Evaluation of the causal framework used for setting National Ambient Air Quality Standards

Julie E. Goodman, Robyn L. Prueitt, Sonja N. Sax, Lisa A. Bailey, and Lorenz R. Rhomberg

Gradient, Cambridge, MA, USA

Abstract

A scientifically sound assessment of the potential hazards associated with a substance requires a systematic, objective and transparent evaluation of the weight of evidence (WoE) for causality of health effects. We critically evaluated the current WoE framework for causal determination used in the United States Environmental Protection Agency’s (EPA’s) assessments of the scientific data on air pollutants for the National Ambient Air Quality Standards (NAAQS) review process, including its methods for literature searches; study selection, evaluation and integration; and causal judgments. The causal framework used in recent NAAQS evaluations has many valuable features, but it could be more explicit in some cases, and some features are missing that should be included in every WoE evaluation. Because of this, it has not always been applied consistently in evaluations of causality, leading to conclusions that are not always supported by the overall WoE, as we demonstrate using EPA’s ozone Integrated Science Assessment as a case study. We propose additions to the NAAQS causal framework based on best practices gleaned from a previously conducted survey of available WoE frameworks. A revision of the NAAQS causal framework so that it more closely aligns with these best practices and the full and consistent application of the framework will improve future assessments of the potential health effects of criteria air pollutants by making the assessments more thorough, transparent, and scientifically sound.

Keywords

Air quality, causal framework, criteria pollutants, risk assessment, systematic review, weight of evidence

Introduction

A scientifically sound assessment of the potential risks associated with a substance requires a systematic, objective and transparent evaluation of the weight of evidence (WoE) for causality of health effects. Many different methods for conducting WoE evaluations have been used by authoritative bodies and reported in the scientific literature (e.g. Adami et al., 2011; Borgert et al., 2011; ECETOC, 2009; IARC, 2006; Rhomberg et al., 2010, 2011a, 2013; Swaen & van Amelsvoort, 2009). The United States Environmental Protection Agency (EPA) uses WoE approaches to assess the risks of health effects from exposures to chemicals in the environment, and these assessments are used to support a wide range of regulatory activities. For example, as part of the process for setting health-based National Ambient Air Quality Standards (NAAQS), EPA periodically evaluates the WoE for potential health risks of exposure to each of six substances that are classified as criteria air pollutants [particulate matter (PM), sulfur dioxide, nitrogen dioxide, carbon monoxide, ozone and lead] because of their scale occurrence and public health significance. These evaluations are presented as an Integrated Science Assessment (ISA), in which
EPA assesses the body of relevant scientific literature to draw conclusions regarding the causal relationships between exposure to each criteria pollutant and human health and welfare effects (US EPA, 2013a). Information and conclusions from the ISA are used by EPA in the Risk and Exposure Assessment (REA) to evaluate human health or environmental risks associated with recent air quality conditions or air quality that is estimated to just meet current or alternative standards (US EPA, 2013b). In its Policy Assessment, EPA bridges the gap between the scientific assessments in the ISA and REA and judgments required by the EPA Administrator regarding the retention or revision of any of the four elements of each NAAQS (i.e., indicator, averaging time, level, and statistical form) (US EPA, 2013b). In the ISA, decisions regarding causality near the level of the standard carry substantial weight in the Administrator’s judgments (Bachmann, 2007; McClellan, 2012).

The current WoE framework for causal determination in the NAAQS process (referred to herein as the ‘‘NAAQS causal framework’’) was first introduced in the ISAs for oxides of nitrogen and sulfur (US EPA, 2008a,b) and updated in the ISA for PM (US EPA, 2009); it has been refined since in the ISAs for carbon monoxide (US EPA, 2010a), ozone (US EPA, 2013a), and lead (US EPA, 2012).

In addition to assessments of air pollutants in the NAAQS process, EPA evaluates potential human health effects from a broad variety of chemicals in its Integrated Risk and Information System (IRIS) Program. An example of this is EPA’s 2010 draft IRIS assessment of formaldehyde (US EPA, 2010b), for which the evidence for potential adverse effects was compiled and evaluated in a similar manner as was the evidence in the criteria pollutant ISAs. In 2011, a panel assembled by the National Research Council (NRC) of the National Academy of Sciences (NAS) published its Review of the Environmental Protection Agency’s Draft IRIS Assessment of Formaldehyde (NRC, 2011). In this report, the NAS panel identified several issues with the WoE methodology used by EPA, including some that were identified by other NRC committees that had reviewed a number of IRIS assessments over the previous decade. The panel called on EPA to undertake a program to develop a transparent and defensible methodology for WoE evaluations to improve the agency’s IRIS assessments.

The NAS panel proposed a ‘‘roadmap’’ for reform and improvement of the risk assessment process, noting that there are many proven models for evidence-based reviews that might provide guidance to EPA, including systematic WoE approaches for hazard identification. The panel recommended that EPA’s WoE evaluation include clear narratives that provide the rationale for conclusions; furthermore, the panel noted that conclusions suggesting an association in the face of mixed results (e.g., positive, weak, and null studies for a given endpoint) should be accompanied by a thorough discussion of the strengths and weaknesses of the underlying studies. The NAS panel stated that EPA’s current process for NAAQS reviews, including the development of an ISA, is a useful example of revising an established process for conducting evaluations in a relatively short period of time. It noted, however, that the current NAAQS process is not an example of a specific approach for evidence review that should be adopted for revision of EPA’s IRIS process. In fact, several issues that the NAS panel noted with the formaldehyde assessment could also be considered applicable to revision and improvement of the NAAQS WoE evaluation methodology.

After describing the NAAQS causal framework below, we discuss best practices for WoE evaluations gleaned from other available WoE frameworks and note approaches that should be stated explicitly in the NAAQS causal framework to decrease the likelihood of biased evaluations. We then present a case study in which we cite several examples from the ozone ISA that demonstrate how EPA’s use of the current NAAQS causal framework leads to an inconsistent evaluation that often leads to conclusions of causality that are not fully supported by the evidence.

### The NAAQS causal framework

The current NAAQS causal framework used by EPA to conduct evaluations of potential health effects associated with criteria air pollutants is described in the Preamble section of the recent lead and ozone ISAs (US EPA, 2012, 2013a). The Preamble states that the NAAQS causal framework draws from language in sources across the federal government and scientific community, particularly the Institute of Medicine (IOM) report Improving the Presumptive Disability Decision-making Process for Veterans (IOM, 2008). Although the IOM (2008) report provides guidance for making decisions regarding health effects in veterans, and not from exposures to criteria air pollutants, the Preamble notes that it is applicable because it is a comprehensive report on the evaluation of causality.

The NAAQS causal framework is used to determine the WoE in support of causation and characterize the strength of any resulting causal classification after a consideration of the evidence. In addition, EPA uses the causality classification to decide if a quantitative risk assessment will be conducted on a particular health effect. Table 1 lists the general steps of the NAAQS causal framework.

In the Preamble of the ozone ISA, EPA first discusses how studies are identified for inclusion in the evaluation and then describes the framework itself. For brevity, and because study identification and selection are necessary steps prior to an evaluation of causality, hereafter we include these initial steps of study identification and selection methods described in the Preamble, as well as study evaluation and integration, when referring to the NAAQS causal framework. As part of the identification step, for each periodic NAAQS review, EPA maintains an ongoing literature search process to identify relevant studies published since the last NAAQS review of a given pollutant. The Preamble states that search strategies are adopted for revision of EPA’s IRIS process. In fact, several issues that the NAS panel noted with the formaldehyde assessment could also be considered applicable to revision and improvement of the NAAQS WoE evaluation methodology.
conducted by EPA using publicly available data. EPA states that all relevant controlled human exposure, epidemiology and toxicology studies are considered.

The Preamble in the ozone ISA states that the selection of studies for inclusion is based on the general scientific quality of the study, with consideration of the extent to which the study is informative and relevant to the NAAQS process. The assessment of study quality and relevance considers factors such as whether the study subjects, populations or animal models are adequately selected and sufficiently well-defined to allow for meaningful comparisons between groups; the statistical analyses are appropriate and properly performed; the exposure data are of adequate quality and representative of ambient conditions; the health effect measurements are meaningful, valid and reliable; and the analytical methods provide adequate sensitivity and precision to support conclusions. Other considerations discussed by EPA include whether potential confounders and effect modifiers are assessed in epidemiology studies; control exposures to filtered air are included in controlled human exposure studies; and there is sufficient statistical power to assess findings in both controlled human exposure and animal toxicology studies. EPA states that of most relevance for study inclusion is whether it provides useful qualitative or quantitative information regarding exposure–response relationships at exposures relevant to ambient conditions that can inform decisions on whether to retain or change the NAAQS.

After the discussion of study selection, the Preamble describes the NAAQS causal framework. This includes a discussion of some of the general strengths and limitations that should be considered when evaluating controlled human exposure, epidemiology, and animal toxicology studies. The framework uses a modified Bradford Hill approach (Hill, 1965) to aid in judgments regarding causal determinations, including consideration of criteria (as they are commonly referred to) such as consistency of observed associations, coherence, biological plausibility, biological gradient, strength of the observed association, experimental evidence, temporality, specificity and analogy (Table 2). The original Bradford Hill criteria were developed mainly for the interpretation of epidemiology results and are not meant to be specific rules to follow. EPA modified them in the NAAQS causal framework for use with the broad array of study types that are considered in the ISAs (e.g. epidemiology, toxicology, mechanistic), to be more consistent with EPA’s Guidelines for Carcinogen Risk Assessment (US EPA, 2005). Regarding how
Table 2. EPA aspects to aid in judging causality.

