure to arsenic. Groundwater quality maps (<https://www.gapmaps.org/Home/Public#>) can provide information on measurements or predictions of contamination around the world (Podgorski and Berg 2020).⁴ Failure to consider the distribution of the key covariates across the exposure groups will result in a “probably high RoB” or “definitely high RoB” depending on the likelihood of those factors affecting the results of the final analyses. Although lead and iodine sufficiency are considered potential key covariates due to their potential to be related to neurodevelopmental effects, there is no expectation that they would in general be related to fluoride exposure so should not be considered “probably high RoB” or “definitely high RoB” unless there is a reasonable expectation that these would occur with fluoride exposure and would potentially impact the results. Furthermore, potential confounding due to individual and parental demographics, health characteristics, nutrition and growth factors, parental IQ, smoking and smoke exposure, and dental and skeletal fluorosis should be given consideration. Lack of direct or indirect information on these potential confounders should be addressed on a study-by-study basis to determine the potential impact on the results. Dental and skeletal fluorosis are highly correlated with fluoride exposure, so careful consideration should be given to how they are handled in the study, since such physical anomalies may impact performance on neurodevelopmental testing independent of fluoride exposure (von Hilsheimer and Kurko 1979). To receive a “definitely low RoB” rating, it will be necessary that studies both provide quantitative summaries of covariate values across exposure groups or the study population and adjust for covariates in statistical analyses, as well as provide evidence that the method to obtain the covariates is valid and reliable. However, a study can receive a “probably low RoB” rating if there is indirect evidence (e.g., statement that characteristics were similar across groups or that the area was not known to have co-exposure to other neurodevelopmental toxins) that confounders were not a concern for the specific study population and outcome measured.

Critical risk-of-bias domains for animal studies

Randomization to treatment group, confidence in outcome assessment (including blinding of the outcome assessors), adequate characterization of the administered chemical, and controlling for litter effects in developmental studies are considered key drivers in determining potential for an overall very serious RoB for individual studies, referred to as a “tier 3” study in the OHAT Handbook (NTP 2015a, 2019a). Studies considered “probably high” or “definitely high” RoB in several of these domains are considered to pose an overall very serious RoB. These studies might be excluded from the analysis,

REVISION: ⁴ If no arsenic measurements are available for a study area, the arsenic groundwater quality predictions from the global arsenic 2020 map will be used to determine if co-exposure to arsenic is of particular concern in a study population. If an area has less than 50% probability of having arsenic levels greater than 10 µg/L (the WHO guideline concentration), the area will be considered not to have an issue with arsenic that would need to be addressed by the study authors.

although sensitivity analyses will be conducted to assess the impact of excluding studies on conclusions. Studies also might be considered tier 3 due to a combination of concern for RoB in a critical domain(s) and very poor reporting quality (e.g., not reporting the number of animals treated, species).

Rationale for critical risk-of-bias domains for animal studies

- A lack of randomization can bias results away from the null toward larger effect sizes. This effect has been empirically assessed in both controlled human trials (reviewed in Higgins *et al.* 2011) and experimental animals (reviewed in Krauth *et al.* 2013). This element is widely recommended to assess RoB for controlled human trials (IOM 2011, Guyatt *et al.* 2011a, Higgins *et al.* 2011, Viswanathan *et al.* 2012) and is included in most RoB instruments for animal studies, reviewed in (Krauth *et al.* 2013, Hooijmans *et al.* 2014).
- A lack of blinding in randomized human subject trials has been shown empirically to be associated with larger estimations of intervention effects (on average a 9% increase in an odds ratio) (Pildal *et al.* 2007). Schulz *et al.* (1995) analyzed 250 controlled trials and found a 17% larger estimation of treatment effect, on average, in studies that were not double-blinded. In trials with more subjective outcomes, lack of blinding was associated with greater bias (Wood *et al.* 2008), indicating a greater impact with subjective evaluations of outcomes. A similar association between lack of blinding at outcome assessment and larger measures of effect has been reported for experimental animal studies (Bebarta *et al.* 2003, Sena *et al.* 2007, Vesterinen *et al.* 2010). Research specifically evaluating the impact of lack of blinding during allocation to treatment groups or during the course of the study in experimental animal studies is absent or minimal (NTP 2015b). In addition, concealment of animal dose information is problematic if exposure results in obvious effects on the normal daily functioning of the animal. Additional steps can be taken in experimental animal studies to reduce the RoB such as counterbalancing critical factors (e.g., sex, observers, apparatus, session, necropsy order) to equally distribute each factor across dose groups for endpoint assessments. Concern for lack of blinding during allocation or the conduct of the study can be attenuated if blinding was implemented at outcome assessment. For these reasons, blinding at outcome assessment was weighed more heavily during RoB assessment than blinding during allocation concealment or during the course of the study. In neurobehavioral studies, concern for lack of blinding at outcome assessment is attenuated if behavioral parameters are measured by an automated, computer-driven system.
- In experimental animal studies, the confirmation of exposure and dose are important for exposure characterization but rarely empirically determined. Ideally, experimental animal studies would include independent verification of purity, dose level confirmation over the exposure period, and internal measure of the compound within the subject. Independent verification of purity would be considered best practice because the identity and purity as listed on the bottle can be inaccurate. Approximately 3% of commercially purchased chemicals are inaccurately labeled for the chemical, increasing to 10% when purity is considered (unpublished, personal communication Brad Collins, NTP chemist). Impurities also might be more toxic than the compound of interest. Studies that do not report source or purity will be considered “probably high RoB” for exposure, although if purity information is not reported but can be inferred from source (e.g., online product description), the rating would be “probably low RoB.” Studies also will be considered “probably low RoB” for

exposure if information on source and purity is not provided, but levels of fluoride measured in biological samples indicated a dose-response gradient across groups.

- In experimental animal studies, the preferred study design for developmental exposure is to consider the litter as the experimental unit for statistical analysis. Failure to adjust statistically or experimentally for litter in an animal study with developmental exposure or for a developmental endpoint is important. Animals generated from the same litter tend to respond more similarly than animals from different litters. The direction of the bias is away from the null toward a larger effect size (Haseman *et al.* 2001)⁵ if the individual pup is considered as the statistical unit rather than the dam or litter from which the pup is derived. This can be due to inflation of the sample size or biological influence of the dam and litter.

Missing information for risk-of-bias assessment

OHAT will attempt to contact authors of included studies by email to obtain missing information considered critical for evaluating risk-of-bias that cannot be inferred from the study. If additional information or data are received from study authors, risk-of-bias judgments will be modified to reflect the updated study information. If OHAT does not receive a response from the authors by one month of the contact attempt, a risk-of-bias response of “NR” for “not reported; probably high risk-of-bias” will be used and a note made in the data extraction files that an attempt to contact the authors was unsuccessful.

⁵In 2000, NTP cosponsored a workshop with EPA, “Low Dose Endocrine Disruptors Peer Review.” As part of the peer review, a group of statisticians reanalyzed several “low” dose studies (Haseman *et al.* 2001) Based on studies that used littermates, they determined that litter or dam effects generally were present such that pups within a litter were found to respond more similarly than pups from different litters. The overall conclusion was “[f]ailure to adjust for litter effects (e.g., to regard littermates as independent observations and thus the individual pup as the experimental unit) can greatly exaggerate the statistical significance of experimental findings.”

Table 4. OHAT Risk-of-bias Tool						
Bias Domains and Questions	Experimental Animal¹	Human Controlled Trials²	Cohort	Case-control	Cross-sectional³	Case Series
Selection Bias						
1. Was administered dose or exposure level adequately randomized?	X	X				
2. Was allocation to study groups adequately concealed?	X	X				
3. Did selection of study participants result in appropriate comparison groups?			X	X	X	
Confounding Bias						
4. Did the study design or analysis account for important confounding and modifying variables?			X	X	X	X
Performance Bias						
5. Were experimental conditions identical across study groups?	X					
6. Were the research personnel and human subjects blinded to the study group during the study?	X	X				
Attrition/Exclusion Bias						
7. Were outcome data complete without attrition or exclusion from analysis?	X	X	X	X	X	
Detection Bias						
8. Can we be confident in the exposure characterization?	X	X	X	X	X	X
9. Can we be confident in the outcome assessment?	X	X	X	X	X	X
Selective Reporting Bias						
10. Were all measured outcomes reported?	X	X	X	X	X	X
Other Sources of Bias						
11. Were there no other potential threats to internal validity (e.g., statistical methods were appropriate and researchers adhered to the study protocol)?	X	X	X	X	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).						

Step 5. Assessment of Confidence in the Body of Evidence

The quality of evidence for each outcome will be graded using the GRADE system for rating the confidence in the body of evidence (Guyatt *et al.* 2011a, Guyatt *et al.* 2011b, Guyatt *et al.* 2011c, Guyatt *et al.* 2011d, NTP 2015a, 2019a). To approach evidence assessment, the framework provides guidance for determining overall certainty in the evidence as “high,” “moderate,” “low,” or “very low” based on five factors for downgrading (e.g., RoB across studies, indirectness, unexplained inconsistency, imprecision, publication bias) and three for upgrading (e.g., large magnitude of the effect, dose response, plausible confounding). OHAT also considers consistency of findings across studies as a potential upgrade factor (Figure 1 and Table 5). More detailed guidance is available in the OHAT Handbook for conducting systematic review (NTP 2015a, 2019a).

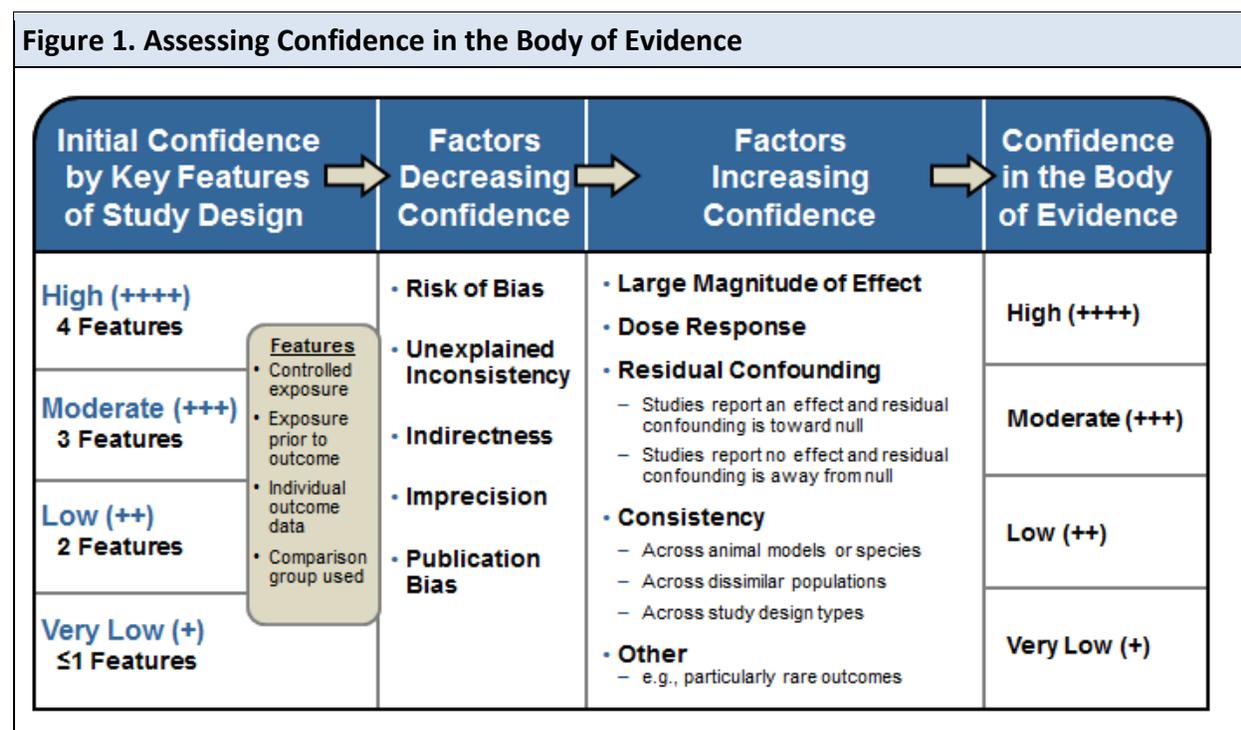


Table 5. Key Factors when Considering Whether to Downgrade or Upgrade Across a Body of Evidence

Downgrade factors	Rationale for potential downgrade
Risk-of-bias	Risk-of-bias across all domains. <ul style="list-style-type: none"> • Human studies: Critical factors include confounding bias, selection bias, exposure assessment, and outcome assessment. • Animal studies: Critical factors include randomization, blinding at outcome assessment, exposure characterization (e.g., reporting source, purity, internal dose level), and control for litter effects (developmental exposure studies).

Table 5. Key Factors when Considering Whether to Downgrade or Upgrade Across a Body of Evidence	
Unexplained inconsistency	Focuses on the presence of unexplained inconsistency in studies of similar population (or experimental model system) and design. Inconsistency can be determined by assessing similarity of point estimates and extent of overlap between confidence intervals or more formally through statistical tests of heterogeneity. Sensitivity analysis can be used to assess the impact of specific variables on the outcome (e.g., variation in RoB at individual study level, species, route of dosing, statistical power). Inconsistency that can be plausibly explained is typically not used to support for a downgrade. If only one study is available, a rating of “none” is used to characterize inconsistency.
Indirectness	<ul style="list-style-type: none"> • Human studies: All exposure levels and scenarios encountered in human studies (e.g., general population, occupational settings) are considered direct and not downgraded. • Animal studies: <ul style="list-style-type: none"> – <i>Within animal models:</i> Are the reported endpoints direct indicators of learning and memory? Can factors that might impact an animal’s performance in learning and memory tests, such as impaired motor or sensory function, be ruled out? Also consider the route of administration; oral is considered most relevant for fluoride. – <i>Extrapolation between mammalian animals and humans:</i> In vivo mammalian model systems have demonstrated utility for examining autonomic, sensory, and motor system functioning as they relate to human health and are considered direct and not downgraded. Although human cognitive function is not easily assessed in such systems, aspects of learning and memory can be evaluated as based on learning theory that translates across species (Crawley 2007). Some neurobehavioral measures (e.g., social behaviors, aggression, risky behaviors), however, have not been demonstrated to translate readily between species, and other behaviors (e.g., verbal learning/performance, gender preferences) cannot be evaluated adequately in a non-human mammalian model system. Studies that report only these endpoints would be downgraded for lack of directness. • In vitro and non-mammalian animal studies: Studies that involve direct treatment of cells or tissues or studies that only measure a biochemical reaction (e.g., receptor binding) are typically downgraded for lack of directness.
Imprecision	Confidence in quantitative measures such as effect sizes, identification of no observed effect level (NOEL) or lowest observed effect level (LOEL) doses, or parameters for benchmark dose analysis. Typically, 95% confidence intervals are used as the primary method to assess imprecision (Guyatt <i>et al.</i> 2011b). OHAT also considers whether studies are adequately powered when considering whether to downgrade.
Publication bias	Downgrade if “strongly detect” publication bias. Publication bias is difficult to assess, especially when multiple endpoints related to the primary outcome are reported in the same study, few studies are available, and papers do not report funding or conflicts of interest. Analytical tools, such as funnel plots or trim-and-fill approaches, can be used to assess publication bias but have substantial limitations and should be interpreted with caution (Guyatt <i>et al.</i> 2011c).

Table 5. Key Factors when Considering Whether to Downgrade or Upgrade Across a Body of Evidence	
Upgrade factors	Rationale for potential upgrade
Large magnitude of effect	Factors to consider include the outcome being measured and the dose or exposure range assessed.
Dose response	Patterns of dose response are evaluated within and across studies.
Consistency	<ul style="list-style-type: none"> • Human studies: Consider whether consistent results were reported across populations that differ in factors such as geographic region, different measures of exposure, nature of the cohort, e.g., occupational, general population. • Animal studies: Consider whether consistent results were reported in multiple experimental animal models or species.

Evidence synthesis

Endpoint grouping

Neurological endpoints will be broadly categorized using the frameworks below for human (Table 6) and animal (Table 7) studies. Evidence synthesis will focus on learning, memory, and intelligence. Studies reporting on other neurodevelopmental and cognitive endpoints will also be identified and summarized but not necessarily assessed for RoB at the individual study level (depending on the extent and nature of the literature).

Table 6. Neurological Outcomes Grouping for Human Studies		
General Domain	Example Tests or Test Batteries	Example Endpoints or Subtests
Learning, Memory, Intelligence, Cognitive Development	Neurobehavioral Core Test Battery (NCTB), Wechsler Intelligence Scale for Children (WISC) – Revised	Digit-Symbol Substitutions, Digit Span
	Wide Range Assessment of Memory and Learning (WRAML)	Verbal Memory Index, Visual Memory Index, Learning Index
	Wechsler Adult Intelligence Scale (WAIS)	Full Scale, Verbal, and Performance IQ
	Wechsler Preschool and Primary Scale of Intelligence (WPPSI), WPPSI-IV	Full Scale and Primary Index Scales (Verbal Comprehension Index, Working Memory Index, etc.)
	Halstead-Reitan Battery	Trail Making Test (Parts A and B)
	Wechsler Memory Scales (WMS)	Design Memory, Symbol Span
	Child and Adolescent Memory Profile (ChAMP)	Verbal Memory (lists), Visual Memory (objects)
	Stanford-Binet Test	Verbal and Non-Verbal Subtests in Visual Spatial Reasoning, Working Memory, etc.
	Raven’s Progressive Matrices	
Combined Raven’s Test for Rural China		

Table 6. Neurological Outcomes Grouping for Human Studies		
General Domain	Example Tests or Test Batteries	Example Endpoints or Subtests
Attention	Neurobehavioral Evaluation System (NES)	Finger Tapping, Continuous Performance Test, Simple Reaction Time
	Neurobehavioral Core Test Battery (NCTB)	Simple Reaction Time
Hyperactivity	Behavioral Assessment System for Children, Child Behavior Checklist	Hyperactivity Symptoms, Attention Problems scale
Motor/Sensory Function or Development	Neurobehavioral Evaluation System (NES)	Simple Reaction Time, Hand-Eye Coordination
	Neurobehavioral Core Test Battery (NCTB)	Santa Ana, Aiming
	Brazelton Neonatal Behavioral Assessment Scale, NICU Network Neurobehavioral Scales	Reflexes, Startle Response, Habituation to Sensory/Light Stimuli, Hand-Mouth Coordination
	Nerve conduction velocity	Motor or Sensory Conduction Velocity
Internalizing behaviors	Beck Depression Inventory, Behavioral Assessment System for Children, Child Depression Inventory, SPENCE Child Anxiety Scale, State-Trait Anxiety Inventory	
	Neurobehavioral Evaluation System (NES)	Profile of Mood States
Visual-Spatial or Visual-Motor Function	Neurobehavioral Core Test Battery (NCTB)	Benton Visual Retention Test
	Wechsler Intelligence Scale for Children (WISC) – Revised or Wechsler Adult Intelligence Scale (WAIS) – Revised	Block Design, Digit Symbol Substitution

Table 7. Neurological Outcomes Grouping for Animal Studies	
Endpoints	Example tests
Learning and Memory	Maze tests (Morris water maze, T-maze, Y-maze, Radial Arm); exploration (novel object recognition, mini-holeboard, activity cage); active and passive avoidance (step-down test, shuttle box); operant behavior
Motor and Sensory Function	Locomotor activity (open field, activity cage); movement coordination (akinesia/catalepsy, plank walking, rotarod, slanted surface, swim test); reflex (auditory startle, negative geotaxis, pain response: tail immersion and Von Frey hair test); developmental motor sensory landmarks (cliff avoidance, surface righting, pivoting/orienting reflex)
Depression	Forced swim; tail suspension test
Anxiety	Elevated plus maze
Other	Grooming; urination/defecation; sexual behavior; territorial behavior

Considerations for pursuing a narrative or quantitative evidence synthesis

Heterogeneity within the available evidence will determine the type of evidence integration that is appropriate: either a quantitative synthesis (meta-analysis) or narrative approach for evidence integration. Where appropriate, a meta-analysis will be conducted to summarize the findings. Summaries of main study design characteristics for each included study will be compiled to determine comparability between studies, identify data transformations necessary to ensure comparability, and determine whether study heterogeneity is a concern. Including a study might not be appropriate when (1) data on exposure or outcome are too different to be combined, (2) concerns about high RoB are present, (3) endpoints or measurement scales are not sufficiently similar, or (4) other circumstances indicate that averaging study results would not produce meaningful results. Topic-specific experts will be consulted to help assess whether studies are too heterogeneous for meta-analysis to be appropriate. When quantitatively combining results is inappropriate or infeasible, findings will be narratively described or visually presented. The main characteristics considered when determining whether to combine studies quantitatively include the following for human studies and animal studies.

Human studies:

- Study design (e.g., cohort, case-control study, cross-sectional, controlled trial, case report)
- Population demographics (sex, race/ethnicity, age or lifestage at exposure and outcome assessment)
- Exposure assessment method or matrix (e.g., blood, urine, hair, air, drinking water, job classification)
- Exposure range
- Neurodevelopmental and cognitive function measurements, methodology, and scale
- Type of data (e.g., continuous, dichotomous), statistics presented in paper, ability to access raw or additional data
- Variation in degree of RoB at individual study level or very serious concern for RoB across studies

Animal studies:

- Animal model used (species, strain, sex, genetic background)
- Age of animals (at start of treatment and outcome assessment, mating, pregnancy status)
- Dose levels, frequency of treatment, timing, duration, and exposure route
- Neurodevelopmental and cognitive function measurements and methodology
- Type of data (e.g., continuous, dichotomous), statistics presented in paper, ability to access raw or additional data
- Variation in degree of RoB at individual study level or very serious concern for RoB across studies

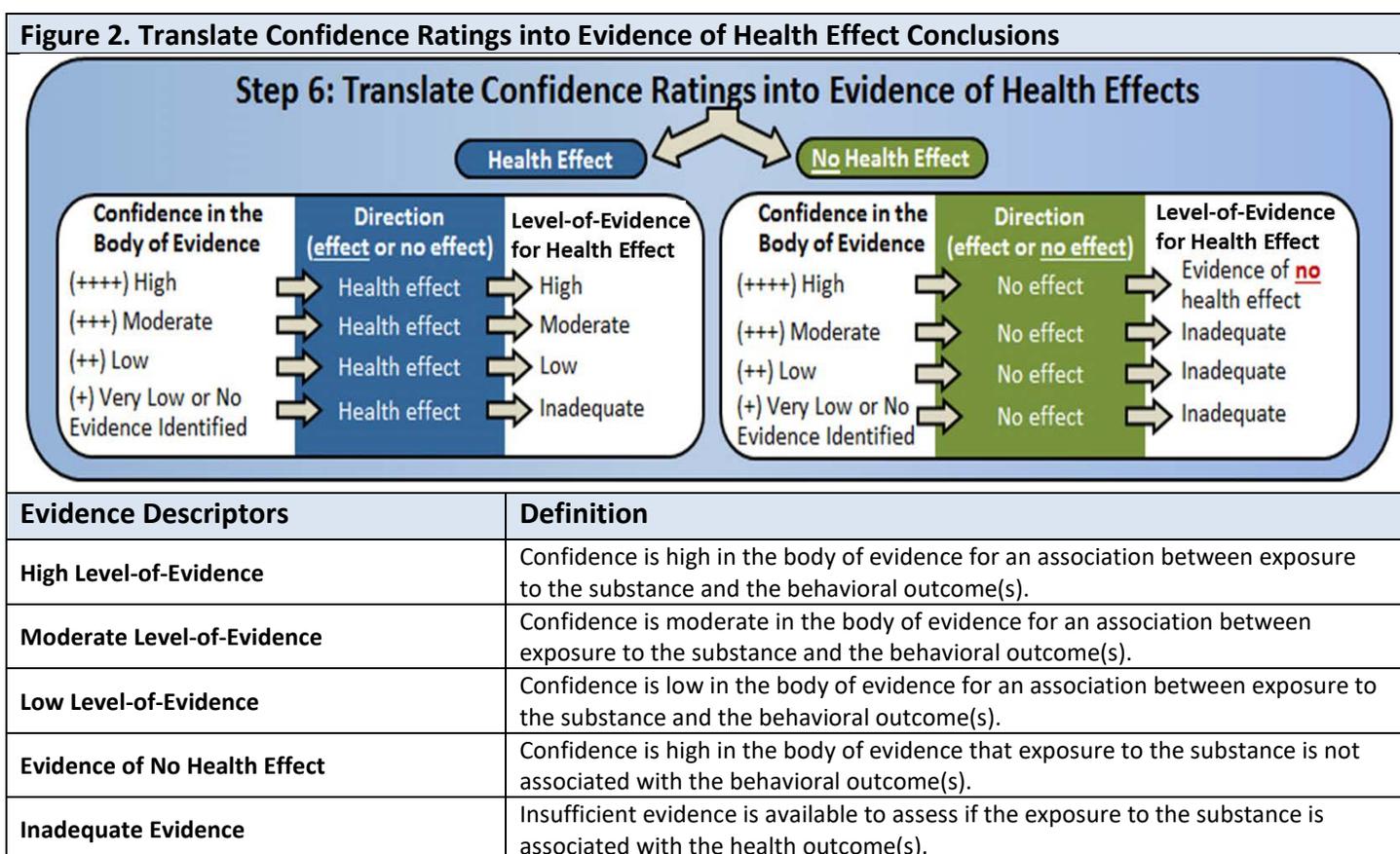
REVISION: Further consideration of meta-analysis

In the November 2019 review of the September 6, 2019 draft monograph (NASEM 2020), NASEM strongly recommended that a meta-analysis be conducted. In response, NTP performed meta-analyses

of IQ studies in children. A focused meta-analysis protocol was developed to add to this overall protocol for conducting the systematic review. The meta-analysis protocol was peer-reviewed separately, revised, and used to conduct meta-analyses included in revisions to the monograph. Appendix 6 contains the detailed protocol for the meta-analyses.

