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

PROTOCOL FOR SYSTEMATIC REVIEW OF

LONG-TERM NEUROLOGICAL EFFECTS FOLLOWING ACUTE EXPOSURE TO THE ORGANOPHOSPHORUS NERVE AGENT SARIN

April 2017

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

TABLE OF CONTENTS

TABLES

FIGURES

PROTOCOL TO EVALUATE THE EVIDENCE OF LONG-TERM NEUROLOGICAL EFFECTS FOLLOWING ACUTE EXPOSURE TO THE ORGANOPHOSPHORUS NERVE AGENT SARIN

Project co-leads: David A. Jett, Ph.D., Director, NIH Countermeasures Against Chemical Threats (CounterACT) Program, NINDS, Office of Translational Research and Andrew A. Rooney, Office of Health Assessment and Translation (OHAT), DNTP

Summary: In partnership with NTP, the NIH CounterACT program is conducting a systematic review to evaluate the evidence for long-term neurological effects[1] in humans and animals following acute exposure[2] to the organophosphorus (OP) nerve agent sarin.

BACKGROUND AND SIGNIFICANCE

Background

Sarin (CAS #: 107-44-8) is a nerve agent developed for chemical warfare during World War II. This highly toxic "nerve gas" is actually liquid at ambient temperatures. It is also known as GB, which is a two-character identifier assigned by NATO. Sarin belongs to a chemically diverse group of organophosphorus chemicals that have at least one carbon atom bound to a phosphorous atom. The group includes other chemical weapons and many agricultural and residential pesticides. It is likely that sarin continues to be used in conflict, as reported by the United Nations in Syria in 2013 (Sellstrom 2013).

The median lethal dose (LD50) of dermal exposure to sarin for a 70 kg person is only 1 - 10 mL (ATSDR 2015). Immediately after exposure, the effects of sarin are well-characterized. Acute effects include a progression from miosis (constriction of the pupils), excessive secretions and muscle fasciculation to seizures that may progress to status epilepticus, muscle paralysis, cardiorespiratory depression, and death due to respiratory failure. The signs and symptoms of acute exposure are generally referred to as cholinergic signs. From experimental studies of humans and animals where dose is known, cholinergic signs are generally understood as "mild, moderate, or severe." When dose is unknown, a reasonable approximation of dose may be estimated by observing clinical signs and symptoms. This is an established approach in which a threshold detection level may be translated consistently into estimates of high-, intermediate- and low-level exposures (Brown and Brix 1998). In humans, cholinergic signs following a single intermediate- to low-level exposure to sarin may last up to 96 hours, when these short-term signs typically resolve (Okumura et al. 2005). Pharmacological and medical treatment may accelerate the time that cholinergic signs subside.

After high-, intermediate- or low-level cholinergic signs have subsided, any *sequelae* are considered long-term. Such long-term neurological health effects may be observed several hours, days, weeks or

[1] Throughout this document, "Long-term" neurological effect is defined as any neuropathological, pathophysiological, or behavioral effect observed that occurs after the cholinergic signs and symptoms caused by an initial sublethal acute exposure have subsided. Long-term neurological effects may occur immediately following this cholinergic crisis or they may be observed over a range of time periods including days, weeks, months or years after the cholinergic symptoms subside. Long-term neurological effects may resolve or may persist.

[2] Throughout this document, an "acute exposure" is defined as a single dose of sarin that causes cholinergic signs and symptoms.

years after the cholinergic crisis subsides. Long-term effects may be pathophysiological and/or behavioral.

There are suggestions that neurological effects of sarin poisoning are observed months to years after the initial acute exposure. For example, there are multiple reports that victims who were exposed to sarin in the Matsumoto and Tokyo subway attacks may have suffered long-term neurological effects, including behavioral abnormalities and alteration of brain morphology (Murata et al. 1997; Yamasue et al. 2007).

Among a number of animal studies reporting long-term neurological effects, a study in Rhesus monkeys observed significant and persistent increases in the relative amount of high frequency beta activity in electroencephalograms (EEG) one year post-sarin exposure (Burchfiel and Duffy 1982). In rats, at three months after exposure to a single non-convulsive symptomatic concentration of sarin for 60 minutes in an inhalation chamber, a significant alteration was observed in the exposed rats as assessed using a functional observatory battery (FOB), characterized as changes in mobile activity and gait, and as an increase in stereotyped behavior (Kassa et al. 2001).

Several literature reviews of the long-term neurological effects following exposure to sarin have been published (SIPRI 1975; Defense Science Board 1994; Brown and Brix 1998; Augerson 2000; IOM 2004; Brown 2009; Binns JH 2004; White et al. 2016). However, many of these reviews have been an assessment of health effects in military personnel during conflicts such as the Gulf War, and are confounded by concurrent mixed exposures to other chemicals including other chemical warfare agents. The current review will focus on a specific data set where sarin nerve agent is the only suspected exposure. Additionally, a systematic review of the evidence has not been performed in which selection criteria were clearly stated and consistently applied; where a broad hierarchy of evidence is considered including all evidence streams (human, animal and mechanistic); where a broad range of human study designs are considered including uncontrolled studies and case-reports or case series; and in which individual studies were assessed for internal validity or risk of bias. This review will focus on neurological outcomes because the overwhelming majority of the evidence for potential long-term health effects of sarin address neurological endpoints.

Impetus for this Systematic Review

Reports in the literature raise the questions of whether long-term neurological effects of acute exposure to OP agents in humans are a common occurrence, particularly in individuals exposed to higher doses. Developing a clear understanding of the level of evidence for long-term neurological effects in both humans and animal models is the first step to determining whether treatments should be considered for development for these effects, much the same as neuroprotectants are being developed to treat patients after stroke, status epilepticus, and other brain injuries. These questions, together with related questions discussed at a workshop convened in February 2014 by the NIH CounterACT program and the NINDS (https://www.ninds.nih.gov/sites/default/files/2014-NIH-Workshop-on-Nerve-Agents_0.pdf), form the impetus for the current effort to conduct a Systematic Review of the literature to date that would evaluate the evidence that long-term neurological effects may follow acute exposure to sarin.

Significance

This evaluation will review the available literature on long-term neurological effects of acute exposure to sarin. If the literature base is sufficient, this review will reach level of evidence conclusions for hazard identification. Even if the database is too limited to support hazard conclusions, the evaluation will include a synthesis of the neurological effects data and will identify areas of consistency and uncertainty

in the evidence of long-term neurological effects of acute exposure to sarin. The evaluation will include critical appraisal of the identified studies that may strengthen the design and conduct of future studies of sarin, as well as data gaps and research needs.

This critical evaluation and clear statement of the level of evidence for long-term neurological health effects in humans and animals is necessary to: (1) inform the need for further study on the development of additional medical countermeasures against the effects of sarin in humans; and (2) help inform the development and refinement of animal models for long-term neurological effects following acute sarin exposure that can be used in therapeutic discovery and development, and are based on available evidence in humans.

OVERALL OBJECTIVE AND SPECIFIC AIMS

Objective

The overall objective of this systematic review is to evaluate the evidence for long-term neurological effects following acute, sublethal exposure to sarin based on integrating levels of evidence from human and animal studies and consideration of the degree of support from mechanistic data.

Specific Aims

- Identify literature that assessed long-term neurological health effects following acute exposure to sarin in human, animals, and *in vitro*/mechanistic studies.
- Extract data on potential long-term neurological health effects from relevant studies.
- Assess the internal validity (risk of bias) of individual studies using pre-defined criteria.
- Summarize the extent and types of health effects evidence available.
- Describe limitations of the systematic review, limitations of the evidence base, identify areas of uncertainty, as well as data gaps and research needs on long-term neurological health effects of sarin.

Dependent on the extent and nature of the available evidence:

- Synthesize the evidence using a narrative approach or meta-analysis (if appropriate) considering limitations on data integrating such as study design heterogeneity.
- Rate confidence in the body of evidence for human and animal studies separately according to one of four statements: High, Moderate, Low, or Very Low/No Evidence Available.
- Translate confidence ratings into level of evidence of health effects for human and animal studies separately according to one of four statements: High, Moderate, Low, or Inadequate.
- Combine the level of evidence ratings for human and animal data to reach one of five possible hazard identification conclusions: Known, Presumed, Suspected, Not classifiable, or Not identified to be a hazard to humans.

The evaluation will integrate evidence of long-term neurological effects associated with acute sarin exposure from human studies across a broad range of study design types along with controlled exposure animal studies and mechanistic/*in vitro* studies. 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 health effects. As a result of our scoping efforts, there is a relatively small number of in vitro mechanistic studies using sarin. Depending on the extent of these studies, we will consult with our technical advisors and subject matter experts to consider their quality and relevance to the PECO.

PECO Statement

PECO (**P**opulation, **E**xposure, **C**omparators and **O**utcomes) statements were developed as an aid to identify search terms and inclusion/exclusion criteria as appropriate for addressing the overall research question (long-term neurological effects of acute sarin exposure) for the systematic review (Higgins and Green 2011).The PECO statements are listed below for human (Table 1), animal (Table 2), and *in vitro*/mechanistic studies (Table 3).

Table 1. Human PECO (Population, Exposure, Comparator and Outcome) Statement

PECO Element	Evidence
Population	Humans without restriction as to age or sex, geographic location, or life stage at exposure or outcome assessment; and no restriction as to whether military or civilian/non-combatant
Exposure	Single acute exposure to sarin based on: • known dose or concentration in an experimental protocol • diagnostic biomonitoring data (e.g., sarin or biomarkers in plasma or urine) • environmental detection (e.g., air, soil) • corroboration by assessment of direct (in hospital, in clinic) or indirect observation of symptoms of acute cholinergic signs (video or reported by patient's family) • dose may be extrapolated from clinical signs and symptoms per Brown and Brix (1998) and as adapted from (Namba et al. 1971) • No restriction on whether exposure is accidental or intentional
Comparators	For controlled and uncontrolled studies, comparable populations not exposed to sarin; and case series-reports, no comparable populations
Outcomes	Neurological outcomes including changes in nervous system function (e.g., cognitive, sensory, motor), and neuropathology (e.g., imaging and post-mortem)

Table 2. Animal PECO (Population, Exposure, Comparator and Outcome) Statement	
PECO Element	**Evidence**
Population	Without restriction as to species, age, or sex, or life stage at exposure or outcome assessment
Exposure	Single acute exposure to sarin based on administered dose or concentration or biomonitoring data (e.g., urine, blood, or other specimens)
Comparators	Comparable untreated animal subjects or animals exposed to vehicle-only treatment
Outcomes	Neurological outcomes including changes in nervous system function (e.g., cognitive, sensory, motor), neuropathology (e.g., imaging and post-mortem), and neurophysiology (e.g., ion channel and receptor function)

Table 3. *In vitro*/mechanistic PECO (Population, Exposure, Comparator and Outcome) Statement	
PECO Element	**Evidence**
Population	Human or animal cells, tissues or model systems with *in vitro* exposure regimens
Exposure	Exposure to sarin based on administered dose or concentration
Comparators	Comparable cells or tissues exposed to vehicle-only treatment or untreated controls
Outcomes	Measurements of the survival and the morphology of neurons or glia, including neurohistochemical and immunohistochemical techniques such as H&E, Nissl, Rapid Golgi, Fluoro-Jade, Silver Stain, HRP, GFAP, neurotransmitter stains, axon/dendrite-specific markers, and others

The overall objective, PECO statements, and strategy to synthesize study results were based on a series of problem formulation steps beginning with detailed input from: (1) scientific and clinical experts with backgrounds in neurotoxicology and systematic review; and (2) deliberation with CounterACT staff and in consultation with scientists at other Federal agencies who contributed to, and/or participated in the CounterACT Workshop in February 2014. More details about problem and process formulation can be found below in the methods section.