<table>
<thead>
<tr>
<th>Aspect</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Consistency of the observed association</td>
<td>An inference of causality is strengthened when a pattern of elevated risks is observed across several independent studies. The reproducibility of findings constitutes one of the strongest arguments for causality. If there are discordant results among investigations, possible reasons such as differences in exposure, confounding factors, and the power of the study are considered.</td>
</tr>
<tr>
<td>Coherence</td>
<td>An inference of causality from one line of evidence (e.g. epidemiologic, controlled human exposure [clinical], or animal studies) may be strengthened by other lines of evidence that support a cause-and-effect interpretation of the association. The coherence of evidence from various fields greatly adds to the strength of an inference of causality. In addition, there may be coherence in demonstrating effects across multiple study designs or related health endpoints within one scientific line of evidence.</td>
</tr>
<tr>
<td>Biological plausibility</td>
<td>An inference of causality tends to be strengthened by consistency with data from experimental studies or other sources demonstrating plausible biological mechanisms. A proposed mechanistic linking between an effect and exposure to the agent is an important source of support for causality, especially when data establishing the existence and functioning of those mechanistic links are available.</td>
</tr>
<tr>
<td>Biological gradient (exposure–response relationship)</td>
<td>A well-characterized exposure–response relationship (e.g. increasing effects associated with greater exposure) strongly suggests cause and effect, especially when such relationships are also observed for duration of exposure (e.g. increasing effects observed following longer exposure times).</td>
</tr>
<tr>
<td>Strength of the observed association</td>
<td>The finding of large, precise risks increases confidence that the association is not likely due to chance, bias or other factors. However, it is noted that a small magnitude in an effect estimate may represent a substantial effect in a population.</td>
</tr>
<tr>
<td>Experimental evidence</td>
<td>Strong evidence for causality can be provided through ‘‘natural experiments’’ when a change in exposure is found to result in a change in occurrence or frequency of health or welfare effects.</td>
</tr>
<tr>
<td>Temporal relationship of the observed association</td>
<td>Evidence of a temporal sequence between the introduction of an agent, and appearance of the effect, constitutes another argument in favor of causality.</td>
</tr>
<tr>
<td>Specificity of the observed association</td>
<td>Evidence linking a specific outcome to an exposure can provide a strong argument for causation. However, it must be recognized that rarely, if ever, does exposure to a pollutant invariably predict the occurrence of an outcome, and that a given outcome may have multiple causes.</td>
</tr>
<tr>
<td>Analogy</td>
<td>Structure activity relationships and information on the agent’s structural analogs can provide insight into whether an association is causal. Similarly, information on mode of action for a chemical, as one of many structural analogs, can inform decisions regarding likely causality.</td>
</tr>
</tbody>
</table>

Source: US EPA (2013a, Table 1).

Although these aspects provide a framework for assessing the evidence, they do not lend themselves to being considered in terms of simple formulas or fixed rules of evidence leading to conclusions about causality (Hill, 1965). For example, one cannot simply count the number of studies reporting statistically significant results or statistically nonsignificant results and reach credible conclusions about the relative weight of the evidence and the likelihood of causality. Rather, these aspects provide a framework for systematic appraisal of the body of evidence, informed by peer and public comment and advice, which includes weighing alternative views on controversial issues. (US EPA, 2013a)

The NAAQS causal framework suggests that evidence be evaluated for major health outcome categories (e.g. respiratory effects) and conclusions be drawn based on the integration of evidence from across health disciplines and across the spectrum of related health endpoints (e.g. health effects ranging from inflammation of the lungs to respiratory disease mortality). In drawing judgments, there is a focus on evidence of effects in the range of relevant human exposures to a given pollutant, with a general consideration of studies with doses or exposures in the range of no more than one or two orders of magnitude above current or ambient conditions. In discussing causal determinations, EPA characterizes the evidence on which the judgment is based, including the strength of evidence for individual endpoints within each major health outcome category. Based on these characterizations, conclusions regarding the WoE for causation are classified using a five-level hierarchy: ‘‘Causal relationship’’; ‘‘Likely to be a causal relationship’’; ‘‘Suggestive of a causal relationship’’; ‘‘Inadequate to confer a causal relationship’’ and ‘‘Not likely to be a causal relationship’’ (Table 1). The framework also discusses several factors for consideration in delineating between adverse and non-adverse health effects resulting from exposure to air pollution.
Causal relationship Evidence is sufficient to conclude that there is a causal relationship with relevant pollutant exposures (i.e. doses or exposures generally within one to two orders of magnitude of current levels). That is, the pollutant has been shown to result in health effects in studies in which chance, bias, and confounding could be ruled out with reasonable confidence. For example: (a) controlled human exposure studies that demonstrate consistent effects; or (b) observational studies that cannot be explained by plausible alternatives or are supported by other lines of evidence (e.g. animal studies or mode of action information). Evidence includes multiple high-quality studies.

Likely to be a causal relationship Evidence is sufficient to conclude that a causal relationship is likely to exist with relevant pollutant exposures, but important uncertainties remain. That is, the pollutant has been shown to result in health effects in studies in which chance and bias can be ruled out with reasonable confidence but potential issues remain. For example: (a) observational studies show an association, but copollutant exposures are difficult to address and/or other lines of evidence (controlled human exposure, animal, or mode of action information) are limited or inconsistent; or (b) animal toxicological evidence from multiple studies from different laboratories that demonstrate effects, but limited or no human data are available. Evidence generally includes multiple high-quality studies.

Suggestive of a causal relationship Evidence is suggestive of a causal relationship with relevant pollutant exposures, but is limited. For example, (a) at least one high-quality epidemiologic study shows an association with a given health outcome but the results of other studies are inconsistent; or (b) a well-conducted toxicological study, such as those conducted in the National Toxicology Program (NTP), shows effects in animal species.

Inadequate to infer a causal relationship Evidence is inadequate to determine that a causal relationship exists with relevant pollutant exposures. The available studies are of insufficient quantity, quality, consistency, or statistical power to permit a conclusion regarding the presence or absence of an effect.

Not likely to be a causal relationship Evidence is suggestive of no causal relationship with relevant pollutant exposures. Several adequate studies, covering the full range of levels of exposure that human beings are known to encounter and considering at-risk populations, are mutually consistent in not showing an effect at any level of exposure.

Source: US EPA (2013a, Table II).

As discussed above, the NAAQS causal framework is largely based on an IOM (2008) WoE framework. The current IOM framework has four categories of causal determination, with only the top two designating evidence that establishes a causal relationship (top category) or a relationship for which causality is “at least as likely as not” (second category) (Table 3). In contrast, as shown in Table 4, the NAAQS causal framework has five categories of causal determination, with the top three designating evidence that establishes a causal relationship (i.e. causal, likely causal, or suggestive of a causal relationship). Further, in contrast to the IOM approach, the NAAQS causal framework states that only one positive study is sufficient to establish a suggestive causal relationship when the results of other studies are inconsistent; this can greatly increase the frequency of positive causal conclusions.
There are many valuable features in the current NAAQS causal framework. These include literature searching that is iterative, consideration of modified Bradford Hill criteria to aid in judgments regarding causal determinations, weighing alternative views on controversial issues, integration of judgments of the evidence across disciplines and endpoints, focusing on evidence of health effects in the range of human-relevant exposures, consideration of the adversity of health outcomes and a hierarchy for categorization of the strength of the evidence.

Some aspects of the NAAQS causal framework could be more specific, however, and the framework does not explicitly include several features that, we would argue, are critical for a thorough and internally consistent WoE evaluation. For example, there is no explicit guidance for literature search strategies or determining study exclusion criteria. There is also no explicit guidance for evaluating the strengths and weaknesses of studies and no information on how to use study quality criteria to assign a quality weight to individual studies. There are no clear statements regarding how the Bradford Hill criteria should be applied (e.g. there is no indication of what constitutes a strong association) or how the causality judgments should consider all criteria jointly. There is also no guidance on how to evaluate alternative hypotheses that are supported by the data (e.g. a given substance is not a causal factor for a health effect and apparent associations are explicable by other factors) or what criteria should be used to determine which hypothesis is most supported by the data. As indicated in the NAS roadmap, EPA needs to state the justification for its judgments explicitly – including the reasoning behind judgments – to aid transparency.

Because of these and other limitations discussed below, EPA’s NAAQS causal framework has not always been applied consistently in evaluations of causality for individual substances – much less across substances – and adequate transparency regarding the justification of specific causal judgments is not achieved in the criteria air pollutant ISAs. We provide specific examples of this with a case study of the ozone ISA (US EPA, 2013a) below.

**WoE best practices**

We recently conducted a survey of existing WoE frameworks, including the NAAQS causal framework, to evaluate WoE best practices (Rhomberg et al., 2013). From these frameworks, we identified four successive phases to the overall WoE process that are consistent with the critical steps for the development of a scientifically sound assessment of the WoE for causation, as identified by the NAS panel that reviewed EPA’s formaldehyde assessment. These critical steps should be considered by EPA in developing an updated NAAQS framework. Below, we describe these phases; Table 5 presents their general features.

**Phase 1. Define the causal question and develop criteria for study selection**

Frame the purpose of the evaluation and the causal questions to be evaluated, and define the criteria for selecting studies relevant to the evaluation to ensure transparency. Prior to this phase, there is a separate role for problem formulation, which could be considered as “Phase 0” and entails the evaluation of what is needed to address the decisions to be made, along with an assessment of the capacity of an analysis of available data to provide answers addressing those needs.

**Phase 2. Develop and apply criteria for review of individual studies**

Conduct and present a systematic and consistent review of available studies relevant to the causal
question. Evaluate the rigor and quality of individual study results using pre-defined criteria applied uniformly across studies.

**Phase 3. Integrate and evaluate evidence.** Make sound and defensible scientific judgments about the existence and nature of causative processes for the health outcome under consideration. This is one of the more challenging phases for any WoE framework; no matter how one lays out procedures and methods for synthesizing across studies, in the end, the question is about how studies in one setting (e.g. animal or *in vitro* assays) should affect our assessment of potential causality or risks in another (e.g. the general human population exposed environmentally).