Step 6. Preparation of Level of Evidence Statement

The confidence in the body of evidence conclusions from Figure 1 will be translated into draft statements of health effects for human studies according to one of four statements: 1. High, 2. Moderate, 3. Low, or 4. Inadequate (Figure 2). The descriptor “evidence of no health effect” is used to indicate confidence that the substance is not associated with a health effect. Because of the inherent difficulty in proving a negative, the conclusion “evidence of no health effect” is only reached when there is high confidence in the body of evidence.



Step 7. Integrate Evidence to Develop Hazard Identification Conclusions

Initial hazard identification conclusions will be reached by integrating the highest level-of-evidence conclusion for neurodevelopmental effect(s) by integrating evidence of each outcome from the human and the animal evidence streams. Owing to ambiguities related to the interpretation of behavior tests in animals, it will likely not be possible to correlate specific outcomes in test animals with those in humans. Similarities in the general patterns of results for specific domains (such as learning and memory) may be considered across species as reviewers and experts consider appropriate, however. Human studies will

provide the primary basis for hazard conclusions, with animal test results providing ancillary and supporting evidence.

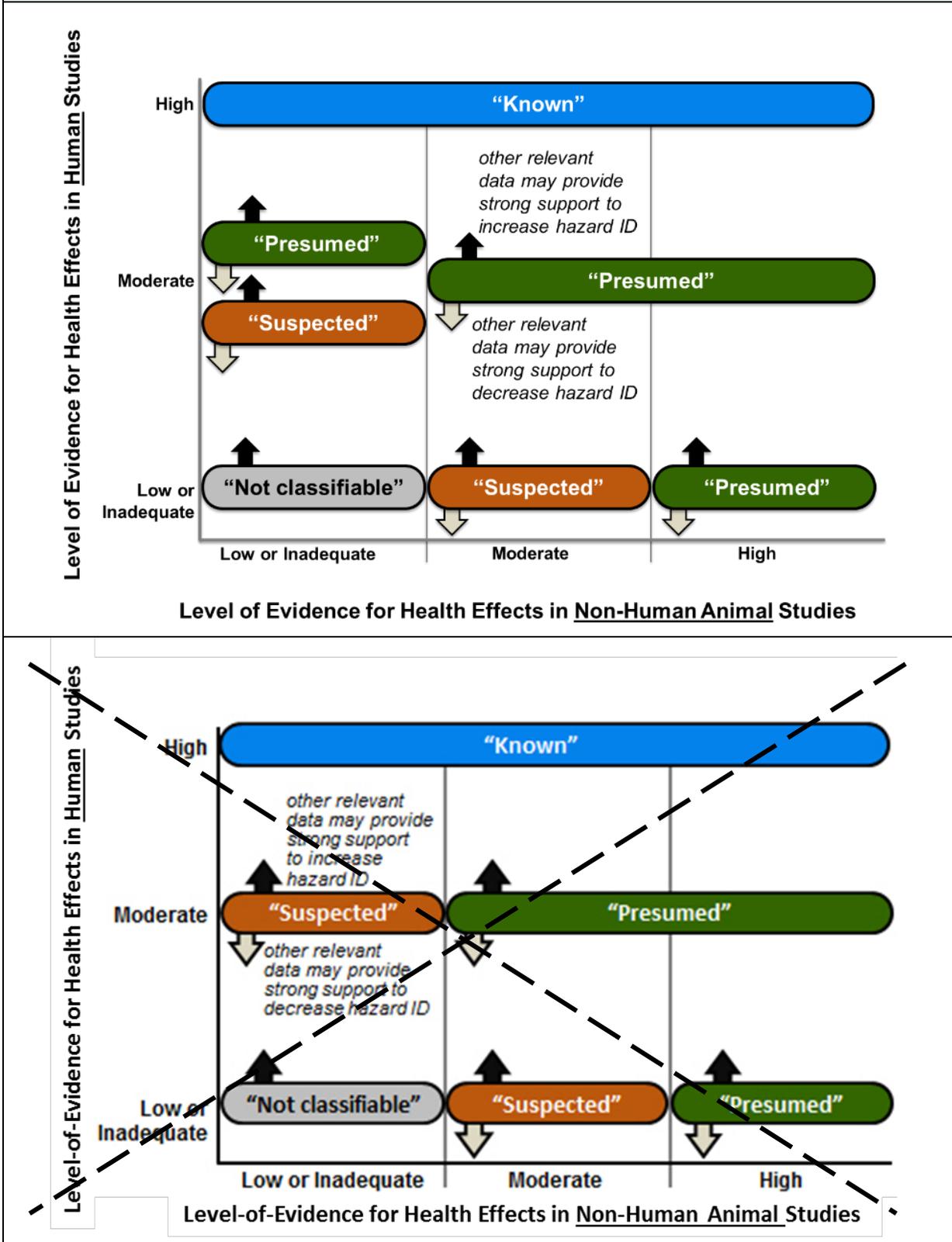
Hazard identification conclusions may be reached on the groups of biologically-related outcomes (using outcome groups identified in Table 6 and Table 7 as well as more specific endpoints if data are available to make more specific conclusions.

For similar/equivalent outcomes:

If the data support a specific neurological effect, the level-of-evidence conclusion for human data from Step 6 for that health outcome will be considered together with the level of evidence for the biologically related or equivalent non-human animal data to reach one of four initial hazard identification conclusions: Known, Presumed, Suspected, or Not classifiable. If either the human or animal evidence stream is characterized as “Inadequate Evidence,” then conclusions are based on the remaining evidence stream alone (which is equivalent to treating the missing evidence stream as “Low” in Figure 3).

For outcomes where the human and animal endpoints are dissimilar, the hazard conclusion may be developed on either the human or animal evidence alone. As shown in Figure 3, if the level of confidence in a health effect is “High” in animals, but evidence is “Low” or “Inadequate” in humans, the overall level of evidence conclusion can be no greater than “Presumed.” That is, animal evidence alone will not be sufficient to support a conclusion of “Known” neurodevelopmental effects in humans. If the human level of evidence rating of “Evidence of no health effect” from Step 6 is supported by a similar level of evidence rating for animal evidence for no health effect, the hazard identification conclusion would be “Not identified to be a hazard to humans.” If a moderate level-of-evidence conclusion for human data were reached with “Inadequate or Low Evidence” for the animal evidence stream, a hazard identification conclusion of either “suspected to be a hazard to humans” or “presumed to be a hazard to humans” could be reached based on scientific judgement as to the robustness of the body of evidence that supports moderate confidence in the human data and consideration of the potential impact of additional studies.

REVISION: Figure 3. Hazard Identification Scheme for Neurodevelopmental or Cognitive Effects

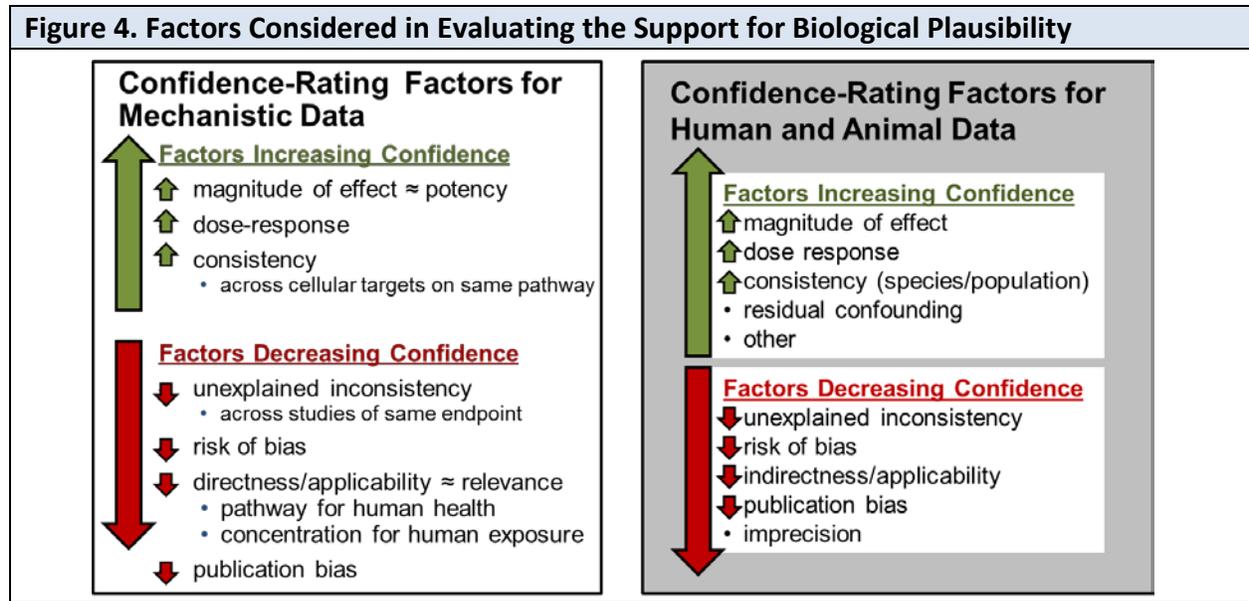


Consideration of mechanistic data

NTP does not require mechanistic or mode-of-action data to reach hazard identification conclusions, although when available, this and other relevant supporting types of evidence may be used to raise (or lower) the category of the hazard identification conclusion. Mechanistic data can come from a wide variety of studies that are not intended to identify a disease phenotype. This source of experimental data includes in vitro and in vivo laboratory studies directed at cellular, biochemical, and molecular mechanisms that explain how a chemical produces particular adverse effects.

The strength of the support or opposition presented by the other relevant data is evaluated using the guidance presented in Figure 4. The factors outlined for increasing or decreasing confidence in that the mechanistic data support biological plausibility are conceptually similar to those used to rate confidence in bodies of evidence for human or animal in vivo studies. Evaluations of the strength of evidence provided by mechanistic data are made on an outcome-specific basis based on discussion by the evaluation team and consultation with technical advisors as needed.

The factors presented in Figures 3 and 4 will be considered in an iterative and effect-specific manner. For example, mechanistic data in animals may affect the level of evidence in animal studies (Figure 4), which can affect the overall hazard identification conclusion based on combined animal and human studies (Figure 3). For example:



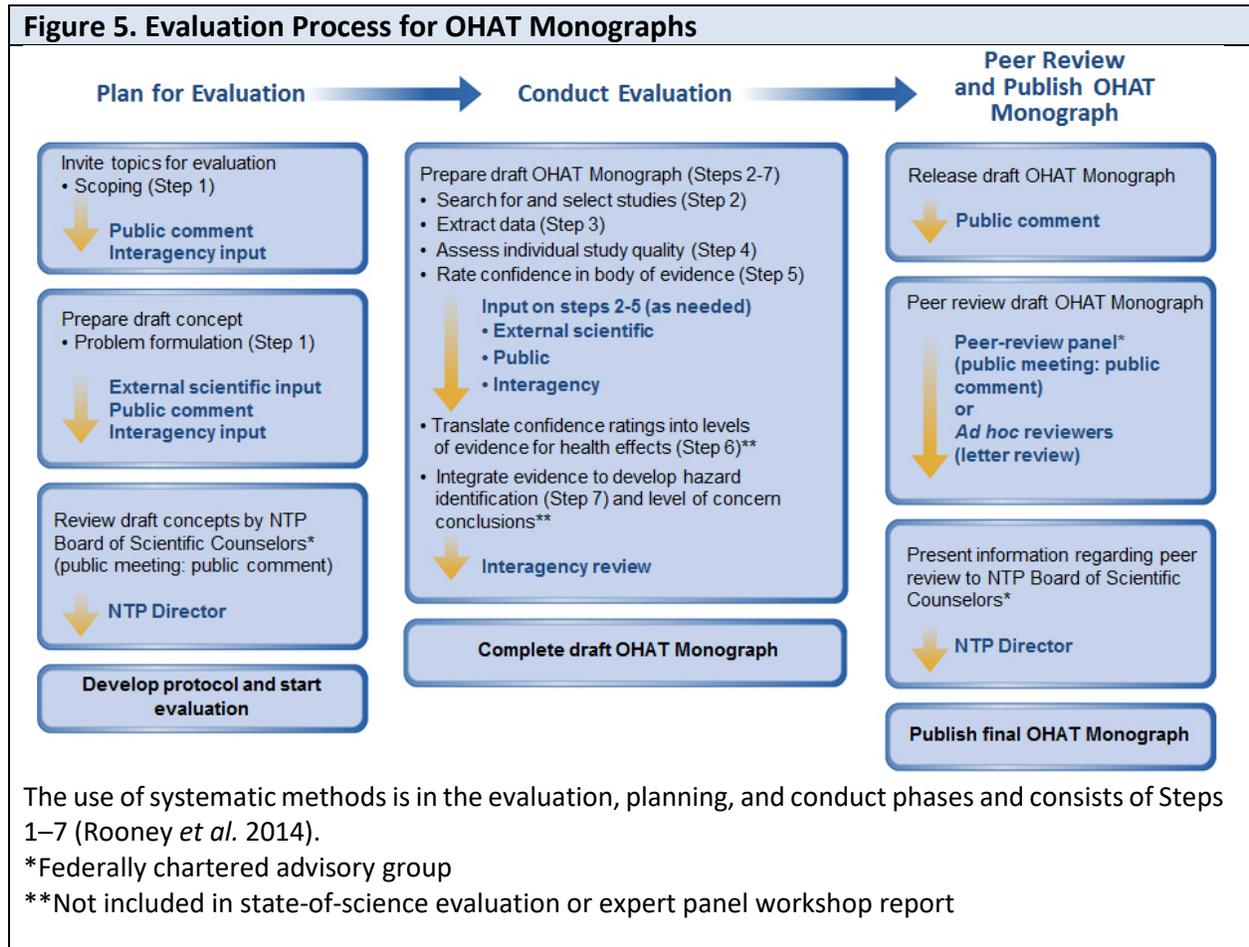
- If mechanistic data provide strong support for biological plausibility of the relationship between exposure and the health effect, the hazard identification conclusion may be upgraded (indicated by black “up” arrows in the Step 7 graphic in Figure 3) from that initially derived by considering the human and non-human animal evidence together.
- **REVISION: If mechanistic data provide strong opposition** ~~If mechanistic data fail to provide support 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 3) from that initially derived by considering the human and non-human animal evidence together.

As mode of action pathways have not been well-established for the neurodevelopmental effects of fluoride, the primary role of mechanistic data will be to inform the biological plausibility of observed outcomes from in vivo data. That is, mechanistic data alone will not be sufficient by itself to support hazard identification conclusions for neurodevelopmental endpoints.

NTP MONOGRAPH

Evaluation Process

The problem formulation and evaluation process of preparing an NTP Monograph includes multiple opportunities for external scientific, public, and interagency inputs and external peer review (Figure 5).



The NTP Monograph will include the methodology, results, discussion, and conclusion.

Methodology

This section will provide a brief overview of the methodologies used in the review process, including:

- the research question (PECO statement);
- the search strategy used to identify and retrieve studies;
- the process for selecting the included studies;
- the quality assessment of included studies;
- the methods of data extraction;

- the methods used to critically appraise for RoB, sensitivity, and synthesize the data of included studies.

Results

REVISION: This section will include the results from the systematic review on the neurotoxicity of fluoride in human **and animal** studies. Results will be presented in tables or figures as appropriate using HAWC. The results from the included studies will be discussed by outcome. This will include a description of:

- The number of studies identified considered relevant to PECO statement;
- The quality of the studies, as assessed using the appropriate tool;
- A data extraction and summary of the results from all studies;
- Quality of evidence and corresponding level of evidence conclusions rated according to one of four statements: 1. High, 2. Moderate, 3. Low, or 4. Very Low/No Evidence Available;
- Hazard identification conclusions based on integrating level of evidence ratings for human and animal data and consider the degree of support from mechanistic data (Known, Presumed, Suspected, Not classifiable, or Not identified to be a hazard to humans).

Discussion

The discussion will provide a summary of the review findings and characterize uncertainty based on describing limitations of the evidence base, limitations of the systematic review, consideration of dose-relevance and pharmacokinetic differences when extrapolating findings from animal studies to human exposure levels, and identifying key data gaps and research needs.

Conclusion

This will present the conclusion of the review.

REFERENCES

- [ATSDR] Agency for Toxic Substances and Disease Registry. 2003. *Toxicological Profile for Fluorides, Hydrogen Fluoride, and Fluorine*. US Department of Health and Human Services. Available: <https://www.atsdr.cdc.gov/toxprofiles/tp11.pdf> [accessed 4 April 2019].
- [IOM] Institute of Medicine. 2011. *Finding what works in health care: Standards for systematic reviews*. Available: http://www.nap.edu/openbook.php?record_id=13059&page=R1 [accessed 13 January 2013].
- REVISION: [NASEM] National Academies of Sciences Engineering and Medicine. 2020. *Review of the draft NTP monograph: Systematic review of fluoride exposure and neurodevelopmental and cognitive health effects*. Washington, DC: The National Academies Press. Available: <https://doi.org/10.17226/25715> [accessed 9 September 2020].**
- [NIOSH] The National Institute for Occupational Safety and Health. 1994. *Fluoride in urine: Method 8308*. NIOSH Manual of Analytical Methods (NMAM). Fourth Edition. Available: <https://www.cdc.gov/niosh/docs/2003-154/pdfs/8308.pdf> [accessed 4 April 2019].
- [NRC] National Research Council. 2006. *Fluoride in drinking water: A scientific review of EPA's standards*. Committee on Fluoride in Drinking Water, Board on Environmental Studies and Toxicology. Washington: National Academies Press. Available: <http://www.nap.edu/catalog/11571/fluoride-in-drinking-water-a-scientific-review-of-epas-standards> [accessed August 23 2015].
- [NTP] National Toxicology Program. 2015a. *Handbook for conducting a literature-based health assessment using OHAT approach for systematic review and evidence integration*. RTP, NC: Office of Health Assessment and Translation, Division of the National Toxicology Program, National Institute of Environmental Health Sciences. Available: https://ntp.niehs.nih.gov/ntp/ohat/pubs/handbookjan2015_508.pdf [accessed 25 January 2016].
- [NTP] National Toxicology Program. 2015b. *OHAT risk of bias tool*. RTP, NC: Office of Health Assessment and Translation. Division of the National Toxicology Program, National Institute of Environmental Health Sciences. Available: https://ntp.niehs.nih.gov/ntp/ohat/pubs/riskofbiastool_508.pdf [accessed 4 April 2019].
- REVISION: [NTP] National Toxicology Program. 2016. *Systematic literature review on the effects of fluoride on learning and memory in animal studies*. Research Report 1. RTP, NC. Available: https://ntp.niehs.nih.gov/ntp/results/pubs/rr/reports/01fluoride_508.pdf [accessed 9 September 2020].**
- [NTP] National Toxicology Program. 2019a. *Handbook for conducting a literature-based health assessment using OHAT approach for systematic review and evidence integration*. RTP, NC: Office of Health Assessment and Translation. Division of the National Toxicology Program, National Institute of Environmental Health Sciences. Available: https://ntp.niehs.nih.gov/ntp/ohat/pubs/handbookmarch2019_508.pdf [accessed 29 March 2019].

- REVISION: [NTP] National Toxicology Program. 2019b. *Protocol for systematic review of effects of fluoride exposure on neurodevelopment*. RTP, NC: U.S. Department of Health and Human Services, Public Health Services, National Toxicology Program. Available: <https://ntp.niehs.nih.gov/whatwestudy/assessments/noncancer/completed/fluoride/index.html> [accessed 19 August 2019].**
- [OEHHA] California Office of Environmental Health Hazard Assessment. 2011. *Meeting synopsis and slide presentations: Carcinogen identification committee meeting held on October 12, 2011*. Available: http://oehha.ca.gov/prop65/public_meetings/cic101211synop.html [accessed 17 September 2015].
- [SCHER] Scientific Committee on Health and Environmental Risks. 2011. *Critical review of any new evidence on the hazard profile, health effects, and human exposure to fluoride and the fluoridating agents of drinking water*. European Commission Directorate-General for Health and Consumers, Scientific Committees. Available: http://ec.europa.eu/health/scientific_committees/environmental_risks/docs/scher_o_139.pdf [accessed 17 September 2015].
- [US DHHS] U.S. Department of Health and Human Services Federal Panel on Community Water Fluoridation. 2015. PHS recommendation for fluoride concentration in drinking water. *Public Health Reports* 130(July-August): 318-331. Available: http://www.publichealthreports.org/documents/PHS_2015_Fluoride_Guidelines.pdf [accessed 17 September 2015].
- [US EPA] U.S. Environmental Protection Agency. 1988. *Recommendations for and documentation of biological values for use in risk assessment*. Washington, DC, EPA/600/6-87/008. Available: <http://cfpub.epa.gov/ncea/risk/recordisplay.cfm?deid=34855> [accessed 4 April 2019].
- [US EPA] U.S. Environmental Protection Agency. 1994. *Methods for derivation of inhalation reference concentrations (RfCs) and application of inhalation dosimetry*. Washington, DC, EPA/600/8-90/066F. Available: <http://cfpub.epa.gov/ncea/cfm/recordisplay.cfm?deid=71993> [accessed 4 April 2019].
- [US EPA] U.S. Environmental Protection Agency. 2010. *Fluoride: Exposure and relative source contribution analysis*. Washington, DC. Available: <http://www.epa.gov/dwstandardsregulations/fluoride-risk-assessment-and-relative-source-contribution> [accessed 29 February 2016].
- [US EPA] U.S. Environmental Protection Agency. 2013. *Basic information about fluoride in drinking water: Review of fluoride drinking water standard*. Available: <http://www2.epa.gov/dwsixyearreview/review-fluoride-drinking-water-regulation> [accessed 16 November 2015].
- [WHO] World Health Organization. 2004. *Fluoride in drinking-water: Background document for development of WHO Guidelines for Drinking-water Quality*. Available: http://www.who.int/water_sanitation_health/dwq/chemicals/fluoride.pdf [accessed 4 April 2019].