METHODS

Step 1. Problem Formulation

Nomination History

The HHS Office of the Assistant Secretary of Preparedness and Response (ASPR) and Department of Defense have large multi-agency programs to develop better medical countermeasures against chemical threat agents. The NIH is charged with basic and translational research in this area. Within the NIH, since 2006, the NINDS has led the NIH CounterACT program.

Within HHS, the ASPR develops Scenario-Based and Product-Specific Requirement Documents to help the HHS set research priorities to develop products to protect the American public from a range of

chemical threat agents. One Requirement Document under consideration is for neuroprotectants to mitigate long-term neurological health effects following exposure to nerve agents.

Protocol Development: CounterACT Workshop; Systematic Review Subcommittee

A subcommittee within CounterACT subsequently developed this project [Systematic Review Subcommittee (SRS) chaired by David A. Jett, Ph.D.; Pamela J. Lein Ph.D., and Mark Kirk M.D.; See Contributors for list of subcommittee members].

The research question and specific aims stated above were developed and refined through a series of problem formulation steps including: (1) consideration of reports in the literature as to whether or not long-term neurological effects of OP agents in humans are a common occurrence; (2) discussion of these reports and related questions on long-term neurological health effects of OP agents at a workshop convened in February 2014 consideration at the CounterACT Workshop in February 2014; (3) development of the Systematic Review Subcommittee (chaired by David A. Jett, Ph.D.; Pamela J. Lein Ph.D., and Mark Kirk M.D). The Systematic Review Subcommittee was convened in June 2015 and began to refine the research question, specific aims, and develop a draft protocol for conducting the systematic review. The focus of the current project on sarin, rather than all OP agents, was selected to aid in reaching conclusions (i.e., equivalent exposure to the same agent could be more directly compared than exposure across multiple agents). Similarly, the focus on neurological health effects was selected as an aid to reaching conclusions and because the majority of data on potential health effects of sarin are for neurological effects. The protocol and literature search strategy were reviewed by technical and medical experts who have backgrounds in toxicology and in the review of toxicological evidence of harm.

Following preliminary literature searches, the committee determined that there is sufficient evidence for a systematic review of the long-term neurological effects of sarin exposure in humans and animals. This evaluation protocol was developed with detailed methods for conducting the systematic review based on problem formulation activities, feedback received, and following the OHAT Handbook for Conducting a Literature-Based Health Assessment (National Toxicology Program 2015).

Consideration of key scientific issues

Several key scientific issues were identified during problem formulation. A summary of the most important issues and how these will be addressed in the evaluation are summarized below.

1. **The relevance of chronic exposures to sarin.**

   There are reports of chronic human exposures to sarin, for example, exposures in military or occupational settings. We have limited this literature review to acute exposures only. The rationale behind this inclusion/exclusion criterion is: (1) it is unknown if the underlying mechanisms and pathophysiological outcomes for multiple or chronic exposures are the same as those for single acute exposures, both in terms of acute or long-term neurological effects; and (2) developing medical countermeasures for multiple or chronic exposures to chemical threat agents is beyond the scope of the NIH CounterACT program. In some cases, multiple exposures may occur over a short period of time, for example, if soldiers are exposed to a cloud of sarin and cannot evacuate the area quickly. We will include these types of studies in the search and use them after the systematic review to determine if they support or refute the findings. We will only evaluate studies where the multiple exposures occur within a 24-hour period.

2. **The lack of definitive dose-exposure data in human studies.**

   Findings from the preliminary evaluation suggest that even when exposure is definitive in human studies, dose is usually unknown. There are no equivalent problems with dose-exposure data in animal studies, since all are experimental studies in which dose and route of administration are known. When possible, this evaluation will extract data for human dose-exposure by utilizing any of the methods listed in Table 1. Data for dose-exposure will be extrapolated by applying the approach established by Brown and Brix (1998), as adapted from Namba et al. (1971), as follows:

   > "We define exposure levels into three categories based empirically on the associated signs and symptoms observed in exposed humans or laboratory mammals:
   >
   > (i) High-level exposure: an exposure to either a single dose or a series of smaller doses leading to a cumulative effect causing acute cholinergic signs and symptoms, which can include miosis, rhinorrhea, apnea, convulsions and death. Humans experiencing such acute signs and symptoms would likely seek hospital evaluation and possibly emergency treatment.
   >
   > (ii) Intermediate-level exposure: a single or multiple exposures causing minimum, threshold acute cholinergic effects with signs and symptoms generally limited to miosis, rhinorrhea or clinically detectable depression in cholinesterase levels. In humans, such effects might result in hospital examination, such as in state-run occupation programs that monitor cholinesterase levels in potentially exposed workers.
   >
   > (iii) Low-level exposure: a single or multiple exposures leading to no clinically detectable cholinergic signs or symptoms, including depressed cholinesterase levels. For example, US federal guidelines for general population exposure to military OP nerve agents are based on values for minimum clinically detectable cholinergic effects (intermediate-level exposure) reduced by a conservative (*e.g*. 100-fold) safety factor to ensure that no such effects will occur in any individual exposed to that level."[3]

3. **Difficulty documenting human neuropathology following exposures that result in seizures.**

   Although it is possible to document seizures following high-level exposure to sarin, it is difficult to definitively link aberrant neuropathology to exposure. We will address this issue in part by analyzing the subset of studies in which seizures were known to have occurred, and/or in which clinical and/or subclinical neuropathology was reported.

Step 2. Search For and Select Studies for Inclusion

Literature Search Strategy

Search terms were developed to identify all relevant published evidence that addresses the research question on long-term neurological health effects potentially associated with acute, sublethal exposure

[3] United States Department of Health and Human Services, Centers for Disease Control, Public Health Service, Final recommendations for protecting the health and safety against potential adverse effects of long-term exposure to low doses of agents: GA, GB, VX, mustard agent (H, HD, T) and Lewisite (L), 53 F.R. 8504 (15 March 1988).

to sarin in humans and animals by (1) using the search term "sarin" and related synonyms "GB" or sarin's IUPAC ID "(RS)-propan-2-yl methylphosphonofluoridate, (2) without restriction by health outcome or key words to identify long-term neurological effects. A test set of relevant studies was used to ensure that the search terms retrieved 100% of the test set. The following eight electronic databases will be searched using a search strategy tailored for each database by an informationist on the evaluation team (details presented in **Appendix 1**). No language restrictions or publication year limits will be imposed and the literature search will be updated for a final time approximately 90-120 days prior to peer-review.

Databases Searched

- Cochrane Library
- DITIC
- EMBASE
- NIOSHTIC
- PubMed
- Scopus
- Toxline
- Web of Science

Searching other resources

We will use the following methods to find studies that would not be identified through the electronic searches. Studies will be evaluated against the same inclusion and exclusion criteria as used for screening records retrieved from the electronic search. Relevant studies identified through these steps will be marked as "provided from other sources" in a study selection flow diagram. Manual searching will be conducted by:

- Reviewing the reference lists of relevant reviews or reports.
- Reviewing commentaries or letters on specific studies to consider whether they contain content that meet inclusion criteria.
- Searching the reference lists of all included studies after the full text review.

Per the expanded hierarchy of evidence for human studies, original papers may include non-peer-reviewed studies, for example, reports from US military observational studies, as well as uncontrolled studies, case series, case reports, or social media. In all instances, the paper or social media source must: (1) document exposure to sarin; (2) confirm both acute symptoms, *i.e.*, cholinergic crisis; and (3) assess and report some long-term neurological health effects from the exposure.

Unpublished data

This evaluation will only include publications that have been publicly disclosed and are available to the public so that they can be transparently reviewed and evaluated.

Screening Process

References retrieved from the literature search will be screened for relevance and eligibility using DistillerSR®, a web-based, systematic-review software program with structured forms and procedures to

ensure standardization of the process[4]. Search results will first be consolidated in Endnote reference management software and duplicate articles will be removed prior to uploading the references into DistillerSR®.

Evidence Selection Criteria

In order to be eligible for inclusion, studies must comply with the type of evidence specified by the PECO statements (Table 1, Table 2, Table 3). Inclusion and exclusion criteria based on the PECO statements are detailed in Table 4. These criteria will be used to screen articles for relevance and eligibility at both the title-and-abstract and full-text screening stages. In addition to criteria defining the relevant population, exposure, comparator, and outcomes, Table 4 defines criteria for relevant publications types (*e.g.*, the report must contain original data). Studies that do not meet these criteria will be excluded. Some articles may be categorized as possible supportive material if they appear inappropriate for inclusion, but appear to contain relevant background information.

Table 4. Inclusion and Exclusion Criteria to Determine Study Eligibility

	Inclusion Criteria	Exclusion Criteria (or blank if none)
Population (Human Studies or Experimental Model Systems)		
Human	• No restrictions on sex, age, or life stage at exposure or outcome assessment	
Animal	• No restrictions on sex, age, species, or life stage at exposure or outcome assessment	• Animal observational/wildlife studies
***In vitro*/ mechanistic**	• Mechanistic studies will be restricted to human or animal cells, tissues or model systems with in vitro exposure regimens that examine the survival and/or morphology of neurons and/or glia	• Studies in non-animal organisms
Exposure		
Human	• Studies with an exposure duration less than 24 hours, which may include several doses within a time period less than 24 hours; studies that examine exposures to sarin greater than 24 hours will be used if the outcome is first measured after 24 hours but before any exposure that occurs after 24 hours Exposure to sarin based on: • known dose or concentration in an experimental protocol • diagnostic biomonitoring data (*e.g.*, sarin or biomarkers in plasma or urine) • environmental detection (e.g., air, soil)	• Studies with an exposure duration greater than or equal to 24 hours, unless the outcome is first measured at least 24 hours after the first dose but before any exposure that occurs after 24 hours • Controlled studies where the purpose was only to apply treatment for acute sarin effects • Exposures where cholinergic signs were monitored but not observed

[4]DistillerSR® (https://www.evidencepartners.com/products/distillersr-systematic-review-software/) is a proprietary project management tool for tracking studies through the screening process and storing data extracted from these studies using user-customized forms.

Table 4. Inclusion and Exclusion Criteria to Determine Study Eligibility		
	Inclusion Criteria	**Exclusion Criteria (or blank if none)**
	• corroboration by assessment of direct (in hospital, in clinic) or indirect observation of symptoms of acute cholinergic signs (video or reported by patient's family) where sarin exposure was suspected • dose may be extrapolated from cholinergic clinical signs and symptoms per Brown and Brix (1998) and as adapted from (Namba et al. 1971) where sarin exposure was suspected	• Acute exposures to several different chemicals • Multiple or chronic exposures
Animal	• Studies with an exposure duration less than 24 hours, which may include several doses within a time period less than 24 hours; studies that administer several doses of sarin during a time period greater than 24 hours, but the outcome is first measured greater than 24 hours after the first dose but before any doses after 24 hours • Exposure to sarin based on known administered dose or concentration	• Studies with an exposure duration greater than or equal to 24 hours, unless if the outcome is first measured at least 24 hours after the first dose but before any doses after 24 hours • Treatment/recovery studies that administer sarin and a treatment, unless if there is a sarin-only control group • Acute or subacute exposures to several different chemicals • Multiple or chronic exposures
***In vitro*/ mechanistic**	• Exposure to sarin based on known administered dose or concentration	• Acute or subacute exposures to several different chemicals
Comparators		
Human	• Humans not exposed to sarin	
Animals	• Comparable untreated animal subjects or animals exposed to vehicle-only treatment	
***In vitro*/ mechanistic**	• Study must include vehicle only control group	
Outcomes		
Human	• Neurological outcomes measured 24 hours after exposure including PTSD, acetylcholinesterase levels, and changes in nervous system function (e.g., cognitive, sensory, motor), and neuropathology (e.g., imaging and post-mortem)	• No assessment of health effect outcomes after cholinergic crisis has subsided • Neurological effects only measured within 24 hours after exposure
Animal	• Neurological outcomes including changes in nervous system function (e.g., cognitive, sensory, motor), and neuropathology (e.g., imaging and post-mortem)	• Neurological effects only measured within 24 hours after exposure, including studies where animals were sacrificed at 24 hours