**Phase 4. Draw conclusions based on inferences.** Apply the results of the WoE evaluation from Phase 3 to make conclusions that can be used to inform regulatory decision-making. Although this phase is not risk management itself, it can be influenced by risk management considerations. In a regulatory setting, decisions about WoE categories (“known” causative agent, “likely” causative agent, etc.) or findings about the science [sufficient evidence for a mode of action (MoA) or to replace a default assumption for developing a toxicity value] are influenced in this stage by policy questions and regulatory consequences for those decisions and, ultimately, by policies and judgments about the sufficiency of evidence to support those decisions.

The most important aspect of a WoE framework is that it provides specific and transparent guidance for how to work through the WoE process. At the same time, the framework needs to be flexible so that it can be applied in a consistent manner across different types of datasets for evaluations of causality. A WoE evaluation is only useful and applicable to constructive scientific debate if the logic behind it is made clear, and, with that, it is often necessary to take the reader through alternative interpretations of the data so that the various interpretations can be compared logically. This approach does not eliminate the need for scientific judgment, and often may not lead to a definitive choice of one interpretation over the other, but it will clearly lay out the logic for how one weighs the evidence for and against each interpretation. Only in this way is it possible to have constructive scientific debate about potential causality that is focused on an organized, logical “weighing” of the evidence.

The NAAQS causal framework provides guidance on many features of the WoE phases noted above. Some of them are explicit and some are clearly included in the framework, but specific guidance is not always provided (e.g. study exclusion criteria). As a result, evaluations are not always consistent either among studies for a particular criteria pollutant or among pollutants, as discussed below in the case study of ozone.

Table 5 outlines WoE best practices generally and Table 6 outlines features of the NAAQS causal framework, as well as additional features that we recommend based on best practices. These additions will enable the framework to include these WoE best practices and help ensure that they are applied consistently. Several of the recommended features may have, in practice, been incorporated into WoE evaluations in the criteria pollutant ISAs; however, if they are not explicitly stated as guidance in the discussion of the NAAQS causal framework in the ISA Preamble, or if they are discussed as occurring in a Phase that does not match where they should occur according to WoE best practices, they are still noted in Table 6 as additional, best practice features. Below, we discuss these features by phase.

**Phase 1: Define the causal question and develop criteria for study selection**

The causal question and breadth of scope should be explicitly stated in Phase 1 of a WoE evaluation to provide a clear purpose for the assessment as well as direction for the remaining steps. Defining the causal question at the beginning of the process helps ensure that the correct question is posed, and it sets the context within which the bearing and utility of studies can be evaluated. Study selection criteria and the literature search design should be articulated clearly (Tables 5 and 6). The end product of Phase 1 is a clearly defined causal question, a documented search strategy with clear inclusion and exclusion criteria, a list of included and excluded studies with reasons for inclusion or exclusion, and a record of any deviations from the original plan.

In the NAAQS causal framework, the definition of the causal question is inferred from the purpose of the ISA in the NAAQS process, which is to provide a “concise review, synthesis, and evaluation of the most policy-relevant science to serve as a scientific foundation for the review of the [NAAQS]” (US EPA, 2013a). EPA does not, but in our view should, explicitly resolve the overarching causal question into specific issues for problem formulation. For example, EPA could specify aspects of the analysis regarding what is relevant to existing human exposures, the basis for what constitutes sensitive subpopulations, whether exposure to other agents affects response, and interactions with other stressors. In other words, by specifying whether the purpose is a “hazard in context” decision or a more general identification of a potential hazard at any exposure level, this would help to guide decisions on whether or not it is found in the actual population. Problem formulation should try to identify the specific sub-judgments that need to be made to support the overall goal, with the aim of enabling assessment of whether the available data speak to and enable sound decisions on them.

With regard to the literature search, the Preamble of the ozone ISA provides an overview of EPA’s general literature search strategy and the various ways in which studies are identified for potential inclusion. Some of the study inclusion criteria are also presented, and an evaluation of study quality is included in the study selection process. However, the Preamble does not provide guidance regarding study exclusion criteria or justification for inclusion or exclusion of studies. As shown in Table 6, the NAAQS causal framework should include explicit guidance regarding literature search strategies for identifying all available evidence, as well as criteria for both study inclusion and exclusion (e.g. study type, types of participants, exposures, outcomes). While relevance to ambient exposures to criteria pollutants is important to consider at this point, allowances may be made for scientific information that does not meet inclusion criteria.
Table 6. Modified NAAQS causal framework.

<table>
<thead>
<tr>
<th>Step</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Phase 1</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Define causal question or hypothesis</strong></td>
<td>Inferred from NAAQS process (as is problem formulation)</td>
</tr>
<tr>
<td>Determine breadth of scope</td>
<td>Developed during the iterative process of the literature search</td>
</tr>
<tr>
<td>Explore review questions not identified a priori</td>
<td></td>
</tr>
<tr>
<td><strong>Define criteria for study inclusion and exclusion</strong></td>
<td>May occur in several stages as new review questions are identified based on literature search</td>
</tr>
<tr>
<td>Determine study types</td>
<td></td>
</tr>
<tr>
<td>Define types of participants</td>
<td></td>
</tr>
<tr>
<td>Define exposures/risk factors/interventions</td>
<td></td>
</tr>
<tr>
<td>Define outcomes</td>
<td></td>
</tr>
<tr>
<td>Consider dose</td>
<td></td>
</tr>
<tr>
<td><strong>Plan literature search</strong></td>
<td></td>
</tr>
<tr>
<td>Involve librarians</td>
<td>Iterative modifications to optimize identification of pertinent publications</td>
</tr>
<tr>
<td>Choose databases</td>
<td>Specialized searches on specific topics</td>
</tr>
<tr>
<td>Identify articles by experts, the public, and advisory committees</td>
<td></td>
</tr>
<tr>
<td>Review reference lists, table of contents of relevant journals, citations in previous assessments</td>
<td></td>
</tr>
<tr>
<td>Search grey literature</td>
<td></td>
</tr>
<tr>
<td>Identify ongoing/unpublished studies</td>
<td></td>
</tr>
<tr>
<td><strong>Design literature search strategies</strong></td>
<td></td>
</tr>
<tr>
<td>Choose key words</td>
<td></td>
</tr>
<tr>
<td>Determine language, date, and document format requirements</td>
<td></td>
</tr>
<tr>
<td><strong>Screen and select studies, with justification for exclusion</strong></td>
<td></td>
</tr>
<tr>
<td>Assess agreement of study selection among investigators</td>
<td></td>
</tr>
<tr>
<td>Include peer-reviewed studies and EPA evaluations of publically available data</td>
<td></td>
</tr>
<tr>
<td>Consider relevance to ambient exposures</td>
<td></td>
</tr>
<tr>
<td>Make allowances for pieces of information from areas outside of inclusion/exclusion criteria that may impact evaluation</td>
<td></td>
</tr>
<tr>
<td>Do not exclude studies based on quality or if study reports null or negative findings at this point</td>
<td></td>
</tr>
<tr>
<td>Do not consider policy-relevance of studies</td>
<td></td>
</tr>
<tr>
<td><strong>Record search strategies, results, and decisions</strong></td>
<td>Sort references in bibliographic software</td>
</tr>
<tr>
<td><strong>Phase 2</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Extract study characteristics</strong></td>
<td>These include participants and setting, methods, interventions, outcome measures, results, and quality aspects</td>
</tr>
<tr>
<td>Develop quality control for data entry</td>
<td></td>
</tr>
<tr>
<td>Determine disagreement resolution</td>
<td></td>
</tr>
<tr>
<td><strong>Extract data</strong></td>
<td>Should be based on relevance to causal question</td>
</tr>
<tr>
<td>Determine data categories</td>
<td>Make sure all relevant data are extracted (i.e., it is not sufficient to extract a subset of relevant data)</td>
</tr>
<tr>
<td>Organize evidence into consistent sets of categories</td>
<td></td>
</tr>
<tr>
<td>Categorize health effect(s) by specificity and latency</td>
<td></td>
</tr>
</tbody>
</table>
Assess study quality

- **Study design**
  - Strengths and weaknesses
  - Exposure/intervention assessment

- **Confounders**
  - Systematic error (bias)
  - Precision of measurements
  - Reliability of data
  - Statistical methods

- **Outliers**
  - Selective outcome reporting
  - Identify fraudulent studies

Discuss general limitations to consider for each study type (e.g., controlled human exposure, epidemiology, toxicology)

Include assessment of counterfactualty of study design

Assign higher weight to studies with direct exposure measurement, absence of co-pollutant exposures, or use of multi-pollutant models

Determine if potential confounders identified and if partially or fully addressed by authors

Assign higher weight to studies that used statistical methods to detect and control for confounding and multiple comparisons, when necessary

Eliminate from analysis.

