NCASI
IMPACT. SCIENCE. SOLUTIONS.

NOVEMBER 22, 2021

TO: USEPA
Docket ID No.
EPA-HQ-ORD-2014-0859-0073

FROM: Giffe Johnson, PhD

SUBJECT: Supplement to the 2019 Integrated Science Assessment for Particulate Matter (External Review Draft)

The National Council for Air and Stream Improvement (NCASI) greatly appreciates the opportunity to submit comments to the United States Environmental Protection Agency (USEPA) on Supplement to the 2019 Integrated Science Assessment for Particulate Matter (External Review Draft). NCASI is a research organization engaged in conducting research on environmental topics relevant to the forest products industry. Over its 75-year history, NCASI has conducted studies in a variety of areas related to air emissions and has worked extensively in developing emissions data used in multiple National Emissions Standards for Hazardous Air Pollutants (NESHAP) rulemakings affecting this industry.

NCASI agrees with the mission of the USEPA under the Clean Air Act to protect public health by setting National Ambient Air Quality Standards (NAAQS). However, this policy should be supported by the best available science, integrated within a reliable systematic review framework that produces an accurate characterization of the relationship between criteria pollutants such as particulate matter and potential health effects. As such, NCASI offers the following technical comments and work products to inform the basis for drawing scientific conclusions in the Integrated Science Assessment. These comments relate to the areas of Systematic Review and Other Important Considerations for Causal Inference.

NATIONAL COUNCIL FOR AIR
AND STREAM IMPROVEMENT, INC.

402 SW 140th Terrace
Newberry, FL 32669-3000

(352) 331-1745
(352) 331-1766
ncasi.org

Systematic Review

Several institutions that support science-based policy development have pursued the adoption of increasingly rigorous systematic review methodologies, including the National Toxicology Program (NTP) under the Toxic Substances Control Act (TSCA), the National Academies of Sciences (NAS), and EPA's Integrated Risk Information System (IRIS) program. The effort undertaken by these institutions to adopt more rigorous systematic review procedures is done in order to more accurately rank, weight, and evaluate quality of individual studies within a framework to more reliably draw conclusions related to exposure/disease relationships.

While the current Particulate Matter Integrated Science Assessment (ISA) does compile a large swath of scientific literature related to the potential health effects from exposure to particulate matter, many, if not most of the critical features of systematic review are absent from the current process. As such, studies presented in the ISA have not been appropriately ranked based on study quality and method veracity. This leads to the reliance on studies that either have disqualifying amounts of uncertainty inherent to their design or are not designed to address the policy relevant question at hand and, in some cases, exclusion of studies from evidence integration that may be extremely informative for evaluating cause-and-effect relationships between particulate matter and health outcomes.

In fact, the review approach relied on in the NAAQS Causal Framework is uniquely antiquated by comparison to other EPA program areas. Table 1 highlights critical elements of systematic review needed for causal analysis and risk assessment found in the approaches the Office of Pesticide Programs (OPP), the Integrated Risk Information System (IRIS), and the Integrated Science Assessment (ISA). The review approaches of these three program areas were surveyed for prescriptive criteria in the relevant information domains to determine if enough detailed guidance was provided for reviewers to ensure the capture of necessary data elements for systematic review and integration of evidence. The table illustrates that the current ISA process either lacks or does not clearly define criteria for more critical elements of systematic review needed for causal analysis and risk assessment than other EPA programs that make similar regulatory determinations. The result of this lack of detailed criteria is that the absence of requisite data for decision making may be ignored or heavily influenced by reviewer subjectivity.

Table 1: Critical elements of systematic review needed for causal analysis found in the approaches used by the Office of Pesticide Programs (OPP), the Integrated Risk Information System (IRIS), and the Integrated Science Assessment (ISA)

	Includes Detailed Criteria		
Systematic Review Elements	**OPP**	**ISA**	**IRIS**
Confirm outcome	Yes	Not detailed	Yes
Confirm exposure	Yes	Not detailed	Yes

Report methods fully and transparently	No	Not detailed	Yes
Include information on shape of the curve	Yes	Not detailed	Yes
Harmonize exposure categories	No	No	No
Describe direction/magnitude of error	Not detailed	No	Yes
Evaluate source-to-intake pathways	Not detailed	No	Not detailed
Describe complete exposure data	No	No	Not detailed
Report on quality assurance/quality control	Yes	Not detailed	Not detailed

In the original 2019 Integrated Science Assessment, cardiovascular mortality was considered one of the strongest lines of evidence for considering the need to potentially lower the annual $PM_{2.5}$ standard. It is indicated in the Supplement that new science is essentially the same as previous work in this area:

> *"Overall, these recent studies support the conclusions in the 2019 PM ISA of consistent positive associations of long-term PM2.5 exposure with cardiovascular mortality, and specifically with IHD- and stroke-related mortality."* 3-39

These studies were also evaluated using the NAAQS Causal Framework, as before, and include the same risk of bias issues that were inherent to the previous literature, but not integrated into the causal conclusions in a systematic approach, as before. Uncertainty and risk of bias of these, even when acknowledged, were not systematically integrated into shaping the conclusions of the ISA.

