National Toxicology Program

Review of the Draft NTP Monograph: Systematic Review of Fluoride Exposure and Neurodevelopmental and Cognitive Health Effects

A Consensus Study Report of the National Academies of Sciences, Engineering, and Medicine (NASEM) 2020

Response to Comments

Main Comments

1) **NASEM recommended that NTP clearly describe the role of the OHAT Handbook in developing the systematic review protocol. In addition, NASEM identified topic areas (e.g., nomination history, problem formulation) of the protocol where the level of detail differed from the general methods outlined the OHAT Handbook.**
    - **RESPONSE**: NTP added a foreword to the monograph and text to the protocol to clarify the relationship between the OHAT Handbook, which outlines general methods for conducting NTP's health effects evaluations, and the protocol, which describes project-specific procedures tailored to the specific systematic review. NTP also added further details to sections of the protocol and the monograph, when appropriate, to address NASEM's recommendations for further information.

2) **NASEM identified a need for clarifications regarding the literature search strategy and screening procedures in the protocol and methods of the monograph**
    a) NASEM indicated that it is unclear how the evaluation of animal data in the NTP 2016 systematic review related to the literature search strategy and assessment of animal studies in the current systematic review.
        - **RESPONSE:** NTP added text to the protocol and monograph to clarify that the literature search strategy in Appendix A that was used for the current systematic review 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.

          RESPONSE: NTP added text to the protocol and monograph to clarify that the assessment of animal data for the current systematic review was an extension of the NTP 2016 systematic review and not a literature search update. As an extension of the NTP 2016 report, this evaluation relied on the NTP 2016 systematic review of animal studies as an assessment of the animal literature published prior to 2015.

    b) NASEM raised a concern that the use of the Fluoride Action Network (FAN) as a source to identify studies for the systematic review could potentially bias the selection of studies, as FAN identified a number of studies published in Chinese language journals but the process by which FAN identified and selected studies was not clear. NASEM suggested the

possibility that the FAN database of studies from Chinese language journals could be incomplete and may have missed null studies (i.e., studies that did not report an association with fluoride exposure). NASEM suggested that NTP conduct an independent search of non-English (specifically Chinese) literature.

- **RESPONSE:** NTP conducted supplemental searches of Chinese databases. NTP added text to describe the development of the search strategy and the selection of studies in the protocol and the methods and results of the monograph. Newly retrieved human references were 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. Null studies that were identified were translated and included.

c) NASEM expressed concern regarding the use of SWIFT-Active Screener to conduct title and abstract screening due to NASEM's understanding that the tool had not been validated and the potential that the tool would result in a large number of missed studies. Specifically, NASEM estimated that 260 studies could have been relevant at the title and abstract screening level but were missed by NTP in the initial screening due to the use of SWIFT-Active Screener.

- **RESPONSE:** The SWIFT-Active Screener paper published in 2020 (Howard *et al*. 2020) provides additional details on the tool including validation of the approach. NTP added text to the protocol and monograph to cite this validation paper and summarize the results.

  RESPONSE: NTP added a new section titled "Evaluation of SWIFT-Active Screener Results" to the monograph to clarify a misunderstanding by NASEM regarding the number of potentially missed studies resulting from the use of SWIFT-Active Screener. In short, the SWIFT-Active statistical algorithm predicted that 10 relevant studies at the title and abstract level were not identified (rather than 260). NTP also evaluated the SWIFT-Active screening results to gain a better understanding of the potential impact of using SWIFT-ACTIVE Screener for this systematic review. Based on this evaluation, NTP estimates that the use of Swift-Active Screener may have resulted in missing 1–2 relevant human studies and 1–2 relevant animal studies with primary neurodevelopmental or cognitive outcomes from the database searches. A summary of this evaluation has been added to this new section in the monograph. Please see Appendix A of this Response to Comments document for a more thorough response to this comment.

3) NASEM identified a need for clarification of several aspects of the risk-of-bias methods.

a) NASEM identified several serious concerns regarding whether NTP's risk-of-bias evaluation adequately captured important threats to internal validity that are specific to neurobehavioral outcomes in animal studies.