Categorize study quality

- Qualitatively or by ranking, grading, or scoring
- Do not eliminate any studies based on weakness (unless fatally flawed and not interpretable)

Assess individual study results

- **Strength of association**
  - Determine a priori the definition of a ‘strong’ association (e.g. risk estimate $\geq 3$). Use the same criteria for judging something as a risk factor vs. a protective factor (e.g. a risk estimate of 0.992 should carry the same weight as a risk estimate of 1.008)

- **Internal consistency**
  - Results using different models should be relatively consistent

- **Biological plausibility**
  - Both data indicating biological plausibility and a lack of biological plausibility should be considered
  - If increasing effects are observed with increasing exposures or duration of exposures, this is evidence for a causal relationship; a lack of this association is evidence against a causal relationship

- **Temporality**
  - It is not only that the exposure must occur before an effect, but that it occurs within an appropriate time frame

- **Dose-response or exposure–response**
  - Determine how dependent results are on values in the control group, particularly if they differ from historical controls

- **Variation/historical control rates**

- **Outcome assessment**

- **Random error (chance)**

Categorize study relevance and adequacy

- Determine whether and to what extent results are relevant to ambient exposures

- Determine appropriateness and usefulness of the data for hazard and risk assessment

Phase 3

Evaluate data within and across realms of evidence (e.g., toxicology, epidemiology, MoA)

- **Strength of association**
  - Integrate data across all lines of evidence so interpretation of one will inform interpretation of the other

- **Consistency of associations**
  - Assess all data, including negative, null, and positive results
  - Assign less weight to results of studies of lower quality
  - Incorporate peer and public comment and advice

- **Coherence**
  - Associations should be consistent both within and across studies, particularly those with different study designs
  - Results across all lines of evidence should be coherent; if not, species closest to humans should be considered to have more relevance to humans

(continued)
Table 6. Continued

<table>
<thead>
<tr>
<th>Step</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biological plausibility</td>
<td>Both data indicating biological plausibility and a lack of biological plausibility should be considered</td>
</tr>
<tr>
<td>Biological gradient (dose-response)</td>
<td>If increasing effects are observed with increasing exposures or duration of exposures, this is evidence for a causal relationship; a lack of this association is evidence against a causal relationship</td>
</tr>
<tr>
<td>Experimental evidence</td>
<td>It is not only that the exposure must occur before an effect, but that it occurs within an appropriate time frame</td>
</tr>
<tr>
<td>Temporality</td>
<td>Determine if consistent findings may be a result of a consistent confounder</td>
</tr>
<tr>
<td>Specificity</td>
<td></td>
</tr>
<tr>
<td>Analogy</td>
<td></td>
</tr>
<tr>
<td>Confounding</td>
<td></td>
</tr>
<tr>
<td>Bias</td>
<td></td>
</tr>
</tbody>
</table>

Assess data relevant to mode of action (MoA)

- Postulate MoA theory
- Describe key events in MoA
- Evaluate MoA based on Bradford Hill criteria
- Consider other possible MoAs
- Consider uncertainties, inconsistencies, and data gaps for MoA
- Determine whether MoA established in animals
- Assess whether key events in animal MoA plausible in humans
- Assess dose-response for key events in MoA

Assess adversity of effects

- Adaptive, compensatory, transient, and reversible effects are less likely to be adverse
- Precursors to apical effects, severe effects, and irreversible effects are more likely to be adverse

Compare alternative accounts of observations, including controversial issues

- Assess whether additional evidence supports or weakens the causal assumption
- Propose new hypothesis/account of evidence if it provides more convincing result
- Iterate assessment based on new evidence if results are ambiguous or if new hypothesis requires support

Formulate WoE conclusions

- Note assumptions, especially when they are ad hoc

Identify data gaps and propose next steps

Phase 4

Propose categories for causal relationships

Propose recommendations for risk assessment

Propose recommendations for research or policies

Current steps in the NAAQS causal framework are presented in non-italic font with additional best practices presented in italic font. Some of the additional best practices may have, in practice, been incorporated into WoE evaluations in the criteria pollutant ISAs, but if they are not explicitly stated as guidance in the discussion of the NAAQS causal framework in the ISA Preamble or are discussed as occurring in a Phase that does not match where they should occur according to WoE best practices, they are presented in italic font as additional best practices.
but may nonetheless affect the assessment through their bearing on interpretation of those studies that do meet narrower criteria. Studies should be screened and selected, with justification provided for inclusion or exclusion. All strategies and decisions should be recorded, particularly those that occur after the evaluation is underway. It is critical that studies should not be excluded at this point in the analysis based on quality, type (e.g. case studies), or if they report null or negative findings. These factors should be considered in subsequent phases (as we discuss further below). In addition, in a scientific evaluation such as in the ISA, no study with potential scientific use should be excluded based on policy-relevance. Detailed instructions for conducting literature searches can be found in the Cochrane Handbook for Systematic Reviews of Interventions (Higgins & Green, 2011). Although this handbook is focused on medical interventions, its guidance for literature searches can be valuable in application to literature searches of any WoE evaluation.

**Phase 2: Develop and apply criteria for review of individual studies**

In Phase 2, study characteristics and data are extracted, and study quality is assessed and categorized. Individual study results are assessed, and study relevance and adequacy are determined (Tables 5 and 6).

It is crucial that all relevant data presented, (and not just a subset of those data) are extracted from each study. This may be an iterative process, as what is considered relevant for inclusion may change as more studies are reviewed. In general, however, the data extracted should be based on the causal question. If positive data are extracted more often than null data from individual studies, the data will appear to be more consistent than they actually are. The amount of data extracted per study is generally quite large; thus, data should be organized into consistent categories so that they can be compared within and across studies. For example, specific health outcomes should be captured for each time period that they are evaluated in each study, based on all relevant statistical models, so that results can be compared for consistency.

The NAAQS causal framework does not provide a clear account of its study characteristic and data extraction process (Table 6). Regarding study quality, as EPA notes, one should consider general limitations for each study type (e.g. controlled human exposure, epidemiology, toxicology) (Table 1). For example, EPA states that it assigns higher weight to studies with direct (i.e. personal) exposure measurement, an absence of co-pollutant exposures, or use of multi-pollutant models, and studies that use statistical methods to detect and control for confounding and multiple comparisons, when necessary. However, there also should be a discussion of each of these factors that is independent of the description of individual studies. EPA should discuss all of the ways in which exposure can be measured, the strengths and limitations of each method, the possibility for exposure measurement error, and which methods carry the most weight. For example, studies that use central-site monitors to estimate personal exposure should be assigned less weight than studies that use more refined methods, such as land-use regression models, or personal monitoring data in the rare cases it is available, to estimate exposure. There should also be a discussion of statistical methods used among all studies evaluated and which specific methods are more robust and why (e.g. Have multiple comparisons been addressed? Are assumptions in Cox proportional hazard models appropriate?). Specific confounders should be addressed (e.g. co-pollutants, socioeconomic status, age) in terms of how they are handled in different studies and their likely impact on results. If confounders are evaluated only at the beginning of a study and not during follow-up, the potential for residual confounding should be considered. Other factors that should be considered in detail include bias, measurement precision, replicability of observations, data reliability, outliers, selective outcome reporting and fraudulent reporting. What is crucial is that the way these quality measures are evaluated is the same across studies. If a particular statistical model is considered a limitation in one study, it must be determined whether it should be considered a limitation in other studies that use it for similar analyses. An evaluation of study quality should be independent of the results and funding source of that study. It should be based purely on the methods, in a consistent manner across studies, and studies with more robust methods should receive more weight in the overall analysis.

Study quality can be categorized in a number of ways; it can be done qualitatively or quantitatively using a ranking, grading, or scoring method. EPA does not state explicitly how it characterizes studies within the NAAQS causal framework. Using a qualitative approach, all studies could be listed in a table or set of tables, with strengths and limitations in each column, so one can look across rows and columns to get an idea of strengths and limitations across studies. When a group of studies is then evaluated together in Phase 3, one can refer to this table to determine what weight to give each study based on its relative strengths and limitations.

There are several quantitative weighting approaches that have been recommended for evaluating studies based on quality. For example, Linkov et al. (2011) presented a quantitative multi-criteria decision analysis framework for evaluating individual lines of evidence in a WoE evaluation. Similarly, Lavelle et al. (2012) proposed a step-wise approach to systematically evaluate human and animal studies, which includes rating and ranking studies based on data quality criteria, and Klimischt et al. (1997) proposed a scheme for evaluating human data into four ranking categories for inclusion in a WoE evaluation (i.e. reliable without restriction, reliable with restrictions, not reliable and not assignable). While all these approaches have merits, any quantitative scheme can be arbitrary in terms of the weight given to different facets (e.g. Should exposure measurement error be given the same weight as an unaccounted for confounder?). Thus, a qualitative approach should also be considered.

Regardless of the approach used, studies should not be eliminated from an analysis simply based on weakness or a low score in a quantitative ranking; rather, they should be given less weight in the final analysis. Low quality studies (e.g. studies with several methodological limitations) need to
be considered in Phase 3, particularly if these studies are the basis of an explanation of the observations at hand that contrasts with another explanation (as discussed further below). Study quality information should be used to incorporate results into the larger arguments, as in Phase 3, but study quality should not drive this process. Studies with fatal flaws (e.g. uninterpretable findings resulting from methodological errors or falsified data) can be eliminated, however.

After study quality is addressed, the results of individual studies should be assessed with consideration of strength of association, internal consistency, biological plausibility, temporality, dose-response, historical control rate, outcome assessment and random error. Regarding strength of association, EPA indicates that larger, precise risks increase confidence that the association is not due to chance, bias or other factors. EPA also notes, however, that "a small magnitude in an effect estimate may represent a substantial effect in a population" (Table 2). While this statement is true, consistent with WoE best practices, EPA should apply the same consideration to the converse: a small effect estimate that is not due to chance, bias or another factor could represent a substantial effect, but, because it is small, it may actually indicate that the association is not causal. A great deal of effect estimates in the air pollution epidemiology literature are extremely small (e.g. <1.1); EPA should acknowledge that associations of small magnitude do not provide strong evidence of causality, as the hypothesis attributing the association to a causal effect is not markedly more plausible than it is attributing the association to one or several of the many factors (e.g. confounders, random chance) that could partially or fully account for the statistical association.

Internal consistency is not always, but always should be, considered when assessing individual studies in the ISAs. While it is not necessary that every statistical model produce the same results (if that were the case, one would not need to assess more than one model), the results should be relatively consistent. The consideration of internal consistency also should extend to the evaluation of results for multiple cities or among regions, as well as for different time periods, as discussed further below. If results are not consistent, the reasons why should be explored. That is, the process of evaluating the consistency and plausibility of support for a causal interpretation should consider such issues as the tentative reasons proposed for why apparent discordance exists; why particular findings among the discordant results should be considered informative about potential human risk while other contrary findings are not; and why the contrary findings should not be viewed as refutations of the causality assertion. If a lack of causality is as likely an explanation as another, the study should not be considered evidence for a causal association.