Table 4. Inclusion and Exclusion Criteria to Determine Study Eligibility		
	Inclusion Criteria	**Exclusion Criteria (or blank if none)**
	• Neurological effects measured more than 24 hours after exposure	
***In vitro/* mechanistic**	• Measurements of the survival and the morphology of neurons or glia, including neurohistochemical and immunohistochemical techniques such as H&E, Nissl, Rapid Golgi, Fluoro-Jade, Silver Stain, HRP, GFAP, neurotransmitter stains, axon/dendrite-specific markers, and others	
Publication Type (e.g., specify any language restrictions, use of conference abstracts, etc.)		
Human, animal, or *in vitro/* mechanistic	• Report must contain original data in whole or in part relevant to the aims of this evaluation • Reference to the original report must be in the public domain, *i.e.* rule out classified documents • May be written, video, or other social media report. Where the source data is not written, there must be a contemporaneous written description of original data by an independent medical expert, which description must also include a description of assessment methodology	• 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		

Outcome measures

Scoping activities have revealed that most of the reported long-term effects from sarin exposure are related to neurological function. Therefore the types of health outcome measures that this evaluation will assess are neurological, including changes in nervous system form and function. In both humans and animals, outcome measures may include measures of cognitive, sensory, and motor function. Measures of neuropathology will include nervous system imaging and post-mortem assessment of neuropathology. For animal studies, although there is within-group homogeneity, there appears to be only low- to moderate homogeneity of data elements common to health outcomes in humans. We will therefore select animal studies that are well-designed based on dose-exposure, sample size and on our definition of "long-term neurological health" outcomes, not necessarily on the degree to which a particular sign or symptom may recapitulate the human condition. The predictive value of outcomes will be considered further in deciding whether or not to downgrade evidence for indirectness when rating the confidence in the body of evidence (Figure 1).

Multiple publications of same data

If we identify multiple publications with overlapping data for the same study (*e.g*., publications reporting subgroups, additional outcomes, or exposures outside the scope of an evaluation, or longer follow-up) by examining author affiliations, study designs, cohort name, enrollment criteria, and enrollment dates, if necessary and if possible, we will contact study authors to clarify any uncertainty about the independence of two or more articles. In the case of multiple publications, this review will include all publications on the study, but will select one study to use as the primary study. The other references will be considered as secondary publications, annotated clearly to show relationship to the primary record

during data extraction. In general, the primary study will be the publication with the longest follow-up, or for studies with equivalent follow-up periods, the most recent study or one with the largest number of cases. Relevant original data from all publications of the study will be included, although if the same outcome is reported in more than one report, duplicate data will be excluded.

Title/Abstract Review

There will be two independent screeners for all of the three streams of human, animal and mechanistic evidence. The screeners will be trained using project-specific written instructions that reflect the criteria outlined in Table 4. The screening process will begin with an initial pilot phase to improve clarity of the inclusion and exclusion instructions and/or to improve accuracy and consistency among screeners. If changes to the inclusion criteria are made based on the pilot phase, they will be documented in a protocol amendment along with the date any modifications are made, and the logic for the changes. The trained screeners will then conduct a title and abstract screen of the search results, including any results of manual searches, to determine whether a reference meets the inclusion or exclusion criteria.

Studies that are not excluded based on the title and abstract will be screened through a full-text review. In case of screening conflicts, screeners will independently review their screening results to confirm the inclusion/exclusion decision and, if needed, discuss discrepancies with the other screener. If a true disagreement exists between screeners, the study will pass to the full-text review.

Full-Text Review

After completion of the title/abstract screen, full-text articles will be retrieved for those studies that either clearly meet the inclusion criteria or where eligibility to meet the inclusion criteria is unclear. Full-text review will be independently conducted by the two screeners who participated in the title/abstract screening. Any disagreements will be resolved by discussion through consultation with members of the Systematic Review Subcommittee team and technical advisors.

Tracking study eligibility and reporting the flow of information

The main reason for exclusion at the full-text-review stage will be annotated and reported in a study selection flow diagram in the final report (using reporting practices outlined in Moher et al. 2009). The following reasons for exclusion will be documented: (1) does not distinguish exposure to sarin from other possible co-exposures; (2) no reliable data on long-term neurological health outcomes; (3) lacks original data describing long-term neurological health effects, *e.g.*, is a review; or (4) is a conference abstract or other brief report lacking detailed methods and results.

Release of the list of included studies

The list of included studies will be posted on the CounterACT website once screening has been completed and prior to completion of the draft Systematic Review.

Step 3. Data Extraction

Data Extraction Process and Data Warehousing

Data extraction will be managed using the Health Assessment Workspace Collaborative (HAWC), an open source and freely available web-based interface application.[5] The data extraction results for

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

included studies will be visualized and made publicly available in Excel format upon publication of the final Systematic Review on the HAWC Project website.

The extracted data will be used to help summarize study designs and findings, facilitate assessment of risk of bias, and/or conduct statistical analyses during evidence synthesis. The number of elements or collection of information on a specific element may be revised following the identification of important study details from individual studies included in the review (**Appendix** 2 and **Appendix 3**). Data extraction will be performed by one member of the evaluation team and checked by a second member for completeness and accuracy. Any discrepancies in data extraction will be resolved by discussion or consultation with a third member of the evaluation team. Information that is inferred, converted, or estimated during data extraction will be marked as annotated, *e.g*. using brackets [n=10]. A member of the review team or contractors will attempt to contact authors of included studies to obtain missing data considered important for evaluating key study findings (e.g., level of data required to conduct a meta-analysis). The evaluation report will note that an attempt to contact study authors was unsuccessful if study researchers do not respond to an email or phone request within one month of the attempt to contact.

Step 4. Quality Assessment of Individual Studies

Internal validity or risk of bias will be assessed for individual studies using a tool developed by OHAT that outlines a parallel approach to evaluating risk of bias from human, animal, and mechanistic studies to facilitate consideration of risk of bias across evidence streams with common terms and categories. The risk-of-bias tool is comprised of a common set of eleven questions that are answered based on the specific details of individual studies to develop risk-of-bias ratings (using the four options in Table 5 and the eleven question in Table 6) for each question. Study design determines the subset of questions that should be used to assess risk of bias for an individual study. For example, risk-of-bias questions applicable to all of the experimental study designs includes a question on randomization of exposure that would not be applicable to observational study designs. Therefore, a similar set of questions will be used across experimental study designs (experimental animal, human uncontrolled, and human controlled studies; there are no human prospective “trials,” only observational studies).

Studies will be independently assessed by two assessors who answer all applicable risk-of-bias questions with one of four options in Table 5 following pre-specified criteria detailed in Appendix 4. The criteria describe aspects of study design, conduct, and reporting required to reach risk-of-bias ratings for each question and specify factors that can distinguish among ratings (*e.g*., what separates “definitely low” from “probably low” risk of bias). The instructions and detailed criteria will be tailored to the specific evidence stream and type of human study designs. Risk of bias will be assessed at the outcome level because study design or method specifics may increase the risk of bias for some outcomes and not others within the same study.

Table 5. Answers to the Risk-of-Bias Questions Result in One of Four Risk-of-Bias Ratings	
++	***Definitely Low*** *risk of bias*: There is direct evidence of low risk of bias practices
+	***Probably Low*** *risk of bias*: There is indirect evidence of low risk of bias practices or it is deemed that deviations from low risk of bias practices for these criteria during the study would not appreciably bias results, including consideration of direction and magnitude of bias

Table 5. Answers to the Risk-of-Bias Questions Result in One of Four Risk-of-Bias Ratings	
– NR	***Probably High*** *risk of bias*: There is indirect evidence of high risk of bias practices (indicated with "-") or there is insufficient information provided about relevant risk of bias practices (indicated with "NR" for not reported). Both symbols indicate probably high risk of bias
––	***Definitely High*** *risk of bias*: There is direct evidence of high risk of bias practices

Risk-of-Bias Assessment Process

Two assessors will be trained using the criteria in Appendix 4 with an initial pilot phase undertaken to improve clarity of criteria that distinguish between adjacent ratings, and to improve consistency among assessors. The two assessors may also be screeners, one analyst and one independent non-expert.

All assessors involved in the risk-of-bias 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 more clearly distinguish between adjacent ratings. If major changes to the risk-of-bias 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 modifications are made, and the logic for the changes. It is also expected that information about confounding, exposure characterization, outcome assessment, and other important issues may be identified during or after data extraction, which can lead to further refinement of the risk-of-bias criteria (Sterne et al. 2014).

Table 6. Questions in Risk-of-Bias Assessment and Applicability by Study Design

	Experimental Animal*	Human Controlled Trials**	Cohort	Case-Control	Cross-Sectional***	Case Series
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 the appropriate comparison groups?			X	X	X	
4. Did study design or analysis account for important confounding and modifying variables?			X	X	X	X
5. Were experimental conditions identical across study groups?	X					
6. Were research personnel blinded to the study group during the study?	X	X				
7. Were outcome data complete without attrition or exclusion from analysis?	X	X	X	X	X	X
8. Can we be confident in the exposure characterization?	X	X	X	X	X	X
9. Can we be confident in the outcome assessment (including blinding of outcome assessors)?	X	X	X	X	X	X
10. Were all measured outcomes reported?	X	X	X	X	X	X
11. Were there no other potential threats to internal validity?	X	X	X	X	X	X

After assessors have independently made risk-of-bias determinations for a study across all risk-of-bias 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 Systematic Review Subcommittee and/or technical advisors. The final risk-of-bias rating for each question will be recorded along with a statement of the basis for that rating. The risk-of-bias assessment of included studies will be part of the study summaries released in materials for the Systematic Review that will be released as part of the final report.

Missing Information for Risk of Bias Assessment

Staff involved with this Systematic Review 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 reviewers do not receive a response from the authors by one month of the contact attempt, a risk-of-bias response of "NR" for "not reported; probable 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.

Step 5. Organizing and Rating Confidence in Bodies of Evidence

The Systematic Review Subcommittee, analysts and advisors will consider the collection of studies on long-term neurological health outcomes as bodies of evidence and will develop overall confidence ratings in these bodies of evidence using a modification of the GRADE framework (Grades of Recommendation, Assessment, Development and Evaluation). Procedures for grouping long-term neurological effects, considering quantitative or narrative synthesis, and developing confidence ratings for this evaluation are described below.

Health Outcome and Endpoint Grouping

The main category for long-term neurological health outcomes includes all neurological effects. After the data are collected, neurophysiological and behavioural data will be grouped for related endpoints. Technical advisors and subject matter experts will be consulted to determine: 1) endpoints that can be grouped as similar or related endpoints, and 2) if downgrades are warranted based on the reliability or quality of specific endpoints or groups of endpoints for determining neurological effects.