- NASEM requested that NTP consider their risk-of-bias concerns but also cautioned that given the poor quality of the animal studies, revising the systematic review to address these concerns might not affect NTP's finding that the animal evidence is inadequate to inform conclusions about fluoride exposure and neurodevelopmental and cognitive effects in humans. Therefore, NASEM acknowledged that NTP would need to decide whether it should reanalyze the animal evidence.
- **RESPONSE**: NTP considered NASEM's risk-of-bias concerns and the overall status of the animal body of evidence as part of the full systematic review and came to the decision to not re-evaluate the animal data. NTP added text to the beginning of the "Animal Learning and Memory Data" section of the monograph as an introductory statement to the section, which acknowledges NASEM's concerns from the peer review and explains NTP's decision to not conduct a re-evaluation of the animal data.

b) NASEM recommended that NTP reconcile apparent discrepancies between the critical confounders that are identified in the protocol and monograph.

- **RESPONSE**: NTP revised text in the protocol and monograph to consistently identify the key confounders that would apply to all study populations.

c) NASEM identified apparent inconsistencies between the description of the methods for considering potential confounders in the protocol and the description of the handling of confounders, especially co-exposures, in the monograph.

- **RESPONSE:** NTP revised the "Rationale for critical risk-of-bias domains for human studies" section of the protocol to more clearly explain how potential confounders (specifically co-exposures) were considered (i.e., NTP more clearly defined key confounders and how it was determined if a co-exposure was reasonably anticipated to be a risk-of-bias concern).

d) NASEM recommended that NTP discuss on a study-by-study basis (when applicable) the impact that potential exposure misclassification or potential confounding may have had on magnitude and direction of effect.

- **RESPONSE**: NTP developed and added a new appendix to the monograph titled "Appendix 4. Details for Lower Risk-of-Bias Studies," which describes the potential impact that risk-of-bias concerns related to potential exposure misclassification or potential confounding may have on magnitude and direction of effect for the lower risk-of-bias studies on neurodevelopmental and cognitive function in humans.

e) NASEM recommended that NTP consider blinding of the outcome assessor more carefully when considering human studies.

- **RESPONSE**: The risk-of-bias rationale for each study discusses the consideration for blinding. The new appendix to the monograph titled "Appendix 4. Details for Lower Risk-of-Bias Studies," also provides the direct and indirect evidence on methods and blinding. In the absence of direct or indirect evidence that the outcome was assessed blind to knowledge of the exposure status, authors were contacted for information.

f) NASEM raised a concern that NTP classified studies as having lower risk of bias when the measure of the neurodevelopmental and cognitive outcome was seriously flawed. NASEM gave one example of a study for which the neurodevelopmental outcome was based on parent- or child-reported diagnosis of learning disability or ADHD and suggested that the concern was serious enough to consider the study to have definitely high risk of bias.

- **RESPONSE**: NTP took this comment into consideration and conducted additional review of the studies but ultimately did not make a change to the monograph. Based on the methods for evaluating risk of bias in the protocol, a rating of "definitely high risk of bias" can be reached for outcome assessment if there is direct evidence that the methods for outcome assessment were imprecise or that the outcome assessors were not blind to the exposure. NTP verified that the lower risk-of-bias studies did not provide direct evidence of imprecision or lack of blinding, which would definitely bias the results.

g) NASEM identified a concern that the human studies may not have gone through a rigorous statistical analysis review. Moreover, NASEM identified several studies with concerns related to internal validity based on statistical analyses.

- **RESPONSE**: For each of the lower risk-of-bias studies specifically identified by NASEM, NTP has added text regarding the statistical analyses to the new appendix to the monograph titled "Appendix 4. Details for Lower Risk-of-Bias Studies." The appropriateness of the statistical approach was evaluated for each study, and the published remarks about the statistical methods in the Green *et al*. (2019) were reviewed by a senior statistician. For the purpose of the risk-of-bias evaluations, care was taken to afford the same amount of weight on the statistical analysis domain for all studies evaluated, regardless of whether they were specifically identified in NASEM comments. Please note that some of the issues identified by NASEM would not fall under the statistical analysis metric for consideration of risk of bias. For example, the large differences in the number of male and female offspring in Valdez Jimenez *et al*. (2017) would have been considered as potential issues for risk of bias in both the subject selection and confounding domains.