REVISION: Adinolfi M. 1985. The development of the human blood-CSF-brain barrier. *Dev. Med. Child Neurol.* 27(4): 532-537.

REVISION: Bagnardi V, Zambon A, Quatto P, Corrao G. 2004. Flexible meta-regression functions for modeling aggregate dose-response data, with an application to alcohol and mortality. *Am. J. Epidemiol.* 159(11): 1077-1086.

REVISION: Bashash M, Thomas D, Hu H, Martinez-Mier EA, Sanchez BN, Basu N, Peterson KE, Ettinger AS, Wright R, Zhang Z, Liu Y, Schnaas L, Mercado-Garcia A, Tellez-Rojo MM, Hernandez-Avila M. 2017. Prenatal fluoride exposure and cognitive outcomes in children at 4 and 6-12 years of age in Mexico. *Environ. Health Perspect.* 125(9): 1-12.

Bebarta V, Luyten D, Heard K. 2003. Emergency medicine animal research: does use of randomization and blinding affect the results? *Acad. Emerg. Med.* 10(6): 684-687.

REVISION: Berlin JA, Longnecker MP, Greenland S. 1993. Meta-analysis of epidemiologic dose-response data. *Epidemiology* 4(3): 218-228.

Broadbent JM, Thomson WM, Ramrakha S, Moffitt TE, Zeng J, Foster Page LA, Poulton R. 2015. Community Water Fluoridation and Intelligence: Prospective Study in New Zealand. *Am. J. Public Health* 105(1): 72-76.

Choi AL, Sun G, Zhang Y, Grandjean P. 2012. Developmental fluoride neurotoxicity: A systematic review and meta-analysis. *Environ. Health Perspect.* 120(10): 1362-1368.

REVISION: Cochran WG. 1954. The combination of estimates from different experiments. *Biometrics* 10(1): 101-129.

Crawley J. 2007. *What's wrong with my mouse?: Behavioral phenotyping of transgenic and knockout mice, 2nd Edition*: John Wiley & Sons, Inc.

REVISION: Crippa A, Orsini N. 2016. Dose-response meta-analysis of differences in means. *BMC Med. Res. Methodol.* 16(1): 91.

REVISION: Crippa A, Discacciati A, Bottai M, Spiegelman D, Orsini N. 2019. One-stage dose-response meta-analysis for aggregated data. *Stat. Methods Med. Res.* 28(5): 1579-1596.

REVISION: Cui Y, Zhang B, Ma J, Wang Y, Zhao L, Hou C, Yu J, Zhao Y, Zhang Z, Nie J, Gao T, Zhou G, Liu H. 2018. Dopamine receptor D2 gene polymorphism, urine fluoride, and intelligence impairment of children in China: A school-based cross-sectional study. *Ecotoxicol. Environ. Saf.* 165: 270-277.

REVISION: DerSimonian R, Laird N. 1986. Meta-analysis in clinical trials. *Control. Clin. Trials* 7(3): 177-188.

REVISION: Ding Y, Sun H, Han H, Wang W, Ji X, Liu X, Sun D. 2011. The relationships between low levels of urine fluoride on children's intelligence, dental fluorosis in endemic fluorosis areas in Hulunbuir, Inner Mongolia, China. *J. Hazard. Mater.* 186: 1942-1946.

- REVISION: Discacciati A, Crippa A, Orsini N. 2017. Goodness of fit tools for dose–response meta-analysis of binary outcomes. *Res Synth Methods* 8(2): 149-160.**
- REVISION: Duan Q, Jiao J, Chen X, Wang X. 2018. Association between water fluoride and the level of children's intelligence: A dose–response meta-analysis. *Public Health* 154: 87-97.**
- REVISION: Duval S, Tweedie R. 2000a. Trim and fill: A simple funnel-plot–based method of testing and adjusting for publication bias in meta-analysis. *Biometrics* 56(2): 455-463.**
- REVISION: Duval S, Tweedie R. 2000b. A nonparametric “trim and fill” method of accounting for publication bias in meta-analysis. *J Am Stat Assoc* 95(449): 89-98.**
- REVISION: Egger M, Davey Smith G, Schneider M, Minder C. 1997. Bias in meta-analysis detected by a simple, graphical test. *BMJ* 315(7109): 629-634.**
- REVISION: Filippini T, Hatch Elizabeth E, Rothman Kenneth J, Heck Julia E, Park Andrew S, Crippa A, Orsini N, Vinceti M. 2019. Association between outdoor air pollution and childhood leukemia: A systematic review and dose–response meta-analysis. *Environ. Health Perspect.* 127(4): 046002.**
- REVISION: 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.**
- REVISION: Grandjean P, Landrigan PJ. 2006. Developmental neurotoxicity of industrial chemicals. *Lancet* 368(9553): 2167-2178.**
- REVISION: Grandjean P. 2019. Developmental fluoride neurotoxicity: An updated review. *Environ. Health* 18(1): 110.**
- REVISION: Green R, Lanphear B, Hornung R, Flora D, Martinez-Mier EA, Neufeld R, Ayotte P, Muckle G, Till C. 2019. Association between maternal fluoride exposure during pregnancy and IQ scores in offspring in Canada. *JAMA Pediatr* 173(10): 940-948.**
- REVISION: Greenland S, Longnecker MP. 1992. Methods for trend estimation from summarized dose-response data, with applications to meta-analysis. *Am. J. Epidemiol.* 135(11): 1301-1309.**
- 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. *J. Clin. Epidemiol.* 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. *J. Clin. Epidemiol.* 64(12): 1283-1293.

Protocol for Systematic Review of Effects of Fluoride Exposure on Neurodevelopment

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. 2011c. GRADE guidelines: 5. Rating the quality of evidence--publication bias. *J. Clin. Epidemiol.* 64(12): 1277-1282.

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. 2011d. GRADE guidelines: 4. Rating the quality of evidence--study limitations (risk of bias). *J. Clin. Epidemiol.* 64(4): 407-415.

Haseman JK, Bailer AJ, Kodell RL, Morris R, Portier K. 2001. Statistical issues in the analysis of low-dose endocrine disruptor data. *Toxicol Sci* 61(2): 201-210.

REVISION: Hedges LV, Olkin I. 1985. *Statistical methods for meta-analysis*, San Diego, CA: Academic Press.

Higgins J, Green S. 2011a. *Cochrane handbook for systematic reviews of interventions*, In: The Cochrane Collaboration. Vol 4, New York, NY: John Wiley & Sons.

Higgins J, Green S. 2011b. *Cochrane handbook for systematic reviews of interventions*, In: The Cochrane Collaboration. Version 5.1.0 (updated March 2011). Available: <http://handbook.cochrane.org/> [accessed 3 February 2013].

Higgins JP, Altman DG, Gotzsche PC, Juni P, Moher D, Oxman AD, Savovic J, Schulz KF, Weeks L, Sterne JA. 2011. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *BMJ* 343: d5928.

REVISION: Higgins JPT, Thompson SG, Deeks JJ, Altman DG. 2003. Measuring inconsistency in meta-analyses. *BMJ* 327(7414): 557-560.

REVISION: Higgins JT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA, eds. 2019. *Cochrane Handbook for systematic reviews of interventions version 6.0 (updated July 2019)*. Cochrane.

Hooijmans CR, Rovers MM, de Vries RB, Leenaars M, Ritskes-Hoitinga M, Langendam MW. 2014. SYRCLE's risk of bias tool for animal studies. *BMC Med. Res. Methodol.* 14: 43.

REVISION: Howard BE, Phillips J, Tandon A, Maharana A, Elmore R, Mav D, Sedykh A, Thayer K, Merrick BA, Walker V, Rooney A, Shah RR. 2020. SWIFT-Active Screener: Accelerated document screening through active learning and integrated recall estimation. *Environ. Int.* 138: 105623.

REVISION: Jackson D, White IR, Thompson SG. 2010. Extending DerSimonian and Laird's methodology to perform multivariate random effects meta-analyses. *Stat. Med.* 29(12): 1282-1297.

Krauth D, Woodruff TJ, Bero L. 2013. Instruments for assessing risk of bias and other methodological criteria of published animal studies: a systematic review. *Environ. Health Perspect.* 121(9): 985-992.

REVISION: Larsson SC, Orsini N. 2018. Coffee consumption and risk of dementia and Alzheimer's disease: A dose-response meta-analysis of prospective studies. *Nutrients* 10(10): 1501.

Miller K HB, Phillips J, Shah M, Mav D, Thayer K, Shah R. 2016. SWIFT-Active Screener: Reducing literature screening effort through machine learning for systematic reviews. Cochrane Colloquium Seoul, Seoul, Korea.

Moher D, Liberati A, Tetzlaff J, Altman DG. 2009. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *J. Clin. Epidemiol.* 62(10): 1006-1012.

REVISION: Orsini N, Bellocco R, Greenland S. 2006. Generalized least squares for trend estimation of summarized dose-response data. *Stat J* 6(1): 40-57.

REVISION: Orsini N, Li R, Wolk A, Khudyakov P, Spiegelman D. 2012. Meta-analysis for linear and nonlinear dose-response relations: Examples, an evaluation of approximations, and software. *Am. J. Epidemiol.* 175(1): 66-73.

REVISION: Palmer TM, Sterne JAC, eds. 2016. *Meta-analysis in Stata: An updated collection from the Stata Journal*. 2nd ed. College Station, TX: Stata Press.

Pildal J, Hrobjartsson A, Jorgensen KJ, Hilden J, Altman DG, Gotzsche PC. 2007. Impact of allocation concealment on conclusions drawn from meta-analyses of randomized trials. *Int. J. Epidemiol.* 36(4): 847-857.

REVISION: Podgorski J, Berg M. 2020. Global threat of arsenic in groundwater. *Science* 368(6493): 845-850.

REVISION: Rocha-Amador D, Navarro ME, Carrizales L, Morales R, Calderón J. 2007. Decreased intelligence in children and exposure to fluoride and arsenic in drinking water. *Cad. Saude Publica* 23: S579-S587.

Rooney AA, Boyles AL, Wolfe MS, Bucher JR, Thayer KA. 2014. Systematic review and evidence integration for literature-based environmental health science assessments. *Environ. Health Perspect.* 122(7): 711-718.

REVISION: Rosenthal R. 1994. Parametric measures of effect size. In: *The Handbook of Research Synthesis*. Cooper H, Hedges LV, eds. New York, NY, Russell Sage Foundation.

REVISION: Rota M, Bellocco R, Scotti L, Tramacere I, Jenab M, Corrao G, La Vecchia C, Boffetta P, Bagnardi V. 2010. Random-effects meta-regression models for studying nonlinear dose-response relationship, with an application to alcohol and esophageal squamous cell carcinoma. *Stat. Med.* 29(26): 2679-2687.

REVISION: Russo GS, Eftim SE, Goldstone AE, Dufour AP, Nappier SP, Wade TJ. 2020. Evaluating health risks associated with exposure to ambient surface waters during recreational activities: A systematic review and meta-analysis. *Water Res.* 176: 115729.

Sabanathan S, Wills B, Gladstone M. 2015. Child development assessment tools in low-income and middle-income countries: how can we use them more appropriately? *Arch Dis Child* 100(5): 482-488.

REVISION: Saxena S, Sahay A, Goel P. 2012. Effect of fluoride exposure on the intelligence of school children in Madhya Pradesh, India. *J. Neurosci. Rural Pract.* 3(2): 144-149.

Schulz KF, Chalmers I, Hayes RJ, Altman DG. 1995. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA* 273(5): 408-412.

REVISION: Sebastian ST, Sunitha S. 2015. A cross-sectional study to assess the intelligence quotient (IQ) of school going children aged 10-12 years in villages of Mysore District, India with different fluoride levels. *J Indian Soc Pedod Prev Dent* 33: 307-311.

Sena E, van der Worp HB, Howells D, Macleod M. 2007. How can we improve the pre-clinical development of drugs for stroke? *Trends Neurosci.* 30(9): 433-439.

REVISION: Seraj B, Shahrabi M, Falahzade M, Falahzade F, Akhondi N. 2006. Effect of high fluoride concentration in drinking water on children's intelligence. *J. Dent. Med.* 19(2): 80-86.

REVISION: Seraj B, Shahrabi M, Shadfar M, Ahmadi R, Fallahzadeh M, Eslamlu HF, Kharazifard MJ. 2012. Effect of high water fluoride concentration on the intellectual development of children in Makoo, Iran. *J Dent (Tehran)* 9(3): 221-229.

REVISION: StataCorp. 2019. *Stata Statistical Software: Release 16*, College Station, TX: StataCorp LP.

Sterne J, Higgins J, Reeves B, on behalf of the development group for ACROBAT-NRSI. 2014. *A Cochrane Risk Of Bias Assessment Tool: for Non-Randomized Studies of Interventions (ACROBAT-NRSI), Version 1.0.0*. Available: <http://www.riskofbias.info> [accessed 28 September 2014].

Sutton M, Kiersey R, Farragher L, Long J. 2015. Health effects of water fluoridation: An evidence review. Report conducted for Republic of Ireland's Department of Health. Available: http://www.hrb.ie/uploads/tx_hrbpublications/Health_Effects_of_Water_Fluoridation.pdf [accessed 9 November, 2015].

Szklo M, F, Nieto J, Reilly T, Parke K. 2014. *Epidemiology: Beyond the basics, 4th edition*: Jones & Bartlett Learning.

REVISION: Till C, Green R, Flora D, Hornung R, Martinez-Mier EA, Blazer M, Farmus L, Ayotte P, Muckle G, Lanphear B. 2020. Fluoride exposure from infant formula and child IQ in a Canadian birth cohort. *Environ. Int.* 134: 105315.

Vesterinen HM, Sena ES, French-Constant C, Williams A, Chandran S, Macleod MR. 2010. Improving the translational hit of experimental treatments in multiple sclerosis. *Mult. Scler.* 16(9): 1044-1055.

Vesterinen HM, Sena ES, Egan KJ, Hirst TC, Churolov L, Currie GL, Antonic A, Howells DW, Macleod MR. 2014. Meta-analysis of data from animal studies: a practical guide. *J. Neurosci. Methods* 221: 92-102.

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: 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].

REVISION: Vlaanderen J, Portengen L, Rothman N, Lan Q, Kromhout H, Vermeulen R. 2010. Flexible meta-regression to assess the shape of the benzene–leukemia exposure–response curve. *Environ. Health Perspect.* 118(4): 526-532.

von Hilsheimer G, Kurko V. 1979. Minor physical anomalies in exceptional children. *J. Learn. Disabil.* 12(7): 462-469.

REVISION: Wang M, Liu L, Li H, Li Y, Liu H, Hou C, Zeng Q, Li P, Zhao Q, Dong L, Zhou G, Yu X, Liu L, Guan Q, Zhang S, Wang A. 2020. Thyroid function, intelligence, and low-moderate fluoride exposure among Chinese school-age children. *Environ. Int.* 134: 105229.

Wood L, Egger M, Gluud LL, Schulz KF, Juni P, Altman DG, Gluud C, Martin RM, Wood AJ, Sterne JA. 2008. Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: Meta-epidemiological study. *BMJ* 336(7644): 601-605.

REVISION: Xiang Q, Liang Y, Chen L, Wang C, Chen B, Chen X, Zhou M. 2003. Effect of fluoride in drinking water on children's intelligence. *Fluoride* 36: 84-94.

REVISION: Xiang Q, Liang Y, Chen B, Chen L. 2011. Analysis of children's serum fluoride levels in relation to intelligence scores in a high and low fluoride water village in China. *Fluoride* 44(4): 191.

REVISION: Yu X, Chen J, Li Y, Liu H, Hou C, Zeng Q, Cui Y, Zhao L, Li P, Zhou Z, Pang S, Tang S, Tian K, Zhao Q, Dong L, Xu C, Zhang X, Zhang S, Liu L, Wang A. 2018. Threshold effects of moderately excessive fluoride exposure on children's health: A potential association between dental fluorosis and loss of excellent intelligence. *Environ. Int.* 118: 116-124.

REVISION: Zhang S, Zhang X, Liu H, Qu W, Guan Z, Zeng Q, Jiang C, Gao H, Zhang C, Lei R, Xia T, Wang Z, Yang L, Chen Y, Wu X, Cui Y, Yu L, Wang A. 2015. Modifying effect of COMT gene polymorphism and a predictive role for proteomics analysis in children's intelligence in endemic fluorosis area in Tianjin, China. *Toxicol Sci* 144: 238-245.

ABOUT THE PROTOCOL

Contributors

Evaluation Team

Evaluation teams are composed of federal staff and contractor staff. Contractor staff members are screened for potential conflicts of interest. Federal staff members should do a self-evaluation. Epidemiologists and toxicologists on OHAT evaluation teams should have at least three years' experience and/or training in reviewing studies, including summarizing studies and critical review (e.g., assessing study quality and interpreting findings). Experience in evaluating occupational or environmental studies is preferred. Team members should have at least a master's degree or equivalent level of experience in epidemiology, toxicology, environmental health sciences, or a related field.

Name	Affiliation
Kyla Taylor, PhD	NIEHS/NTP, Project Lead
John Bucher, PhD	NIEHS/NTP
Andrew Rooney, PhD	NIEHS/NTP
Vickie Walker	NIEHS/NTP

Contract support

Contractors listed below are anticipated to provide support necessary to complete the literature searches, study selection, data extraction, and risk of bias assessment

Name	Affiliation
Anna Engstrom, PhD	ICF
Robyn Blain, PhD	ICF
Pamela Hartman	ICF
William Mendez, PhD	ICF
Cara Henning, PhD	ICF
Johanna Rochester, PhD	ICF
Kristin Bornstein, PhD	ICF
Ali Goldstone	ICF
Camden Byrd	ICF
Anna Stamatogiannakis	ICF
Whitney Mitchell	ICF
Penelope Kellar	ICF
REVISION: Christopher Sibrizzi	ICF

Protocol for Systematic Review of Effects of Fluoride Exposure on Neurodevelopment

Name	Affiliation
Kristen Magnuson	ICF
Jeremy Frye	ICF

Technical Advisors

Technical advisors are outside experts retained on an as-needed basis to provide individual advice to the NTP for a specific topic. Technical advisors selected for this project were selected for their experience with fluoride exposure and neurotoxicity. Technical advisors were screened for conflict of interest prior to their service. Service as a technical advisor does not necessarily indicate that the advisor has read the entire protocol or endorses the final state-of-the-science document.

Name	Affiliation
Joseph Braun, PhD	Brown University
Marie Sutton, PhD	Health Research Board
Thomas Zoeller, PhD	University of Massachusetts, Amherst
Thomas Webster, PhD	Boston University
Gail Wasserman, PhD	Columbia University
REVISION: Suril Mehta, MPH	NIEHS/NTP
Tianjing Li, PhD	University of Colorado Denver

Sources of Support

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

Protocol History and Revisions

Date	Activity or revision
December 14, 2016	Draft evaluation protocol reviewed: sent to technical advisors for comment/review
April 10, 2017	Draft human risk of bias protocol reviewed; sent to technical advisors for comment/review
May 2, 2017	Draft animal risk of bias protocol reviewed; sent to technical advisors for comment/review
June 2017	Draft finalized
May 29, 2019	Revised protocol posted to reflect principal updates made during evaluation with justifications noted (date implemented): <ol style="list-style-type: none"> 1. Revision of terminology used throughout the protocol from “neurobehavioral effects” to “neurodevelopmental and cognitive effects” to better reflect the focus of the evaluation on neurodevelopmental effects in children and cognitive effects in adults (February 2019); 2. Removal of aluminum fluoride and aluminum fluoride complexes from the PECO statement because these were not specifically considered in

Protocol for Systematic Review of Effects of Fluoride Exposure on Neurodevelopment

Date	Activity or revision
	<p>the evaluation due to the complexity of the aluminum and fluoride interaction (September 2018);</p> <ol style="list-style-type: none"> 3. Due to the large number of excluded studies based on SWIFT Active exclusion during title and abstract screening, decision that only the list of included studies (i.e., studies that comply with the PECO criteria for inclusion) will be posted at the NTP website for this project (May 2019); 4. Updating of Figure 3 to reflect a clarification and update to the OHAT approach for systematic review and evidence integration which clarified how hazard conclusions are reached when there is a moderate level-of-evidence for human data with low or inadequate level of evidence for the animal evidence stream (March 2019); and 5. Updating of text in results section to reflect that results of animal studies would be included in the monograph. Note that this is not a procedural change to the evaluation but an edit to the protocol text to better reflect NTP's evaluation process (May 2019).
September 15, 2020	<p>In the November 2019 review of the September 6, 2019 draft monograph (NASEM 2020), NASEM recommended that a meta-analysis be conducted and the protocol be updated to clarify a number of issues. The protocol was revised in response to those recommendations with the principal updates as follows:</p> <ol style="list-style-type: none"> 1. A footnote was added to clarify the role of the OHAT handbook in developing systematic review protocols; 2. Additional information was provided on problem formulation steps including opportunities for and consideration of public comments; 3. A supplemental search strategy was included for non-English-language databases to systematically search for studies that were previously identified from other resources; 4. A summary of a SWIFT-Active Screener validation study was added (Howard <i>et al.</i> 2020); 5. Text was updated to clarify risk-of-bias criteria for determining risk of bias from improper adjustment for confounders, providing clarification of key covariates for all study populations, further consideration of co-exposures, and further details describing rationales for distinguishing between different risk-of-bias ratings; 6. A separate protocol was developed to guide the conduct of meta-analyses of IQ studies in children. The meta-analysis protocol was peer-reviewed, revised, and added as Appendix 6; 7. Text was updated to clarify and further describe the training of the evaluation team (including contractor staff) and NTP's role in supervision/oversight of screening and data extraction; and 8. A table outlining inclusion and exclusion criteria to determine study eligibility (Table 2) was added.