EPA discusses the importance of biological plausibility, as a known biologically plausible MoA increases the likelihood that an association is causal. If there is evidence suggesting an association is not biologically plausible (e.g. the body’s defense mechanisms prevent an adverse effect up to a certain point), however, this must be considered as well. Similarly, as discussed by EPA (Table 2), a well-characterized exposure–response relationship increases confidence for a cause-effect association, but EPA should also consider that a lack of exposure–response in a study with the statistical power to evaluate it is evidence against causation.

Temporality is also considered by EPA, but the NAAQS causal framework is missing a discussion (or evaluation) of whether effects occur within an appropriate time frame. Effects that occur at time points that cannot be linked to an exposure in a biologically plausible manner cannot be considered evidence of an effect. If a biologically plausible time frame is not known a priori, one should look for consistency among the time frames during which effects occur across studies. Another important factor is the consideration of controls. A control population might not be truly representative of unexposed individuals, and this may influence the outcome. It is essential that the control population be evaluated to determine if and how it influences results. Another point regarding the assessment of individual study results is the consideration of whether results are likely a result of random error or chance. If these are just as likely explanations as a causal association, one should not consider an association indicative of causation.

The final step in Phase 2 is to determine whether and to what extent studies are relevant to ambient exposures and the relevance (i.e. appropriateness) and adequacy (i.e. usefulness) of data for hazard and risk assessment. The NAAQS causal framework indicates this should be done, but this is often not explicitly discussed in ISAs.

**Phase 3: Integrate and evaluate evidence**

Phase 3 involves integrating and evaluating data across realms of evidence (Tables 5 and 6). The NAAQS causal framework includes modified Bradford Hill criteria (Table 2) to aid in judgments regarding causality, which is consistent with Phase 3 best practices, but it should also update its guidance on how the criteria should be applied to the available evidence and how all criteria should be considered jointly (i.e. as a whole) and consistently. As discussed above, descriptions for strength of association, coherence, consistency, biological plausibility and temporality are missing several key components that should always be considered in a WoE evaluation (Table 6).

The EPA NAAQS causal framework looks separately at the collective human studies and animal toxicology evidence, first coming to a synthesized WoE judgment for each of these realms and then integrating these separate judgments for different lines of evidence into an overall qualitative statement about causality and integration of human and animal toxicity data with data on exposure levels (US EPA, 2013a). It is preferable that the data evaluation be integrated across all lines of evidence based on their mutual illumination of potential biological processes that may underlie toxicity. In this way, interpretation of each line of evidence informs the interpretation of the others. This approach emphasizes the evaluation of how the results of particular studies can be applied more generally to inform the potential for similar causal processes in other studies, including studies in other realms of investigation. It is the potential for such commonality of causal processes that makes each particular result constitute evidence about the causality question being evaluated, and it is the reason that animal data can be
considered as evidence for potential effects in humans. EPA’s method of integrating judgments at the end of the evaluation does not allow data from one realm of evidence to influence conclusions from another. For example, if an epidemiology analysis can be interpreted two ways, and animal studies can shed light on whether one is more plausible than the other, this should be considered when making judgments about the epidemiology analysis. An empirical finding of, say, a dose-rate effect or a sex difference may be plausible without any direct evidence among human data, but animal or mechanistic studies may provide insight into whether such a phenomenon has a plausible biological basis in humans. A related aspect is that one should ensure that added assumptions—even reasonable and biologically plausible ones—that are invoked to account for variations within a realm of investigation (say, among human studies) are not contradicted by assumptions made when coming to integrated evaluation within another realm (say, to explain variations among animal biosassay outcomes in different rodent species). This cannot be done if judgments are made separately within realms, and only the conclusions (and not their basis) are integrated at the end of the process.

An assessment of MoA also plays a crucial role in data integration because it provides a means of understanding the degree of parallelism to be expected in underlying causal processes in human and animal responses, and it can give insight into the potential reasons behind differences in these processes at different exposure levels, exposure rates or study populations. The NAAQS causal framework acknowledges the importance of considering MoA, but it does not state explicitly how data relevant to MoA are assessed. To the degree possible, a proposed MoA and its key events, as well as any alternative MoAs and their key events, should be described and evaluated. Uncertainties, inconsistencies and data gaps should be considered. It should be determined whether a MoA has been established in animals and whether key events in animals are plausible in humans, although less than fully established MoAs can still be valuable tools in assessing the plausibility of tentative understandings about the bearing of different kinds of data on the overall causality question. Finally, exposure–response relationships for particular MoAs should be evaluated, as some MoAs are dependent upon exposure rate or cumulative exposure. While it is not necessary to know the specific MoA to establish causality, it certainly adds to the WoE of the likelihood of a causal association. It can also help bound exposure levels at which a MoA occurs. In addition, a known MoA may provide evidence against causality if it can be shown that certain key events in the MoA are not biologically plausible in humans at relevant exposures.

The NAAQS causal framework includes an evaluation of the adversity of effects, but this occurs after the WoE decisions are made and causal relationships are categorized (in Phase 4). It could also occur in Phase 3, however, when making judgments about the WoE for causality. This would be appropriate if the causal question is to establish whether an agent can cause effects that are specifically deemed adverse, rather than whether it can cause any effect at all. As Phase 4 is an assessment of the sufficiency of evidence to make decisions, and if a decision depends on an adversity question (as is often the case, such as in the NAAQS process), then the adequacy of the Phase 3 evaluation to support a Phase 4 conclusion on such matters requires a discussion of adversity, and of the scientific basis for distinguishing adverse from non-adverse effects. In general, if effects are determined to be adaptive, compensatory, transient, or reversible, they are less likely to be adverse; if they are precursors to apical effects, severe or irreversible, they are more likely to be adverse (Goodman et al., 2010). As there are exceptions, these should not be considered as hard and fast rules; rather, they may serve as guidelines for the evaluation of the adversity of effects.

The NAAQS causal framework includes the weighing of alternative views on controversial issues, but this should be expanded to an evaluation of alternative explanations of the same sets of results, regardless of whether the data are controversial. The evidence for and against a hypothesis should be discussed, and a clear illustration of how one explanation compares to another should be provided. Conclusions regarding causality should be presented not just as the result of judgments but with their context of reasons for choosing them over competing conclusions. Assumptions in the evaluation should be noted, especially when they are ad hoc (i.e. introduced to explain some phenomenon already seen). One should also assess whether additional evidence supports or weakens the causal assumption and provide new hypotheses or accounts of the evidence if it provides a more convincing result. The assessment should be updated based on new evidence if results are ambiguous or a new hypothesis requires support. At this point, one can formulate a WoE conclusion, identify data gaps, and propose next steps. The NAAQS causal framework could be strengthened by utilizing this final step—particularly identifying data gaps—in Phase 3.

Phase 4: Draw conclusions based on inferences

The NAAQS causal framework proposes categories for causal relationships (Table 4); based on these relationships, EPA determines which health effects will be evaluated in quantitative risk assessments. An important issue is that EPA’s guidance for causality categorization is not congruent with the judgments based on the Bradford Hill criteria. The framework claims to rely heavily on the criterion of consistency across studies in its categorization scheme, but, in practice, it does not fully evaluate consistency or incorporate other criteria such as coherence, biological plausibility, biological gradient or strength of association (discussed below in the case study of ozone).

EPA states that evidence is sufficient to conclude a causal relationship if “chance, bias, and confounding [can] be ruled out with reasonable confidence” (US EPA, 2013a), yet there is no guidance on what constitutes “reasonable confidence”. Based on the current framework, EPA cannot make that determination reliably because there is no guidance for assessing chance, bias, or confounding in a consistent manner. Although such an assessment is inevitably a scientific judgment, the basis for coming to and justifying the judgment should be included in the evaluation. Moreover, an evaluation of how confident one is in ruling out chance, bias and confounding should include an examination of the evidence.
for or against competing explanations of the results being evaluated. This will help determine whether the results are attributable to a causal effect of the agent or solely the results of, or at least severely skewed by, the influence of chance, bias or confounding. This evaluation should involve a discussion of the possible scope for the chance, bias and confounding, as well as the evidence for or against the actual (rather than merely possible) influence of these factors on the reported outcomes.

EPA suggests “controlled human exposure studies that demonstrate consistent effects” constitute evidence for a causal relationship (US EPA, 2013a), but it should indicate that this is only true if the results are coherent with other lines of evidence for particular exposure scenarios. EPA also indicates that “observational studies that cannot be explained by plausible alternatives” constitute evidence for a causal relationship (US EPA, 2013a). Yet, EPA does not fully explore alternative explanations for study results in practice, as discussed below in the case study of the ozone ISA. Currently, EPA sets forth a hypothesis (i.e. a criteria pollutant causes a particular health effect) and determines whether the data may support that hypothesis. As discussed above, EPA does not, but should, fully explore whether and to what degree the data support other hypotheses (e.g. a confounder, rather than the criteria pollutant, causes a particular health effect). It is only in this manner that alternative hypotheses can truly be explored.

EPA states that evidence is sufficient to conclude a likely causal relationship if “copollutant exposures are difficult to address and/or other lines of evidence (controlled human exposure, animal, or mode of action information) are limited or inconsistent” or if “animal toxicological evidence from multiple studies from different laboratories...demonstrate effects, but limited or no human data are available” (US EPA, 2013a). EPA concludes evidence is suggestive of a causal relationship if “at least one high-quality epidemiologic study shows an association with a given health outcome but the results of other studies are inconsistent” or if “a well-conducted toxicological study, such as those conducted in the National Toxicology Program (NTP), shows effects in animal species” (US EPA, 2013a).