4) NASEM recommended that NTP revise the reporting of the number of studies in the monograph so that multiple publications on the same cohort of individuals are not counted as independent studies.

- **RESPONSE:** NTP revised the monograph to report the number of studies and the number of study populations, when appropriate, to distinguish between the number of

publications and independent study populations. Multiple publications on the same population were described together when feasible or noted to be on the same population. In addition, when conducting the meta-analysis, only a single publication was selected from a given study population, and details on the study selection were provided. Note that study authors did not always clearly identify the initial study population in a publication, and in some cases NTP had to make assumptions regarding the independence or lack thereof of publications based on the information available including authors, study area, and year of cohort recruitment.

5) **NASEM strongly recommended that NTP reconsider its decision not to perform a meta- analysis**
   - **RESPONSE**: NTP concurred with NASEM's recommendation and decided to conduct a meta-analysis to evaluate the association between fluoride exposure and children's intelligence. NTP developed a meta-analysis protocol and had it peer reviewed, which now appears as Appendix 6 in the revised systematic review protocol. NTP's meta-analysis included two specific aims: 1) to update existing meta-analyses (Choi *et al*. 2012, Duan *et al*. 2018) with additional studies that compared mean IQ scores between areas with high and low fluoride exposure groups; and 2) to conduct a new meta-analysis using individual-level exposure data. NTP's meta-analysis also conducted a formal dose-response analysis as part of the meta-analysis.

     RESPONSE: NTP conducted subgroup analyses under aims 1 and 2 of the meta-analysis described above to address heterogeneity in the data and to further analyze the consistency of the data. Subgroup analyses were conducted by risk of bias, country, gender, age, method for measuring IQ, and exposure type. In addition, dose-response analyses were conducted using all fluoride levels and lower fluoride exposure levels only, evaluating water and urine separately.

References

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

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.

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.

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.

Valdez Jimenez L, Lopez Guzman OD, Cervantes Flores M, Costilla-Salazar R, Calderon Hernandez J, Alcaraz Contreras Y, Rocha-Amador DO. 2017. In utero exposure to fluoride and cognitive development delay in infants. Neurotox 59: 65-70.

Appendix A: Additional Comment Response Details

Comment from Section 2, "Important Details," of NASEMs Report:

Screening for inclusion. Studies were screened for inclusion by using a structured form in SWIFT-Active Screener, a machine-learning software program used to rank studies for screening. The National Academies has stated that automated screening procedures can facilitate efficiencies in the process and that incorporation of software tools, such as SWIFT-Active Screener, can help to achieve that goal (NRC 2014; NASEM 2018). However, those tools are relatively new and have not undergone rigorous evaluation or validation. Specifically, to the committee's knowledge, they have not been validated for screening studies for inclusion in systematic reviews. Furthermore, screening up to 98% inclusion means that as many as 2% of the 13,023 studies excluded on the basis of the SWIFT algorithm in this systematic review—260 studies— could be relevant according to title and abstract screening but missed in the initial screening. Given the large number of studies screened for this systematic review, that is not an insignificant number, although the committee notes that not all the studies would likely be deemed relevant in the full-text screening step. The OHAT handbook mentions the SWIFT text-mining and machine-learning tools but does not justify or cite why 98% estimated recall is considered sufficient. The committee recommends that the protocol discuss the basis of that decision and potentially conduct a sensitivity analysis to determine the effect of that cutoff on the overall findings (for example, by reviewing a random subset of the studies excluded on the basis of the SWIFT algorithm to identify the number of potentially missed references).