Considerations for Pursuing a Narrative or Quantitative Evidence Synthesis

Preliminary findings of the human evidence suggest that within- and between-group heterogeneity is so high that only a narrative, not a quantitative or meta-analysis, is appropriate for evidence integration. Summaries of main characteristics for each included study will be compiled and reviewed by two reviewers to determine comparability between studies, and to identify data transformations necessary to ensure comparability. The main characteristics considered across all eligible studies will include the following:

Human Studies

- Study design (*e.g.*, cross-sectional, cohort; age and gender in study group and comparators)
- Details on how participants were classified into exposure groups
- Details on source of exposure data (*e.g.*, questionnaire, area monitoring, biomonitoring)
- Health outcome(s) reported, whether self-reported or evaluated using independent physiological, functional, or cognitive tests

- Subset of health outcomes reported using independent tests of post-exposure pathology, whether clinically observable or not
- Conditioning variables in the analysis (*e.g.*, variables considered confounders)
- Type of data (*e.g.*, continuous or dichotomous), statistics presented in paper, access to raw data
- Variation in degree of risk of bias at individual study level

Animal Studies

- Experimental design (*e.g.,* acute, chronic)
- Animal model used (*e.g.,* species, strain, sex, genetic background)
- Age of animals (*e.g.,* at start of treatment, mating, and/or pregnancy status)
- Developmental stage of animals at treatment and outcome assessment
- Dose levels, frequency of treatment, timing, duration, and exposure route
- Health outcome(s) reported
- Type of data (*e.g.,* continuous or dichotomous), statistics presented in paper, access to raw data
- Variation in degree of risk of bias at individual study level

More detailed guidance on evaluating heterogeneity, transforming or normalizing data to ensure comparability, and the process for determining whether a meta-analysis will be pursued is provided in the OHAT Handbook for Conducting a Literature-Based Health Assessment (http://ntp.niehs.nih.gov/go/38673, see STEP 5). In addition to screening and assessment by at least two independent personnel, we expect to solicit input from topic-specific experts to help assess whether studies are too heterogeneous for meta-analysis. Situations where it may not be appropriate to include a study are those in which: (1) data on exposure or outcome are too different to be combined; (2) there are concerns about high risk of bias; or (3) other circumstances may indicate that averaging study results would not produce meaningful results.

Stratified Analyses, Meta-Regression, and Publication Bias

If there is significant study-level heterogeneity, we may conduct stratified analyses or multivariate meta-regression to assess how much heterogeneity can be explained by taking into account both within- and between-study variance (Vesterinen et al. 2014). Multivariate meta-regression approaches are especially useful for assessing the significance of associations in between-study design characteristics. These approaches are considered most suitable if there are at least six to ten studies for a continuous variable and at least four studies for a categorical variable.

If there are sufficient studies to conduct a meta-analysis, we will assess potential publication bias by developing funnels and performing Egger regression on the estimates of effect size. In addition, if these methods suggest that publication bias is present, we will use trim and fill methods to predict the impact of the hypothetical "missing" studies (Vesterinen et al. 2014) .

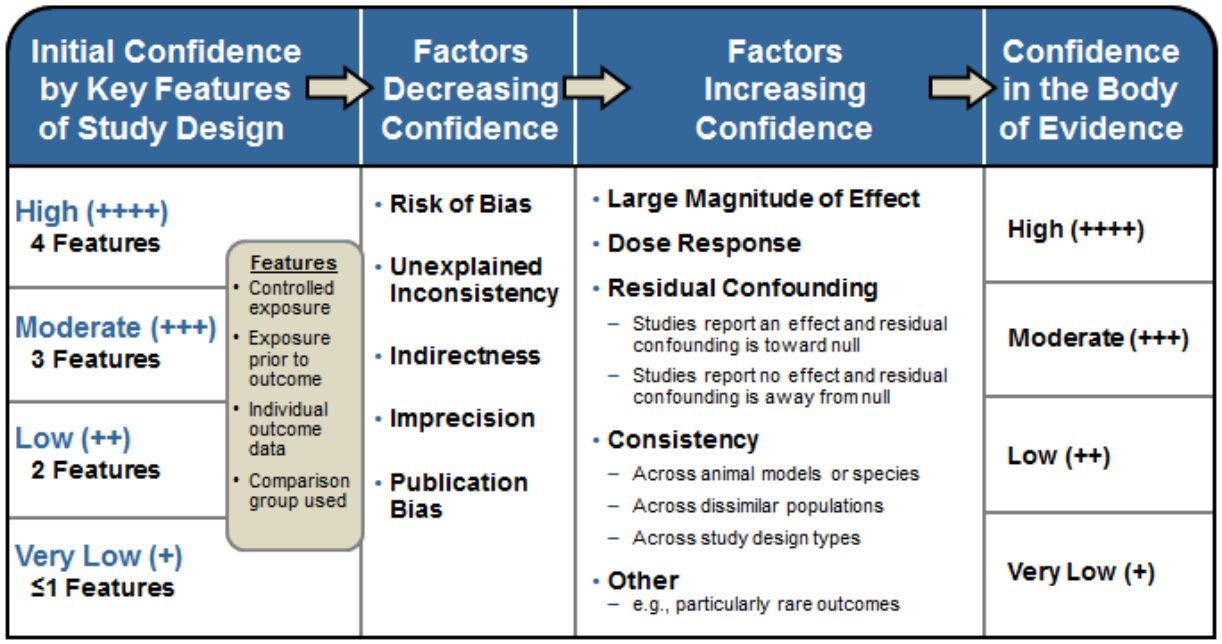
Confidence Rating: Assessment of Body of Evidence

The certainty of evidence within groups of human studies will be graded using the GRADE system for rating the confidence in the body of evidence (Guyatt et al. 2011) as used by OHAT (Rooney et al. 2014). More detailed guidance on reaching confidence ratings in the body of evidence as "high," "moderate," "low," or "very low" is provided in the OHAT Handbook for Conducting a Literature-Based Health Assessment (http://ntp.niehs.nih.gov/go/38673, see STEP 5). (National Toxicology Program 2015)

In brief, available studies on a particular outcome will be initially grouped by key study design features, and each grouping of studies will be given an initial confidence rating by those features. This initial rating (column 1 of Figure 1) will be downgraded for factors that decrease confidence in the results [column 2 of Figure 1 (risk of bias, unexplained inconsistency, indirectness or lack of applicability, imprecision, and publication bias)] and upgraded for factors that increase confidence in the results [column 3 of Figure 1 (large magnitude of effect, dose response, consistency across study designs/populations/animal models or species, consideration of residual confounding, and other factors that increase our confidence in the association or effect)].

The reasons for downgrading (or upgrading) confidence may not be due to a single domain of the body of evidence. If a decision to downgrade is borderline for two domains, the body of evidence will be downgraded once in a single domain to account for both partial concerns based on considering the key drivers of the strengths or weaknesses. Similarly, the body of evidence will not be downgraded twice for what is essentially the same limitation (or upgraded twice for the same asset) that could be considered applicable to more than one domain of the body of evidence. Consideration of consistency across study designs, human populations, or animal species is not included in the GRADE guidance (Guyatt et al. 2011); however, it will be considered here in the modified version of GRADE used by OHAT (Rooney et al. 2014).

Figure 1. Assessing Confidence in the Body of Evidence



Confidence ratings will be independently assessed by CounterACT personnel, the analyst-contractors, and the Systematic Review Subcommittee. Any discrepancies will be resolved by consensus and in consultation with technical advisors as needed. Confidence ratings will be summarized in evidence profile tables (see Table 7 for general format).

Relevance of Animal Models to Long-term neurological Effects on Human Health

- *Rats, mice, and other mammalian model systems:* Noting differences in human and animal metrics to observe symptoms and test for effects, we will consider several alternatives to extract physiological and cognitive data from animal studies so that, to the extent possible, animal data can be integrated with human data into a single evidence stream. Due to known similarities in acetylcholine function, studies conducted in mammalian model systems will be assumed to be relevant for humans (*i.e.*, not downgraded for indirectness) unless compelling data to the contrary is identified during the course of the evaluation.
- *Birds, reptiles, amphibians, fish, and other non-mammalian vertebrate model systems:* Most cell types are relatively consistent across vertebrate systems. However, use of these model systems to address human health is not as well established as use of the mammalian model systems (WHO 2012). For this reason, any studies meeting inclusion criteria conducted in non-mammalian vertebrates will be downgraded one level for indirectness.

Table 7. Evidence Profile Table Format

Example of the type of information that will be in an evidence profile for long-term neurological health outcomes

Body of Evidence	Risk of Bias	Unexplained Inconsistency	Indirectness	Imprecision	Publication Bias	Magnitude	Dose Response	Residual Confounding	Consistency Across Species/ Model	FINAL RATING
Evidence stream (human or animal)	Serious or not serious	Serious or not serious	Serious or not serious	Serious or not serious	Detected or undetected	Large or not large	Yes or no	Yes or no	Yes or no	**Final Rating**
(# Studies) Initial Rating	• Describe trend • Describe key questions • Describe issues	• Describe results in terms of consistency • Explain apparent inconsistency (if it can be explained)	• Discuss use of upstream indicators or populations with less relevance	• Discuss ability to distinguish treatment from control • Describe confidence intervals	• Discuss factors that might indicate publication bias (*e.g.*, funding, lag)	• Describe magnitude of response	• Outline evidence for or against dose response	• Address whether there is evidence that confounding would bias toward null	• Describe cross-species, model, or population consistency	**High, Moderate, or Low**

Long-term Neurological Health Outcomes

For the evaluation of long-term neurological health effects, all outcomes and effects related to neurological form and function at the individual and population level will be considered.

Exposure to sarin

- *Human studies:* All exposure levels and scenarios encountered in the human studies (*e.g.*, general population, occupational settings, *etc.)* will be considered direct and not downgraded. As noted above, a key inclusion criterion is that exposure at any level must be documented.

- *Dose levels used in animal studies:* There will be no downgrade for dose level used in experimental animal studies. We recognize that the level of dose or exposure is an important factor when considering the relevance of animal findings to human health. In addition, in OHAT's process the relevance of the dose or exposure level occurs after hazard identification as part of reaching a "level of concern" conclusion.

- *Route of administration in animal studies:* All of the most commonly used routes of administration will be considered direct for the purposes of establishing confidence ratings. We recognize that some of these exposure routes may only be relevant for certain human sub-populations. However, for this review, this consideration occurs after identifying long-term neurological health effects as part of reaching a "level of concern" conclusion.

    - Oral (no downgrade for indirectness) – Gavage, drinking water, or feeding studies are considered relevant because oral exposure is expected to be relatively rare, but is a possible source of exposure in humans (ATSDR 2015).
    - Dermal skin or eye contact (no downgrade for indirectness) – Dermal exposure studies are considered relevant because skin or eye contact is a primary route of exposure to sarin (ATSDR 2015).
    - Subcutaneous injection (no downgrade for indirectness) – These studies are not relevant to the nature of human exposure. However, because of the difficulty in modeling human exposures, and the common use and accepted standard for using parenteral methods of administration for challenging agents in rigorous well-controlled animal toxicology studies, studies using this method will not be downgraded for indirectness.
    - Inhalation (no downgrade for indirectness) – Inhalation studies are considered relevant to human exposure to sarin (ATSDR 2015).
    - Intraperitoneal injection (no downgrade for indirectness) – These studies are not relevant to the nature of human exposure. However, because of the difficulty in modeling human exposures, and the common use and accepted standard for using parenteral methods of administration for challenging agents in rigorous well-controlled animal toxicology studies, studies using this method will not be downgraded for indirectness.
    - Intramuscular injection (no downgrade for indirectness) – These studies are not relevant to the nature of human exposure. However, because of the difficulty in modeling human exposures, and the common use and accepted standard for using parenteral methods of administration for challenging agents in rigorous well-controlled animal toxicology studies, studies using this method will not be downgraded for indirectness.
    - Intravenous injection (no downgrade for indirectness) – These studies are not relevant to the nature of human exposure. However, because of the difficulty in modeling human exposures, and the common use and accepted standard for using parenteral methods of

administration for challenging agents in rigorous well-controlled animal toxicology studies, studies using this method will not be downgraded for indirectness.