Any WoE evaluation, by definition, involves a consideration of all lines of evidence in a consistent manner. It is not about resolving all uncertainty but, rather, determining whether the evidence as a whole supports causation more than it supports a lack of effect. If co-pollutants cannot be addressed or studies are inconsistent, the impact of the study results is ambiguous and the WoE might equally well be interpreted to be consistent with a lack of causality. In other words, evaluating the base of data “as a whole” means that one attends not simply to how much evidence can be adduced to support (or to counter) the hypothesized causal effect, but also how separate lines of evidence support (or contradict) one another. How discrepancies are to be accounted for and why apparent counterexamples to the generality of the proposed causal process should not be regarded as refutations of its applicability to the human risk question being evaluated should also be assessed.

The NAAQS causal framework requires only one “high-quality” study for evidence of causal relationship to be deemed “suggestive”. However, EPA does not provide explicit guidance on what constitutes a high-quality study, and, under this requirement, high-quality studies that are inconsistent may not be considered as long as one high-quality study exists that demonstrates an effect. Instead, all studies should be reviewed using the same criteria, and only if the WoE indicates that a causal association is more likely than not, based on all the available data, should one conclude a suggestive causal association. In a similar vein, because EPA groups health outcomes into broad categories, one high-quality study of a single endpoint (e.g. wheezing) can be used as evidence for a suggestive causal relationship for a much broader category that includes more severe outcomes (e.g. respiratory effects).

An assessment of the adversity of health effects is included in the NAAQS causal framework as a step to consider after categorization of causal relationships, but, as noted above, such an assessment can also be considered prior to making judgments regarding causality (i.e. in Phase 3). Regardless, it is key that EPA be more critical when defining an adverse effect, and the assessment should differentiate between homeostatic and biological changes that can adversely affect health (Goodman et al., 2010). Another issue that is worth pursuing in the NAAQS causal framework is the determination of the portion of an adverse health effect burden that could be plausibly attributed to the particular criteria pollutant under review. Even if the WoE indicates causality for a given health effect, the estimated magnitude of the health burden contributed by that pollutant may be very small (or even negligible) in comparison to other causes of or contributors to the same health effect. Failure to place the total burden of an effect into its proper context is a step that is currently lacking in regulatory decision-making.

Many of the issues noted above could be resolved by updating EPA’s categories for causal determination (Table 4) to be more consistent with the IOM framework (Table 3) on which it was based originally, and we recommend such an update to the NAAQS causal framework. This would not only be in line with WoE best practices, but also would improve the use of the causal determination categories for risk communication. In the end, EPA should evaluate all of the data in a consistent manner using well-specified criteria and determine whether, as a whole, they constitute evidence for causation or are more likely indicative of an alternative explanation. It is by considering these alternatives that one sees most directly how the evidence at hand relates to possible conclusions.

Case study: Ozone ISA

The lack of explicit guidance in the NAAQS causal framework has led to it being applied in a manner that does not consistently meet its objectives to provide and communicate a sound scientific basis for causality judgments. As a case study, below, we provide several examples to demonstrate how this lack of explicit guidance led to an overall unbalanced evaluation of health effects in the ozone ISA, which yielded causality conclusions that were not fully supported by the evidence (US EPA, 2013a). The purpose of this case study is not to indicate every instance in which the NAAQS causal
framework is applied inconsistently in the ISA, nor is the aim to reargue evaluations of specific causality questions. Rather, the purpose is to provide several concrete examples of the systematic challenges that arise with the use of the NAAQS causal framework in practice, as a way to indicate that the general issues with the framework, that we describe above, have real occurrences.

Phase 1

The NAAQS causal framework describes a general literature search strategy and inclusion criteria for study selection, but does not provide explicit guidance for this and does not discuss exclusion criteria. EPA provides a database of the studies considered for inclusion, noting which studies were cited in the ozone ISA and which were not; however, it is unclear why particular studies were excluded and if excluded studies were relevant or informative.

For example, in the ozone ISA, EPA (2013a, p 2-2) states, “[l]iterature searches have been conducted routinely since then to identify studies published since the last review, focusing on studies published from 2005 (closing date for the previous scientific assessment) through July 2011”. EPA included the study by Zanobetti & Schwartz (2011) in the ozone ISA but omitted a study by Lipsitt et al. (2011) that was published online the same day (23 June 2011). EPA also omitted a study by Spencer-Hwang et al. (2011), which was published online on 21 July 2011. In addition, there were several studies of both ozone and PM that were not included in the ozone ISA but played a prominent role in EPA’s PM evaluation (e.g. Jerrett et al., 2005; Miller et al., 2007). This indicates that not all relevant studies were captured by the literature search strategy used.

Study quality is a basis for study selection in the NAAQS causal framework, but, as discussed above, studies of lower quality could still provide information and should be included and weighed in later phases of the evaluation. Some lower quality studies that, despite their shortcomings, may have contributed usefully to the WoE were likely excluded early on from EPA’s analysis. The overall study selection process in the ozone ISA builds on past analyses and public comments, so it is possible that some relevant studies that may have been excluded were considered later in the process. However, without a systematic method for identifying and selecting studies up front, the overall assessment may be biased.

Phase 2

Evaluation of study quality

The NAAQS causal framework does not describe a specific method for evaluating the strengths and limitations of individual studies and using these to determine the weight of each study in the evaluation of causality. This led to an inconsistent evaluation of individual studies throughout the ozone ISA. As discussed below, there are many examples in the ozone ISA where studies with positive associations were emphasized over studies with null associations, as opposed to studies of greater quality being emphasized over those of lesser quality. This provided a false perception that most of the reliable evidence supported a positive causal association.

In the discussion of respiratory effects in adult day-hikers in the ozone ISA, positive associations with lung function decrements reported in one study (Korrick et al., 1998) were emphasized over the null associations for the same endpoints reported in another study (Girardot et al., 2006), and there was no discussion of the strengths and limitations of either study. In the evaluation of studies examining short-term ozone exposure and cause-specific mortality, EPA stated that there is evidence of associations with cardiovascular (CV) and respiratory-specific mortality, yet all risk estimates for respiratory mortality and almost all for CV mortality were null in analyses of these endpoints year-round; results were mixed for both endpoints in analyses of the summer months, when ambient ozone concentrations are highest.

In its summary of epidemiology data for short-term effects of ozone on pulmonary inflammation and oxidative stress in the ozone ISA, EPA stated that many recent studies reported positive associations. Yet throughout its discussion, EPA noted that the results were mixed and inconsistent. EPA did not provide an explanation for why studies with positive associations should carry more weight than those reporting null associations. Finally, in the discussion of short-term effects of ozone on respiratory symptoms, EPA stated there is a “strong” body of evidence demonstrating associations between ozone and increased respiratory symptoms (e.g. wheezing) and asthma medication (e.g. bronchodilator) use in asthmatic children, yet almost all risk estimates were null for these outcomes. This also suggests that EPA’s evaluation of the epidemiology data did not fully consider evidence from studies with null results.

EPA’s evaluation of lung function decrements in asthmatic children is another example of how the lack of methods for evaluating the strengths and limitations of studies leads to an overall inconsistent evaluation of individual studies across pollutants. EPA relied on the study by Mortimer et al. (2002) as a key study to indicate that increasing ozone concentrations are associated with decreasing lung function in children with asthma, as determined by self-reported peak expiratory flow rate (PEFR) measurements (i.e. the maximum speed at which an individual can breathe out air, a measure of the degree of airway obstruction). By contrast, in the ISA for nitrogen dioxide (US EPA, 2008a), EPA determined that the same study was unreliable because of the use of self-reported PEFR data, which has been demonstrated to have an unacceptably high rate of error (Kamps et al., 2001).

If the NAAQS causal framework provided explicit guidance on evaluating the strengths and limitations of each study, it would allow for weight to be assigned to the studies in a consistent manner based on quality. It could then be determined whether the large number of studies with null results carry equal or greater weight than the number with positive results, strengthening the evidence against a causal relationship, or whether the positive studies carry more weight and EPA’s emphasis on these results is justified.

Consideration of study limitations

Measurement error. Exposure and outcome measurement biases are potentially significant sources of uncertainty in effect estimates derived from epidemiology studies of ozone
exposure. Most epidemiology studies rely on data from central ambient monitoring sites to provide community-average ambient ozone exposure concentrations (e.g. Gent et al., 2003; Katsouyanni et al., 2009; Mortimer et al., 2002; Naehler et al., 1999; Neas et al., 1999; Stieb et al., 2009), and the interpretation of statistical associations is predicated upon the assumption that these ambient measurements reflect personal exposures. Studies have shown that large discrepancies exist in the ambient/personal ozone relationships, however, and these can vary across cities (e.g. Sarnat et al., 2001, 2005). Thus, the use of fixed ambient monitoring data as surrogates for personal ozone exposures can result in exposure measurement error, which can bias the results of an epidemiology analysis in either direction (Jurek et al., 2005, 2008; Wacholder et al., 1995). Rhomberg et al. (2011b) reviewed exposure measurement errors that prevail in environmental epidemiology studies, demonstrating that the degree of bias known to apply to these studies is sufficient to produce a false low-dose linear result. In the ozone ISA, EPA acknowledged that exposure measurement error is a potential source of variability in the ozone epidemiology study results, but it did not give weight to the fact that it is also a potentially large source of bias that could account for the small statistical associations reported in many of the studies on which EPA relies for evidence of causality.

Outcome measurement error results from inaccurate reporting of health effects, and this was also not always considered a limitation in the ozone ISA. For example, self-measured and self-reported lung function measurements are an important limitation of many ozone respiratory morbidity studies. These measurements can be unreliable, particularly if a population is young and/or not highly motivated to comply and when a long time period of self-reporting is required. Despite this limitation, EPA relied on a number of studies using self-reported measures of lung function as evidence of causality in the ozone ISA. In one key study on which EPA relied for lung function effects in asthmatic children, O’Connor et al. (2008) evaluated lung function changes that were self-measured by the children; EPA did not acknowledge the high frequency of unreliable values in this study. Some other studies on which EPA relied also included self-reporting of symptoms and, as a result, were subject to recall bias. For example, in a study of children with asthma, Gent et al. (2003) used a non-standard questionnaire (a daily calendar) and relied on subjective symptom reporting by participants’ mothers, which could have biased the effect estimates. Overall, in the ozone ISA, EPA did not give appropriate weight to the potential bias from outcome measurement error.