In the 2020 rulemaking, the decision was made to retain the current particulate matter standards based on the conclusion that substantial uncertainties are associated with the current evidence base that evaluates potential health impacts below the current NAAQS. The principal areas of uncertainty that USEPA identified in the evidence base that precluded reliance on these studies for altering the standard include: 1) the various methods used to estimate $PM_{10-2.5}$ concentrations have not been systematically evaluated, contributing to uncertainty regarding the spatial and temporal correlations in $PM_{10-2.5}$ concentrations across methods and in the $PM_{10-2.5}$ exposure estimates used in epidemiologic studies; 2) beyond the uncertainty associated with $PM_{10-2.5}$ exposure estimates in epidemiologic studies, the limited information on the potential for confounding by co-pollutants and other unmeasured confounders also broadly contributes uncertainty to the evidence base; and 3) uncertainty related to the biological plausibility of serious effects caused by $PM_{10-2.5}$ exposures results from the small number of controlled human exposure and animal toxicology studies that have evaluated the health effects of experimental $PM_{10-2.5}$ inhalation exposures.

NCASI agrees that these uncertainties exist in the current evidence base and substantially limit the

interpretation of these studies regarding the relationship between $PM_{10-2.5}$ exposures at policy relevant concentrations and adverse health effects. This was demonstrated by NCASI, in collaboration with subject matter experts, wherein we developed a proposed systematic review protocol to evaluate this evidence base in order to more specifically characterize the uncertainties and study quality issues that exist within currently available studies. This proposed systematic review protocol, based on the Office of Health Assessment and Translation (OHAT) framework was applied to six (6) articles highlighted in the 2019 Integrated Science Assessment for Particulate Matter (External Review Draft) as important in determining cause and effect relationships between policy relevant concentrations of $PM_{2.5}$ and adverse health effects. The core findings of the work indicated that if risk of bias was integrated systematically into the evaluation, no articles were of sufficient quality to serve as primary lines of evidence for a causal relationship between policy relevant concentrations of $PM_{2.5}$ and cardiovascular mortality.

In order to evaluate if the findings of new research considered in the Supplement to the 2019 Integrated Science Assessment for Particulate Matter (External Review Draft) were also associated with similar risk of bias issues, NCASI in collaboration with subject matter experts applied this systematic review protocol to three (3) large cohort studies in the Supplement indicated as continuing evidence in the area of cardiovascular mortality (see Attachment 1).

The risk of bias analysis, performed using the NCASI proposed systematic review protocol, ranks studies as Tier 1, 2, or 3 through an in-depth analysis of study features and methods. Tier 1 studies are those that directly contribute to the evidence base to support an exposure/disease relationship. Tier 2 studies are not sufficient on their own to evidence an exposure/disease relationship but may support Tier 1 studies in evidence integration. Tier 3 studies are considered to have a degree of risk of bias that disqualify them from contributing to the evidence base. In our analysis of six (3) studies featured in the ISA and PA, one (1) study ranked Tier 2 and two (2) studies ranked Tier 3. The results of this analysis indicate that none of the reviewed studies are of sufficient quality to directly contribute to the evidence base as primary sources of evidence. Key risk of bias domains that prevented these articles from being reliable as primary lines of evidence were the lack of individual level control for confounding, poor control for exposure misclassification, inappropriate model specification, and poor control for external sources of bias. While a full evidence integration and complete systematic review were outside the scope and resources of this study, the articles selected for review were highlighted in the Supplement as important pieces of evidence, indicating that it is unlikely that higher quality studies exist in the current body of literature.

On the basis of applying a systematic review framework to sentinel studies featured in the ISA and PA, NCASI agrees with the scientific defensibility of the USEPA decision to not lower the current NAAQS standards for particulate matter. The attached analysis details sources of uncertainty, bias, some which are identified by USEPA and some additional sources that were not. The conclusion of our analysis is that no body of evidence of sufficient quality is available to demonstrate a clear cardiovascular mortality risk from particulate matter exposures at current levels of the NAAQS.