APPENDICES

Appendix 1. Electronic Database Search Strategies

REVISION: Main literature databases search

BIOSIS

Date of search: 11/28/2016; 6,743 results REVISION: date and results of original search listed only

BIOSIS search terms		
#1	Fluoride	TOPIC: ((fluorid* OR flurid* OR fluorin* OR florin* OR fluorosis) NOT (f-labeled OR "fluorine-18" OR radioligand* OR 18F OR F-18 OR "fluorine-18" OR 19F OR F-19 OR "fluorine-19" OR (PET AND scan)))
#2	Neurological and Thyroid outcomes	TOPIC: (Academic-performance OR active-avoidance OR ADHD OR alzheimer* OR amygdala OR antisocial OR anxiety OR anxious OR asperger* OR attention-deficit OR auditory OR autism OR autistic OR behavioral OR behaviors OR behavioural OR behaviours OR bipolar OR cerebellum OR cognition OR cognitive OR communication-disorder* OR comprehension OR cortical OR cranial OR delayed-development OR dementia OR dendrit* OR dentate-gyrus OR depression OR developmental-impairment OR Developmental-delay* OR developmental-disorder* OR dextrothyroxine OR diiodothyronine* OR diiodotyrosine OR down-syndrome OR dyslexia OR entorhinal-cortex OR epilep* OR euthyroid OR gait OR gangli* OR glia* OR gliogenesis OR goiter OR graves-disease OR hearing OR hippocamp* OR human-development OR hyperactiv* OR hyperthyroid* OR hypothalam* OR hypothyroid* OR impulse-control OR impulsiv* OR Intellectual-disability OR intelligence OR iodide-peroxidase OR IQ OR ischemi* OR language OR learning OR lewy-bod* OR locomotor OR long-term-potential OR long-term-synaptic-depression OR memory OR mental-deficiency OR mental-development OR mental-disorder* OR mental-illness OR mental-recall OR mental-deficit OR mobility OR monoiodotyrosine OR mood OR morris-maze OR morris-water OR motor-abilit* OR Motor-activities OR Motor-activity OR Motor-performance OR Motor-skill* OR Multiple-sclerosis OR myxedema OR nerve OR Nervous-system OR neural OR neurit* OR neurobehav* OR Neurocognitive-impairment OR neurodegenerat* OR Neurodevelopment* OR neurodisease* OR neurologic* OR neuromuscular OR neuron* OR neuropath* OR obsessive-compulsive OR OCD OR olfaction OR olfactory OR open-field-test OR optic OR palsy OR panic OR parahippocamp* OR paranoia OR paranoid OR parkinson* OR passive-avoidance OR perception OR perforant* OR personality OR phobia OR plasticity OR problem-solving OR proprioception OR psychomotor OR reflex OR risk-taking OR schizophrenia OR seizure* OR senil* OR sensation* OR sleep OR smell OR sociab* OR spatial-behavior OR speech* OR spelling OR stereotypic-movement* OR stroke OR substantia-nigra OR synap* OR taste OR tauopath* OR Thyroglobulin OR Thyroid-disease* OR Thyroid-gland OR Thyroid-hormone* OR thyroiditis OR thyronine* OR thyrotoxicosis OR

Protocol for Systematic Review of Effects of Fluoride Exposure on Neurodevelopment

BIOSIS search terms		
		Thyrotropin OR thyroxine OR triiodothyronine OR vision OR visual-motor OR Visuospatial-processing OR Water-maze)
#3	Final Search	#1 AND #2 Indexes=BCI, Timespan=All years Refined by: RESEARCH AREAS: (BIOCHEMISTRY MOLECULAR BIOLOGY OR NEUROSCIENCES NEUROLOGY OR ENDOCRINOLOGY METABOLISM OR PHARMACOLOGY PHARMACY OR TOXICOLOGY OR CELL BIOLOGY OR PHYSIOLOGY OR PUBLIC ENVIRONMENTAL OCCUPATIONAL HEALTH OR DENTISTRY ORAL SURGERY MEDICINE OR ENVIRONMENTAL SCIENCES ECOLOGY OR DEVELOPMENTAL BIOLOGY OR UROLOGY NEPHROLOGY OR PEDIATRICS OR LIFE SCIENCES BIOMEDICINE OTHER TOPICS OR GENETICS HEREDITY OR PSYCHIATRY OR REPRODUCTIVE BIOLOGY OR PATHOLOGY)

Protocol for Systematic Review of Effects of Fluoride Exposure on Neurodevelopment

EMBASE

Date of search: 11/28/2016; 9,426 results REVISION: date and results of original search listed only

EMBASE search terms		
#1	Fluoride	((Fluoride/exp OR fluorid* OR flurid* OR fluorin* OR florin* OR fluorosis OR fluorosis/exp OR 'fluorosis, dental'/exp) NOT (f-labeled OR "fluorine-18" OR radioligand* OR 18F OR F-18 OR "fluorine-18" OR 19F OR F-19 OR "fluorine-19" OR (PET AND scan)))
#2	Neurological and Thyroid outcomes	('Academic performance':ti,ab OR 'active-avoidance':ti,ab OR 'ADHD':ti,ab OR 'alzheimer*':ti,ab OR 'amygdala':ti,ab OR 'antisocial':ti,ab OR 'anxiety':ti,ab OR 'anxious':ti,ab OR 'asperger*':ti,ab OR 'attention deficit':ti,ab OR 'auditory':ti,ab OR 'autism':ti,ab OR 'autistic':ti,ab OR 'behavioral':ti,ab OR 'behaviors':ti,ab OR 'behavioural':ti,ab OR 'behaviours':ti,ab OR 'bipolar':ti,ab OR 'cerebellum':ti,ab OR 'cognition':ti,ab OR 'cognitive':ti,ab OR 'communication-disorder*':ti,ab OR 'comprehension':ti,ab OR 'cortical':ti,ab OR 'cranial':ti,ab OR 'delayed development':ti,ab OR 'dementia':ti,ab OR 'dendrit*':ti,ab OR 'dentate-gyrus':ti,ab OR 'depression':ti,ab OR 'developmental impairment':ti,ab OR 'developmental-delay*':ti,ab OR 'developmental-disorder*':ti,ab OR 'dextrothyroxine':ti,ab OR 'diiodothyronine*':ti,ab OR 'diiodotyrosine':ti,ab OR 'down syndrome':ti,ab OR 'dyslexia':ti,ab OR 'entorhinal cortex':ti,ab OR 'epilep*':ti,ab OR 'euthyroid':ti,ab OR 'gait':ti,ab OR 'gangli*':ti,ab OR 'glia*':ti,ab OR 'gliogenesis':ti,ab OR 'goiter':ti,ab OR 'graves-disease':ti,ab OR 'hearing':ti,ab OR 'hippocamp*':ti,ab OR 'human development':ti,ab OR 'hyperactiv*':ti,ab OR 'hyperthyroid*':ti,ab OR 'hypothalam*':ti,ab OR 'hypothyroid*':ti,ab OR 'impulse-control':ti,ab OR 'impulsiv*':ti,ab OR 'Intellectual disability':ti,ab OR 'intelligence':ti,ab OR 'iodide peroxidase':ti,ab OR 'IQ':ti,ab OR 'ischemi*':ti,ab OR 'language':ti,ab OR 'learning':ti,ab OR 'lewy bod*':ti,ab OR 'locomotor':ti,ab OR 'long-term potentiation':ti,ab OR 'long-term synaptic depression':ti,ab OR 'memory':ti,ab OR 'mental deficiency':ti,ab OR 'mental development':ti,ab OR 'mental disorder*':ti,ab OR 'mental illness':ti,ab OR 'mental recall':ti,ab OR 'mental-deficit':ti,ab OR 'mobility':ti,ab OR 'monoiodotyrosine':ti,ab OR 'mood':ti,ab OR 'morris-maze':ti,ab OR 'morris-water':ti,ab OR 'motor abilit*':ti,ab OR 'Motor activities':ti,ab OR 'Motor activity':ti,ab OR 'motor performance':ti,ab OR 'motor skill*':ti,ab OR 'multiple sclerosis':ti,ab OR 'myxedema':ti,ab OR 'nerve':ti,ab OR 'Nervous system':ti,ab OR 'nervous-system':ti,ab OR 'neural':ti,ab OR 'neurit*':ti,ab OR 'neurobehav*':ti,ab OR 'Neurocognitive impairment':ti,ab OR 'neurodegenerat*':ti,ab OR 'Neurodevelopment*':ti,ab OR 'neurodisease*':ti,ab OR 'neurologic*':ti,ab OR 'neuromuscular':ti,ab OR 'neuron*':ti,ab OR 'neuropath*':ti,ab OR 'obsessive compulsive':ti,ab OR 'OCD':ti,ab OR 'olfaction':ti,ab OR 'olfactory':ti,ab OR 'open-field-test':ti,ab OR 'optic':ti,ab OR 'palsy':ti,ab OR 'panic':ti,ab OR 'parahippocamp*':ti,ab OR 'paranoia':ti,ab OR 'paranoid':ti,ab OR 'parkinson*':ti,ab OR 'passive avoidance':ti,ab OR 'perception':ti,ab OR 'perforant*':ti,ab OR

EMBASE search terms	
	<p>'personality':ti,ab OR 'phobia':ti,ab OR 'plasticity':ti,ab OR 'problem solving':ti,ab OR 'proprioception':ti,ab OR 'psychomotor':ti,ab OR 'reflex':ti,ab OR 'risk taking':ti,ab OR 'schizophrenia':ti,ab OR 'seizure*':ti,ab OR 'senil*':ti,ab OR 'sensation*':ti,ab OR 'sleep':ti,ab OR 'smell':ti,ab OR 'sociab*':ti,ab OR 'spatial behavior':ti,ab OR 'speech*':ti,ab OR 'spelling':ti,ab OR 'stereotypic-movement*':ti,ab OR 'stroke':ti,ab OR 'substantia-nigra':ti,ab OR 'synap*':ti,ab OR 'taste':ti,ab OR 'tauopath*':ti,ab OR 'Thyroglobulin':ti,ab OR 'Thyroid disease*':ti,ab OR 'thyroid gland':ti,ab OR 'thyroid hormone*':ti,ab OR 'thyroiditis':ti,ab OR 'thyronine*':ti,ab OR 'thyrotoxicosis':ti,ab OR 'Thyrotropin':ti,ab OR 'thyroxine':ti,ab OR 'triiodothyronine':ti,ab OR 'vision':ti,ab OR 'visual motor':ti,ab OR 'Visuospatial processing':ti,ab OR 'water maze':ti,ab OR 'Alzheimer disease'/exp OR 'amygdala'/exp OR 'antisocial behavior'/exp OR 'anxiety'/exp OR 'Asperger syndrome'/exp OR 'attention deficit disorder'/exp OR 'autism'/exp OR 'behavior'/exp OR 'behavior disorder'/exp OR 'bipolar disorder'/exp OR 'cognition'/exp OR 'cognitive defect'/exp OR 'communication disorder'/exp OR 'communication disorders'/exp OR 'comprehension'/exp OR 'Constitutive androstane receptor'/exp OR 'dementia'/exp OR 'depression'/exp OR 'developmental delay'/exp OR 'developmental disorder'/exp OR 'dextrothyroxine'/exp OR 'diiodothyronine'/exp OR 'diiodotyrosine'/exp OR 'disorders of higher cerebral function'/exp OR 'disruptive behavior'/exp OR 'dissociative disorder'/exp OR 'dyslexia'/exp OR 'gait'/exp OR 'gait disorder'/exp OR 'Glucuronosyltransferase'/exp OR 'goiter'/exp OR 'graves-disease'/exp OR 'hearing'/exp OR 'high risk behavior'/exp OR 'hyperthyroidism'/exp OR 'impulse control disorder'/exp OR 'impulsiveness'/exp OR 'Intellectual disability'/exp OR 'intelligence'/exp OR 'intelligence quotient'/exp OR 'iodide peroxidase'/exp OR 'ischemia'/exp OR 'language'/exp OR 'learning'/exp OR 'locomotion'/exp OR 'Malate Dehydrogenase'/exp OR 'malate dehydrogenase (decarboxylating)'/exp OR 'memory'/exp OR 'mental deficiency'/exp OR 'mental development'/exp OR 'mental disease'/exp OR 'mood'/exp OR 'Motor activity'/exp OR 'motor dysfunction'/exp OR 'motor performance'/exp OR 'myxedema'/exp OR 'nerve cell'/exp OR 'nerve cell differentiation'/exp OR 'Nervous system'/exp OR 'neurobehavioral'/exp OR 'neurodegeneration'/exp OR 'Neurogranin'/exp OR 'neurologic disease'/exp OR 'neuromuscular disease'/exp OR 'neuropathic'/exp OR 'neuropathology'/exp OR 'neuropathy'/exp OR 'neurotox'/exp OR 'obsessive compulsive disorder'/exp OR 'olfactory system'/exp OR 'panic'/exp OR 'paralysis'/exp OR 'paranoia'/exp OR 'Parkinson disease'/exp OR 'patient mobility'/exp OR 'perception'/exp OR 'perception disorder'/exp OR 'personality'/exp OR 'phobia'/exp OR 'Pregnane X Receptor'/exp OR 'proprioception'/exp OR 'psychomotor activity'/exp OR 'psychomotor disorder'/exp OR 'recall'/exp OR 'thyroid hormone receptor'/exp OR 'thyrotropin receptor'/exp OR 'reflex'/exp OR 'Retinoid X Receptor'/exp OR 'seizure'/exp OR 'senile dementia'/exp OR 'senility'/exp OR 'sensation'/exp OR 'sleep'/exp OR 'social behavior'/exp OR 'speech'/exp</p>

Protocol for Systematic Review of Effects of Fluoride Exposure on Neurodevelopment

EMBASE search terms		
		OR 'Spelling'/exp OR 'taste'/exp OR 'tauopathy'/exp OR 'Thyroglobulin'/exp OR 'Thyroid disease'/exp OR 'thyroid diseases'/exp OR 'thyroid gland'/exp OR 'thyroid hormone'/exp OR 'thyroiditis'/exp OR 'thyronine'/exp OR 'thyrotoxicosis'/exp OR 'Thyrotropin'/exp OR 'thyroxine'/exp OR 'vision'/exp OR 'visuomotor coordination'/exp)
#3	Final Search	#1 AND #2 Embase OR Embase Classic

PsycINFO

Date of search: 11/28/2016; 181 results REVISION: date and results of original search listed only

PsycINFO search terms		
#1	Fluoride	Title OR Abstract: ((fluorid* OR flurid* OR fluorin* OR florin* OR fluorosis) NOT (f-labeled OR "fluorine-18" OR radioligand* OR 18F OR F-18 OR "fluorine-18" OR 19F OR F-19 OR "fluorine-19" OR (PET AND scan)))
#2	Neurological and Thyroid outcomes	Title OR Abstract: (Academic-performance OR active-avoidance OR ADHD OR alzheimer* OR amygdala OR antisocial OR anxiety OR anxious OR asperger* OR attention-deficit OR auditory OR autism OR autistic OR behavioral OR behaviors OR behavioural OR behaviours OR bipolar OR cerebellum OR cognition OR cognitive OR communication-disorder* OR comprehension OR cortical OR cranial OR delayed-development OR dementia OR dendrit* OR dentate-gyrus OR depression OR developmental-impairment OR Developmental-delay* OR developmental-disorder* OR dextrothyroxine OR diiodothyronine* OR diiodotyrosine OR down-syndrome OR dyslexia OR entorhinal-cortex OR epilep* OR euthyroid OR gait OR gangli* OR glia* OR gliogenesis OR goiter OR graves-disease OR hearing OR hippocamp* OR human-development OR hyperactiv* OR hyperthyroid* OR hypothalam* OR hypothyroid* OR impulse-control OR impulsiv* OR Intellectual-disability OR intelligence OR iodide-peroxidase OR IQ OR ischemi* OR language OR learning OR lewy-bod* OR locomotor OR long-term-potential OR long-term-synaptic-depression OR memory OR mental-deficiency OR mental-development OR mental-disorder* OR mental-illness OR mental-recall OR mental-deficit OR mobility OR monoiodotyrosine OR mood OR morris-maze OR morris-water OR motor-abilit* OR Motor-activities OR Motor-activity OR Motor-performance OR Motor-skill* OR Multiple-sclerosis OR myxedema OR nerve OR Nervous-system OR neural OR neurit* OR neurobehav* OR Neurocognitive-impairment OR neurodegenerat* OR Neurodevelopment* OR neurodisease* OR neurologic* OR neuromuscular OR neuron* OR neuropath* OR obsessive-compulsive OR OCD OR olfaction OR olfactory OR open-field-test OR optic OR palsy OR panic OR parahippocamp* OR paranoia OR paranoid OR parkinson* OR passive-avoidance OR perception OR perforant* OR personality OR phobia OR plasticity OR problem-solving OR proprioception OR psychomotor OR reflex OR risk-taking OR schizophrenia OR seizure* OR senil* OR sensation* OR sleep OR smell OR

Protocol for Systematic Review of Effects of Fluoride Exposure on Neurodevelopment

PsycINFO search terms		
		sociab* OR spatial-behavior OR speech* OR spelling OR stereotypic-movement* OR stroke OR substantia-nigra OR synap* OR taste OR tauopath* OR Thyroglobulin OR Thyroid-disease* OR Thyroid-gland OR Thyroid-hormone* OR thyroiditis OR thyronine* OR thyrotoxicosis OR Thyrotropin OR thyroxine OR triiodothyronine OR vision OR visual-motor OR Visuospatial-processing OR Water-maze)
#3	Final Search	#1 OR #2

PubMed

Date of search: 12/19/2016; 7,264 results REVISION: date and results of original search listed only

PubMed search terms		
#1	Fluoride	((Fluorides[mh:noexp] OR fluorides, topical[mh] OR sodium fluoride[mh] OR Fluorosis, Dental[mh] OR fluorosis[tiab] OR fluorid*[tiab] OR flurid*[tiab] OR fluorin*[tiab] OR florin*[tiab]) NOT (18F[tiab] OR f-18[tiab] OR 19F[tiab] OR f-19[tiab] OR f-labeled[tiab] OR "fluorine-18"[tiab] OR "fluorine-19"[tiab] OR pet-scan[tiab] OR radioligand*[tiab]))
#2	Neurological and Thyroid outcomes	AND ((Aryl Hydrocarbon Hydroxylases[mh] OR Aryl Hydrocarbon Receptor Nuclear Translocator[mh] OR Behavior and Behavior Mechanisms[mh] OR Gene Expression Regulation[mh] OR Glucuronosyltransferase[mh] OR Intelligence tests[mh] OR Malate Dehydrogenase[mh] OR Mediator Complex Subunit 1[mh] OR Mental disorders[mh] OR Mental processes[mh] OR Monocarboxylic Acid Transporters[mh] OR Myelin Basic Protein[mh] OR nervous system[mh] OR nervous system diseases[mh] OR nervous system physiological phenomena[mh] OR Neurogranin[mh] OR Oligodendroglia[mh] OR Peroxisome Proliferator-Activated Receptors[mh] OR Psychological Phenomena and Processes[mh] OR Receptors, thyroid hormone[mh] OR Receptors, thyrotropin[mh] OR Retinoid X Receptors[mh] OR thyroid diseases[mh] OR thyroid hormones[mh] OR Thyrotropin-releasing hormone[mh] OR Thyroxine-Binding Proteins[mh] OR Pregnane X Receptor[supplementary concept] OR thyroid-hormone-receptor interacting protein[supplementary concept] OR Constitutive androstane receptor[supplementary concept] OR Academic performance[tiab] OR auditory[tiab] OR cortical[tiab] OR delayed development[tiab] OR developmental impairment[tiab] OR developmental-delay*[tiab] OR developmental-disorder*[tiab] OR euthyroid[tiab] OR gait[tiab] OR glia*[tiab] OR gliogenesis[tiab] OR hyperactiv*[tiab] OR impulse-control[tiab] OR iodide peroxidase[tiab] OR IQ[tiab] OR ischemi*[tiab] OR locomotor[tiab] OR mental deficiency[tiab] OR mental development[tiab] OR mental illness[tiab] OR mental-deficit[tiab] OR mobility[tiab] OR mood[tiab] OR morris-maze[tiab] OR morris-water[tiab] OR motor abilit*[tiab] OR Motor activities[tiab] OR motor performance[tiab] OR nerve[tiab] OR neural[tiab] OR neurobehav*[tiab] OR Neurocognitive impairment[tiab] OR neurodegenerat*[tiab] OR

Protocol for Systematic Review of Effects of Fluoride Exposure on Neurodevelopment

		<p>Neurodevelopment*[tiab] OR neurodisease*[tiab] OR neurologic*[tiab] OR neuromuscular[tiab] OR neuron*[tiab] OR neuropath*[tiab] OR obsessive compulsive[tiab] OR OCD[tiab] OR olfaction[tiab] OR olfactory[tiab] OR open-field-test[tiab] OR passive avoidance[tiab] OR plasticity[tiab] OR senil*[tiab] OR sociab*[tiab] OR speech*[tiab] OR spelling[tiab] OR stereotypic-movement*[tiab] OR synap*[tiab] OR tauopath*[tiab] OR Thyroglobulin[tiab] OR Thyroid disease*[tiab] OR thyroid gland[tiab] OR thyroid hormone*[tiab] OR thyronine*[tiab] OR visual motor[tiab] OR Visuospatial processing[tiab] OR water maze[tiab]) OR ((active-avoidance[tiab] OR ADHD[tiab] OR alzheimer*[tiab] OR amygdala[tiab] OR antisocial[tiab] OR anxiety[tiab] OR anxious[tiab] OR asperger*[tiab] OR attention deficit[tiab] OR autism[tiab] OR autistic[tiab] OR behavioral[tiab] OR behaviors[tiab] OR behavioural[tiab] OR behaviours[tiab] OR bipolar[tiab] OR cerebellum[tiab] OR cognition[tiab] OR cognitive[tiab] OR communication-disorder*[tiab] OR comprehension[tiab] OR cranial[tiab] OR dementia[tiab] OR dendrit*[tiab] OR dentate-gyrus[tiab] OR depression[tiab] OR dextrothyroxine[tiab] OR diiodothyronine*[tiab] OR diiodotyrosine[tiab] OR down syndrome[tiab] OR dyslexia[tiab] OR entorhinal cortex[tiab] OR epilep*[tiab] OR gangli*[tiab] OR goiter[tiab] OR graves-disease[tiab] OR hearing[tiab] OR hippocamp*[tiab] OR human development[tiab] OR hyperthyroid*[tiab] OR hypothalam*[tiab] OR hypothyroid*[tiab] OR impulsiv*[tiab] OR Intellectual disability[tiab] OR intelligence[tiab] OR language[tiab] OR learning[tiab] OR lewy bod*[tiab] OR long-term potentiation[tiab] OR long-term synaptic depression[tiab] OR memory[tiab] OR mental disorder*[tiab] OR mental recall[tiab] OR monoiodotyrosine[tiab] OR Motor activity[tiab] OR motor skill*[tiab] OR multiple sclerosis[tiab] OR myxedema[tiab] OR Nervous system[tiab] OR nervous-system[tiab] OR neurit*[tiab] OR optic[tiab] OR palsy[tiab] OR panic[tiab] OR parahippocamp*[tiab] OR paranoia[tiab] OR paranoid[tiab] OR parkinson*[tiab] OR perception[tiab] OR perforant*[tiab] OR personality[tiab] OR phobia[tiab] OR problem solving[tiab] OR proprioception[tiab] OR psychomotor[tiab] OR reflex[tiab] OR risk taking[tiab] OR schizophrenia[tiab] OR seizure*[tiab] OR sensation*[tiab] OR sleep[tiab] OR smell[tiab] OR spatial behavior[tiab] OR stroke[tiab] OR substantia-nigra[tiab] OR taste[tiab] OR thyroiditis[tiab] OR thyrotoxicosis[tiab] OR Thyrotropin[tiab] OR thyroxine[tiab] OR triiodothyronine[tiab] OR vision[tiab])) NOT medline[<i>sb</i>]))</p>
#3	Final Search	#1 AND #2

Web of Science

Date of search: 11/28/2016; 3,336 results REVISION: date and results of original search listed only