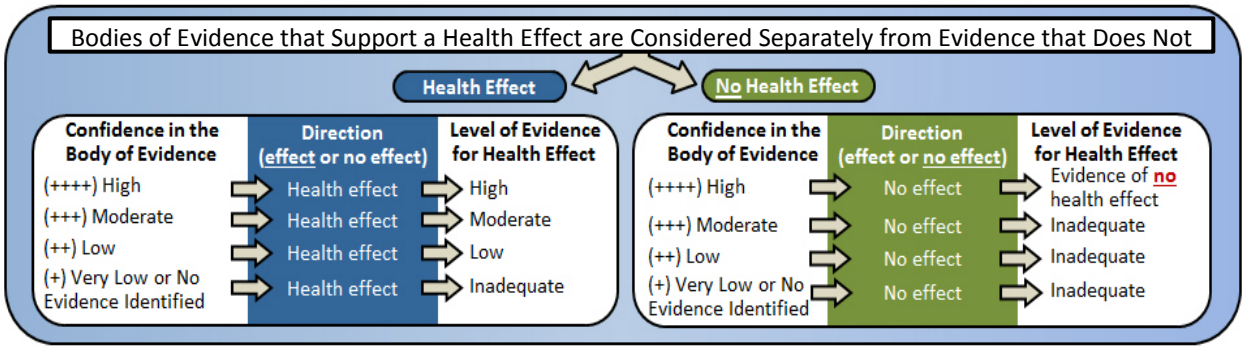
In Vitro/Mechanistic Studies

As noted by OHAT (Rooney et al 2014), the framework described above to develop confidence ratings only applies to human and animal studies. Although there is no analogous model to develop confidence ratings for other relevant data such as outcomes from *in vitro*, mechanistic, cellular or genomic studies, we will group our findings as to "established" and "emerging." The proposed approach to consider other relevant data including *in vitro* studies is described separately in a later section of this the document in Step 7 when integrating other relevant data (see "Consideration of Mechanistic Data").

Step 6. Preparation of Draft Level of Evidence Statement

The confidence ratings will be translated into draft level of evidence of health effects for each type of health outcome separately according to one of four statements: (1) High; (2) Moderate; (3) Low; or (4) Inadequate (Figure 2). The descriptor "evidence of no health effect" will be 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" will only be reached when there is high confidence in the body of evidence.

Figure 2. Translate Confidence Ratings into Evidence of Health Effect Conclusions



Evidence Descriptors	Definition
High Level of Evidence	There is high confidence in the body of evidence for an association between exposure to sarin and the health outcome(s)
Moderate Level of Evidence	There is moderate confidence in the body of evidence for an association between exposure to sarin and the health outcome(s)
Low Level of Evidence	There is low confidence in the body of evidence for an association between exposure to sarin and the health outcome(s), or no data are available
Inadequate Evidence	There is insufficient evidence available to assess if exposure to sarin is associated with the health outcome(s)
Evidence of No Health Effect	There is high confidence in the body of evidence that exposure to sarin is not associated with the health outcome(s)

Step 7. Integrate Evidence to Develop Conclusions of Long-term neurological Health Effects

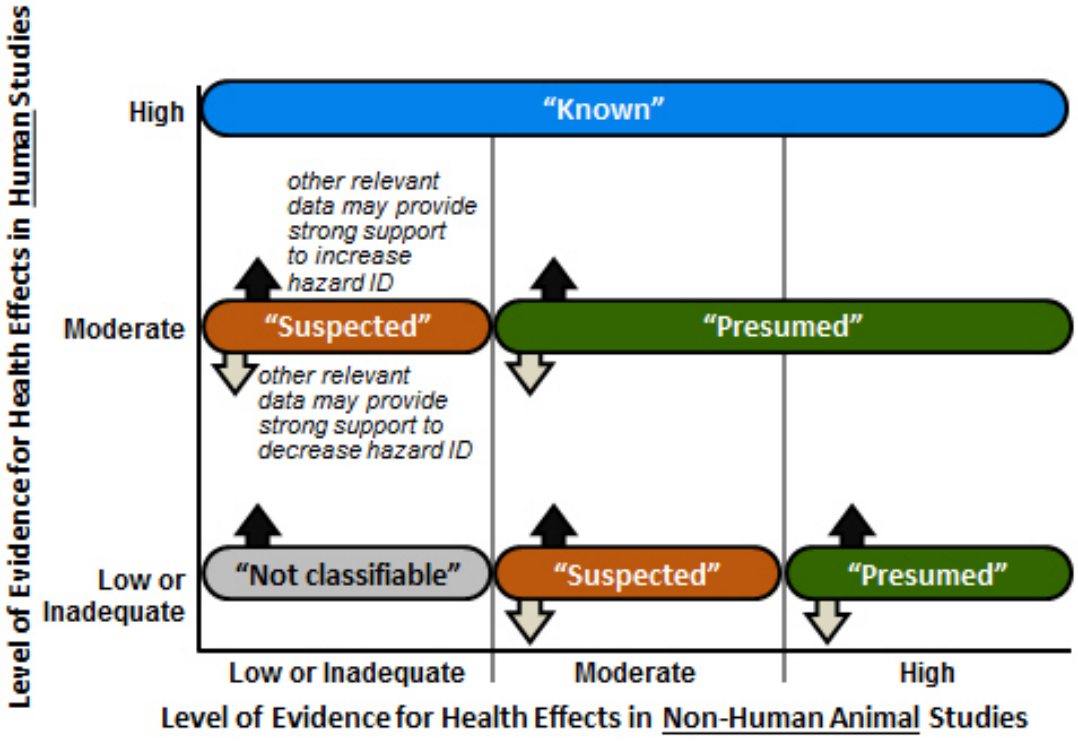
Finally, the levels of evidence ratings for human and animal data will be integrated with consideration of in vitro/mechanistic data to reach one of five possible categories of evidence of long-term neurological health effect: (1) Known, (2) Presumed, (3) Suspected, (4) Not classifiable, or (5) Not identified to be a long-term neurological effect in humans (Figure 3).

Consideration of Human and Animal Data

Initial hazard identification conclusions will be reached by integrating the highest level-of-evidence conclusion for long-term neurological health effect(s) on an outcome basis for the human and the animal evidence streams. Hazard identification conclusions may be reached on groups of biologically related outcomes or functionally related outcomes as well as more specific endpoints if data are available to make more specific conclusions. If the data support a long-term neurological health 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 animal data to reach one of four initial hazard identification conclusions as to the evidence of long-term neurological effects in humans: Known, Presumed, Suspected, or Not classifiable. If either the human or animal evidence stream is characterized as "Inadequate Evidence," then conclusions will be based on the remaining evidence stream alone (which is equivalent to treating the missing evidence stream as "Low" in Figure 3).

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 long-term neurological effect observed in humans."

Figure 3. Hazard Identification Scheme for Long-term neurological Effects



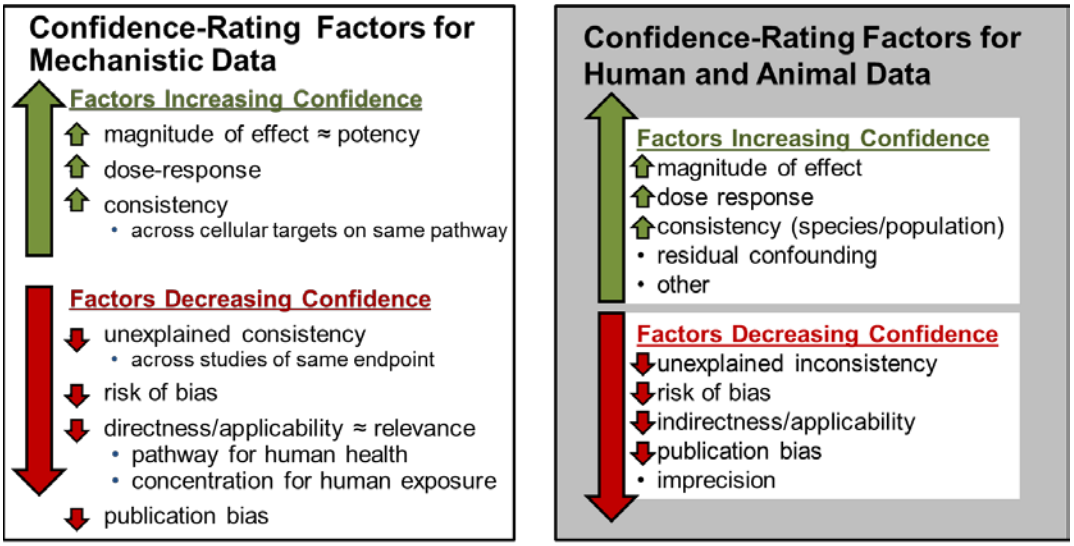
Consideration of Mechanistic Data

There is no requirement to consider mechanistic or mode-of-action data to reach a hazard identification conclusion regarding long-term neurological health effects. However, 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, genetic, and molecular mechanisms that explain how a chemical produces particular adverse effects.

For the evaluation of long-term neurological health effects associated with acute exposure to sarin, we are interested in mechanistic or *in vitro* measures that may support the biological plausibility of corresponding neurological outcomes reported from *in vivo* studies in animals or humans.

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 including the SRS, and in consultation with technical advisors as needed.

Figure 4. Factors Considered in Evaluating the Support for Biological Plausibility



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 **Figure 3**) from that initially derived by considering the human and non-human animal evidence together.

If mechanistic data provide strong opposition for biological plausibility of the relationship between exposure and the health effect, the hazard identification conclusion may be downgraded (indicated by gray "down" arrows in **Figure 3**) from that initially derived by considering the human and non-human animal evidence together.

SYSTEMATIC REVIEW: OUTLINE

This Systematic Review of the association between acute sublethal exposure to sarin and long-term neurological health effects will include the following information:

Introduction

This section will provide a brief background on the topic.

Methodology

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

- Research question
- Search strategy used to identify and retrieve studies
- Process for selecting studies
- Methods of data extraction
- Methods used to assess risk of bias of included studies
- Methods used to synthesize the data of included studies
- Methods used to evaluate confidence in the body of evidence
- Methods used to reach hazard identification conclusions for evidence of long-term neurological health effects

Results

This section will include the results from the Systematic Review of the evidence of long-term neurological health effects following acute exposure to sarin. Results will be presented in tables or figures if possible. The results from the included studies will be discussed by outcome. This will include a description of:

- The number of studies identified that examined neurological effects
- A summary of the results and risk-of-bias assessment for each included study (including files in downloadable format)
- Description of results and ratings for confidence in the bodies of evidence for long-term neurological effects where there are data linked to acute exposure to sarin using GRADE as performed by OHAT
- Evidence profiles for long-term neurological health effects where there are data linked to acute exposure to sarin
- Presentation of level of evidence and draft hazard identification conclusions for long-term neurological health effects where there are data linked to acute exposure to sarin

Discussion

The discussion will provide a summary of the review findings, including a discussion of any gaps identified in the evidence and any suggestions of areas for further research.

Any important limitations of the review will be described and their impact on the available evidence will be discussed.

Conclusion

This will present the conclusion of the review.

REFERENCES

ATSDR. 2015. "Toxic Substances Portal: Nerve Agents (GA, GB, GD, VX) In, *Agency for Toxic Substances and Disease Registry*." In, edited by Division of Toxicology and Environmental Medicine/Applied Toxicology Branch Centers for Disease Control, 404. Atlanta, CA: US Department of Health and Human Services.

Augerson, W.S. 2000. "A Review of the Scientific Literature as it Pertains to Gulf War Illnesses, Volume 5: Chemical and Biological Warfare Agents." In.: RAND Corporation.

Binns JH, Cherry N, Golomb BA, Graves JC, Haley RW, et al. . 2004. "Scientific Progress in Understanding Gulf War Veterans' Illnesses: Report and Recommendations." In. Washington, D.C.: Research Advisory Committee on Gulf War Veterans' Illnesses.

Brown, M. 2009. 'Military chemical warfare agent human subjects testing: part 2--long-term health effects among participants of U.S. military chemical warfare agent testing', *Mil Med*, 174: 1049-54.