Other Important Considerations for Causal Inference

a. Strength of Association in the Context of Uncertainty, Bias, and Confounding

In his sentinel work *The Environment and Disease: Association or Causation?*[1], Sir Bradford Hill describes important features in the body of evidence to consider when attempting to draw causal inference between an exposure and a disease. A key element of this discussion is the strength of the association found between the exposure and disease across multiple studies. In general, the stronger the association calculated across studies, the more likely one is to conclude that a causal relationship exists. It is important to note that the strength of association is not informed simply by the presence of statistical significance, but rather the magnitude of impact that an exposure has on the amount of disease in a population. While it may seem at face value a very simple proposition that the stronger the association the more likely it is for an association to be causal, the element of associative strength actually provides more nuanced information for causal analysis than just that.

All epidemiological studies are plagued with some degree or combination of uncertainty, bias, and confounding. What makes a study suitable for causal analysis is that the results of the study survive the amount uncertainty, bias, and confounding present so that the directionality of the results (e.g. positive, negative, or null association) remain reliable even if the quantifiable exposure/outcome relationship remains imprecise.

A robust strength of association provides insulation against subtle amounts uncertainty, bias, and confounding. For instance, in the well-established report by the Surgeon General on smoking and health[2] in 1964, studies at the time indicated a 9- to 20- fold increased risk of lung cancer from smoking cigarettes. These studies were not free of uncertainty, bias, and confounding. While there were robust data on exposure (cigarettes counted per day) and outcome (diagnosis), researchers were not able to rule out the influence of other environmental and genetic risk factors for lung cancer. Despite this, having robust (yet imprecise) measures of association supported a conclusion of causality as even moderate amounts of uncertainty, bias, and confounding were unlikely to revert these associations to the null if fully measured and controlled for. The robust measures of association between cigarette smoking and lung cancer allowed for a conclusion of causality, even in the presence of some uncertainty, bias, and confounding. In other words, if a study finds an association of 15, but the presence uncertainty, bias, and confounding mean that the real measure of association is really 7, this wouldn't prevent the conclusion of causality.

The body of evidence that evaluates the potential association between particulate matter and health effects at current NAAQS standards wholly lacks the insulation of robust measures of association to protect against the impact of uncertainty, bias, and confounding completely altering the directionality of association in these studies.

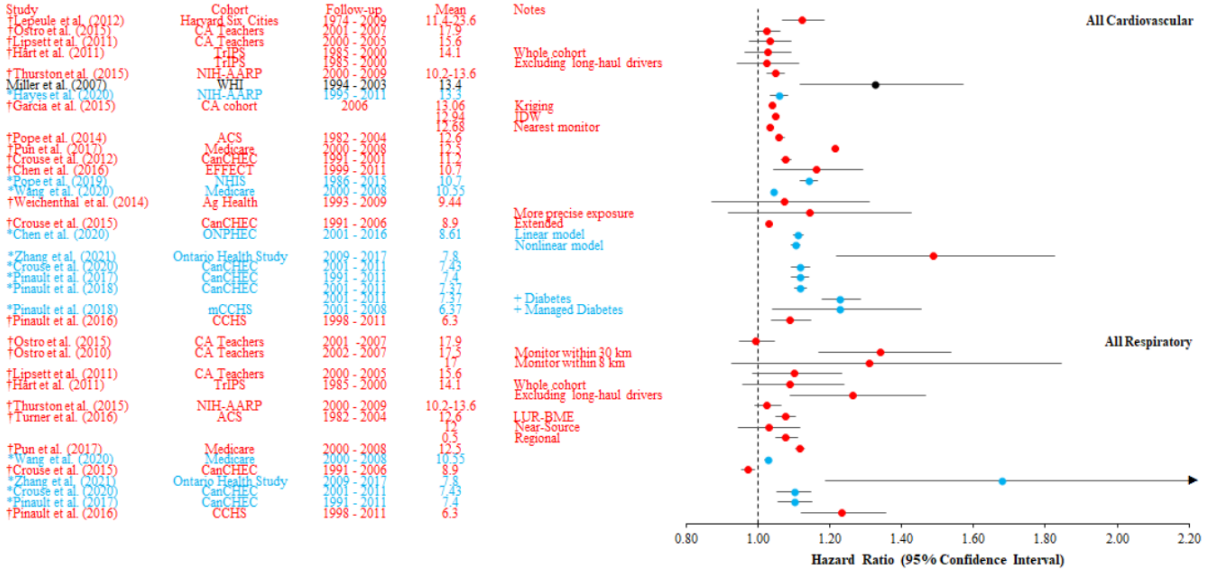
As an example, consider this forest plot of associations (Figure 1; found in the Supplement to the ISA) relevant to cardiovascular and respiratory outcomes related to policy relevant concentrations of $PM_{2.5}$.