Web of Science search terms		
#1	Fluoride	TOPIC: ((fluorid* OR flurid* OR fluorin* OR florin* OR fluorosis) NOT (f-labeled OR "fluorine-18" OR radioligand* OR 18F OR F-18 OR "fluorine-18" OR 19F OR F-19 OR "fluorine-19" OR (PET AND scan)))
#2	Neurological and Thyroid outcomes	TOPIC: (Academic-performance OR active-avoidance OR ADHD OR alzheimer* OR amygdala OR antisocial OR anxiety OR anxious OR asperger* OR attention-deficit OR auditory OR autism OR autistic OR behavioral OR behaviors OR behavioural OR behaviours OR bipolar OR cerebellum OR cognition OR cognitive OR communication-disorder* OR comprehension OR cortical OR cranial OR delayed-development OR dementia OR dendrit* OR dentate-gyrus OR depression OR developmental-impairment OR Developmental-delay* OR developmental-disorder* OR dextrothyroxine OR diiodothyronine* OR diiodotyrosine OR down-syndrome OR dyslexia OR entorhinal-cortex OR epilep* OR euthyroid OR gait OR gangli* OR glia* OR gliogenesis OR goiter OR graves-disease OR hearing OR hippocamp* OR human-development OR hyperactiv* OR hyperthyroid* OR hypothalam* OR hypothyroid* OR impulse-control OR impulsiv* OR Intellectual-disability OR intelligence OR iodide-peroxidase OR IQ OR ischemi* OR language OR learning OR lewy-bod* OR locomotor OR long-term-potential OR long-term-synaptic-depression OR memory OR mental-deficiency OR mental-development OR mental-disorder* OR mental-illness OR mental-recall OR mental-deficit OR mobility OR monoiodotyrosine OR mood OR morris-maze OR morris-water OR motor-abilit* OR Motor-activities OR Motor-activity OR Motor-performance OR Motor-skill* OR Multiple-sclerosis OR myxedema OR nerve OR Nervous-system OR neural OR neurit* OR neurobehav* OR Neurocognitive-impairment OR neurodegenerat* OR Neurodevelopment* OR neurodisease* OR neurologic* OR neuromuscular OR neuron* OR neuropath* OR obsessive-compulsive OR OCD OR olfaction OR olfactory OR open-field-test OR optic OR palsy OR panic OR parahippocamp* OR paranoia OR paranoid OR parkinson* OR passive-avoidance OR perception OR perforant* OR personality OR phobia OR plasticity OR problem-solving OR proprioception OR psychomotor OR reflex OR risk-taking OR schizophrenia OR seizure* OR senil* OR sensation* OR sleep OR smell OR sociab* OR spatial-behavior OR speech* OR spelling OR stereotypic-movement* OR stroke OR substantia-nigra OR synap* OR taste OR tauopath* OR Thyroglobulin OR Thyroid-disease* OR Thyroid-gland OR Thyroid-hormone* OR thyroiditis OR thyronine*)

Protocol for Systematic Review of Effects of Fluoride Exposure on Neurodevelopment

Web of Science search terms		
		OR thyrotoxicosis OR Thyrotropin OR thyroxine OR triiodothyronine OR vision OR visual-motor OR Visuospatial-processing OR Water-maze)
#3	Final Search	#1 AND #2 Indexes=SCI-EXPANDED, SSCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, CCR-EXPANDED, IC Timespan=All years Refined by: RESEARCH AREAS: (DEVELOPMENTAL BIOLOGY OR RESPIRATORY SYSTEM OR DENTISTRY ORAL SURGERY MEDICINE OR BIOCHEMISTRY MOLECULAR BIOLOGY OR PHARMACOLOGY PHARMACY OR LIFE SCIENCES BIOMEDICINE OTHER TOPICS OR ENVIRONMENTAL SCIENCES ECOLOGY OR TOXICOLOGY OR PUBLIC ENVIRONMENTAL OCCUPATIONAL HEALTH OR PSYCHIATRY OR PATHOLOGY OR NEUROSCIENCES NEUROLOGY OR BEHAVIORAL SCIENCES OR VETERINARY SCIENCES OR NUTRITION DIETETICS OR ENDOCRINOLOGY METABOLISM OR PSYCHOLOGY OR MARINE FRESHWATER BIOLOGY OR CELL BIOLOGY OR PHYSIOLOGY OR REPRODUCTIVE BIOLOGY OR PEDIATRICS)

SCOPUS

Date of search: 11/28/2016; 5,222 results REVISION: date and results of original search listed only

SCOPUS search terms		
#1	Fluoride	Title OR Abstract: ((fluorid* OR flurid* OR fluorin* OR florin* OR fluorosis) AND NOT (f-labeled OR "fluorine-18" OR radioligand* OR 18F OR F-18 OR "fluorine-18" OR 19F OR F-19 OR "fluorine-19" OR (PET AND scan)))
#2	Neurological and Thyroid outcomes	Title OR Abstract: (Academic-performance OR active-avoidance OR ADHD OR alzheimer* OR amygdala OR antisocial OR anxiety OR anxious OR asperger* OR attention-deficit OR auditory OR autism OR autistic OR behavioral OR behaviors OR behavioural OR behaviours OR bipolar OR cerebellum OR cognition OR cognitive OR communication-disorder* OR comprehension OR cortical OR cranial OR delayed-development OR dementia OR dendrit* OR dentate-gyrus OR depression OR developmental-impairment OR Developmental-delay* OR developmental-disorder* OR dextrothyroxine OR diiodothyronine* OR diiodotyrosine OR down-syndrome OR dyslexia OR entorhinal-cortex OR epilep* OR euthyroid OR gait OR gangli* OR glia* OR gliogenesis OR goiter OR graves-disease OR hearing OR hippocamp* OR human-development OR hyperactiv* OR hyperthyroid* OR hypothalam* OR hypothyroid* OR impulse-control OR impulsiv* OR Intellectual-disability OR intelligence OR iodide-peroxidase OR IQ OR ischemi* OR language OR learning OR lewy-bod* OR locomotor OR long-term-potential OR long-term-synaptic-depression OR memory

SCOPUS search terms		
		OR mental-deficiency OR mental-development OR mental-disorder* OR mental-illness OR mental-recall OR mental-deficit OR mobility OR monoiodotyrosine OR mood OR morris-maze OR morris-water OR motor-abilit* OR Motor-activities OR Motor-activity OR Motor-performance OR Motor-skill* OR Multiple-sclerosis OR myxedema OR nerve OR Nervous-system OR neural OR neurit* OR neurobehav* OR Neurocognitive-impairment OR neurodegenerat* OR Neurodevelopment* OR neurodisease* OR neurologic* OR neuromuscular OR neuron* OR neuropath* OR obsessive-compulsive OR OCD OR olfaction OR olfactory OR open-field-test OR optic OR palsy OR panic OR parahippocamp* OR paranoia OR paranoid OR parkinson* OR passive-avoidance OR perception OR perforant* OR personality OR phobia OR plasticity OR problem-solving OR proprioception OR psychomotor OR reflex OR risk-taking OR schizophrenia OR seizure* OR senil* OR sensation* OR sleep OR smell OR sociab* OR spatial-behavior OR speech* OR spelling OR stereotypic-movement* OR stroke OR substantia-nigra OR synap* OR taste OR tauopath* OR Thyroglobulin OR Thyroid-disease* OR Thyroid-gland OR Thyroid-hormone* OR thyroiditis OR thyronine* OR thyrotoxicosis OR Thyrotropin OR thyroxine OR triiodothyronine OR vision OR visual-motor OR Visuospatial-processing OR Water-maze)
#3	Final Search	#1 AND #2 AND (LIMIT-TO(SUBJAREA,"MEDI") OR LIMIT-TO(SUBJAREA,"BIOC") OR LIMIT-TO(SUBJAREA,"ENVI") OR LIMIT-TO(SUBJAREA,"PHAR") OR LIMIT-TO(SUBJAREA,"AGRI") OR LIMIT-TO(SUBJAREA,"NEUR") OR LIMIT-TO(SUBJAREA,"MULT") OR LIMIT-TO(SUBJAREA,"PSYC") OR LIMIT-TO(SUBJAREA,"Undefined"))

REVISION: *Supplemental Chinese database literature search*

REVISION: CNKI

Date of search: 05/11/2020; 438 results

CNKI search terms ^{1,2}		
#1	Fluoride	(KY=fluoride OR KY=fluorides OR KY=fluoridation OR KY=fluorine OR KY=fluorines OR KY=fluorosis OR KY=fluoridate OR SU=fluoride OR SU=fluorides OR SU=fluoridation OR SU=fluorine OR SU=fluorines OR SU=fluorosis OR SU=fluoridate OR AB=fluoride OR AB=fluorides OR AB=fluoridation OR AB=fluorine OR AB=fluorines OR AB=fluorosis OR AB=fluoridate)
#2	Outcome ¹	KY=nervous OR KY=iq OR KY=nerve OR KY=adhd OR KY=behavioral OR KY=behaviors OR KY=Behaviours OR KY=Behavioural OR KY=cognition OR KY=cognitive OR KY=dementia OR KY=intelligence OR KY=learning OR KY=memory OR KY=psychomotor OR KY=hyperactivity OR KY=neurobehavioral OR

CNKI search terms ^{1,2}		
		KY=neurodevelopment OR KY=neurodevelopmental OR KY=neurological OR KY=neuronal OR KY=Behavior OR KY=behaviour OR KY=Intellectual OR KY=antagonism OR KY=mental OR KY=performance OR KY=Test OR KY=mental OR KY=symptom OR KY=development
#3	Final Search ²	#1 AND #2

¹ The CNKI search engine does not allow for compound words, such as “nervous system,” “Mental work,” and “schooling performance.” Therefore, in these cases, the search terms were limited to a single term designed to catch the compound term (i.e., “nervous,” “mental,” and “performance”).

² Due to character limits, the outcome terms were searched within the results of the exposure terms. This effectively is the same as using the Boolean phrase AND.

REVISION: Wanfang

Date of search: 05/11/2020; 853 results

Wanfang search terms ¹		
#1	Fluoride	题名或关键词:(fluoride + fluorides + fluoridation + fluorine + fluorines + fluorosis + fluoridate + NaF)
#2	Outcomes	题名或关键词:(“motor activity” + Visuospatial + brain + neurobehavioral + memory + learning + intelligence + IQ + “nervous system” + neurodevelopment + ADHD + psychomotor + neurological + cognitive)
#3	Final Search	#1 AND #2

¹ 题名或关键词 = “Title/Abstract”

Appendix 2. Data Extraction Elements for HAWC: Human Studies

Data extraction elements for human studies	
Funding	Funding source(s)
	Reporting of COI by authors and/or translators (*reporting bias)
Subjects	Study population name/description
	Dates of study and sampling time frame
	Geography (country, region, state, etc.)
	Demographics (sex, race/ethnicity, age or lifestage and exposure and outcome assessment)
	Number of subjects (target, enrolled, n per group in analysis, and participation/follow-up rates) (*missing data bias)
	Inclusion/exclusion criteria/recruitment strategy (*selection bias)
	Description of reference group (*selection bias)
Methods	Study design (e.g., prospective or retrospective cohort, nested case-control study, cross-sectional, population-based case-control study, intervention, case report, etc.)
	Length of follow-up (*information bias)
	Health outcome category, e.g., neurodevelopment
	Health outcome, e.g., memory (*reporting bias)

Protocol for Systematic Review of Effects of Fluoride Exposure on Neurodevelopment

Data extraction elements for human studies	
	Diagnostic or methods used to measure health outcome (*information bias)
	Confounders or modifying factors and how considered in analysis (e.g., included in final model, considered for inclusion but determined not needed (*confounding bias)
	Substance name and CAS number
	Exposure assessment (e.g., blood, urine, hair, air, drinking water, job classification, residence, administered treatment in controlled study, etc.) (*information bias)
	Methodological details for exposure assessment (e.g., HPLC-MS/MS, limit of detection) (*information bias)
	Statistical methods (*information bias)
Results	Exposure levels (e.g., mean, median, measures of variance as presented in paper, such as SD, SEM, 75th/90th/95th percentile, minimum/maximum); range of exposure levels, number of exposed cases
	Statistical findings (e.g., adjusted β , standardized mean difference, adjusted odds ratio, standardized mortality ratio, relative risk, etc.) or description of qualitative results. When possible, OHAT will convert measures of effect to a common metric with associated 95% confidence intervals (CI). Most often, measures of effect for continuous data are expressed as mean difference, standardized mean difference, and percent control response. Categorical data are typically expressed as odds ratio, relative risk (RR, also called risk ratio), or β values, depending on what metric is most commonly reported in the included studies and on OHAT's ability to obtain information for effect conversions from the study or through author query.
	If not presented in the study, statistical power can be assessed during data extraction using an approach that can detect a 10% to 20% change from response by control or referent group for continuous data, or a relative risk or odds ratio of 1.5 to 2 for categorical data, using the prevalence of exposure or prevalence of outcome in the control or referent group to determine sample size. For categorical data where the sample sizes of exposed and control or referent groups differ, the sample size of the exposed group will be used to determine the relative power category. Recommended sample sizes to achieve 80% power for a given effect size, i.e., 10% or 20% change from control, will be compared to sample sizes used in the study to categorize statistical power as "appears to be adequately powered" (sample size for 80% power met), somewhat underpowered (sample size is 75% to < 100% of number required for 80% power), "underpowered" (sample size is 50% to < 75% of number required for 80% power), or "severely underpowered" (sample size is < 50% of number required for 80% power).
	Observations on dose response (e.g., trend analysis, description of whether dose-response shape appears to be monotonic, non-monotonic)
Other	Documentation of author queries, use of digital rulers to estimate data values from figures, exposure unit, and statistical result conversions, etc.

Appendix 3. Data Extraction Elements for HAWC: Animal Studies

Data extraction elements for animal studies	
Funding	Funding source(s)
	Reporting of COI by authors and/or translators (*reporting bias)
Animal Model	Sex
	Species
	Strain
Treatment	Chemical name and CAS number
	Source of chemical
	Purity of chemical (*information bias)
	Dose levels or concentration (as presented and converted to mg/kg bw/d when possible)
	Other dose-related details, such as whether administered dose level was verified by measurement, information on internal dosimetry (*information bias)
	Vehicle used for exposed animals
	Route of administration (e.g., oral, inhalation, dermal, injection)
	Age or lifestage at start of dosing and at health outcome assessment
	Duration and frequency of dosing (e.g., hours, days, weeks when administration was ended, days per week)
Methods	Study design (e.g., single treatment, acute, subchronic (e.g., 90 days in a rodent), chronic, multigenerational, developmental, other)
	Guideline compliance (i.e., use of EPA, OECD, NTP or another guideline for study design, conducted under GLP guideline conditions, non-GLP but consistent with guideline study, non-guideline peer-reviewed publication)
	Number of animals per group (and dams per group in developmental studies) (*missing data bias)
	Randomization procedure, allocation concealment, blinding during outcome assessment (*selection bias)
	Method to control for litter effects in developmental studies (*information bias)
	Use of negative controls and whether controls were untreated, vehicle-treated, or both
	Endpoint health category (e.g., reproductive)
	Endpoint (e.g., infertility)
	Diagnostic or method to measure endpoint (*information bias)
	Statistical methods (*information bias)

Protocol for Systematic Review of Effects of Fluoride Exposure on Neurodevelopment

Data extraction elements for animal studies	
Results	Measures of effect at each dose or concentration level (e.g., mean, median, frequency, measures of precision or variance) or description of qualitative results. When possible, OHAT will convert measures of effect to a common metric with associated 95% confidence intervals (CI). Most often, measures of effect for continuous data will be expressed as percent control response, mean difference, or standardized mean difference. Categorical data will be expressed as relative risk (RR, also called risk ratio).
	No observed effect level (NOEL), lowest observed effect level (LOEL), benchmark dose (BMD) analysis, statistical significance of other dose levels, or other estimates of effect presented in paper. Note: The NOEL and LOEL are highly influenced by study design, give no quantitative information about the relationship between dose and response, and can be subject to author’s interpretation (e.g., a statistically significant effect might not be considered biologically important). Also, a NOEL does not necessarily mean zero response. Ideally, the response rate or effect size at specific dose levels is used as the primary measure to characterize the response.
	If not presented in the study, statistical power can be assessed during data extraction using an approach that assesses the ability to detect a 10% to 20% change from control group’s response for continuous data, or a relative risk or odds ratio of 1.5–2 for categorical data, using the outcome frequency in the control group to determine sample size. Recommended sample sizes to achieve 80% power for a given effect size, i.e., 10% or 20% change from control, will be compared to sample sizes used in the study to categorize statistical power. Studies will be considered adequately powered when sample size for 80% power is met.
	Observations on dose response (e.g., trend analysis, description of whether dose-response shape appears to be monotonic, nonmonotonic)
	Data on internal concentration, toxicokinetics, or toxicodynamics (when reported)
Other	Documentation of author queries, use of digital rulers to estimate data values from figures, exposure unit, statistical result conversions, etc.

Appendix 4. Data Extraction Elements for HAWC: In vitro Studies

Data extraction elements for in vitro studies	
Funding	Funding source(s)
	Reporting of COI by authors and/or translators (*reporting bias)
Cell/Tissue Model	Cell line, cell type, or tissue
	Source of cells/tissues (and validation of identity)
	Sex of human/animal origin
	Species
	Strain
Treatment	Chemical name and CAS number
	Concentration levels (as presented and converted to μM when possible)
	Source of chemical
	Purity of chemical (*information bias)
	Vehicle used for experimental/control conditions
	Duration and frequency of dosing (e.g., hours, days, weeks when administration was ended, days per week)
Methods	Guideline compliance (i.e., use of EPA, OECD, NTP or another guideline for study design, conducted under GLP guideline conditions, non-GLP but consistent with guideline study, non-guideline peer-reviewed publication)
	Randomization procedure, allocation concealment, blinding during outcome assessment (*selection bias)
	Number of replicates per group (*information bias)
	Percent serum/plasma in medium
	Use of negative controls and whether controls were untreated, vehicle-treated, or both
	Report on data from positive controls – was expected response observed? (*information bias)
	Endpoint health category (e.g. neurological and thyroid)
	Endpoint or assay target (e.g., T3, T4, TSH levels).
	Name and source of assay kit
	Diagnostic or method to measure endpoint (e.g., reporter gene)(*information bias)
	Statistical methods (*information bias)

Protocol for Systematic Review of Effects of Fluoride Exposure on Neurodevelopment

Data extraction elements for in vitro studies	
Results	Measures of effect at each dose or concentration level (e.g., mean, median, frequency, and measures of precision or variance) or description of qualitative results. When possible, OHAT will convert measures of effect to a common metric with associated 95% confidence intervals (CI). 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).
	No Observed Effect Concentration (NOEC), Lowest Observed Effect Concentration (LOEC), statistical significance of other concentration levels, AC50, or other estimates of effect presented in paper. Note: The NOEC and LOEC are highly influenced by study design, do not give any quantitative information about the relationship between dose and response, and can be subject to author's interpretation (e.g., a statistically significant effect may not be considered biologically important). Also, a NOEC does not necessarily mean zero response.
	Observations on dose response (e.g., trend analysis, description of whether dose-response shape appears to be monotonic, non-monotonic)
Other	Documentation of author queries, use of digital rulers to estimate data values from figures, exposure unit, statistical result conversions, etc.

Appendix 5. Risk-of-Bias Criteria

The OHAT risk-of-bias tool for human and animal studies (version date January 2015 and available at <https://ntp.niehs.nih.gov/go/38673>) reflects OHAT’s current best practices and provides the detailed discussion and instructions for the risk-of-bias practices used in this evaluation. The OHAT tool uses a single set of questions (also called “elements” or “domains”) to assess risk-of-bias across various study types to facilitate consideration of conceptually similar potential sources of bias across the human and animal evidence streams with a common terminology. Individual risk-of-bias questions are designated as only applicable to certain study designs (e.g., cohort studies or experimental animal studies), and a subset of the questions apply to each study design (Table 4).

The specific criteria used to assess risk-of-bias for this evaluation are outlined below for human/observational studies and experimental animal studies. Based on literature searches, we do not expect any controlled exposure studies in humans (i.e., human controlled trials) or case-control studies and therefore have not included risk-of-bias criteria for these study designs. If relevant human controlled trials of fluoride are identified, the criteria from the January 2015 OHAT risk-of-bias tool will be used to evaluate risk-of-bias.

REVISION: In general, in order to have definitely low or definitely high risk of bias, there has to be direct evidence that bias was or was not addressed. Direct evidence is typically quantitative information or details about method validation. However, qualitative statements can provide indirect evidence that there was not (probably low) or was (probably high) potential for bias. When there is insufficient information provided to determine potential bias and information cannot be ascertained from the authors, this is interpreted as indirect evidence of potential bias and considered “probably high risk of bias.”

Observational Studies (Human studies)

Cross Sectional Studies

- 1. Was administered dose or exposure level adequately randomized? [NA]**
- 2. Was allocation to study groups adequately concealed? [NA]**
- 3. Did selection of study participants result in the appropriate comparison groups? [NA to Case series]**

Q3 Cross-sectional - Definitely Low Risk-of-bias (++)
<p>Direct evidence that subjects (both exposed and non-exposed) were similar (e.g., recruited from the same eligible population, recruited using the same inclusion and exclusion criteria, and were of similar age, socioeconomic, and health status), recruited within the same time frame, and had similar participation/response rates.</p> <p>Note: A study will generally be considered low risk-of-bias if baseline characteristics of groups differed but these differences were considered as potential confounding or stratification variables (see question #4).</p>

Q3 Cross-sectional - Probably Low Risk-of-bias (+)
<p>Indirect evidence that subjects (both exposed and non-exposed) were similar (e.g., recruited from the same eligible population, recruited using the same inclusion and exclusion criteria, and were of similar age and health status) recruited within the same time frame, and had similar participation/response rates,</p>

Q3 Cross-sectional - Probably Low Risk-of-bias (+)

OR there is indirect evidence that differences between groups were not likely to substantively bias results.

Note: Includes studies where the authors state that characteristics of exposed and referent groups were similar (as above), but do not provide quantitative information on covariates.

Q3 Cross-sectional - Probably High Risk-of-bias (-) or (NR)

Indirect evidence that subjects (both exposed and non-exposed) were not similar, recruited within very different time frames, or had the very different participation/response rates,

OR there is insufficient information provided about the comparison group to determine similarity to exposed groups (record “NR” as basis for answer).

Q3 Cross-sectional - Definitely High Risk-of-bias (--)

Direct evidence that subjects (both exposed and non-exposed) were not similar (e.g., recruited from the different eligible populations, recruited using different inclusion and exclusion criteria, or were significantly different in terms of age, socioeconomic, or health status), recruited within very different time frames, or had the very different participation/response rates.

4. Did study design or analysis account for important confounding and modifying variables?

Q4 Cross-sectional - Definitely Low Risk-of-bias (++)

Direct evidence that appropriate adjustments or explicit considerations were made for the variables listed below as potential confounders and/or effect measure modifiers in the final analyses through the use of statistical models to reduce research-specific bias including standardization, matching, adjustment in multivariate model, stratification, or other methods that were appropriately justified. Acceptable consideration of appropriate adjustment factors includes cases when the factor is not included in the final adjustment model because the author conducted analyses that indicated it did not need to be included,

AND there is direct evidence that primary covariates and confounders (including known neurodevelopmental toxicants lead and arsenic) were appropriately measured (using valid and reliable methods) and adjusted for,

OR there is direct evidence that certain covariates and cofounders that are anticipated to bias results were not present.

Note: The following variables should be considered as potential confounders and/or effect measure modifiers for the relationship between fluoride exposure and neurodevelopmental and cognitive function outcomes: age, child’s sex, race/ethnicity, maternal demographics (e.g., maternal age, BMI), parental behavioral and mental health disorders (e.g., ADHD, depression), socioeconomic status (e.g., maternal education, household income, marital status, crowding), smoking (e.g., maternal smoking status, secondhand tobacco smoke exposure), reproductive factors (e.g., parity), nutrition (e.g., BMI, growth, anemia), iodine deficiency/excess, minerals and other chemicals in water associated with neurotoxicity (e.g., arsenic, and lead), maternal (and paternal) IQ, quantity and quality of caregiving environment (e.g., HOME score).

Note: Many studies report incidence of dental and/or skeletal fluorosis, and sometimes stratify results by fluorosis severity. Because fluorosis is highly correlated with fluoride exposure, one should consider how the fluorosis is handled in the study, especially if the study authors adjusted for fluorosis.