RESPONSE: NASEM estimated that 260 studies could be relevant at the title and abstract level but were missed in the initial screening due to the use of SWIFT-Active Screener. This estimate by NASEM is based on the assumption that screening up to 98% inclusion means that as many as 2% of the 13,023 studies that were not screened could be relevant; however, this is not an accurate interpretation of what the 2% represents in the SWIFT-Active statistical algorithm. The SWIFT-Active statistical algorithm predicted that 10 relevant studies at the title and abstract level were not identified, rather than 260 studies. The following paragraph further explains how the algorithm works and how it applies to our dataset.

SWIFT-Active Screener, which is used by other U.S. Government agencies including the Environmental Protection Agency and the Department of Agriculture, employs active learning to continually incorporate user feedback during title and abstract screening to predict the total number of included studies, and title and abstract screening is stopped once the statistical algorithm in SWIFT-Active Screener estimates that 98% of the predicted number of relevant studies were identified. In our assessment, SWIFT-Active predicted that there were 739 relevant studies during the initial title and abstract screening. NTP did in fact continue screening past the 98% relevancy threshold and screened until 98.6% of the predicted number of relevant studies were identified, which equated to 729 studies being included during the initial title and abstract screening. The 2% that NASEM cites in their comment (though in our specific case is 1.4%) is meant to mean 1.4% of the 739 predicted relevant studies (instead of 1.4% of the 13,023 unscreened references). Therefore, the SWIFT-Active statistical algorithm predicted that ***10 relevant studies*** at the title and abstract level (1.4% X 739 predicted relevant studies;

OR 739 predicted relevant studies minus 729 identified relevant studies during screening) were not identified by not screening the remaining 13,023 studies.

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 datasets 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. In the document sets with 5,000 or more references, 95% of the relevant articles were identified after screening 34% of the available references, on average, using SWIFT-Active Screener. Please note that for this NTP systematic review, there were 20,883 available references in Swift-Active Screener , and 37.6% of these references were screened.

To further consider the impact of using SWIFT-Active Screener for this systematic review, NTP evaluated the SWIFT-Active screening results to gain a better understanding of the relevancy of the last group of studies that were screened before 98% predicted recall was satisfied. The goal was to determine the likelihood of having missed important studies by not screening all of the literature.

To do this, NTP evaluated two subsets of studies screened in SWIFT-Active for trends and followed those studies through to full-text review. The first subset included the last 50 studies screened in SWIFT-Active regardless of inclusion or exclusion. Of these last 50 studies, only 2 were included at the title and abstract level. During full-text review, both of these studies were determined to not be relevant and were excluded. Therefore, all of the last 50 studies that were screened using Swift-Active Screener were ultimately excluded.

The second subset of studies included the last 50 studies screened as relevant in SWIFT-Active at the title and abstract level. NTP determined the status of these studies after full-text review. Of these 50 studies, approximately 14% were screened as relevant human studies with primary neurodevelopmental or cognitive outcomes (learning, memory, and/or intelligence), and 14% were screened as relevant animal studies with primary neurodevelopmental or cognitive outcomes.

Next, NTP estimated the number of relevant human and animal studies with primary neurodevelopmental or cognitive outcomes that were potentially missed by not screening the remaining 13,023 studies in SWIFT-Active Screener. Based on the rates of inclusion from the last group of 50 included studies (14% relevant human studies; 14% relevant animal studies) and the estimated number of missed relevant studies at the title and abstract screening level (n = 10 studies), NTP estimates that the use of Swift-Active Screener may have resulted in missing 1–2 relevant human studies and 1–2 relevant animal studies with primary neurodevelopmental or cognitive outcomes.

Please note that the identification of relevant studies through title and abstract screening using SWIFT-Active Screener is only one approach that was utilized to identify relevant literature during the assessment. Systematic reviews also consist of hand-searching relevant studies and searching for relevant reviews. In addition, the peer review process and public comments provide opportunities for NTP to learn of relevant studies that were missed.

Based on the estimates outlined in this response, the SWIFT-Active Screener validation study by Howard *et al*. (2020), and considering the opportunity for the identification of relevant literature through other sources throughout the systematic review process, NTP is confident that the studies included in this review represent the body of evidence.