---

[1] Hill, Austin Bradford (1965). "The Environment and Disease: Association or Causation?". Proceedings of the Royal Society of Medicine. 58 (5): 295–300. doi:10.1177/003591576505800503

[2] United States Public Health Service. Smoking and Health: Report of the Advisory Committee to the Surgeon General of the Public Health Service. Washington, DC: US Department of Health, Education, and Welfare; 1964.

Figure 1. Cardiovascular and Respiratory outcomes related to $PM_{2.5}$



Very few measures of association found in this evidence exceed 1.2. As a whole, even subtle amounts of unmeasured uncertainty, bias, and confounding present in this evidence base could potentially alter the directionality of these associations if accounted for. Given that much of this body of literature relies on similar subject selection, statistical methods, data insufficiency for confounding, and the high potential for exposure misclassification (see Attachment 1 for additional discussion), there is not a reliable conclusion of causation that can be reached. There is simply no insulation provided by robust measures of association in this body of evidence to survive even subtle amounts of uncertainty, bias, and confounding, which to some degree are endemic to observational epidemiological studies in general, but in this evidence base, we have a body of evidence with no individual exposure measurement and no consistent adjustment of confounding with individual confounder data.

b. Presence of Unmeasured Confounding

On page 3-95 in the Supplement to the ISA, there begins discussion on the evidence for unmeasured confounding that may be present in large secondary data-set study designs, such as those that rely on the medicare data set that occur frequently in the evidence base. It is noted:

> *"Similarly, Pun et al. (2017) completed a sensitivity analyses as part of their Medicare cohort study for the years 2000–2008 in which they decomposed PM2.5 into "temporal" and "spatiotemporal" variation, analogous to what was done in Greven et al. (2011). The purpose of this sensitivity analysis was to determine the presence or absence of bias due to unmeasured*

confounding. Pun et al. (2017) observed positive associations for the "temporal" variation model and approximately null associations for the spatiotemporal" variation model for all causes of death except for COPD mortality. The difference in the results of these two models for most causes of death suggests the presence of unmeasured confounding…"

However, the evidence of the presence of this unmeasured confounding is not accounted for in the causal conclusions regarding the potential health effects of $PM_{2.5}$ and mortality. In the study noted above (Greven et al. 2011), residual confounding remained even after attempting demographic level adjustment for confounding with the BFRSS dataset. In reviewing this literature, a systematic approach would need to recognize that demographic level confounding adjustment is not adequate to remove this confounding and would require a robust measure of association that could survive the potential existence of residual confounding that has been demonstrated to be present. A systematic review approach could address this issue through formal risk of bias analysis.

c. Use of epidemiological data for quantitative exposure-response analysis.

In the Supplement to the ISA it is noted that (emphasis added):

Although studies evaluated in the 2019 PM ISA have used many different statistical methods to examine the shape of the C-R relationship and generally provided evidence for a linear, no-threshold relationship, ***many of these studies have not systematically evaluated alternatives to a linear relationship****. -2-29*

EPA has correctly noted that the conclusion of a linear, non-threshold relationship between PM and health outcomes is largely the product of methodological approach. Given the previous discussion regarding the findings that if viewed through risk of bias analysis, there are no evident primary lines of evidence to conclude causality between PM and health effects such as cardiovascular mortality, it is equally challenging, if not more so to rely on these studies to develop concentration-response curves from these data. In addition to previously discussed issues of uncertainty, bias, and confounding that may be present, model specification, that is the selection of statistical models that rely on the underlying assumption of linearity or proportionality along the exposure-response curve may overestimate risk at the upper and lower ends of the curve, where in traditional toxicology dose response models, the risk rapidly accelerates (from the lower end) or rapidly attenuates (at the upper end). Models such as the Cox Proportional Hazard Model, which is popular among epidemiologists for it's ability to manage censored data, suffers from this underlying assumption. The result is that concentration-response outcomes from areas of the curve where the true relationship is sub-linear or attenuates may be over-estimated with this approach. Criteria regarding methods appropriate for threshold detection and non-linearity are needed to evaluate primary lines of evidence that can be useful for generating reliable, non-linear, concentration response curves before conclusions regarding concentration response can be made with certainty.

We thank the EPA staff and the for reviewing and considering these technical comments. Please feel free to contact NCASI with questions or request for further information.