Q4 Cross-sectional - Probably Low Risk-of-bias (+)

Indirect evidence that appropriate adjustments were made,
AND there is indirect evidence that potential covariates and confounders [age, child’s sex, race/ethnicity, maternal demographics (e.g., maternal age, mother’s cohort, BMI), parental behavioral and mental health disorders (e.g., ADHD, depression), socioeconomic status (e.g., maternal education, household income, marital status, crowding), smoking (e.g., maternal smoking status, secondhand tobacco smoke exposure), reproductive factors (e.g., parity), nutrition (e.g., BMI, growth, anemia), iodine deficiency/excess, mineral and other chemicals in water associated with neurotoxicity (e.g., arsenic and lead), maternal (and paternal) IQ, quantity and quality of caregiving environment (e.g., HOME score)] were appropriately measured and adjusted for,
OR it is deemed that not considering or only considering a partial list of covariates or confounders in the final analyses would not appreciably bias results,
AND there is evidence (direct or indirect) that covariates and confounders considered were assessed using valid and reliable measurements,
OR it is deemed that the measures used would not appreciably bias results (i.e., the authors justified the validity of the measures from previously published research),
AND there is evidence (direct or indirect) that other co-exposures anticipated to bias results were not present or were appropriately adjusted for,
OR it is deemed that co-exposures present would not appreciably bias results.

Q4 Cross-sectional - Probably High Risk-of-bias (-) or (NR)

Indirect evidence that the distribution of important covariates and known confounders differed between the groups and was not appropriately adjusted for in the final analyses,
OR there is insufficient information provided about the distribution of known confounders (record “NR” as basis for answer),
OR there is indirect evidence that there was an unbalanced distribution of co-exposures that could affect neurological development across the primary study groups, which were not appropriately adjusted for,
OR there is indirect evidence that covariates and confounders considered were assessed using measurements of unknown validity,
OR there is insufficient information provided about the measurement techniques used to assess covariates and confounders considered (record “NR” as basis for answer).

Q4 Cross-sectional - Definitely High Risk-of-bias (--)

Direct evidence that the distribution of important covariates, known confounders, and co-exposures differed between the groups and was not accounted for,
OR confounding was demonstrated or likely to be present but not appropriately adjusted for in the final analyses,
OR there is direct evidence that covariates and confounders considered were assessed using non valid measurements,
OR there is indirect evidence of co-exposure to high levels of lead and arsenic (or other agents associated with negative effects on cognition) but these co-exposures were not appropriately measured and adjusted for.

5. Were experimental conditions identical across study groups? [NA]

6. Were the research personnel blinded to the study group during the study? [NA]

7. Were outcome data complete without attrition or exclusion from analysis?

Q7 Cross-sectional - Definitely Low Risk-of-bias (++)
Direct evidence that exclusion of subjects from analyses was adequately addressed, and reasons were documented when subjects were removed from the study or excluded from analyses.
Q7 Cross-sectional - Probably Low Risk-of-bias (+)
Indirect evidence that exclusion of subjects from analyses was adequately addressed, and reasons were documented when subjects were removed from the study or excluded from analyses.
Q7 Cross-sectional - Probably High Risk-of-bias (-) or (NR)
Indirect evidence that exclusion of subjects from analyses was not adequately addressed, OR there is insufficient information provided about why subjects were removed from the study or excluded from analyses (record "NR" as basis for answer).
Q7 Cross-sectional - Definitely High Risk-of-bias (--)
Direct evidence that exclusion of subjects from analyses was not adequately addressed. Note: Unacceptable handling of subject exclusion from analyses includes: reason for exclusion likely to be related to true outcome, with either imbalance in numbers or reasons for exclusion across study groups.

8. Can we be confident in the exposure/intake characterization?

Q8 Cross-sectional - Definitely Low Risk-of-bias (++)
Direct evidence that exposure or intake was consistently assessed (i.e., using the same method and during the same time-frame) using well-established methods that directly characterize exposure or intake, OR fluoride intake is estimated from exposure sources, such as drinking water, that are well-characterized, OR exposure was assessed using less-established methods that directly measure exposure and are validated against well-established methods, AND there is sufficient range or variation in exposure measurements across groups to potentially identify associations with health outcomes, AND there is evidence that most of the exposure data measurements are above the limit of quantitation for the assay such that different exposure groups can be distinguished, AND there is evidence (direct or indirect) that the study used appropriate quality control (including blanks and spiked samples) or all analytical methods. Note: Includes studies that measure fluoride in subjects' household drinking water or urine because the relationship between levels of fluoride in drinking water and urinary fluoride levels is generally well-characterized (ATSDR, 2003). Note: Includes studies with direct intake estimates (e.g., studies that use high-quality measurements and validated estimation techniques to estimate fluoride intake from individual (e.g., water) or multiple (e.g., water and diet) exposure media)). Note: The preferred analytical method to measure total fluoride levels in liquid samples is the ion selective electrode (ISE) method, with appropriate QA and calibration (e.g., standardized ionic strength buffer and control pH to ≤ 5). The ISE method is simple, sensitive, and rapid, and it is the most commonly used analytical method to measure fluoride in environmental and biological samples (ATSDR, 2003, WHO, 2004). The ISE method is reliable to about 0.019 mg/L, and it is

Q8 Cross-sectional - Definitely Low Risk-of-bias (++)

the method recommended by NIOSH for measuring fluoride in urine (level of detection of 0.1 mg/L) (NIOSH, 1994, NRC, 2006).

Note: For fluoride levels in urine, a study needs to specify whether spot urine or 24 hour urine was used. For spot urine samples, a study should include an explanation for how urinary dilution was examined (e.g., specific gravity or creatinine).

Note: May include other less commonly used methods, such as gas or liquid chromatography or colorimetric methods, as long as appropriate QA is employed and calibration is documented.

Q8 Cross-sectional - Probably Low Risk-of-bias (+)

Indirect evidence that the exposure or intake was consistently assessed using well-established methods that directly measure exposure or intake,

OR exposure was assessed using indirect measures (e.g., drinking water levels and residency, questionnaire or occupational exposure assessment by a certified industrial hygienist) that have been validated or empirically shown to be consistent with methods that directly measure exposure (i.e., inter-methods validation: one method vs. another),

AND there is sufficient range or variation in exposure measurements across groups to potentially identify associations with health outcomes (at a minimum from high exposure or ever exposed from low exposure or never exposed),

AND there is evidence that most of the exposure data measurements are above the limit of quantitation for the assay such that different exposure groups can be distinguished,

AND there is evidence (direct or indirect) that the study used appropriate calibration and QA procedures.

Note: Includes studies that measure fluoride in subjects' blood, serum, plasma, or fingernails because the relationship between levels of fluoride in drinking water and blood, plasma, or serum levels is less well-established (ATSDR, 2003).

Note: Includes studies that report intake based on some self-reported elements.

Q8 Cross-sectional - Probably High Risk-of-bias (-) or (NR)

Indirect evidence that the exposure or intake was assessed using poorly validated methods that directly measure exposure,

OR there is evidence that the exposure was assessed using indirect measures that have not been validated or empirically shown to be consistent with methods that directly measure exposure (e.g., a job-exposure matrix or self-report without validation) (record "NR" as basis for answer),

OR there is insufficient information provided about the exposure assessment, including validity and reliability, but no evidence for concern about the method used (record "NR" as basis for answer).

Note: Includes studies with ecological exposure metrics (e.g., few measurements from a large geographic area) for which there is evidence (indirect or direct) about migration between different geographic areas. If no information on migration is reported, then the geographic setting (rural vs. urban) should be considered. Insufficient information about migration may not be a large risk-of-bias concern for studies conducted in rural (low mobility/migration) areas, in contrast to studies conducted in urban (high migration/mobility) areas.

Q8 Cross-sectional - Definitely High Risk-of-bias (--)

Direct evidence that the exposure or intake was assessed using methods with poor validity,

OR evidence of exposure misclassification (e.g., differential recall of self-reported exposure, evidence

Q8 Cross-sectional - Definitely High Risk-of-bias (--)

for high population mobility that is not accounted for).

9. Can we be confident in the outcome assessment?

Q9 Cross-sectional - Definitely Low Risk-of-bias (++)

Direct evidence that the neurodevelopmental or cognitive function outcome was assessed using well-established, validated assessment methods (well-established test methods are listed in Table 6),

AND there is direct evidence that the outcome assessors (including study subjects, if outcomes were self-reported or reported by a parent or guardian) were adequately blinded to the exposure level, and it is unlikely that they could have broken the blinding prior to reporting outcomes,

AND there is direct evidence that the test methods are appropriate to the population being studied. Evidence can include: (1) the use of tests previously tested and validated in similar populations (e.g., the Raven’s Test for Rural China applied in a study of Chinese schoolchildren) or (2) the authors provide direct evidence that the chosen methods had been specifically adapted for the study subjects and that results were valid and reproducible).

Q9 Cross-sectional - Probably Low Risk-of-bias (+)

Indirect evidence (see Note) that the outcome was assessed using instruments that were valid and reliable in the study population,

OR it is deemed that the outcome assessment methods used would not appreciably bias results,

AND there is indirect evidence that the outcome assessors were adequately blinded to the exposure level, and it is unlikely that they could have broken the blinding prior to reporting outcomes,

OR it is deemed that lack of adequate blinding of outcome assessors would not appreciably bias results (including that subjects self-reporting outcomes were likely not aware of reported links between the exposure and outcome lack of blinding is unlikely to bias a particular outcome).

Note: Indirect evidence includes: (1) the authors specify that they used methods listed in Table 6 or (2) the authors use instrument(s) not listed in Table 6 but indicate that they have been designed, tested, calibrated, or validated for measurement of relevant outcomes in the test subjects or a similar population.

Q9 Cross-sectional - Probably High Risk-of-bias (-) or (NR)

Indirect evidence (see Note) that the outcome assessment method is an insensitive or imprecise instrument,

OR there is indirect evidence that it was possible for outcome assessors to infer the exposure level prior to reporting outcomes (including that subjects self-reporting outcomes were likely aware of reported links between the exposure and outcome),

OR there is insufficient information provided about blinding of outcome assessors (record “NR” as basis for answer).

Note: Indirect evidence includes: (1) the authors specify that they used methods not listed in Table 6 and (2) the authors do not indicate (NR) that they have been designed, tested, calibrated, or validated for measurement of relevant outcomes in the test subjects or a similar population.

Q9 Cross-sectional - Definitely High Risk-of-bias (--)

Direct evidence (see Note) that the outcome assessment method is an insensitive or imprecise instrument,

OR there is direct evidence that the test method had not been previously calibrated or validated in similar populations,

OR there is direct evidence that outcome assessors were aware of the exposure level prior to reporting outcomes (including that subjects self-reporting outcomes were aware of reported links between the exposure and outcome).

Note: Direct evidence would include a previous demonstration that the instrument was not reliable in the study subjects or similar population or internal inconsistencies in the outcome assessment results or interpretation.

10. Were all measured outcomes reported?

Q10 Cross-sectional - Definitely Low Risk-of-bias (++)

Direct evidence that all of the study's measured outcomes (primary and secondary) outlined in the protocol, methods, abstract, and/or introduction (that are relevant for the evaluation) have been reported. This would include outcomes reported with sufficient detail to be included in meta-analysis or fully tabulated during data extraction and analyses had been planned in advance.

Q10 Cross-sectional - Probably Low Risk-of-bias (+)

Indirect evidence that all of the study's measured outcomes (primary and secondary) outlined in the protocol, methods, abstract, and/or introduction (that are relevant for the evaluation) have been reported,

OR analyses that had not been planned in advance (i.e., retrospective unplanned subgroup analyses) are clearly indicated as such and deemed that unplanned analyses were appropriate and selective reporting would not appreciably bias results (e.g., appropriate analyses of an unexpected effect). This would include outcomes reported with insufficient detail such as only reporting that results were statistically significant (or not).

Q10 Cross-sectional - Probably High Risk-of-bias (-) or (NR)

Indirect evidence that all of the study's measured outcomes (primary and secondary) outlined in the protocol, methods, abstract, and/or introduction (that are relevant for the evaluation) have not been reported,

OR and there is indirect evidence that unplanned analyses were included that may appreciably bias results,

OR there is insufficient information provided about selective outcome reporting ("NR" as basis for answer).

Note: Includes studies that report

Q10 Cross-sectional - Definitely High Risk-of-bias (--)

Direct evidence that all of the study's measured outcomes (primary and secondary) outlined in the protocol, methods, abstract, and/or introduction (that are relevant for the evaluation) have not been reported. In addition to not reporting outcomes, this would include reporting outcomes based on composite score without individual outcome components or outcomes reported using measurements, analysis methods or subsets of the data (e.g., subscales) that were not pre-

Q10 Cross-sectional - Definitely High Risk-of-bias (--)

specified or reporting outcomes not pre-specified, or that unplanned analyses were included that would appreciably bias results.

11. Were there no other potential threats to internal validity?

There are no fluoride-specific additions to the risk-of-bias questions for this evaluation. This question will be used to examine individual studies for appropriate statistical methods (e.g., confirmation of homogeneity of variance for ANOVA and other statistical tests that require normally distributed data). It will also be used for risk-of-bias considerations that do not fit under the other questions.

Cohort studies

1. Was administered dose or exposure level adequately randomized? [NA]

2. Was allocation to study groups adequately concealed? [NA]

3. Did selection of study participants result in the appropriate comparison groups?

Q3 Cohort - Definitely Low Risk-of-bias (++)

Direct evidence that subjects (both exposed and non-exposed) were similar (e.g., recruited from the same eligible population, recruited using the same inclusion and exclusion criteria, and were of similar age, socioeconomic and health status), and recruited within the same time frame.

Note: A study will be generally be considered low risk-of-bias if baseline characteristics of groups differed but these differences were considered as potential confounding or stratification variables (see question #4).

Q3 Cohort - Probably Low Risk-of-bias (+)

Indirect evidence that subjects (both exposed and non-exposed) were similar (e.g., recruited from the same population, recruited with the same method of ascertainment using the same inclusion and exclusion criteria, and were of similar age and health status), recruited within the same time frame, and had similar participation/response rates,

OR differences between groups were not likely to substantively bias results.

Note: Includes studies where the authors state that characteristics of exposed and referent groups were similar (as above), but do not provide quantitative information on covariates.

Q3 Cohort - Probably High Risk-of-bias (-) or (NR)

Indirect evidence that subjects (both exposed and non-exposed) were not similar, were recruited within very different time frames, or had the very different participation/response rates,

OR there is insufficient information provided about the comparison group including a different rate of non-response without an explanation (record "NR" as basis for answer).

Q3 Cohort - Definitely High Risk-of-bias (--)

Direct evidence that subjects (both exposed and non-exposed) were not similar (e.g., recruited from the different eligible populations, recruited using different inclusion and exclusion criteria, or were significantly different in terms of age, socioeconomic, or health status), recruited within very different time frames, or had the very different participation/response rates.

4. Did study design or analysis account for important confounding and modifying variables?

Q4 Cohort - Definitely Low Risk-of-bias (++)
<p>Direct evidence that appropriate adjustments or explicit considerations were made for the variables listed below as potential confounders and/or effect measure modifiers in the final analyses through the use of statistical models to reduce research-specific bias including standardization, matching, adjustment in multivariate models, stratification, or other methods that were appropriately justified. Acceptable consideration of appropriate adjustment factors includes cases when the factor is not included in the final adjustment model because the author conducted analyses that indicated it did not need to be included,</p> <p>AND there is direct evidence that primary covariates and confounders (including known neurodevelopmental toxicants lead and arsenic) were appropriately measured (using valid and reliable methods) and adjusted for,</p> <p>OR there is direct evidence that certain covariates and cofounders that are anticipated to bias results were not present.</p> <p>Note: The following variables should be considered as potential confounders and/or effect measure modifiers for the relationship between fluoride exposure and neurodevelopmental and cognitive function outcomes: age, child’s sex, race/ethnicity, maternal demographics (e.g., maternal age, mother’s cohort, BMI parental behavioral and mental health disorders (e.g., ADHD, depression), socioeconomic status (e.g., maternal education, household income, marital status, crowding), smoking (e.g., maternal smoking status, secondhand tobacco smoke exposure), reproductive factors (e.g., parity), nutrition (e.g., BMI, growth, anemia), iodine deficiency/excess, mineral and other chemicals in water associated with neurotoxicity, maternal (and paternal) IQ, quantity and quality of caregiving environment (e.g., HOME score).</p> <p>Note: Many studies report incidence of dental and/or skeletal fluorosis, and sometimes stratify results by fluorosis severity. Because fluorosis is highly correlated with fluoride exposure, one should consider how the fluorosis is handled in the study, especially if the study authors adjusted for fluorosis.</p>

Q4 Cohort - Probably Low Risk-of-bias (+)
<p>Indirect evidence that appropriate adjustments were made,</p> <p>AND there is indirect evidence that potential covariates and confounders [age, child’s sex, race/ethnicity, maternal demographics (e.g., maternal age, mother’s cohort, BMI), parental behavioral and mental health disorders (e.g., ADHD, depression), socioeconomic status (e.g., maternal education, household income, marital status, crowding), smoking (e.g., maternal smoking status, secondhand tobacco smoke exposure), reproductive factors (e.g., parity), nutrition (e.g., BMI, growth, anemia), iodine deficiency/excess, mineral and other chemicals in water associated with neurotoxicity (e.g., arsenic and lead), maternal (and paternal) IQ, quantity and quality of caregiving environment (e.g., HOME score)] were appropriately measured and adjusted for,</p> <p>AND there is evidence (direct or indirect) that covariates and confounders considered were assessed using valid and reliable measurements,</p> <p>OR it is deemed that the covariate measures used would not appreciably bias results (i.e., the authors justify the validity of the measures from previously published research),</p> <p>AND there is evidence (direct or indirect) that other covariates and confounders considered were not present or were appropriately adjusted for,</p> <p>OR it is deemed that not considering or only considering a partial list of covariates or confounders in the final analyses would not appreciably bias results.</p>

Q4 Cohort - Probably High Risk-of-bias (-) or (NR)

Indirect evidence that the distribution of important covariates and known confounders differed between the groups and was not appropriately adjusted for in the final analyses,
OR there is insufficient information provided about the distribution of known confounders (record “NR” as basis for answer),
OR there is indirect evidence that there was an unbalanced distribution of co-exposures that could affect neurological development across the primary study groups, which was not appropriately adjusted for,
OR there is indirect evidence that covariates and confounders considered were assessed using measurements of unknown validity,
OR there is insufficient information provided about the measurement techniques used to assess covariates and confounders considered (record “NR” as basis for answer).

Q4 Cohort - Definitely High Risk-of-bias (--)

Direct evidence that the distribution of important covariates, known confounders, and co-exposures differed between the groups and was not accounted for,
OR confounding was demonstrated or likely to be present but not appropriately adjusted for in the final analyses,
OR there is direct evidence that covariates and confounders considered were assessed using non valid measurements,
OR there is indirect evidence of co-exposure to high levels of lead and arsenic (or other agents associated with negative effects on cognition) but these co-exposures were not appropriately measured and adjusted for.
Note: Includes studies that report high levels of skeletal fluorosis in the study population but do not adjust for it.

5. Were experimental conditions identical across study groups? [NA]

6. Were the research personnel blinded to the study group during the study? [NA]

7. Were outcome data complete without attrition or exclusion from analysis?

Q7 Cohort - Definitely Low Risk-of-bias (++)

Direct evidence that loss of subjects (i.e., incomplete outcome data) was adequately addressed and reasons were documented when human subjects were removed from a study.
Note: Acceptable handling of subject attrition includes: very little missing outcome data; reasons for missing subjects unlikely to be related to outcome (for survival data, censoring unlikely to be introducing bias); missing outcome data balanced in numbers across study groups, with similar reasons for missing data across groups,
OR missing data have been imputed using appropriate methods and characteristics of subjects lost to follow up or with unavailable records are described in identical way and are not significantly different from those of the study participants.

Q7 Cohort - Probably Low Risk-of-bias (+)

Indirect evidence that loss of subjects (i.e., incomplete outcome data) was adequately addressed and reasons were documented when human subjects were removed from a study,

Q7 Cohort - Probably Low Risk-of-bias (+)

OR it is deemed that the proportion lost to follow-up would not appreciably bias results. This would include reports of no statistical differences in characteristics of subjects lost to follow up or with unavailable records from those of the study participants. Generally, the higher the ratio of participants with missing data to participants with events, the greater potential there is for bias. For studies with a long duration of follow-up, some withdrawals for such reasons are inevitable.

Q7 Cohort - Probably High Risk-of-bias (-) or (NR)

Indirect evidence that loss of subjects (i.e., incomplete outcome data) was unacceptably large, or substantially different across groups, and not adequately addressed,
OR there is insufficient information provided about numbers of subjects lost to follow-up (record "NR" as basis for answer).

Q7 Cohort - Definitely High Risk-of-bias (--)

Direct evidence that loss of subjects (i.e., incomplete outcome data) was unacceptably large and not adequately addressed.
Note: Unacceptable handling of subject attrition includes: reason for missing outcome data likely to be related to true outcome, with either imbalance in numbers or reasons for missing data across study groups; or potentially inappropriate application of imputation.

8. Can we be confident in the exposure/intake characterization?

Q8 Cohort - Definitely Low Risk-of-bias (++)

Direct evidence that exposure or intake was consistently assessed (i.e., using the same method and during the same time-frame) using well-established methods that directly characterize exposure or intake,
OR fluoride intake is estimated from exposure sources, such as drinking water, that are well-characterized,
OR exposure was assessed using less-established methods that directly measure exposure and are validated against well-established methods,
AND there is sufficient range or variation in exposure measurements across groups to potentially identify associations with health outcomes,
AND there is evidence that most of the exposure data measurements are above the limit of quantitation for the assay such that different exposure groups can be distinguished,
AND there is evidence (direct or indirect) that the study used appropriate quality control (including blanks and spiked samples) for all analytical methods.
Note: Includes studies that measure fluoride in subjects' household drinking water or urine because the relationship between levels of fluoride in drinking water and urinary fluoride levels is generally well-characterized (ATSDR, 2003).
Note: Includes studies with direct intake estimates (e.g., studies that use high-quality measurements and validated estimation techniques to estimate fluoride intake from individual (e.g., water) or multiple (e.g., water and diet) exposure media)).
Note: The preferred analytical method to measure total fluoride levels in liquid samples is the ion selective electrode (ISE) method, with appropriate QA and calibration (e.g., standardized ionic strength buffer and control pH to ≤ 5). The ISE method is simple, sensitive, and rapid, and it is the most commonly used analytical method to measure fluoride in environmental and biological samples (ATSDR, 2003, WHO, 2004). The ISE method is reliable to about 0.019 mg/L, and it is the method recommended by NIOSH for measuring fluoride in urine (level of detection of 0.1 mg/L) (NIOSH, 1994, NRC, 2006).

Q8 Cohort - Definitely Low Risk-of-bias (++)

Note: For fluoride levels in urine, a study needs to specify whether spot urine or 24 hour urine was used. For spot urine samples, a study should include an explanation for how urinary dilution was examined (e.g., specific gravity or creatinine).

Note: May include other less commonly used methods, such as gas or liquid chromatography or colorimetric methods, as long as appropriate QA is employed and calibration is documented.

Q8 Cohort - Probably Low Risk-of-bias (+)

Indirect evidence that the exposure or intake was consistently assessed using well-established methods that directly measure exposure or intake,

OR exposure was assessed using indirect measures (e.g., drinking water levels and residency, questionnaire or occupational exposure assessment by a certified industrial hygienist) that have been validated or empirically shown to be consistent with methods that directly measure exposure (i.e., inter-methods validation: one method vs. another),

AND there is sufficient range or variation in exposure measurements across groups to potentially identify associations with health outcomes (at a minimum from high exposure or ever exposed from low exposure or never exposed),

AND there is evidence that most of the exposure data measurements are above the limit of quantitation for the assay such that different exposure groups can be distinguished,

AND there is evidence (direct or indirect) that the study used appropriate calibration and QA procedures.

Note: Includes studies that measure fluoride in subjects' blood, serum, plasma, or fingernails because the relationship between levels of fluoride in drinking water and blood, plasma, or serum levels is less well-established (ATSDR, 2003).