Brown, M. A., and K. A. Brix. 1998. 'Review of health consequences from high-, intermediate- and low-level exposure to organophosphorus nerve agents', *J Appl Toxicol*, 18: 393-408.

Burchfiel, J. L., and F. H. Duffy. 1982. 'Organophosphate neurotoxicity: chronic effects of sarin on the electroencephalogram of monkey and man', *Neurobehav Toxicol Teratol*, 4: 767-78.

Defense Science Board, Department of Defense Office of the Under Secretary of Defense for Acquisition and Technology. 1994. "Report of the Defense Science Board Task Force on Persian Gulf War Health Effects." In, 105. Washington, D.C.: Department of Defense.

Guyatt, G. H., A. D. Oxman, E. A. Akl, R. Kunz, G. Vist, J. Brozek, S. Norris, Y. Falck-Ytter, P. Glasziou, H. Debeer, R. Jaeschke, D. Rind, J. Meerpohl, P. Dahm, and H. J. Schunemann. 2011. 'GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables', *Journal of Clinical Epidemiology*, 64: 383-94.

Higgins, JPT, and S Green. 2011. "Cochrane Handbook for Systematic Reviews of Interventions." In.: The Cochrane Collaboration.

IOM. 2004. "Gulf War and Health: Updated Literature Review of Sarin." In, edited by Institute of Medicine, 132. Washington, D.C.: National Academies Press.

Kassa, J., M. Pecka, M. Tichy, J. Bajgar, M. Koupilova, J. Herink, and Z. Krocova. 2001. 'Toxic effects of sarin in rats at three months following single or repeated low-level inhalation exposure', *Pharmacol Toxicol*, 88: 209-12.

Moher, D., A. Liberati, J. Tetzlaff, and D. G. Altman. 2009. 'Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement', *Journal of Clinical Epidemiology*, 62: 1006-12.

Murata, K., S. Araki, K. Yokoyama, T. Okumura, S. Ishimatsu, N. Takasu, and R. F. White. 1997. 'Asymptomatic sequelae to acute sarin poisoning in the central and autonomic nervous system 6 months after the Tokyo subway attack', *J Neurol*, 244: 601-6.

Namba, T., C. T. Nolte, J. Jackrel, and D. Grob. 1971. 'Poisoning due to organophosphate insecticides. Acute and chronic manifestations', *Am J Med*, 50: 475-92.

National Toxicology Program, NTP. 2015. "Handbook for Conducting a Literature-Based Health Assessment Using OHAT Approach for Systematic Review and Evidence Integration - January 9 2015." By National Toxicology Program. Research Triangle Park, NC: Office of Health Assessment and Translation.

Okumura, T., T. Hisaoka, T. Naito, H. Isonuma, S. Okumura, K. Miura, H. Maekawa, S. Ishimatsu, N. Takasu, and K. Suzuki. 2005. 'Acute and chronic effects of sarin exposure from the Tokyo subway incident', *Environ Toxicol Pharmacol*, 19: 447-50.

Rooney, A.A., A.L. Boyles, M.S. Wolfe, J.R. Bucher, and K.A. Thayer. 2014. 'Systematic Review and Evidence Integration for Literature-Based Environmental Health Science Assessments', *Environmental Health Perspectives*, 122: 711-18.

Sellstrom, A. (Head of Mission). 2013. "United Nations Mission to Investigate Allegations of the Use of Chemical Weapons in the Syrian Arab Republic." In *Report on the Alleged Use of Chemical Weapons in the Ghouta Area of Damascus on 21 August 2013*, 41. United Nations.

SIPRI, Swedish Peace Institute. 1975. 'Delayed Toxic Effects of Chemical Warfare Agents'.

Sterne, JAC, JPT Higgins, BC Reeves, and 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." In.

Vesterinen, H. M., E. S. Sena, K. J. Egan, T. C. Hirst, L. Churolov, G. L. Currie, A. Antonic, D. W. Howells, and M. R. Macleod. 2014. 'Meta-analysis of data from animal studies: a practical guide', *J Neurosci Methods*, 221: 92-102.

White, R. F., L. Steele, J. P. O'Callaghan, K. Sullivan, J. H. Binns, B. A. Golomb, F. E. Bloom, J. A. Bunker, F. Crawford, J. C. Graves, A. Hardie, N. Klimas, M. Knox, W. J. Meggs, J. Melling, M. A. Philbert, and R. Grashow. 2016. 'Recent research on Gulf War illness and other health problems in veterans of the 1991 Gulf War: Effects of toxicant exposures during deployment', *Cortex*, 74: 449-75.

WHO. 2012. "Guidance for Immunotoxicity Risk Assessment for Chemicals." In *IPCS Harmonization Project No. 10*. Geneva: International Programme on Chemical Safety, World Health Organization.

Yamasue, H., O. Abe, K. Kasai, M. Suga, A. Iwanami, H. Yamada, M. Tochigi, T. Ohtani, M. A. Rogers, T. Sasaki, S. Aoki, T. Kato, and N. Kato. 2007. 'Human brain structural change related to acute single exposure to sarin', *Ann Neurol*, 61: 37-46.

ABOUT THE PROTOCOL

Contributors

Systematic Review Subcommittee Team

The Systematic Review Subcommittee teams will be composed of federal staff and contractor staff. Federal staff members will do a self-evaluation for conflict of interest. Epidemiologists and toxicologists on this evaluation team should have at least three years of experience and/or training in reviewing studies, including summarizing studies and critical review (*e.g.*, assessing study quality and interpreting findings). Experience in evaluating neurotoxicological or clinical studies involving neurological harms will be preferred. Team members should have at least a master's degree or equivalent in epidemiology, neurotoxicology, translational development of novel compounds, or a related field.

Name	Affiliation
David A. Jett, Ph.D.	CounterACT, CounterACT, Project co-Lead
Andrew A. Rooney Ph.D.	NIEHS, DNTP, OHAT, Project co-lead
Pamela J. Lein, Ph.D.	UC Davis, Workshop Co-Chair
Mark Kirk, M.D.	DHS, Workshop Co-Chair
Jonathan Newmark, M.D.	DHS
Alicia Livinski	NIH Library
Elaine Wencil, Ph.D	HHS/ASPR

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
Louise Assem, Ph.D.	ICF
Robyn Blain, Ph.D.	ICF
Susan B. Goldhaber, M.P.H	ICF
Ali Goldstone, M.P.H.	ICF
Pamela Hartman, M.E.M.	ICF
Kaedra Jones, M.P.H.	ICF
Kristen Magnuson, M.E.S.M.	ICF
Maureen Malloy	ICF
Constance McKee, M.B.A.	i2 Grants Associates, LLC
Christina Niemeyer, Ph.D.	i2 Grants Associates, LLC
Pam Ross, M.S.P.H.	ICF
Robert Shin, M.H.S.	ICF
Christopher Sibrizzi, M.P.H.	ICF
Courtney Lemeris	ICF
Nicole Vetter, M.L.S.	ICF
Ashley R. Williams, M.S.E.E.	ICF

Technical Advisors

Technical advisors will be outside experts retained on an as-needed basis to provide individual advice to the Systematic Review Subcommittee and analysts for a specific topic. The technical advisors will be selected for their experience with the health effects of sarin, its neurotoxicity, and Systematic Review procedures. Technical advisors will be screened for conflict of interest (COI) prior to their service and will be asked to report any conflicts of interest. Service as a technical advisor will not necessarily indicate that an advisor has read the entire protocol or endorses the final document.

Name	Affiliation
Roberta Scherer, PhD	Johns Hopkins University
Jonathan Newmark	DOD/DHS (retired)

Sources of Support

[NIH CounterACT Program, NIH Office of the Director, National Institute of Neurological Disorders and Stroke (NINDS)]

Protocol History and Revisions

Date	Activity or revision
February 26-27, 2014:	**Problem formulation:** outcome of CounterACT workshop. Posted online and circulated internally for comment/review – impetus for this systematic review
January 5, 2016	**Draft evaluation protocol reviewed:** sent to technical advisors for peer reviewer by Drs. Madsen and Scherer

APPENDICES

Appendix 1. Literature Search Strategy

This should only include the search strings for each database.

Database	Search Terms
Cochrane Library	(sarin or zarin or "o Isopropylmethyl Phosphonofluoridate" or "ortho Isopropylmethyl Phosphonofluoridate" or "ortho-Isopropylmethyl Phosphonofluoridate" or "Isopropyl methylphosphonofluoridate" or "Isopropyl Methylfluorophosphonate" or "(RS)-propan-2-yl methylphosphonofluoridate" or (GB and organophos*) or (GB and nerve)):ti,ab,kw
Embase	'sarin':ab,ti OR 'sarin'/exp OR 'o isopropylmethyl phosphonofluoridate':ab,ti OR 'ortho isopropylmethyl phosphonofluoridate':ab,ti OR 'ortho-isopropylmethyl phosphonofluoridate':ab,ti OR 'isopropyl methylphosphonofluoridate'/exp OR 'isopropyl methylphosphonofluoridate':ab,ti OR 'isopropyl methylfluorophosphonate':ab,ti OR '(rs)-propan-2-yl methylphosphonofluoridate':ab,ti OR (gb:ab,ti AND organophos*:ab,ti) OR (gb:ab,ti AND nerve:ab,ti)
NIOSHTIC-2 Publications	Sarin[title] OR sarin[abstract] OR zarin[title] OR zarin[abstract]
NIOSHTIC-2 Publications	GB[abstract] AND nerve[abstract]
NIOSHTIC-2 Publications	GB[title] AND nerve[title]
NIOSHTIC-2 Publications	GB[abstract] AND organophos*[abstract]
NIOSHTIC-2 Publications	GB[title] AND organophos*[title]
PubMed/MEDLINE	(sarin[tiab] OR sarin[mesh] OR zarin[tiab] OR “o Isopropylmethyl Phosphonofluoridate”[tiab] OR “ortho Isopropylmethyl Phosphonofluoridate”[tiab] OR “ortho-Isopropylmethyl Phosphonofluoridate”[tiab] OR “Isopropyl methylphosphonofluoridate”[tiab] OR “Isopropyl Methylfluorophosphonate”[tiab] OR “(RS)-propan-2-yl methylphosphonofluoridate”[tiab] OR (GB[tiab] AND organophos*[tiab]) OR (GB[tiab] AND nerve[tiab]))
Scopus	Title-Abs-Key((sarin OR zarin OR {o Isopropylmethyl Phosphonofluoridate} OR {ortho Isopropylmethyl Phosphonofluoridate} OR {ortho-Isopropylmethyl Phosphonofluoridate} OR {Isopropyl methylphosphonofluoridate} OR {Isopropyl Methylfluorophosphonate} OR {(RS)-propan-2-yl methylphosphonofluoridate} OR (GB AND organophos*) OR (GB AND nerve)))
TOXLINE	Title: sarin[ti] OR sarin[mh] OR zarin[ti] OR “o Isopropylmethyl Phosphonofluoridate”[ti] OR “ortho Isopropylmethyl

Database	Search Terms
	Phosphonofluoridate"[ti] OR "ortho-Isopropylmethyl Phosphonofluoridate"[ti] OR "Isopropyl methylphosphonofluoridate"[ti] OR "Isopropyl Methylfluorophosphonate"[ti] OR "(RS)-propan-2-yl methylphosphonofluoridate"[ti] OR (GB[ti] AND organophos*[ti]) OR (GB[ti] AND nerve[ti])
Toxline	Abstract: sarin[ab] OR sarin[mh] OR zarin[ab] OR "o Isopropylmethyl Phosphonofluoridate"[ab] OR "ortho Isopropylmethyl Phosphonofluoridate"[ab] OR "ortho-Isopropylmethyl Phosphonofluoridate"[ab] OR "Isopropyl methylphosphonofluoridate"[ab] OR "Isopropyl Methylfluorophosphonate"[ab] OR "(RS)-propan-2-yl methylphosphonofluoridate"[ab] OR (GB[ab] AND organophos*[ab]) OR (GB[ab] AND nerve[ab])
Web of Science	TS=(sarin OR zarin OR "o Isopropylmethyl Phosphonofluoridate" OR "ortho Isopropylmethyl Phosphonofluoridate" OR "ortho-Isopropylmethyl Phosphonofluoridate" OR "Isopropyl methylphosphonofluoridate" OR "Isopropyl Methylfluorophosphonate" OR "(RS)-propan-2-yl methylphosphonofluoridate" OR (GB AND organophos*) OR (GB AND nerve))

Appendix 2. Data Extraction Elements for Human Studies

HUMAN	
Funding	Funding source(s)
	Reporting of conflict of interest (COI) by authors (*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 life stage at exposure and at 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., cardiovascular
	Health outcome, e.g., blood pressure (*reporting bias)
	Diagnostic or methods used to measure health outcome (*information bias)

HUMAN	
	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.
Items marked with an asterisk (*) are examples of items that can be used to assess internal validity/risk of bias	

Appendix 3. Data Extraction Elements for Animal Studies

ANIMAL	
Funding	Funding source(s)
	Reporting of COI by authors (*reporting bias)
Animal Model	Sex
	Species
	Strain
	Source of animals
	Age or lifestage at start of dosing and at health outcome assessment
	Diet and husbandry information (e.g., diet name/source)
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)
	Duration of dosing (e.g., minutes, hours)
Methods	Study design (e.g., single acute treatment)
	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
	Report on data from positive controls – was expected response observed? (*information bias)
	Endpoint health category (e.g., reproductive)
	Endpoint (*e.g.*, infertility)
	Diagnostic or method to measure endpoint (*information bias)
	Statistical methods (*information bias)
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).