Submitted respectfully,

Giffe Johnson, PhD
Principal Scientist/Program Manager
Toxicology, Epidemiology, and Risk Assessment
National Council for Air and Stream Improvement (NCASI)

Attachment 1

Protocol for a Systematic Review to Evaluate the Potential Association between Cardiovascular-Related Mortality and Exposure to Policy Relevant Concentrations of PM2.5

EXECUTIVE SUMMARY

In September 2021 EPA published the External Review Draft of the Supplement to the 2019 Integrated Science Assessment (ISA) for Particulate Matter to respond to the petition to reconsider the National Ambient Air Quality Standard for Particulate Matter. In this effort, the EPA reviewed potentially relevant scientific articles related to PM exposure and the potential for human health impact that were published after the cut-off date of the 2019 ISA.

While the PM ISA successfully summarizes a substantial amount of literature related to the potential human health impacts of PM exposure, the approach used to critically evaluate and integrate lines of evidence produced in the literature search lacks several critical features of modern systematic review practices. Areas of the PM ISA that demonstrate potential for improvement include selection of policy-relevant research questions, evaluation of risk of bias, and prescribed approaches to the integration of evidence. Use of modern systematic review approaches in these areas will improve both the transparency of how scientific conclusions are arrived at in the PM ISA process, and the reliability of conclusions drawn in the PM ISA that will ultimately be relied on for policy decision making.

In an effort to further the dialog with EPA staff scientists on efforts to continually improve upon the ISA process and ensure that best science is used to inform the Policy Assessment, NCASI staff, in collaboration with subject matter experts, has developed an example of a protocol that demonstrates how modern systematic review approaches would improve and potentially alter the conclusions drawn by a contemporary analysis of the PM human health literature.

This document presents a proposed systematic review protocol for evaluating the impact of PM2.5 exposure on the outcome of and cardiovascular-related mortality. In the appendices, scientific articles highlighted in the PM ISA supplement as being salient to conclusions drawn regarding the relationship between PM2.5 exposure and cardiovascular-related mortality are treated with the approach outlined in the protocol as a demonstration of the impact of these modern systematic review methodologies on interpretation of scientific evidence.

Reviewers
Linda Dell, MS
Carol Burns, PhD
Giffe Johnson, PhD

TABLE OF CONTENTS

Protocol for a Systematic Review to Evaluate the Potential Association between Cardiovascular-Related Mortality and Exposure to Policy Relevant Concentrations of PM2.5

1.0 INTRODUCTION

1.1 Background and Rationale

In developing the Integrated Science Assessment (ISA) for Particulate Matter (PM), EPA reviews and summarizes evidence from studies on atmospheric sciences, dosimetry, human exposure, animal toxicology, mode of action, controlled human exposure, epidemiology, biogeochemistry, and/or terrestrial and aquatic ecology and other welfare effects in order to inform risk management and policy decisions under the National Ambient Air Quality Standards (NAAQS). Two principal documents, the NAAQS Integrated Review Plan (USEPA 2016) and the NAAQS Preamble (USEPA 2015), discuss overarching principles and provide guidance for conducting reviews under the ISA process and suggest incorporation of some elements of systematic review, but fall short of prescribing the use of a fully featured systematic review framework to address key policy questions. In addition, the Clean Air Act places strict deadlines on EPA to complete the ISA process, making it an impracticable endeavor to evaluate the numerous lines of scientific evidence that inform the health effects of PM exposures within a robust systematic review framework while adhering to modern practices of literature selection, study quality evaluation, and evidence integration.

However, in the absence of such a framework the process becomes less robust and reliable conclusions cannot be reached regarding the relationship between exposure and health outcomes. This is largely a result of the inherent limitations of epidemiological studies in describing well-defined exposure-response relationships at low levels of exposure amid the substantial methodological issues of controlling uncertainty under those research conditions. A body of evidence of this nature requires detailed and well-defined criteria of evaluation in order to reach scientifically defensible conclusions. EPA's recent evaluation of potential health impacts from exposures to decreasing concentrations of particulate matter ≤2.5 microns in diameter (PM2.5) within the framework of the ISA by using traditional methods has this drawback.

Formal systematic review protocols are one way to balance the need for efficient, yet detailed and transparent, literature reviews within ISAs, while conforming to modern approaches for literature selection, study quality assessment, and evidence integration. These protocols should address key policy-relevant research questions within the broader context of scientific literature as a whole. The purpose of the systematic review protocol presented herein is to provide a framework for evaluating the relationship between policy-relevant exposure concentrations of PM2.5 and cardiovascular-related mortality, which can serve as a model that EPA can apply under the NAAQS ISA framework. The goal is to provide EPA with the necessary tools for review of research questions that are targeted and narrower in scope, but still require robust conclusions with a high degree of certainty to inform policy decisions.