Confounding. Confounding is a major limitation of a number of key ozone respiratory morbidity studies that is not adequately accounted for in the ozone ISA. For example, Mortimer et al. (2002) reported associations between ozone and asthma exacerbation in children in single-pollutant models, but no statistically significant associations were observed in multi-pollutant models. This indicates that the effect may have been influenced partially or fully by other pollutants. Korrick et al. (1998) also adjusted for several confounding factors in their study of ozone effects on lung function in adult hikers, showing that confounding by PM₂₅ and aerosol acidity could not be ruled out. Similarly, in a study by Neas et al. (1999) of associations between decreased lung function in children and ozone exposure, effects were attenuated in a co-pollutant model with sulfate. EPA tends to downplay the effect of co-pollutants, and generally considers only study results from single-pollutant models as positive evidence for causality.

Temperature and other environmental factors can also confound the relationship between ozone and respiratory morbidity, and it is not clear if this is considered by EPA. For example, the longitudinal study of children with asthma by Gent et al. (2003) only considered same-day maximum temperature in the statistical models, while meteorological variables such as relative humidity may have been potential confounders of respiratory symptoms. Air conditioning use and exposure to tobacco smoke are also important potential confounders of the associations with respiratory effects that were not accounted for in some of the key studies on which EPA relied for its causality judgments (Mortimer et al., 2002; Stieb et al., 2009).

Confounding by co-pollutants, meteorology, and other factors is also a major source of uncertainty in time-series analyses of ozone-mortality associations. This issue is complicated by the fact that air pollutants can be highly correlated, and it is often dismissed because of this. Some studies have reported risk estimates that appear to be robust to inclusion of co-pollutants (Bell et al., 2007), but findings from other studies affirm that this issue remains relevant and important. Specifically, many studies have investigated confounding of PM up to 10 micrometers in diameter (PM₁₀), PM₂₅, and some PM components, and found evidence of confounding effects (Franklin & Schwartz, 2008; Katsouyanni et al., 2009; Smith et al., 2009), and a few studies have shown confounding effects of temperature (Smith et al., 2009). For example, Smith et al. (2009) reported that inclusion of PM₁₀ with ozone in a two-pollutant model resulted in a 22–33% decrease in the ozone-mortality association. Similarly, Franklin & Schwartz (2008) reported a 31% decrease in the ozone-mortality risk estimate when sulfate PM was included in the model. Smith et al. (2009) also considered alternative models to control for meteorology that resulted in reduced mortality estimates, suggestive of additional confounding by temperature. EPA should consider both the studies that show confounding as well as those that do not in its assessments.

Although confounding is discussed as part of the evaluation of respiratory morbidity and mortality in the ozone ISA, it does not appear to play a key role in EPA’s ultimate conclusions regarding the causal relationship between ozone and these outcomes.

Model specification. The ozone ISA noted that, in selecting epidemiology studies, consideration was given to methodological issues related to the interpretation of the evidence (including model selection), but this was often not the case. The results of many studies, including meta-analyses of ozone mortality time-series studies, have indicated that model selection has a key role in the determination of results.

The significance of model selection was brought to bear in 2002 when National Morbidity, Mortality, and Air Pollution Study (NMMAPS) researchers discovered an issue with the
way the commonly used statistical model, the Generalized Additive Model (GAM), was implemented by the statistical software, S-plus. Specifically, these issues led to an investigation of how differences in the level of control for temporal trends and weather can affect the magnitude of effect estimates. When compared to other statistical models or different criteria for controlling these important confounding factors, the GAM model in S-plus was generating much larger health effects estimates (HEI, 2003). Similarly, studies have also shown that ozone mortality estimates can vary depending on the degrees of freedom selected for smoothing long-term trend (HEI, 2003; Ito et al., 2005; Katsouyanni et al., 2009).

In the Air Pollution and Health: a European and North American Approach (APHENA) study, which included datasets from US, Canadian, and European multi-city studies, ozone mortality estimates were sensitive to the smoothing function type applied (Katsouyanni et al., 2009). Despite extensive sensitivity analyses comparing a number of different models, Katsouyanni et al. (2009) were unable to identify a model deemed most appropriate for comparing health effect estimates across the different study locations they evaluated. They reported large differences with penalized versus natural splines, as results were negative when penalized splines were used and positive when natural splines were used. In the ozone ISA, EPA only presented the positive associations that were reported from the use of natural splines, because ‘‘alternative spline models have been previously shown to result in similar effect estimates’’. Although EPA provided a justification for why it did not present the APHENA results from both smoothing functions, this justification does not make sense when the large APHENA study (upon which EPA relied heavily in the ozone ISA for its causal determinations) indicates that there is sensitivity of risk estimates to the type of smoothing function used in the model.

These examples highlight the prevalence of model selection bias in the epidemiology literature. Despite findings from several researchers that have shown the substantial impact of model selection bias (e.g. Clyde, 2000; Ito, 2003; Koop & Tole, 2004; Koop et al., 2010; Moolgavkar et al., 2013), EPA has not systematically considered this issue in its causality determinations for ozone.

Phase 3
Consideration of modified Bradford Hill criteria

The NAAQS causal framework indicates that modified Bradford Hill criteria should be used to aid in causal judgments, but it does not provide explicit guidance regarding how these criteria should be applied to the available evidence and how all criteria should be considered jointly. Furthermore, several criteria were not fully considered by EPA in the ozone ISA, including biological plausibility, biological gradient and strength of association. Several examples are described below.

EPA’s evaluation of lag times (i.e. the period of time between the measurement of an exposure and an effect) is an example of not fully considering biological plausibility. Many epidemiology studies utilized statistical models that examined health effects occurring at several time points after measured ozone exposures. Extensive human clinical and mechanistic data on ozone indicate that the respiratory effects of ozone occur soon after exposure. Findings for 0- to 1-day lags or for cumulative ozone exposure over a few days are most likely to be biologically plausible, while findings for longer lags are less likely to be so. Therefore, effects (or a lack thereof) reported at the earlier time points should have carried more weight than those reported at later times, but, in the ozone ISA, effects at any time point were generally considered evidence for causality.

EPA also often overlooked the criterion of biological gradient (i.e. exposure–response relationship) in drawing conclusions in the ozone ISA. For example, EPA noted that there is no exposure–response relationship for lung function changes across short-term exposure studies of outdoor workers (Brauer et al., 1996; Chan & Wu, 2005; Hoppe et al., 1995; Romieu et al., 1998). This provides evidence against a causal association. Yet, in the ozone ISA’s summary and causal conclusion for short-term respiratory effects, this was not discussed, indicating it carried little, if any, weight in EPA’s final evaluation.

As a final example, EPA did not fully consider the strength of associations reported in the ozone epidemiology studies and did not factor this into the overall judgments of causality in the ozone ISA. EPA appears to have considered any risk estimate above the null as supportive of a causal relationship, regardless of the magnitude of the association. Many epidemiology studies report very weak findings, and this is true for studies of ozone. As reviewed by Taubes (1995), epidemiologists generally consider risk estimates greater than 3–4 to reflect strong associations and to be supportive of a causal link, while smaller risk estimates (1.5–3) are considered to be weak and require other lines of evidence to demonstrate causality. In addition, many small, statistically significant associations are found to be not statistically significant when confounders are accounted for. Boffetta et al. (2008) noted that the importance of unmeasured and residual confounding as a source of bias has been sometimes downplayed in the literature, but a recent statistical simulation study showed that, with plausible assumptions, such confounding can generate effect sizes in the range of 1.5–2.0. EPA relied on many studies with risk estimates even lower than this. For example, in EPA’s evaluation of associations between ozone exposure and respiratory symptoms in children with asthma, 35 of 37 risk estimates above the null that were presented as evidence of effects were between 1.01 and 1.39, with two additional risk estimates of 1.59 and 2.13. EPA did not fully consider that the positive associations on which it relied for causal associations are often weak and within the range of magnitude that can be attributed to confounding. If these weak associations are to be considered as evidence of causality, consistent with WoE best practices, EPA should also give equal weight to associations of equal magnitude in the opposite direction (i.e. 0.72 to 0.99) as evidence for a protective effect of ozone. As a protective effect of ozone on the respiratory system is not plausible, associations of this magnitude (either positive or negative) are likely not associated with ozone exposure and instead can be attributed to confounding or other factors.

The above examples indicate some of the specific criteria that were not fully considered by EPA. It is also apparent that
not all of the criteria are considered jointly. For each health effect, EPA should systematically consider each of the criteria and determine where the WoE lies. While every criterion does not have to be met per se, each needs to be considered. For those that are not met, EPA should provide an explanation of why that criterion does not affect the interpretation of the results. With no clear guidance on how to apply the criteria or how to consider them all jointly, EPA’s evaluations of individual studies tend to be biased toward findings of effect.

Weighing alternative views

The NAAQS causal framework indicates one should weigh alternative views on controversial issues, but it does not provide guidance on how to do this. The framework should be more explicit in this regard, e.g. by noting that evidence for and against competing explanations should be discussed and clear illustrations of how one explanation compares to the other should be provided. There are several instances in the ozone ISA where EPA did not discuss or give due weight to alternative views.

For example, EPA noted that consideration of the limitations of epidemiology studies, such as potential confounding and exposure measurement error, must be taken into account to properly inform the interpretation of epidemiology evidence. Yet, in its final evaluation, EPA did not consider that another factor may have caused the health effects associated with ozone in certain epidemiology studies. EPA did not discuss the reasons why this view (i.e. ozone is not causal) is less likely to be true than the view that ozone is the causal factor.