ANIMAL	
	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, 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 NOEL does not necessarily mean zero response. Ideally, the response rate 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 to 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 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)
	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, and statistical result conversions, etc.
Items marked with an asterisk (*) are examples of items that can be used to assess internal validity/risk of bias	

Appendix 4. Risk of Bias Criteria

The OHAT risk of bias tool (version date January 2015 and available at http://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 to apply to each study design (Table 6).

The specific criteria used to assess risk of bias for this evaluation are outlined below for human/observational studies and for experimental animal studies.

All comments below on bias refer to a single agent, sarin, in human, animal, and mechanistic studies.

Observational Studies (Human studies)

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?

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 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, • **Note:** A study will 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)
Probably Low Risk of Bias (+)
• Indirect evidence that subjects (both exposed and non-exposed) were similar (*e.g.*, recruited from the same eligible 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 would not appreciably bias results
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 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)
Definitely High Risk of Bias (--)
• Direct evidence that subjects (both exposed and non-exposed) were not similar, recruited within very different time frames, or had very different participation/response rates

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

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, propensity scoring, 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 were assessed using valid and reliable measurements, • **AND** there is direct evidence that other exposures anticipated to bias results were not present or were appropriately measured and adjusted for. In occupational studies or studies of contaminated sites, other chemical exposures known to be associated with those settings were appropriately considered.

Definitely Low Risk of Bias (++) *continued*
• **Note:** The following variables should be considered as potential confounders and/or effect measure modifiers for the relationship between exposure and outcomes: age, sex, race/ethnicity, smoking, body mass index, alcohol consumption, and variables that represent socioeconomic status (*e.g.*, educational level, household income) based on prior reports of associations with exposure levels and outcomes involving sarin.
Probably Low Risk of Bias (+)
• Indirect evidence that appropriate adjustments were made, • **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. • **Note:** this includes insufficient information provided on co-exposures in general population studies.
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 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), • **OR** there is indirect evidence that there was an unbalanced provision of additional co-exposures across study groups, which were not appropriately adjusted for, • **OR** there is insufficient information provided about co-exposures in occupational studies or studies of contaminated sites where high exposures to other chemical exposures would have been reasonably anticipated (record "NR" as basis for answer).
Definitely High Risk of Bias (--)
• Direct evidence that the distribution of important covariates and known confounders differed between the groups, confounding was demonstrated, and was 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 direct evidence that there was an unbalanced provision of additional co-exposures across study groups, which were not appropriately 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?

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.
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, • **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.
Probably High Risk of Bias (-) or (NR)
• Indirect evidence that loss of subjects (*i.e.*, incomplete outcome data) was unacceptably large and not adequately addressed, • **OR** there is insufficient information provided about numbers of subjects lost to follow-up (record "NR" as basis for answer).
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 characterization?

Definitely Low Risk of Bias (++)
• Direct evidence that exposure was consistently assessed (*i.e.*, under the same method and time-frame) using well-established methods that directly measure exposure in blood, serum, or plasma), • **OR** exposure was assessed using less-established methods that directly measure exposure and are validated against well-established methods, • **OR** exposure where sarin release is established and exposure of the individual established by clinical signs and symptoms per Brown and Brix (1998) with corroboration by assessment of direct (in hospital, in clinic) observation of symptoms of acute cholinergic signs • **AND** exposure was assessed in a relevant time-window for development of the outcome, • **AND** there is sufficient range or variation in exposure measurements across groups to potentially identify associations with health outcomes,

Definitely Low Risk of Bias (++) *continued*
• **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.
Probably Low Risk of Bias (+)
• Indirect evidence that the exposure was consistently assessed using well-established methods that directly measure exposure, • **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) • **OR** exposure where sarin release is established and exposure of the individual established by clinical signs and symptoms per Brown and Brix (1998) with corroboration by assessment of indirect observation of symptoms of acute cholinergic signs (video or reported by patient's family), • **AND** exposure was assessed in a relevant time-window for development of the outcome. Current exposure measures will be considered relevant for any outcome. • **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.
Probably High Risk of Bias (-) or (NR)
• Indirect evidence that the exposure 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.*, questionnaire, job-exposure matrix or self-report without validation), • **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).
Definitely High Risk of Bias (--)
• Direct evidence that the exposure was assessed using methods with poor validity, • **OR** evidence of exposure misclassification (*e.g.*, differential recall of self-reported exposure).

9. Can we be confident in the outcome assessment?

Definitely Low Risk of Bias (++)
• Direct evidence that the outcome was assessed using well-established methods (*e.g.*, gold standard at the time or best-available method) • **AND** subjects had been followed for the same length of time in all study groups, • **AND** there is direct evidence that the outcome assessors (including study subjects, if outcomes were self-reported) were adequately blinded to the study group or exposure level, and it is unlikely that they could have broken the blinding prior to reporting outcomes.

Probably Low Risk of Bias (+)
• Indirect evidence that the outcome was assessed using acceptable methods (*i.e.*, deemed valid and reliable but not the gold standard), • **AND** subjects had been followed for the same length of time in all study groups • **OR** it is deemed that the outcome assessment methods used would not appreciably bias results, • **AND** there is indirect evidence that the outcome assessors (including study subjects, if outcomes were self-reported) 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:** Acceptable, but not ideal assessment methods will depend on the outcome, but examples of such methods may include proxy reporting of outcomes such as and mining of data collected for other purposes. Proxy reporting (*e.g.*, parental reporting of days sick or doctor-diagnosis) of other types of disease, colds, *etc.* should be considered on a case-by-case basis with consideration of whether or not there is empirical evidence as to the reliability of proxy reporting for that outcome.
Probably High Risk of Bias (-) or (NR)
• Indirect evidence that the outcome assessment method is an insensitive instrument (*e.g.*, a questionnaire used to assess outcomes with no information on validation), • **OR** the length of follow up differed by study group, • **OR** there is indirect evidence that it was possible for outcome assessors (including study subjects if outcomes were self-reported) to infer the study group prior to reporting outcomes, • **OR** there is insufficient information provided about blinding of outcome assessors (record "NR" as basis for answer).
Definitely High Risk of Bias (--)
• Direct evidence that the outcome assessment method is an insensitive instrument, • **OR** the length of follow up differed by study group, • **OR** there is direct evidence for lack of adequate blinding of outcome assessors (including study subjects if outcomes were self-reported), including no blinding or incomplete blinding.

10. Were all measured outcomes reported?

Definitely Low Risk of Bias (++)
• Direct evidence that all of the study's measured outcomes 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.
Probably Low Risk of Bias (+)
• Indirect evidence that all of the study's measured outcomes 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).

Probably High Risk of Bias (-) or (NR)
• Indirect evidence that all of the study's measured outcomes outlined in the protocol, methods, abstract, and/or introduction (that are relevant for the evaluation) have 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).
Definitely High Risk of Bias (--)
• Direct evidence that all of the study's measured outcomes 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 sarin-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.

Cross Sectional and Case Series 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]

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 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, • **Note:** A study will 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).
Probably Low Risk of Bias (+)
• Indirect evidence that subjects (both exposed and non-exposed) were similar, *e.g.*, recruited from the same eligible 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 would not appreciably bias results.
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 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).
Definitely High Risk of Bias (--)
• Direct evidence that subjects (both exposed and non-exposed) were not similar, recruited within very different time frames, or had very different participation/response rates.

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

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, propensity scoring, 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 were assessed using valid and reliable measurements, • **AND** there is direct evidence that other exposures anticipated to bias results were not present or were appropriately measured and adjusted for. In occupational studies or studies of contaminated sites, other chemical exposures known to be associated with those settings were appropriately considered.

Probably Low Risk of Bias (+)
• Indirect evidence that appropriate adjustments were made, • **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. • **Note:** this includes insufficient information provided on co-exposures in general population studies.
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 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), • **OR** there is indirect evidence that there was an unbalanced provision of additional co-exposures across study groups, which were not appropriately adjusted for, • **OR** there is insufficient information provided about co-exposures in occupational studies or studies of contaminated sites where high exposures to other chemical exposures would have been reasonably anticipated (record "NR" as basis for answer).
Definitely High Risk of Bias (--)
• Direct evidence that the distribution of important covariates and known confounders differed between the groups, confounding was demonstrated, and was 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 direct evidence that there was an unbalanced provision of additional co-exposures across study groups, which were not appropriately 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?

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.
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.

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).
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 characterization?

Definitely Low Risk of Bias (++)
• Direct evidence that exposure was consistently assessed (*i.e.*, under the same method and time-frame) using well-established methods that directly measure exposure, • **OR** exposure was assessed using less-established methods that directly measure exposure and are validated against well-established methods, • **OR** exposure where sarin release is established and exposure of the individual established by clinical signs and symptoms per Brown and Brix (1998) with corroboration by assessment of direct (in hospital, in clinic) observation of symptoms of acute cholinergic signs, • **AND** exposure was assessed in a relevant time-window for development of the outcome. Current exposure measures will be considered relevant for any outcome, • **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 •
Probably Low Risk of Bias (+)
• Indirect evidence that the exposure was consistently assessed using well-established methods that directly measure exposure), • **OR** exposure where sarin release is established and exposure of the individual established by clinical signs and symptoms per Brown and Brix (1998) with corroboration by assessment of indirect observation of symptoms of acute cholinergic signs (video or reported by patient's family), • **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** exposure was assessed in a relevant time-window for development of the outcome. Current exposure measures will be considered relevant for any outcome, • **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.

Probably High Risk of Bias (-) or (NR)
• Indirect evidence that the exposure was assessed using poorly validated methods that directly measure exposure, • **OR** there is direct 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).
Definitely High Risk of Bias (--)
• Direct evidence that the exposure was assessed using methods with poor validity, • **OR** evidence of exposure misclassification (*e.g.*, differential recall of self-reported exposure).