1.2 Objectives

The primary objectives of this document are to: (1) provide a protocol for addressing policy-relevant research questions using contemporary approaches to systematic review; and (2)

demonstrate applicability of this protocol to address a specific area in the NAAQS PM ISA. This is consistent with stated goals of the Integrated Review Plan (USEPA 2019, 3.1 Scope of the PM ISA):

> In order to provide a more focused evaluation of the scientific evidence for health and non-ecological welfare related effects, the PM ISA will discuss the most important topics that address policy-relevant questions. Therefore, the PM ISA will more fully evaluate those health and non-ecological welfare effects for which the evidence in the 2009 PM ISA was less certain (i.e., effects where the causal determination was "likely to be causal", "suggestive", or "inadequate" as detailed below in section 3.4.3) and where there is now a larger body of evidence (e.g., diabetes, nervous system effects, etc.). For those health and non-ecological welfare effects where the 2009 PM ISA concluded that the evidence was sufficient to infer a causal relationship (i.e., health: short- and long-term PM2.5 exposures and cardiovascular effects; short- and long-term PM2.5 exposures and mortality; and welfare: PM exposures and effects on visibility, climate, and materials), the PM ISA will focus more specifically on policy-relevant considerations, such as the level at which effects are observed, and on characterizing the extent to which new studies address key uncertainties and limitations identified in the previous review or provide insight on new issues [p. 3-2]

The specific aims of the study protocol are to provide a systematic review and evidence integration framework to address this research question:

1. Among the adult population, are cardiovascular mortality effects, observed in studies accounting for confounding and other biases, related to exposures to annual average PM2.5 <12 $\mu g/m^3$ (equivalent to the current PM2.5 NAAQS)?

While previous ISAs (e.g., USEPA 2009) have described the potential causal link between PM2.5 exposure and the outcomes of mortality, these links were not focused on exposure concentrations relevant to current PM2.5 NAAQS standards. Therefore, while EPA has heretofore classified these outcomes as being causally associated with PM2.5 exposures, the approach has largely ignored exposure concentration. The protocol proposed here specifically addresses this issue, seeking to study whether effects are observed at levels below the current NAAQS in studies in which confounding, bias, and chance are not likely to explain the observed association.

2.0 METHODS

2.1 Eligibility Criteria

Participants, Exposure, Comparator, Outcomes, and Study Design (PECOS) statements are developed for each health outcome included in the framework. Studies are selected for inclusion according to the criteria specified in the PECOS statements shown herein for the outcome of cardiovascular (Table 1). Only human studies are included for mortality outcomes, whereas animal and mechanistic studies are included for the IHD outcome.

Excluded publications include reviews and secondary research, editorials, and studies that analyze surrogates of exposure to PM2.5 (e.g., distance between residence and roadways, traffic counts), irrelevant exposures or outcomes, or that are absent stated requirements in the PECOS tables.

Additional studies used to inform the analysis but separate from the formal systematic review include studies that inform on the potential for uncertainty, confounding, or publication bias among the selected studies. These studies are incorporated into the evidence integration process to assist in evaluating the overall degree of uncertainty present in the body of evidence.

Table 1. PECOS Statement for Outcome of Cardiovascular Mortality

Population	Adult population (18 years or older)
Exposure	Particulate matter ≤2.5 μm in diameter (PM2.5) through inhalation route at or below exposure concentrations ≤12 μg/m^3; studies that rely on surrogate or indirect exposure measurements (e.g., distance from exposure source) excluded
Comparator	Population of same demographics as target population with PM2.5 exposures that include discrete concentration ranges below the current NAAQS standards to be compared; for time series studies, other days exposed will be the comparator
Outcome	Cardiovascular-related mortality as indicated by hospital death certificates
Study Design	Studies that compare discrete categories of exposure at and below the current NAAQS standard in a prospective cohort, retrospective cohort, case cross-over, or time series designs; only studies with estimates of relative risk in relation to quantitative estimates of exposure included

2.2 Search Strategy

The search for relevant articles is conducted using the EPA HERO database, PubMED, SCOPUS, and BASE and the keywords PM2.5, particulate matter, mortality, cardiovascular disease, and cardiovascular mortality with duplicates removed. The flow of record search to final inclusion proceeds as shown in Figure S1 found in Appendix S.

2.3 Data Extraction

Data extraction is conducted using the appropriate template by study type as shown in the appendices.

2.4 Assessment of Quality of Evidence/Risk of Bias

The risk of bias for each included study is assessed using a modified instrument developed based on the National Toxicology Program's Office of Health Assessment and Translation (OHAT) Risk of Bias tool for the generally accepted risk of bias domains: selection bias; confounding; performance bias; attrition bias; detection bias; and reporting bias (NTP 2019). The tool is modified to address methods for assessing PM exposure to include considerations of modeling, monitoring, and other methodologies.