Note: Includes studies that report intake based on some self-reported elements

Q8 Cohort - Probably High Risk-of-bias (-) or (NR)

Indirect evidence that the exposure or intake was assessed using poorly validated methods that directly measure exposure,

OR there is evidence that the exposure was assessed using indirect measures that have not been validated or empirically shown to be consistent with methods that directly measure exposure (e.g., a job-exposure matrix or self-report without validation) (record "NR" as basis for answer),

OR there is insufficient information provided about the exposure assessment, including validity and reliability, but no evidence for concern about the method used (record "NR" as basis for answer).

Note: Includes studies with ecological exposure metrics (e.g., few measurements from a large geographic area) for which there is evidence (indirect or direct) about migration between different geographic areas. If no information on migration is reported, then the geographic setting (rural vs. urban) should be considered. Insufficient information about migration may not be a large risk-of-bias concern for studies conducted in rural (low mobility/migration) areas, in contrast to studies conducted in urban (high migration/mobility) areas.

Q8 Cohort - Definitely High Risk-of-bias (--)

Direct evidence that the exposure or intake was assessed using methods with poor validity,

OR evidence of exposure misclassification (e.g., differential recall of self-reported exposure, evidence for high population mobility that is not accounted for).

9. Can we be confident in the outcome assessment?

Q9 Cohort - Definitely Low Risk-of-bias (++)
<p>Direct evidence that the neurodevelopmental or cognitive function outcome was assessed using well-established, validated assessment methods (well-established test methods are listed in Table 6),</p> <p>AND there is direct evidence that the outcome assessors (including study subjects, if outcomes were self-reported or reported by a parent or guardian) were adequately blinded to the exposure level, and it is unlikely that they could have broken the blinding prior to reporting outcomes,</p> <p>AND there is direct evidence that the test methods are appropriate to the population being studied. Evidence can include: (1) the use of tests previously tested and validated in similar populations (e.g., the Raven’s Test for Rural China applied in a study of Chinese schoolchildren) or (2) the authors provide direct evidence that the chosen methods had been specifically adapted for the study subjects and that results were valid and reproducible).</p>

Q9 Cohort - Probably Low Risk-of-bias (+)
<p>Indirect evidence (see Note) that the outcome was assessed using instruments that were valid and reliable in the study population,</p> <p>OR it is deemed that the outcome assessment methods used would not appreciably bias results,</p> <p>AND there is indirect evidence that the outcome assessors were adequately blinded to the exposure level, and it is unlikely that they could have broken the blinding prior to reporting outcomes,</p> <p>OR it is deemed that lack of adequate blinding of outcome assessors would not appreciably bias results (including that subjects self-reporting outcomes were likely not aware of reported links between the exposure and outcome lack of blinding is unlikely to bias a particular outcome).</p> <p>Note: Indirect evidence includes: (1) the authors specify that they used methods listed in Table 6 or (2) the authors use instrument(s) not listed in Table 6 but indicate that they have been designed, tested, calibrated, or validated for measurement of relevant outcomes in the test subjects or a similar population.</p>

Q9 Cohort - Probably High Risk-of-bias (-) or (NR)
<p>Indirect evidence (see Note) that the outcome assessment method is an insensitive or imprecise instrument,</p> <p>OR there is indirect evidence that it was possible for outcome assessors to infer the exposure level prior to reporting outcomes (including that subjects self-reporting outcomes were likely aware of reported links between the exposure and outcome),</p> <p>OR there is insufficient information provided about blinding of outcome assessors (record “NR” as basis for answer).</p> <p>Note: Indirect evidence includes: (1) the authors specify that they used methods not listed in Table 6 and (2) the authors do not indicate (NR) that they have been designed, tested, calibrated, or validated for measurement of relevant outcomes in the test subjects or a similar population.</p>

Q9 - Cohort Definitely High Risk-of-bias (--)
<p>Direct evidence (see Note) that the outcome assessment method is an insensitive or imprecise instrument, OR there is direct evidence that the test method had not been previously calibrated or validated in similar populations,</p>

Q9 - Cohort Definitely High Risk-of-bias (--)

OR there is direct evidence that outcome assessors were aware of the exposure level prior to reporting outcomes (including that subjects self-reporting outcomes were aware of reported links between the exposure and outcome).

Note: Direct evidence would include a previous demonstration that the instrument was not reliable in the study subjects or similar population or internal inconsistencies in the outcome assessment results or interpretation.

10. Were all measured outcomes reported?

Q10 Cohort - Definitely Low Risk-of-bias (++)

Direct evidence that all of the study's measured outcomes (primary and secondary) outlined in the protocol, methods, abstract, and/or introduction (that are relevant for the evaluation) have been reported. This would include outcomes reported with sufficient detail to be included in meta-analysis or fully tabulated during data extraction and analyses had been planned in advance.

Q10 Cohort - Probably Low Risk-of-bias (+)

Indirect evidence that all of the study's measured outcomes (primary and secondary) outlined in the protocol, methods, abstract, and/or introduction (that are relevant for the evaluation) have been reported,

OR analyses that had not been planned in advance (i.e., retrospective unplanned subgroup analyses) are clearly indicated as such and deemed that unplanned analyses were appropriate and selective reporting would not appreciably bias results (e.g., appropriate analyses of an unexpected effect). This would include outcomes reported with insufficient detail such as only reporting that results were statistically significant (or not).

Q10 Cohort - Probably High Risk-of-bias (-) or (NR)

Indirect evidence that all of the study's measured outcomes (primary and secondary) outlined in the protocol, methods, abstract, and/or introduction (that are relevant for the evaluation) have not been reported,

OR and there is indirect evidence that unplanned analyses were included that may appreciably bias results,

OR there is insufficient information provided about selective outcome reporting (record "NR" as basis for answer).

Q10 Cohort - Definitely High Risk-of-bias (--)

Direct evidence that all of the study's measured outcomes (primary and secondary) outlined in the protocol, methods, abstract, and/or introduction (that are relevant for the evaluation) have not been reported. In addition to not reporting outcomes, this would include reporting outcomes based on composite score without individual outcome components or outcomes reported using measurements, analysis methods or subsets of the data (e.g., subscales) that were not pre-specified or reporting outcomes not pre-specified, or that unplanned analyses were included that would appreciably bias results.

11. Were there no other potential threats to internal validity?

There are no fluoride-specific additions to the risk-of-bias questions for this evaluation. This question will be used to examine individual studies for appropriate statistical methods (e.g., confirmation of

homogeneity of variance for ANOVA and other statistical tests that require normally distributed data). It will also be used for risk-of-bias considerations that do not fit under the other questions.

Experimental Animal Studies

1. Was administered dose or exposure level adequately randomized?

<p>Q1 Experimental Animal - Definitely Low Risk-of-bias (++)</p>
<p>Study provides: Direct evidence that animals were allocated to all study groups, including concurrent controls, using a method with a random component, AND there is direct evidence that the study used a concurrent control group as an indication that randomization covered all study groups, Note: Acceptable methods of randomization include: referring to a random number table, using a computer random number generator, coin tossing, or shuffling cards (Higgins and Green 2011a). Note: Restricted randomization (e.g., blocked randomization) to ensure particular allocation ratios will be considered low bias. Similarly, stratified randomization approaches that attempt to minimize imbalance between groups on important prognostic factors (e.g., body weight) will be considered acceptable.</p>

<p>Q1 Experimental Animal - Probably Low Risk-of-bias (+)</p>
<p>Study provides: Indirect evidence that animals were allocated to all study groups including concurrent controls using a method with a random component (i.e., authors state random allocation, without description of method), AND either:</p> <ul style="list-style-type: none"> • Evidence that the study used a concurrent control group as an indication that randomization covered all study groups, <p>OR</p> <ul style="list-style-type: none"> • It is deemed that allocation without a clearly random component would not appreciably bias results.

<p>Q1 Experimental Animal - Probably High Risk-of-bias (-) or (NR)</p>
<p>Study provides either:</p> <ul style="list-style-type: none"> • Indirect evidence that animals were allocated to all study groups using a method with a non-random component, <p>OR</p> <ul style="list-style-type: none"> • Indirect evidence that there was a lack of a concurrent control group, <p>OR</p> <ul style="list-style-type: none"> • There is insufficient information provided about how animals were allocated to study groups (record "NR" as basis for answer).

<p>Q1 Experimental Animal - Definitely High Risk-of-bias (--)</p>
<p>Study provides either:</p> <ul style="list-style-type: none"> • Direct evidence that animals were allocated to all study groups using a non-random method including judgment of the investigator, the results of a laboratory test, or a series of tests, <p>OR</p> <ul style="list-style-type: none"> • Direct evidence that there was a lack of a concurrent control group.

2. Was allocation to study groups adequately concealed?

Q2 Experimental Animal - Definitely Low Risk-of-bias (++)
Study provides: Direct evidence that at the time of assigning study groups the research personnel did not know what group animals were allocated to, and it is unlikely that they could have broken the blinding of allocation until after assignment was complete and irrevocable.
Note: Acceptable methods used to ensure allocation concealment include sequentially numbered treatment containers of identical appearance or equivalent methods.

Q2 Experimental Animal - Probably Low Risk-of-bias (+)
Study provides either:

- Indirect evidence that at the time of assigning study groups the research personnel did not know what group animals were allocated to and it is unlikely that they could have broken the blinding of allocation until after assignment was complete and irrevocable,

OR

- It is deemed that lack of adequate allocation concealment would not appreciably bias results.

Q2 Experimental Animal - Probably High Risk-of-bias (-) or (NR)
Study provides either:

- Indirect evidence that at the time of assigning study groups it was possible for the research personnel to know what group animals were allocated to, or it is likely that they could have broken the blinding of allocation before assignment was complete and irrevocable,

OR

- There is insufficient information provided about allocation to study groups (record “NR” as basis for answer).

Q2 Experimental Animal - Definitely High Risk-of-bias (--)
Study provides: Direct evidence that at the time of assigning study groups it was possible for the research personnel to know what group animals were allocated to, or it is likely that they could have broken the blinding of allocation before assignment was complete and irrevocable.

3. Did selection of study participants result in the appropriate comparison groups? [NA]

4. Did study design or analysis account for important confounding and modifying variables? [NA]

5. Were experimental conditions identical across study groups?

Q5 Experimental Animal - Definitely Low Risk-of-bias (++)
Study provides: Direct evidence that same vehicle was used in control and experimental animals, **AND** direct evidence that non-treatment-related experimental conditions were identical across study groups (i.e., the study explicitly states that animals were all in the same room or provides other details to indicate that the conditions were identical).
Note: In many cases the vehicle may just be drinking water

Q5 Experimental Animal - Probably Low Risk-of-bias (+)
Study provides either:

- Indirect evidence that the same vehicle was used in control and experimental animals,

OR

- It is deemed that the vehicle used would not appreciably bias results,

Q5 Experimental Animal - Probably Low Risk-of-bias (+)

AND the authors do not explicitly state that non-treatment-related experimental conditions were identical (e.g., experimental conditions were provided, but there is no statement or demonstration that conditions were the same across study groups)

Q5 Experimental Animal - Probably High Risk-of-bias (-) or (NR)

Study provides either: Indirect evidence that the vehicle differed between control and experimental animals,

OR

- Authors did not report the vehicle used (record “NR” as basis for answer),

OR

- Indirect evidence that non-treatment-related experimental conditions were not comparable between study groups.

Q5 Experimental Animal - Definitely High Risk-of-bias (--)

Study provides either:

- Direct evidence from the study report that control animals were untreated, or treated with a different vehicle than experimental animals,

OR

- Direct evidence that non-treatment-related experimental conditions were not comparable between study groups.

6. Were the research personnel blinded to the study group during the study?

Q6 Experimental Animal - Definitely Low Risk-of-bias (++)

Study provides: Direct evidence that the research personnel were adequately blinded to study group, and it is unlikely that they could have broken the blinding during the study.

Note: Methods used to ensure blinding include: central allocation, sequentially numbered treatment containers of identical appearance, sequentially numbered animal cages; or equivalent methods.

Q6 Experimental Animal - Probably Low Risk-of-bias (+)

Study provides either:

- Indirect evidence that the research personnel were adequately blinded to study group, and it is unlikely that they could have broken the blinding during the study,

OR

- It is deemed that lack of adequate blinding during the study would not appreciably bias results. This would include cases where blinding was not possible but research personnel took steps to minimize potential bias, such as restricting the knowledge of study group to veterinary or supervisory personnel monitoring for overt toxicity, or randomized husbandry or handling practices (e.g., placement in the animal room, necropsy order, etc.).

Q6 Experimental Animal - Probably High Risk-of-bias (-) or (NR)

Study provides either:

- Indirect evidence that the research personnel were not adequately blinded to study group,

OR

- There is insufficient information provided about blinding to study group during the study (record “NR” as basis for answer).

Q6 Experimental Animal - Definitely High Risk-of-bias (--)

Direct evidence that the research personnel were not adequately blinded to study group.

7. Were outcome data complete without attrition or exclusion from analysis?

Q7 Experimental Animal - Definitely Low Risk-of-bias (++)

Study provides either:

- Direct evidence that loss of animals was adequately addressed and reasons were documented when animals were removed from a study.

OR

- Direct evidence that missing data have been imputed using appropriate methods (insuring that characteristics of animals are not significantly different from animals retained in the analysis).

Note: Acceptable handling of attrition includes: very little missing outcome data; reasons for missing animals unlikely to be related to outcome (or for survival data, censoring unlikely to be introducing bias); missing outcome data balanced in numbers across study groups, with similar reasons for missing data across groups; missing outcomes is not enough to impact the effect estimate.

Q7 Experimental Animal - Probably Low Risk-of-bias (+)

Study provides either:

- Indirect evidence that loss of animals was adequately addressed and reasons were documented when animals were removed from a study,

OR

- It is deemed that the proportion lost would not appreciably bias results. This would include reports of no statistical differences in characteristics of animals removed from the study from those remaining in the study.

Q7 Experimental Animal - Probably High Risk-of-bias (-) or (NR)

Study provides either:

- Indirect evidence that loss of animals was unacceptably large and not adequately addressed,

OR

- There is insufficient information provided about loss of animals (record "NR" as basis for answer).

Q7 Experimental Animal - Definitely High Risk-of-bias (--)

Study provides: Direct evidence that loss of animals was unacceptably large and not adequately addressed.

Note: Unacceptable handling of attrition or exclusion includes: reason for loss is likely to be related to true outcome, with either imbalance in numbers or reasons for loss across study groups.

8. Can we be confident in the exposure characterization?

Q8 Experimental Animal - Definitely Low Risk-of-bias (++)

Study provides: Direct evidence that the exposure to fluoride was independently characterized and purity confirmed generally as $\geq 98\%$, (and compliance with the treatment, if applicable)

AND exposure was consistently administered (i.e., with the same method and time-frame) across treatment groups,

Q8 Experimental Animal - Definitely Low Risk-of-bias (++)

AND for dietary or drinking water studies information is provided on consumption or internal dose metrics to confirm expected exposure levels sufficiently to allow discrimination between exposure groups,

AND if internal dose metrics are available, there is direct evidence that most of the exposure data measurements are above the limit of quantitation for the assay such that different exposure groups can be distinguished,

AND if internal dose metrics are available, the study used spiked samples or a dilution curve to confirm assay performance,

AND the analytical methods used to independently characterize fluoride are described or referenced.

Note: If controls are administered tap water, the level of fluoride in the drinking water should be provided.

Note: The preferred analytical method to measure total fluoride levels is the ion selective electrode (ISE) method, with appropriate QA and calibration (e.g., standardized ionic strength buffer and control pH to ≤ 5). The ISE method is simple, sensitive, and rapid, and it is the most commonly used analytical method to measure fluoride in environmental and biological samples (ATSDR 2003, WHO 2004). The ISE method is reliable to about 0.019 mg/L, and it is the method recommended by NIOSH for measuring fluoride in urine (level of detection of 0.1 mg/L) (NIOSH 1994, NRC 2006).

Note: Includes other less-established (or not commonly used) methods, such as gas or liquid chromatography, colorimetric methods, or the acid-hexamethyldisiloxane diffusion method, as long as appropriate QA and calibration were well-documented.

Note: If internal dose measurements are made, measurement of fluoride in blood, serum, plasma, bone, or in urine are the standard accepted biomarkers of exposure.

Note: For internal dose metrics, the timing of the fluoride exposure assessment (sample collection) in relation to treatment (e.g., at end of period, mid-way, at outcome assessment) should be provided. The internal dose should be measured at a time to represent the exposure, therefore, measuring internal dose in close temporal proximity to the outcome assessment may not be appropriate if the outcome is measured months after the exposure.

Q8 Experimental Animal - Probably Low Risk-of-bias (+)

Study provides: Indirect evidence that the exposure to fluoride was appropriately characterized and purity confirmed generally as $\geq 98\%$ (i.e., the supplier of the chemical provides documentation of the purity of the chemical) **OR** direct evidence that purity was independently confirmed as $\geq 95\%$ and it is deemed that impurities of up to 5% would not appreciably bias results,

AND exposure was consistently administered (i.e., with the same method and time-frame) across treatment groups,

AND for dietary or drinking water studies, no information is provided on consumption or internal dose metrics,

AND if internal dose metrics are available, there is indirect evidence that most of the exposure data measurements are above the limit of quantitation for the assay such that different exposure groups can be distinguished.

Note: Studies without purity, stability, or consumption information can still be considered to be probably low risk-of-bias if there are internal measurements that indicate there is low concern for bias.

Q8 Experimental Animal - Probably Low Risk-of-bias (+)

Note: Fluorosilicic acid is generally provided as a 20% weight volume solution in water. This is acceptable because it is assumed that they used a compound of appropriate purity to create the solution and the dissociation of the fluoride ion is complete.

Q8 Experimental Animal - Probably High Risk-of-bias (-) or (NR)

Study provides either:

- Indirect evidence that the exposure (including purity of the test substance and compliance with the treatment, if applicable) was assessed using poorly validated methods,

OR

- There is insufficient information provided about the validity of the exposure assessment method, but no evidence for concern (record “NR” as basis for answer),

AND if internal dose metrics are available, there is indirect evidence that most of the exposure data measurements are below the limit of quantitation for the assay such that different exposure groups cannot be distinguished.

Q8 Experimental Animal - Definitely High Risk-of-bias (--)

Study provides: Direct evidence that the exposure (including purity of the test substance and compliance with the treatment, if applicable) was assessed using poorly validated methods.

9. Can we be confident in the outcome assessment?

Q9 Experimental Animal - Definitely Low Risk-of-bias (++)

Study provides: Direct evidence that the outcome was assessed using well-established methods and assessed at the same length of time after initial exposure in all study groups,

AND either:

- Direct evidence that the outcome assessors were adequately blinded to the study group, and it is unlikely that they could have broken the blinding prior to reporting outcomes,

OR

- Outcomes were assessed with a fully automated method (e.g., automatic video recording and scoring of behavioral performance), which removes the potential bias of outcome assessors (note: a semi-automated method such as video recording without the automated scoring is not considered fully automated)

Note: Well-established methods will depend on the outcome, but examples of such methods may include: Morris water maze, T-maze, Y-maze, novel object recognition, mini-holeboard activity, activity cage, step-down test, shuttle box, operant behavior tests, open field, plank walking, rotarod, slanted surface, auditory startle, negative geotaxis, tail immersion, Von Frey hair test, cliff avoidance, surface righting, pivoting/orienting reflex, forced swim test, tail suspension test, and the elevated plus maze.

Note: There are standard protocols for each of these well-established methods. For example, the general protocol for the Morris water maze includes a task acquisition phase, during which the animal learns the location of a hidden platform over successive training sessions with multiple trials per day, followed by a probe test to measure spatial memory for the hidden platform location. Studies that use the Morris water maze should report performance in both the task acquisition phase and the probe test, and large deviations from this general protocol should be documented and supported by previously published studies.

Q9 Experimental Animal - Probably Low Risk-of-bias (+)

Study provides either:

- Indirect evidence that the outcome was assessed using acceptable methods (i.e., deemed valid and reliable but not the gold standard) **AND** indirect evidence that the outcome was assessed at the same length of time after initial exposure in all study groups,

OR

- It is deemed that the outcome assessment methods used would not appreciably bias results,

AND either:

- Indirect evidence that the outcome assessors were adequately blinded to the study group, and it is unlikely that they could have broken the blinding prior to reporting outcomes,

OR

- It is deemed that lack of adequate blinding of outcome assessors would not appreciably bias results, which is more likely to apply to objective outcome measures.

Note: For some outcomes, particularly histopathology assessment, outcome assessors are not blind to study group as they require comparison to the control to appropriately judge the outcome, but additional measures such as multiple levels of independent review by trained pathologists can minimize potential bias.

Q9 Experimental Animal - Probably High Risk-of-bias (-) or (NR)

Study provides either:

- Indirect evidence that the outcome assessment method is an insensitive instrument,

OR

- Indirect evidence that the length of time after initial exposure differed by study group,

OR

- Indirect evidence that it was possible for outcome assessors to infer the study group prior to reporting outcomes without sufficient quality control measures,

OR

- There is insufficient information provided about blinding of outcome assessors (record "NR" as basis for answer).

Q9 Experimental Animal - Definitely High Risk-of-bias (--)

Study provides either:

- Direct evidence that the outcome assessment method is an insensitive, or internally or externally invalid instrument,

OR

- Direct evidence that the length of time after initial exposure differed by study group,

OR

- Direct evidence for lack of adequate blinding of outcome assessors, including no blinding or incomplete blinding without quality control measures.

10. Were all measured outcomes reported?

Q10 Experimental Animal - Definitely Low Risk-of-bias (++)

Study provides: Direct evidence that all of the study's measured outcomes (primary and secondary) outlined in the protocol, methods, abstract, and/or introduction (that are relevant for the evaluation) have been reported. This would include outcomes reported with sufficient detail to

Q10 Experimental Animal - Definitely Low Risk-of-bias (++)

be included in meta-analysis or fully tabulated during data extraction and analyses had been planned in advance.

Q10 Experimental Animal - Probably Low Risk-of-bias (+)

Study provides either:

- Indirect evidence that all of the study’s measured outcomes (primary and secondary) outlined in the protocol, methods, abstract, and/or introduction (that are relevant for the evaluation) have been reported,

OR

- Indirect evidence that analyses that had not been planned in advance (i.e., retrospective unplanned subgroup analyses) are clearly indicated as such and deemed that unplanned analyses were appropriate and selective reporting would not appreciably bias results (e.g., appropriate analyses of an unexpected effect). This would include outcomes reported with insufficient detail such as only reporting that results were statistically significant (or not).

Q10 Experimental Animal - Probably High Risk-of-bias (-) or (NR)

Study provides either:

- Indirect evidence that all of the study’s measured outcomes (primary and secondary) outlined in the protocol, methods, abstract, and/or introduction (that are relevant for the evaluation) have not been reported,

OR

- Indirect evidence that unplanned analyses were included that may appreciably bias results,

OR

- There is insufficient information provided about selective outcome reporting (record “NR” as answer basis).

Q10 Experimental Animal - Definitely High Risk-of-bias (--)

Study provides: Direct evidence that all of the study’s measured outcomes (primary and secondary) outlined in the protocol, methods, abstract, and/or introduction (that are relevant for the evaluation) have not been reported. In addition to not reporting outcomes, this would include reporting outcomes based on composite score without individual outcome components or outcomes reported using measurements, analysis methods or subsets of the data (e.g., subscales) that were not pre-specified or reporting outcomes not pre-specified, or that unplanned analyses were included that would appreciably bias results.

11. Were there no other potential threats to internal validity?

Internal validity refers to whether the methods and modes of analysis used in a study can be interpreted to reflect a potential causal relationship between specific factor(s) and observed outcomes. It can be interpreted to mean, generally, whether a study has been done “right” so that the results are “valid” in the specific study setting and is free of bias and confounding. This question will be used to examine individual studies for appropriate statistical methods (e.g., confirmation of homogeneity of variance for ANOVA and other statistical tests that require normally distributed data). It will also be used for risk-of-bias considerations that do not fit under the other questions.