9. Can we be confident in the outcome assessment?

Definitely Low Risk of Bias (++)
• Direct evidence that the outcome was assessed using well-established methods (the gold standard), • **AND** there is direct evidence that the outcome assessors (including study subjects, if outcomes were self-reported) were adequately blinded to the exposure level, and it is unlikely that they could have broken the blinding prior to reporting outcomes.
Probably Low Risk of Bias (+)
• Indirect evidence that the outcome was assessed using acceptable methods, • **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).
Probably High Risk of Bias (-) or (NR)
• Indirect evidence that the outcome assessment method is an insensitive 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).
Definitely High Risk of Bias (--)
• Direct evidence that the outcome assessment method is an insensitive instrument, • **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).

10. Were all measured outcomes reported?

Definitely Low Risk of Bias (++)
• Direct evidence that all of the study's measured outcomes 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.

Probably Low Risk of Bias (+)
• Indirect evidence that all of the study's measured outcomes 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).
Probably High Risk of Bias (-) or (NR)
• Indirect evidence that all of the study's measured outcomes outlined in the protocol, methods, abstract, and/or introduction (that are relevant for the evaluation) have 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).
Definitely High Risk of Bias (--)
• Direct evidence that all of the study's measured outcomes 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 sarin-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.

Case Control 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?

Definitely Low Risk of Bias (++)
• Direct evidence that cases and controls were similar (*e.g.*, recruited from the same eligible population including being of similar age, gender, ethnicity, and eligibility criteria other than outcome of interest as appropriate), recruited within the same time frame, and controls are described as having no history of the outcome, • **Note:** A study will 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).
Probably Low Risk of Bias (+)
• Indirect evidence that cases and controls were similar (*e.g.*, recruited from the same eligible population, recruited with the same method of ascertainment using the same inclusion and exclusion criteria, and were of similar age), recruited within the same time frame, and controls are described as having no history of the outcome, • **OR** it is deemed differences between cases and controls would not appreciably bias results.
Probably High Risk of Bias (-) or (NR)
• Indirect evidence that controls were drawn from a very dissimilar population than cases or recruited within very different time frames, • **OR** there is insufficient information provided about the appropriateness of controls including rate of response reported for cases only (record "NR" as basis for answer).
Definitely High Risk of Bias (--)
• Direct evidence that controls were drawn from a very dissimilar population than cases or recruited within very different time frames.

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

Definitely Low Risk of Bias (++)
• Direct evidence that appropriate adjustments 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 of cases and controls, adjustment in multivariate model, stratification, propensity scoring, or other methods were appropriately justified, • **AND** there is direct evidence that primary covariates and confounders were assessed using valid and reliable measurements, • **AND** there is direct evidence that other exposures anticipated to bias results were not present or were appropriately measured and adjusted for.

Probably Low Risk of Bias (+)
• Indirect evidence that appropriate adjustments were made, • **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. • **Note:** this includes insufficient information provided on co-exposures in general population studies.
Probably High Risk of Bias (-) or (NR)
• Indirect evidence that the distribution of important covariates and known confounders differed between cases and controls and was not investigated further, • **OR** there is insufficient information provided about the distribution of known confounders in cases and controls (record "NR" as basis for answer), • **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), • **OR** there is indirect evidence that there was an unbalanced provision of additional co-exposures across cases and controls, which were not appropriately adjusted for, • **OR** there is insufficient information provided about co-exposures in occupational studies or studies of contaminated sites where high exposures to other chemical exposures would have been reasonably anticipated (record "NR" as basis for answer).
Definitely High Risk of Bias (--)
• Direct evidence that the distribution of important covariates and known confounders differed between cases and controls, confounding was demonstrated, but was 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 direct evidence that there was an unbalanced provision of additional co-exposures across cases and controls, which were not appropriately 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?

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.
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.

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).
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 characterization?

Definitely Low Risk of Bias (++)
• Direct evidence that exposure was consistently assessed (*i.e.*, under the same method and time-frame) using well-established methods that directly measure exposure, • **OR** exposure was assessed using less-established methods that directly measure exposure and are validated against well-established methods' • **OR** exposure where sarin release is established and exposure of the individual established by clinical signs and symptoms per Brown and Brix (1998) with corroboration by assessment of direct (in hospital, in clinic) observation of symptoms of acute cholinergic signs, • **AND** exposure was assessed in a relevant time-window for development of the outcome. Current exposure measures will be considered relevant for any outcome, • **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 •
Probably Low Risk of Bias (+)
• Indirect evidence that the exposure was consistently assessed using well-established methods that directly measure exposure, • **OR** exposure where sarin release is established and exposure of the individual established by clinical signs and symptoms per Brown and Brix (1998) with corroboration by assessment of indirect observation of symptoms of acute cholinergic signs (video or reported by patient's family),, • **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** exposure was assessed in a relevant time-window for development of the outcome. Current exposure measures will be considered relevant for any outcome, • **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.

Probably High Risk of Bias (-) or (NR)
• Indirect evidence that the exposure was assessed using poorly validated methods that directly measure exposure, • **OR** there is direct 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).
Definitely High Risk of Bias (--)
• Direct evidence that the exposure was assessed using methods with poor validity, • **OR** evidence of exposure misclassification (*e.g.*, differential recall of self-reported exposure).

9. Can we be confident in the outcome assessment?

Definitely Low Risk of Bias (++)
• Direct evidence that the outcome was assessed in cases (*i.e.*, case definition) and controls using well-established methods (the gold standard), • **AND** subjects had been followed for the same length of time in all study groups, • **AND** there is direct evidence that the outcome assessors (including study subjects, if outcomes were self-reported) were adequately blinded to the exposure level when outcome was assessed in cases (*i.e.*, case definition) and controls. • **NOTE** Well-established methods will depend on the outcome, but examples of such methods may include doctor diagnosis of health dysfunction or doctor diagnosis obtained from medical records.
Probably Low Risk of Bias (+)
• Indirect evidence that the outcome was assessed in cases (*i.e.*, case definition) and controls using acceptable methods), • **AND** subjects had been followed for the same length of time in all study groups, • **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 when 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 or lack of blinding is unlikely to bias a particular outcome). • **NOTE** Acceptable, but not ideal assessment methods will depend on the outcome, but examples of such methods may include proxy reporting of outcomes and mining of data collected for other purposes. Proxy reporting disease should be considered on a case-by-case basis with consideration of whether or not there is empirical evidence as to the reliability of proxy reporting for that outcome.
Probably High Risk of Bias (-) or (NR)
• Indirect evidence that the outcome was assessed in cases (*i.e.*, case definition) using an insensitive instrument, • **OR** there is insufficient information provided about how cases were identified (record "NR" as basis for answer), • **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).

Definitely High Risk of Bias (--)
• Direct evidence that the outcome was assessed in cases (*i.e.,* case definition) using an insensitive instrument, • **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).

10. Were all measured outcomes reported?

Definitely Low Risk of Bias (++)
• Direct evidence that all of the study's measured outcomes 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.
Probably Low Risk of Bias (+)
• Indirect evidence that all of the study's measured outcomes 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).
Probably High Risk of Bias (-) or (NR)
• Indirect evidence that all of the study's measured outcomes outlined in the protocol, methods, abstract, and/or introduction (that are relevant for the evaluation) have 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).
Definitely High Risk of Bias (--)
• Direct evidence that all of the study's measured outcomes 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 sarin-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?

Definitely Low Risk of Bias (++)
• Direct evidence that animals were allocated to any study group including 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 2011). • **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.
Probably Low Risk of Bias (+)
• Indirect evidence that animals were allocated to any study group including controls using a method with a random component (*i.e.,* authors state random allocation, without description of method), • **AND** evidence that the study used a concurrent control group as an indication that randomization covered all study groups, • **OR** it is deemed that allocation without a clearly random component would not appreciably bias results.
Probably High Risk of Bias (-) or (NR)
• Indirect evidence that animals were allocated to study groups using a method with a non-random component, • **OR** indirect evidence that there was a lack of a concurrent control group, • **OR** there is insufficient information provided about how cells were allocated to study groups (record "NR" as basis for answer).
Definitely High Risk of Bias (--)
• Direct evidence that animals were allocated to study groups using a non-random method including judgment of the investigator, the results of a laboratory test or a series of tests, • **OR** direct evidence that there was a lack of a concurrent control group.

2. Was allocation to study groups adequately concealed?

Definitely Low Risk of Bias (++)
• 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.
Probably Low Risk of Bias (+)
• 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.

Probably High Risk of Bias (-) or (NR)
• 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).
Definitely High Risk of Bias (--)
• 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?

Definitely Low Risk of Bias (++)
• 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 report explicitly provides this level of detail).
Probably Low Risk of Bias (+)
• 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, • **AND** identical non-treatment-related experimental conditions are assumed if authors did not report differences in housing or husbandry.
Probably High Risk of Bias (-) or (NR)
• 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** there is indirect evidence that non-treatment-related experimental conditions were not comparable between study groups.
Definitely High Risk of Bias (--)
• Direct evidence from the study report that control animals were untreated, or treated with a different vehicle than experimental animals, • **OR** there is 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?

Definitely Low Risk of Bias (++)
• 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. Methods used to ensure blinding include central allocation; sequentially numbered treatment containers of identical appearance; sequentially numbered animal cages; or equivalent methods.

Probably Low Risk of Bias (+)
• 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.*).
Probably High Risk of Bias (-) or (NR)
• 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).
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?

Definitely Low Risk of Bias (++)
• Direct evidence that loss of animals was adequately addressed and reasons were documented when animals were removed from a study. • **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. • **OR** missing data have been imputed using appropriate methods (insuring that characteristics of animals are not significantly different from animals retained in the analysis).
Probably Low Risk of Bias (+)
• 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.
Probably High Risk of Bias (-) or (NR)
• 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).
Definitely High Risk of Bias (--)
• 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?

Definitely Low Risk of Bias (++)
• Direct evidence that the exposure to sarin was independently characterized and purity confirmed generally as ≥98%, • **AND** that exposure was consistently administered (*i.e.*, with the same method and time-frame) across treatment groups, • **AND** for dietary or drinking water studies that 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 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 to confirm assay performance.
Probably Low Risk of Bias (+)
• Indirect evidence that the exposure to sarin was appropriately characterized and purity confirmed generally as ≥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 ≥95% and it is deemed that impurities of up to 5% would not appreciably bias results, • **AND** that 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.
Probably High Risk of Bias (-) or (NR)
• 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.
Definitely High Risk of Bias (--)
• 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?

Definitely Low Risk of Bias (++)
• Direct evidence that the outcome was assessed using well-established methods (*e.g.*, gold standard) • **AND** assessed at the same length of time after initial exposure in all study groups, • **AND** there is 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.
Probably Low Risk of Bias (+)
• Indirect evidence that the outcome was assessed using acceptable methods (*i.e.,* deemed valid and reliable but not the gold standard), • **AND** 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** there is 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. • **NOTE** Acceptable assessment methods will depend on the outcome, but examples of such methods may include: objectively measured antibody or cytokine concentrations with diagnostic methods using commercial kits with some variation, but ability to discriminate between the high dose treatment and control group (or indirect evidence that the assay could have detected a difference based on responses to a positive control).
Probably High Risk of Bias (-) or (NR)
• Indirect evidence that the outcome assessment method is an insensitive instrument, • **OR** the length of time after initial exposure differed by study group, • **OR** there is 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).
Definitely High Risk of Bias (--)
• Direct evidence that the outcome assessment method is an insensitive instrument, • **OR** the length of time after initial exposure differed by study group, • **OR** there is 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?

Definitely Low Risk of Bias (++)
• Direct evidence that all of the study's measured outcomes 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.

Probably Low Risk of Bias (+)
• Indirect evidence that all of the study's measured outcomes 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).
Probably High Risk of Bias (-) or (NR)
• Indirect evidence that all of the study's measured outcomes outlined in the protocol, methods, abstract, and/or introduction (that are relevant for the evaluation) have 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 answer basis).
Definitely High Risk of Bias (--)
• Direct evidence that all of the study's measured outcomes 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 sarin-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.