In addition, EPA reviewed several recent controlled human exposure studies of ozone and decrements in lung function over a total of 6.6 hours of exposure (Adams, 2006; Schelegle et al., 2009; Kim et al., 2011). The mean lung function decrement for the subjects in each study after exposure to 60 parts per billion (ppb) ozone was only statistically significant in the study by Kim et al. (2011). Because of concern that the statistical test in the original publication was not sufficiently powerful to detect an effect on lung function, EPA reanalyzed the data for the final time point (but not at other interim hourly time points) in the study by Adams (2006) using several statistical methods that differed from those of the study author. This reanalysis, reported by Brown (2007) and Brown et al. (2008), indicated statistically significant lung function decrements at 60 ppb ozone. Other investigators have also reanalyzed the data from the study by Adams (2006) using various statistical methods and reported no statistically significant changes in lung function (Lefohn et al., 2010; Nicolich, 2007). It has been argued that EPA’s approach produced statistically significant results that can be attributed to the majority of the data being selectively omitted from the analysis (Goodman & Sax, 2012). EPA appeared to weigh the findings of its own reanalyses over the original analyses conducted by the study authors and reanalyses by other investigators without fully considering the strengths and limitations of each analytical method.

As a final example, EPA examined the concentration-response (C-R) relationship between short-term ozone exposure and mortality, but it did not weigh alternative views regarding the shape of the C-R curve and the possible existence of a threshold (i.e. the minimum ozone concentration necessary to produce an effect). It is well recognized that epidemiology analyses such as those that investigate associations between mortality and air pollution exposures are complicated by the noise in the observational data and uncertainties in the epidemiology models, both of which may mask a threshold. Despite the challenges in identifying ozone thresholds in ozone-mortality evaluations, several studies have applied various statistical techniques to evaluate the nature of the ozone-mortality C-R function and found evidence suggestive of a threshold for ozone-mortality effects (Smith et al., 2009; Stylianou & Nicolich, 2009; Xia & Tong, 2006). Similarly, Jerrett et al. (2009) reported evidence of improved model fit with a threshold model. EPA appeared to weigh the view that there is no clear threshold over the alternative view that a threshold exists, even though there is mounting evidence to support the latter.

Overall, while EPA often discusses alternative views, it does not appear to weigh them appropriately. To avoid this, the NAAQS causal framework should provide explicit guidance regarding how to discuss and compare the evidence for and against competing explanations.

Phase 4

The NAAQS causal framework claims to rely heavily on the consistency of results across studies in its categorization of causal relationships between criteria pollutants and different health effects, but, in practice, the consistency of results is not evaluated fully. An example of this can be seen by comparing the classification of causal relationships for ozone-related mortality in the ozone ISA to those in the previous NAAQS review of ozone, as summarized in the 2006 AQCD for ozone (US EPA, 2006).

EPA’s causal determination for short-term ozone exposure and total all-cause mortality was upgraded from “highly suggestive” in the previous review to “likely causal” in the ozone ISA. In the 2006 AQCD, a large body of epidemiology studies, including single- and multi-city studies, was evaluated. Only a limited number of additional studies have been published since then, including several that are follow-up studies or reanalyses of previous studies. Importantly, EPA stated in the ozone ISA that the new evidence is supportive of a robust association between short-term ozone exposures and mortality, but the studies showed very small pooled mortality effects that are inconsistent across studies. Furthermore, multi-city studies have demonstrated that there is substantial heterogeneity in city-specific findings, which are dominated by ozone associations that are not statistically significant and close to null, or even negative (i.e. indicating reduced mortality with increased ozone exposure). The pooled “national” ozone-mortality effect estimates are therefore misleading, as they do not reflect this heterogeneity. In addition, EPA did not fully consider many of the important uncertainties that are still present in the new studies, including confounding, model selection, appropriate lag times, non-threshold assumptions in the C-R relationship, and exposure measurement error.
With respect to long-term ozone exposure and mortality, EPA relied on very limited new evidence since the last ozone review in upgrading its classification from “insufficient” to “suggestive” evidence of a causal relationship in the ozone ISA. Specifically, EPA appears to have relied solely on one new study that provided some findings suggesting an increased risk of respiratory-related mortality for long-term ozone exposure (but no consistent associations for all-cause mortality) (Jerrett et al., 2009), and did not fully consider other key studies that suggest no effects of ozone on total and cause-specific mortality or counterintuitive geographical heterogeneity in mortality risks (e.g. Abbey et al., 1999; Dockery et al., 1993; Jerrett et al., 2005; Lipfert et al., 2000; Miller et al., 2007). In addition, new studies published since the last review support a lack of association between long-term ozone exposure and mortality from specific causes, including respiratory causes (e.g. Lipsett et al., 2011). This example illustrates that requiring only one positive (which may be high-quality, but still may have issues) study to provide “suggestive” evidence can yield conclusions toward a causal determination even if the majority of the evidence does not support this.

Because the NAAQS causal framework does not fully evaluate the consistency of results across studies or fully incorporate other Bradford Hill criteria such as coherence, biological plausibility, biological gradient and strength of association, the resulting causality categorization is not congruent with judgments based on these criteria. To address this, EPA could redefine these categories so that they more accurately reflect the WoE. Further, in harmony with best practices, EPA should evaluate all of the data in a consistent manner, using well-specified criteria that can be applied consistently across evaluations over time, to determine whether they constitute evidence for causation.

Conclusions

The current NAAQS framework for causal determination has several valuable features, but there are certain elements that are missing or not stated explicitly. To improve the NAAQS framework, EPA could add explicit guidance regarding study selection criteria. These criteria should include specifying that no studies be excluded based on their quality or results; rather, weak studies should be given less weight in the final analysis. Also, to reduce bias, analyses conducted by EPA should not be included unless they are peer reviewed. In addition, guidance for data extraction, clearly stating that all data relevant to the causal question should be extracted from each study and not just a subset of the data, should be added to the NAAQS framework. In terms of evaluating study quality and results, this should be done in the same manner across all studies to ensure that the assessment is thorough and consistent. Data suggesting a lack of biological plausibility or a lack of an exposure–response relationship for a specific effect should be considered as evidence against causation.

The NAAQS causal framework should provide guidance on how the Bradford Hill criteria should be applied to the available evidence and how all criteria should be considered jointly; the lack of such guidance in the current framework can lead to an assessment that unduly favors a causal determination. Indeed, in following the current framework, EPA sets forth a hypothesis and determines whether the data support that hypothesis. The framework should include guidance regarding the weighing of alternative hypotheses, explicitly directing EPA to fully explore whether and to what degree the data support other hypotheses. The current framework’s method of integrating judgments for different realms of evidence at the end the evaluation also does not allow data from one realm of evidence to inform the interpretation of the others. The framework should state explicitly that the data evaluation be integrated across all realms of evidence to provide insight into potential biological processes underlying the effects under consideration for causality.

Lastly, there should be guidance for causality categorization that is congruent with the judgments based on the Bradford Hill criteria, and the use of categories for causal determination should be based on a consistent evaluation of all of the data using well-specified criteria for causation. Although the current NAAQS causal framework includes several of the elements discussed above, it does not always contain explicit guidance for them, and this can lead to conclusions of causality that are not fully supported by the evidence.

We provided some specific examples of how the limitations of the current NAAQS causal framework led to an evaluation of ozone-related health effects that was not always consistent and yielded causality conclusions that were not fully supported by the evidence. This is also apparent in other assessments that used the NAAQS causal framework. For example, in a 9 December 2011, letter to EPA, the Clean Air Scientific Advisory Committee review panel for the first draft ISA for lead recommended a “more rigorous and transparent ‘weight of the evidence’ analysis” that should “devote more attention to the limitations of the existing studies with respect to consistency, reproducibility, bias, control for confounders, and shortcomings in statistical methodology” (Frey & Samet, 2011). In addition, many of the same issues were also identified in EPA’s draft IRIS assessment of formaldehyde, prompting the NAS panel that reviewed that assessment to propose a roadmap for improving the risk assessment process and to recommend that EPA use existing models of systematic WoE approaches for hazard identification.

If EPA undertakes a program to revise the NAAQS causal framework to be more consistent with the NAS roadmap, it is our hope that the suggestions in this paper will be considered in the process. The NAAQS causal framework could be revised so that it is consistent with WoE best practices, and the causal determination categories should be updated to reflect these revisions. The full and consistent application of the revised framework is needed to ensure that future assessments of the potential health effects of criteria air pollutants will be thorough, transparent, and scientifically sound.

Acknowledgements

The authors thank the anonymous reviewers whose comments helped to improve this paper.
Declaration of interest

The authors are employed by Gradient, a private environmental consulting firm. The Gradient staff has strong expertise in assessing human, experimental animal, and mechanistic data in WoE analyses (as is evident in recent evaluations conducted for bisphenol A, naphthalene, formaldehyde, chlorpyrifos, methanol, styrene, nickel, and toluene diisocyanate) and has presented several of these analyses to regulatory bodies. In addition, the Gradient staff, including the authors of this paper, have carefully evaluated the science underlying EPA’s review of various NAAQS and offered both oral and written testimony to EPA. Gradient has also addressed issues on systematic review and integration of evidence for a number of clients. The work reported in this paper was conducted by the authors during the normal course of employment at Gradient with financial support provided by the American Petroleum Institute (API). API is the major trade association of corporations in the petroleum sector, from discovery through production and refining. Drafts of this paper were reviewed by members or affiliates of API. The authors have the sole responsibility for the writing, contents, and conclusions in this paper. The conclusions are not necessarily those of API.

References

Adams WC. (2006). Comparison of chamber 6.6-h exposures to 0.04–0.08 ppm ozone via square-wave and triangular profiles on pulmonary responses. Inhal Toxicol, 18, 127–36.
Jurek AM, Greenland S, Maldonado G. (2008). How far from non-differential does exposure or disease misclassification have to be to bias measures of association away from the null? Int J Epidemiol, 37, 382–5.


McClellan RO. (2012). Role of science and judgment in setting national ambient air quality standards: how low is low enough? Air Qual Atmos Health, 5, 243–58.