OHAT criteria, modified to address PM ecological studies and specific issues related to PM health effects evaluations, is used for each domain. Table 2 shows specific domains, domain questions, and judgement criteria. The risk of bias is assessed for each study question using a rating system with four categories: low risk of bias; probably low risk of bias; probably high risk of bias; and high risk of bias. Although general criteria are provided for each category, each reviewer is expected to provide supporting rationale for each rating given to each study.

Three reviewers independently rate each study for each of the domains, and the reviewers discuss and resolve any discrepancies.

Templates for Risk of Bias profiles are presented separately for human (Table S2) and animal/mechanistic studies (Table S3) in Appendix S.

Risk of Bias Profile and Judgements for each Study

Not all risk of bias domains should be considered of equal importance in the overall evaluation of study quality. Studies can be placed into tiers that reflect the two most important domains: quality of the exposure assessment, and consideration of confounding and effect modification.

For example, to be considered a higher quality study (Tier 1), a human study must be rated as "definitely low" or "probably low" for both these risk of bias domains:

- Exposure assessment (can we be confident in the exposure characterization?)
- Confounding (does the study design or analysis account for important confounding and modifying variables?)

Similarly, an animal study is considered a higher quality study (Tier 1) when these risk of bias domains are judged "definitely low" or "probably low":

- Exposure relevance to humans (are exposures relevant to exposures humans would experience?)
- Controls (are animal controls adequate?)

In contrast, a human study is considered a lower quality study (Tier 3) when the risk of bias domains are rated as "definitely high" or "probably high" for:

- Exposure assessment
- Confounding

Similarly, an animal study is considered a lower quality study (Tier 3) when the risk of bias domains are judged "definitely high" or "probably high" because:

- Controls are inadequate

Other studies are considered Tier 2 studies when either one of the above critical domains is "probably high" or "definitely high" or other domain considerations help inform the reviewer that additional sources of bias exist in a study.

With regards to exposure characterization, health impacts must be presented in relation to quantitative metrics of PM2.5; higher quality studies rely upon better measures of exposure. For example, personal exposure monitoring is preferred; exposure estimated using personal activity records (for example) to modify measurements from centrally located monitors or modeling are also acceptable. Studies in which exposure is measured ecologically (i.e., estimated using measurements from one or more central monitors) are considered to be of the lowest tier (Tier 3).

With regards to confounding/effect modification, Tier 1 studies include studies where confounding was considered and effect modification was evaluated and effectively controlled or adjusted for, such as those including age, gender, socioeconomic status, meteorological parameters, respiratory or influenza patterns, and co-pollutant exposures. Reviewers also evaluate and note evidence of unmeasured confounding or uncertainty relevant to a particular study from the broader literature search in order to bring this finding into the evidence integration process.

2.5 Structure for Body of Evidence

The structure of an evaluation of overall body of evidence includes a heat map based on the risk of bias profiles for each study (see appendices). In particular, risk of bias considerations determine which studies fall into each study quality Tier. Only studies that fall into Tiers 1 and 2 are

considered for evaluation in the evidence integration (Section 2.6) process to evaluate the relationship between policy-relevant exposures to PM2.5 and target health endpoints.

Table 3 illustrates the organization of evidence obtained from literature selection and risk of bias analysis. This organization depicts relevant domains of study quality and relevance in order to draw conclusions regarding the confidence of available evidence for addressing the proposed research questions. A brief rationale is provided in the table, with more detailed discussion in the text. A similar table would be developed for IHD and include evidence from epidemiological, animal, and mechanistic studies.

Table 3. Synthesis of Evidence for Epidemiological Studies

Outcome	Cardiovascular mortality from long-term exposure
Number of epidemiology studies	X studies List type of study design
Overall risk of bias	Severe OR not severe Include rationale based on heat map, other external evidence of confounding, or other bias (discordant findings)
Inconsistency of evidence	Severe OR not severe Include heterogeneity of results
Indirectness of evidence	Severe OR not severe Include rationale for rating; e.g., outcome data not assessed individually, and/or exposure not assessed
Publication bias	Detected OR not detected OR not assessed Include rationale; e.g., meta-analyses that report evidence of bias OR no evaluations conducted
Overall conclusion	Increases in cardiovascular mortality below 12 μg/m^3 … Include strength of effect?
Confidence in conclusion	Slight/Moderate/Robust/Indeterminate/Compelling evidence of no effect Overall confidence in conclusion based on risk of bias, consistency, strength, publication bias

2.6 Evidence Integration and Conclusions

When results of individual studies have been assessed in the context of methodological strengths and limitations, they should be integrated both within and across evidence realms. For mortality outcomes, only epidemiological studies are assessed; thus, evidence integration across realms will not be conducted. Table 4 shows how these latter studies could be evaluated to determine whether they support causation and human relevance for cardiovascular-related mortality. Table 5 shows how this should be applied to determine biological plausibility in humans.