Q11 Experimental Animal - Definitely Low Risk-of-bias (++)

Study provides: Direct evidence that homogeneity of variance was tested for any statistical test that requires normally distributed data (e.g., t-test, ANOVA).

AND Direct evidence that repeated measures statistical analyses were used for any experiments that repeatedly measured outcomes in the same animals,

AND Direct evidence that the litter was considered the basic unit of analysis for any study that used littermates in an experiment.

Note: Due to differences between males and females, it is preferable that results for males and females be reported separately. Reporting the results together can bias the results towards the null if an effect was observed in only one sex.

Q11 Experimental Animal - Probably Low Risk-of-bias (+)

Study provides: Indirect evidence that the litter was considered the basic unit of analysis for any study that used littermates in an experiment.

AND indirect evidence that repeated measures statistical analyses were used for any experiments that repeatedly measured outcomes in the same animals.

Q11 Experimental Animal - Probably High Risk-of-bias (-) or (NR)

Study provides either:

- Indirect evidence that homogeneity of variance was not tested for any statistical test that requires normally distributed data (e.g., t-test, ANOVA),

OR

- Indirect evidence that repeated measures statistical analyses were not used for any experiments that repeatedly measured outcomes in the same animals,

OR

- Indirect evidence that the litter was not considered the basic unit of analysis for any study that used littermates in an experiment.

OR

- There is insufficient information provided about statistical methods including if litter was used as the basic unit (record "NR" as basis for answer).

Q11 Experimental Animal - Definitely High Risk-of-bias (--)

Study provides either:

- Direct evidence that homogeneity of variance was not tested for any statistical test that requires normally distributed data (e.g., t-test, ANOVA),

OR

- Direct evidence that repeated measures statistical analyses were not used for any experiments that repeatedly measured outcomes in the same animals,

OR

- Direct evidence that the litter was not considered the basic unit of analysis.

Note: Includes studies that considered each littermate an independent observation and the individual pup as the experimental unit.

REVISION: Appendix 6. Meta-analysis

This protocol was developed in response to comments from the NASEM committee peer review in November 2019 (NASEM 2020). The committee strongly recommended conducting a meta-analysis. NTP developed this protocol for meta-analyses to evaluate the association between fluoride exposure and children's intelligence. A draft protocol was peer reviewed in June of 2020 and document below reflects consideration of comments received.

Background

A 2006 evaluation by the National Research Council (NRC) concluded that consumption of high levels of naturally occurring fluoride in drinking water is associated with neurological effects in humans and recommended further investigation (NRC 2006). Specifically, the NRC concluded that the consistency of the results of IQ deficits in children exposed to fluoride at 2.5 to 4 mg/L in drinking water from a few epidemiological studies of Chinese populations appeared significant enough to warrant additional research on the effects of fluoride on intelligence. Moreover, it has been suggested that fluoride can cross the blood-brain barrier, induce neurotoxicity, and affect children's cognitive abilities and mental development (Adinolfi 1985, ATSDR 2003, Grandjean and Landrigan 2006).

Since the NRC review, two meta-analyses have provided support for an association between fluoride exposure and neurotoxicity in children (Choi *et al.* 2012, Duan *et al.* 2018). Details of these meta-analyses are briefly described below and are summarized in Table A6-1.

The first meta-analysis (Choi *et al.* 2012) assessed 27 studies comparing mean IQ scores between areas with high and low fluoride exposure groups. Most of these study populations were from China and exposed to fluoride through drinking water. Choi *et al.* (2012) found that high fluoride exposure was associated with significantly lower IQ in children (Table A6-1). In meta-regression analyses, they found that year of publication (but not mean age of the study children) was a significant source of heterogeneity. When the analyses were restricted to the 16 studies that used the Combined Raven's Test—The Rural edition in China (CRT-RC), the authors found that the mean age of the study children (but not year of publication) was a significant predictor of the estimated effect (i.e., standardized weighted mean difference in IQ). While the authors did determine a risk ratio for living in an endemic fluorosis area, they excluded studies with individual-level measures of exposure and were not able to perform a formal dose-response analysis.

A more recent meta-analysis (Duan *et al.* 2018) assessed 26 studies that evaluated intelligence levels in children exposed to high or low drinking water fluoride, 15 of which were also included in the 2012 meta-analysis and 7 that were published after the 2011 inclusion period. Thirteen studies included in the Choi *et al.* (2012) analysis were not considered in the Duan *et al.* (2018) evaluation for reasons that are not apparent from the text of the manuscript. The 2018 update included four studies from Iran, four from India, and the remaining from China. Duan *et al.* (2018) found that high water fluoride exposure was associated with lower intelligence levels in children (Table A6-1). In meta-regression analyses, the mean age of study children significantly affected the relationship between high water fluoride levels and children's intelligence levels. Subgroup analyses included country, age (<10 or ≥10 years), water fluoride dosage, type of intelligence assessment, and sex.

Duan *et al.* (2018) also performed a dose-response meta-analysis which suggested a significant association between increased water fluoride exposure and lower intelligence levels. However, it is

unclear which studies were included in this analysis. The authors reported both linear and nonlinear relationships between fluoride exposure and intelligence levels.

Overall, the two meta-analyses present consistent results indicating that high water fluoride exposure is associated with lower intelligence levels. These meta-analyses either included studies “with the high and low-exposure groups only” (Choi *et al.* 2012), or in which “the experimental group comprised [of] individuals from areas with high fluoride levels in water, and the control group comprised individuals with normal levels in water”; however, neither meta-analysis defines a priori the fluoride exposure levels considered high, low or normal.⁶

Analysis	Number of Studies	Model	Exposure	Outcome	Age Range	SMD ¹ or RR ² (95% CI)	Heterogeneity (I ² , p-value)
Choi <i>et al.</i> (2012)	27	All studies	High vs. low fluoride areas	IQ score	4–16 ³	SMD: –0.45 (–0.56, –0.34)	80%, p = 0.000
	23	Excluding nonstandardized tests				SMD: –0.44 (–0.54, –0.33)	78%, p < 0.001
	16	Excluding non-CRT-RC tests				SMD: –0.36 (–0.48, –0.25) RR: 1.93 (1.46, 2.55)	78%, p < 0.001 59%, (p NR)
	9	Excluding studies with coexposures (iodine, arsenic) or non-drinking-water fluoride exposure				SMD: –0.29 (–0.44, –0.14) RR: 1.75 (1.16, 2.65)	82%, p < 0.001 71%, (P NR)
Duan <i>et al.</i> (2018)	26	All studies	High vs. low water fluoride	Intelligence level	4–16 ³	SMD: –0.52 (–0.62, –0.42)	69%, p = 0.040
	15	Excluding non-CRT-RC tests				SMD: –0.53 (–0.67, –0.38)	77%, p < 0.001
	NR ⁴	Dose-response: absolute fluoride dosage				Nonlinear Linear	p < 0.001 p = 0.976
	NR	Dose-response: relative fluoride dosage				Nonlinear Linear	p < 0.001 p < 0.001

¹SMD = standardized mean difference.

²RR = risk ratio for low/marginal IQ (<80) vs. normal IQ using the Combined Raven’s Test–The Rural edition in China (CRT-RC).

³One study (Seraj *et al.* 2006) did not specify the age range of subjects.

⁴NR = not reported.

Since Duan *et al.* (2018), there have been several new publications addressing the association between exposure to fluoride and neurodevelopmental and cognitive effects, including two cohorts from Canada

⁶ Fluoride levels in the high exposure groups ranged from 0.57-11.5 mg/L drinking water, based on the “high” ranges reported in the Tables 1 of Choi *et al.* (2012) and in Duan *et al.* (2018). The low or reference groups levels ranged from 0 to 2.35 mg/L drinking water, based on the “high” ranges reported in the Tables 1 of Choi *et al.* (2012) and in Duan *et al.* (2018).

and one cohort from Mexico with individual-level exposure data (Bashash *et al.* 2017, Green *et al.* 2019, Till *et al.* 2020). A recent integrated literature review, with a focus on studies in children (Grandjean 2019), did not develop a dose-response curve but concluded that the association between fluoride exposure and neurotoxicity appeared to be dose-dependent.

Because no previous meta-analysis has focused on studies with individual-level data, there is a need for an updated evaluation of the strength of association between fluoride exposure and children's intelligence, including a deeper understanding of the shape of the dose-response curve.

Objectives

The key objectives that will be addressed by the proposed meta-analysis are as follows:

- What is the magnitude and precision of the association between exposure to fluoride and children's intelligence?
- What is the shape of the relationship between exposure to fluoride and children's intelligence?

Specific aims

1. To update existing meta-analyses with additional studies

There are at least eight potentially relevant studies published since the November 2016 inclusion cutoff of the Duan *et al.* (2018) analysis that will be considered. All studies included in Duan *et al.* (2018) and Choi *et al.* (2012) analyses will also be considered. Selection criteria (see Method section) will likely yield a different set of relevant studies than in the previous meta-analyses. Based on the last literature update (May 2020), an updated meta-analysis comparing studies with exposed and reference groups would include at least 10 more studies than were included in Choi *et al.* (2012) and Duan *et al.* (2018), combined.

It is expected that not all these studies will provide effect estimates that can be used in a quantitative analysis.

2. To conduct new meta-analyses using individual-level exposure data

The proposed new meta-analysis would include studies with more precise individual-level exposures, including prenatal exposures, and better address potential confounders. In addition, while the majority of the available studies compare populations with high fluoride exposure to those with lower fluoride exposure, one of our sub-analyses will assess studies with fluoride exposure levels comparable to drinking water fluoridation levels in the United States.

There are several studies reporting IQ score as a function of increased levels of fluoride in water (Saxena *et al.* 2012, Sebastian and Sunitha 2015, Yu *et al.* 2018, Green *et al.* 2019, Till *et al.* 2020, Wang *et al.* 2020) or urine (Xiang *et al.* 2003, Ding *et al.* 2011, Xiang *et al.* 2011, Saxena *et al.* 2012, Zhang *et al.* 2015, Bashash *et al.* 2017, Cui *et al.* 2018, Till *et al.* 2020, Wang *et al.* 2020). However, some studies would not be appropriate for inclusion in a meta-analysis because of missing measures of variability for the effect estimates (Rocha-Amador *et al.* 2007, Seraj *et al.* 2012).

Methods

General methods

The purpose of this meta-analysis is to explore the associations between exposure to fluoride and children's intelligence levels. The available evidence will be evaluated with respect to several selection criteria to ensure results included in the meta-analysis will be sufficiently similar. The following selection criteria will be considered:

- **Outcomes:** Children's intelligence levels reflected by IQ scores. The most comprehensive IQ score reported will be used (e.g., Full scale IQ versus Performance or Verbal IQ) for studies using the Wechsler Primary and Preschool Scale of Intelligence-III).
- **Exposure:** Exposure to fluoride based on environmental measures or biomonitoring data
 1. Exposed versus reference fluoride exposure groups. If there are more than two exposure groups, the highest exposure will be the "exposed" group and the lowest exposure will be the "reference" group. (Aim 1); however, all the exposure groups reporting exposure ranges will be used in the Aim 3 analysis.
 2. Individual-level fluoride levels. The study outcomes will be evaluated with respect to a 1-mg/L unit increase in exposure. When necessary, to ensure consistent units across studies, units of exposure will be transformed to 1 mg/L. For studies reporting multiple measures of fluoride exposure, in the analysis of all studies combined, results associated with measured or estimated individual-level exposures, biomarker levels (such as urinary fluoride), or fluoride intake will be prioritized over water fluoride concentrations (NTP 2019b). (Aim 2)

For dose-response analyses, typically, at least three exposure groups or a dose-response curve are required.
- **Effect estimates:**
 - 1: Mean outcome measure with measures of uncertainty (95% confidence interval [CI], standard error [SE], and N) for IQ levels for exposed versus reference fluoride exposure groups. As a first step, the standardized mean difference and corresponding 95% CI will be calculated. (Aim 1, Aim 3). When only mean effects, Ns, and p-values for differences between groups are reported, SDs will be calculated using the SE and t-statistic (assuming equal variances) following the approach outlined in the Cochrane Handbook for Systematic Reviews (Higgins *et al.* 2019).
 - 2: Studies that report effects as coefficients from regression models, odds ratios (ORs), or relative risks (RR) and include 95% CI or SE. A crude RR and 95% CI will be calculated when a study did not report effect estimates but reported sufficient data to calculate one (i.e., a slope and associated 95% CI). When available, adjusted effect estimates will be used in the meta-analysis. However, crude estimates will be included in the analysis when adjusted results were not reported. If results from multiple models were reported within a single study, the most adjusted results will be selected. (Aim 2)

For studies with overlapping populations (i.e., multiple studies that use the same cohort), results will be selected considering the following factors: most appropriate exposure metric, exposure range, exposure period, number of subjects, and statistical adjustment for potential confounders. A table will be provided to summarize these selections.

Protocol for Systematic Review of Effects of Fluoride Exposure on Neurodevelopment

Data extraction was completed a priori using HAWC. To minimize potential bias in the selection of the results for inclusion in the meta-analysis, results will be selected by an independent epidemiologist following the criteria outlined above.

The meta-analysis will describe study results and study quality assessed using the risk-of-bias approach outlined in the systematic review protocol (NTP 2019b). Data will be presented in tabular and graphical formats. STATA software will be used to provide summary outputs and figures, such as forest plots that are easily interpretable. Forest plots will be used to illustrate results for subgroups of studies that have the same effect estimate (such as RR), similar dose levels, similar exposure and outcome measurements, and that adjust for comparable sets of confounders.

Proposed analyses

Primary analyses

- Pooled effect estimates for Aims 1, 2 and 3
- Subanalyses by risk-of-bias assessment, country, gender, children's age in years, fluoride exposure measures, type of intelligence assessment test

Sensitivity analyses

- Using any exposure group versus the reference group
- Excluding potential outlier studies
- Including studies using Performance and Verbal IQ rather than Full-scale IQ when studies use the Wechsler Primary and Preschool Scale of Intelligence-III
- Performing additional sensitivity analyses as suggested by the data collected for the primary analyses

Statistical Analyses

Estimating a pooled effect estimate

For the analysis using mean outcome measures (IQ levels), the pooled effect will be estimated as the weighed standardized mean difference (SMD) and corresponding 95% CI. Typically, meta-analyses use the weighted mean difference to combine the differences between the means of the exposed and control groups to estimate the overall mean difference. A prerequisite of this method is that the response is measured in the same units using comparable devices in all studies; however, in our analysis, studies evaluate children's intelligence levels using various assessments including the Chinese standardized Raven's test, Raven's intelligence test, Wechsler Intelligence test, Binet-Simon test, Japan IQ test, and Raymond B Cattell test.

Given the heterogeneity in assessments used in the studies considered for this meta-analysis, the SMD measure will be used, calculated as the ratio of the observed difference in means to an estimate of the standard deviation of the response. This approach is especially appropriate when studies measure the same outcome (e.g., intelligence levels) but use different measurement instruments (Hedges and Olkin 1985). By standardization, the study results are transformed to a common scale (standard deviation

units) that facilitates pooling. An inverse variance weighting method is used in all the approaches (Hedges and Olkin 1985, Rosenthal 1994).

Pooled SMD estimates will be considered statistically significant if their 95% CIs do not overlap zero. A pooled SMD less than zero indicates that children in an exposed fluoride group have lower IQ scores than those in the reference fluoride group. Different pooled SMD estimates are considered significantly different from one another if their 95% CIs do not overlap. Similarly, pooled relative risk estimates will be considered significant if their 95% CIs do not overlap one, and we will consider pooled risk estimates significantly different from one another if their 95% CIs do not overlap. Our interpretation of the results of the meta-analysis will not rely solely on statistical significance to determine whether a result indicates a positive or negative association. The direction and magnitude of the associations will also be evaluated.

Random-effects models with the DerSimonian and Laird method (DerSimonian and Laird 1986) will be used to obtain the summary risk estimates, and heterogeneity of the included studies will be assessed by Cochran's Q test (Cochran 1954) and the I^2 statistic. The significance of heterogeneity will be described by Cochran Q test associated p-value with a significance level of 0.1. However, this test does not measure the degree of heterogeneity present. Alternatively, the I^2 statistic will be used to describe the percentage of total variation across studies that is caused by heterogeneity rather than chance (Higgins *et al.* 2003). A value of 0% indicates no observed heterogeneity, and larger values show increasing heterogeneity. I^2 values less than 25% are generally viewed as low heterogeneity, while values of 25–50% are considered moderate and values over 80% are considered high (Higgins *et al.* 2003).

Forest plots will be used to display results and to examine possible heterogeneity between studies. Potential publication bias will be assessed by developing funnel plots and performing Egger regression on the estimates of effect size (Egger *et al.* 1997). In addition, if these methods suggest that publication bias is present, trim and fill methods (Duval and Tweedie 2000b, a) will be used to estimate the number of missing studies in a meta-analysis affected by publication bias and the effect of those studies on the effect estimate to predict the impact of the hypothetical "missing" studies. Trim-and-fill methods using both linear and "run" estimators, which were found to perform best in terms of mean squared error (MSE) (Duval and Tweedie 2000b), will be used.

If there is significant study-level heterogeneity (i.e., Q test p-value < 0.1, I^2 > 80%;), subgroup analyses may be conducted to determine how much heterogeneity can be explained by accounting for both within- and between-study variance.

There is no universally accepted optimal minimum number of studies that are required for a subanalysis. The size of the studies and the distribution of subgroup variables are also important considerations. In general, meta-analysis approaches are considered most suitable if there are at least six to ten results for a continuous variable and at least four results for a categorical variable in studies considered of moderate to large size (Fu *et al.* 2011). Given the understanding that any recommended number has an arbitrary element, subanalyses will be performed regardless of the number of studies; however, other study characteristics will be considered and results will be interpreted with caution when evaluating subanalyses derived from a very small number of studies (Russo *et al.* 2020).

Conducting subanalyses

Following up on the sources of heterogeneity identified in Choi *et al.* (2012) and Duan *et al.* (2018), subgroup analyses will be considered by country, gender, age in years, fluoride exposure measures, type of intelligence assessment test, and potential confounding variables. In addition, subanalyses will be considered using:

- Risk-of-bias determination (low risk-of-bias studies only, high risk-of-bias studies only). This subanalysis would evaluate whether the overall meta-analysis result represents spurious findings because of the presence of within study bias and confounding due to the observational design (e.g., selection bias, information bias, and confounding).
- Studies with lower levels of fluoride exposure to assess the issue of generalizability of the results to the United States. The decision on which levels of fluoride exposure are considered “lower” would be determined by the data as opposed to an a priori decision.

To perform these subanalyses, the following variables will be extracted: study name; effect estimate (mean outcome measure, mean difference, beta, or risk ratio; SEs or 95% CIs); study design; country; children’s age and gender; intelligence assessment test; total number of subjects; number of subjects for each dose level; and overall risk-of-bias determination.

Statistical analyses will be performed using STATA version 16.1 (StataCorp 2019) using the *metan*, *metareg*, *metainf*, *metafunnel*, *metabias*, and *metatrim* packages (Palmer and Sterne 2016).

Estimating a dose-response relationship

To study dose-response relationships, different studies often use different exposure levels (doses) to report the effects of the same exposure, so estimates across levels need to be pooled to get one estimate in common units for each study. Since estimates for separate exposure levels depend on the same reference group, they are not independent, but are correlated to some degree. The meta-analysis must take this correlation into account.

Several methodological issues need to be considered in conducting a dose-response meta-analysis. For studies reporting at least three exposure-specific RRs and CIs, the standard method of analysis is to first fit a linear regression through the origin (reference category) weighted by the estimated inverse variance of the log RR. Alternative methods are used when the assumption that all RRs in each study are independent is not met (Greenland and Longnecker 1992, Berlin *et al.* 1993, Orsini *et al.* 2006, Orsini *et al.* 2012). In addition to linear dose-response meta-analyses, methods have been developed to evaluate nonlinear dose-response relationships by incorporating flexible splines, such as natural splines or restricted-cubic splines (Vlaanderen *et al.* 2010, Orsini *et al.* 2012). Those flexible methods will provide excellent opportunities to evaluate the shape of the dose-response curve in epidemiologic studies across a wide range of fluoride exposure concentrations; however, uncertainties may become greater at lower fluoride exposures, hence careful consideration will be given when applying the methods and interpreting the results.

For the dose-response meta-analysis using the RR estimates, the methods proposed by (Greenland and Longnecker 1992) and Orsini *et al.* (2012) will be used to compute the trend from the correlated log RR estimates across exposure levels. For each study, the median or mean fluoride intake for each exposure group will be assigned to each corresponding RR. If the median or mean intake in each exposure group were not provided, the midpoint of the upper and lower boundaries in every category will be assigned

as the average intake. If the upper boundary for the highest exposure group was not reported, the boundary will be assumed to have the same amplitude as the nearest category. To examine a potential nonlinear relationship between exposure to fluoride and children's intelligence levels, restricted cubic splines will be created and a potential departure from a linear trend will be assessed by testing the coefficient of the second spline equal to zero (Orsini *et al.* 2012). The restricted cubic spline model will be fit with a generalized least-squares regression taking into account the correlation within each set of published RRs and will combine the study-specific estimates using the restricted maximum likelihood method in a multivariable random-effects meta-analysis (Greenland and Longnecker 1992, Orsini *et al.* 2006, Jackson *et al.* 2010, Orsini *et al.* 2012).

For the dose-response meta-analysis using the mean effect estimates, the approach proposed by Crippa and Orsini (2016) will be used. This approach includes dose-response models estimated within each study (first stage) and an overall curve obtained by pooling study-specific dose-response coefficients (second stage) (Crippa and Orsini 2016). The covariance among study specific mean differences is accounted for in the first stage of analysis using generalized least square estimators, while statistical heterogeneity across studies is allowed by multivariate random-effects model in the second stage. This method allows different modelling strategies, using splines, which do not require a priori information about the shape of the dose-response curve.

Typically, for a dose-response meta-analysis, at least three exposure groups are needed per study. However, a newer one-step approach has been recently developed (Crippa *et al.* 2019) which uses linear mixed models and can use the entire available data (including studies with one nonreference group). A one-stage approach can be formulated as a single statistical model, and it has been shown to yield identical results to a two-stage approach in a fixed-effects analysis (Discacciati *et al.* 2017). It is generally referred to as equivalent in a random-effects model (Bagnardi *et al.* 2004, Rota *et al.* 2010). The main advantage of the one-stage approach is that it allows estimation of multiple coefficients needed to parameterize flexible curves. In addition, study-specific dose-response analyses in a two-stage approach are often limited by the small number of observations, typically two or three nonreference exposure groups. The method had been successfully applied in published meta-analyses using the STATA package (Larsson and Orsini 2018, Filippini *et al.* 2019).

A dose-response meta-analysis will be performed that includes only studies that examined lower fluoride exposure levels for which a dose-response could be assessed. This subanalysis could be used to assess the issue of generalizability of the results to the United States. The decision as to what levels of fluoride exposure are considered "lower" will be determined by the data as opposed to an a priori decision.

To perform the dose-response analysis, the following variables will be extracted: study name, median of each category in common units across studies (dose), risk ratio and associated 95% CIs, study design, country, children's age and gender, intelligence assessment test, total number of subjects, number of subjects for each dose level, and overall risk-of-bias evaluation determination.

Statistical analyses will be performed using STATA version 16.1 (StataCorp 2019) using the *glst*, *mkspline*, and *drmeta* packages (Palmer and Sterne 2016).

Meta-analysis Protocol History and Revisions

Date	Activity or revision
June 2020	Draft meta-analysis protocol reviewed; sent to technical advisors for comment/review
August 2020	Meta-analysis protocol revised and added to systematic review protocol