Table 4 is based on both the International Programme on Chemical Safety (IPCS) Mode-of-Action/Human Relevance framework (Boobis et al. 2008; Meek et al. 2014) and the OHAT framework for evaluating the confidence in the toxicology body of literature (NTP 2019). The IPCS framework is intended to evaluate key events from mechanistic and MoA studies; however, NCASI modified the framework so that it can also be used to evaluate toxicity studies that assess apical effects and whether reported effects in these studies (or lack thereof) are consistent and coherent with key events identified in mechanistic and MoA studies. This framework is quite similar to the criteria OHAT uses to evaluate confidence in the evidence (i.e., risk of bias, unexplained

inconsistency, indirectness, imprecision, magnitude, dose-response, consistency across models/species/related outcomes). All these criteria are incorporated into Table 4 except risk of bias, because risk of bias should be determined at the individual study level and should be incorporated in the initial evaluation of study results (i.e., results of studies with moderate or high risk of bias should not be considered reliable and, as such, should not be considered when evidence is being integrated).

Table 4. Criteria for Evaluating Experimental Studies

Criterion	Considerations
Causation	Consistency: repeatability of key events and effects across species/study designs Magnitude: large, considering type of effect, background prevalence, species and dose range, exposure pattern Essentiality: reversibility of effects if exposure is stopped or a key event prevented Specificity: apical effect likely to occur following a key event Temporality: observation of key events in a hypothesized order, before toxicity is apparent Exposure-Response: key events observed at exposures below or similar to those associated with adverse effect Biological Concordance: proposed mode of action is consistent with current biological knowledge of toxicological outcome Analogy: proposed mode of action is consistent with what is known for other related chemicals with a well-defined mode of action
Human Relevance	Relevant groups and life stages Comparative developmental processes and their relative timing Differences in ontogeny that affect dose metrics (e.g., placental or lactational transfer, key metabolic enzymes) Consequences of interaction of chemical with cells, tissues, and organs Magnitude of exposure differences for observation of key events or apical outcome

[Sources: adapted from Boobis et al. 2008; Meek et al. 2014; NTP 2019]

Table 5. Confidence in Biological Plausibility

	Human Relevance	Inadequate Evidence for Human Relevance	No Human Relevance
Supports effects at low levels	High	Moderate	Inadequate
Inadequate evidence for effects at low levels	Inadequate	Inadequate	Inadequate
Supports lack of effects at low levels	High (not plausible)	Moderate (not plausible)	Inadequate

If all aspects of causation in Table 4 are met, the evidence supports causation in these experimental systems. If not all are met, it must be determined whether it is more likely that the evidence as a whole supports causation (i.e., likely explanations for any aspect not being met must be provided), that the evidence supports no causation, or that the evidence is inadequate to determine causation. Human relevance is evaluated in a similar manner. Once this is complete, high, moderate, or

inadequate confidence in biological plausibility is determined, or whether the evidence indicates high or moderate confidence for a lack of biological plausibility, as shown in Table 5.

This biological plausibility assessment is then incorporated into the evaluation of epidemiology evidence using modified Bradford Hill aspects. The ISA Preamble notes that the Bradford Hill aspects provide a framework for assessing evidence but should not be considered as fixed rules of evidence (i.e., a checklist) for developing causal conclusions (USEPA 2015). Rather, they provide a framework for systematic evaluation of the weight of evidence for inferring causality. As such, not meeting one or more of the aspects does not necessarily preclude a judgment of causality.

The aspects as listed in the Preamble are shown in Table 6. NCASI modified the explanations of each aspect to be more succinct. The experimental evidence provides information on the likelihood of biological plausibility, as demonstrated in the table, and is informative regarding coherence (i.e., whether all evidence fits together). Similar to the framework for experimental studies in Table 4, if all Bradford Hill aspects are met the evidence as a whole supports causation, or in the case of the protocol described herein, supports a conclusion of effect induction at target exposure concentrations. If not all are met, an explanation of whether it is most likely that the evidence as a whole supports causation (i.e., likely explanations for any aspect that is not met must be provided), is suggestive of causation, supports no causation, or is inadequate to determine causation is required.

Table 6. Criteria for Evidence Integration

Aspect	Explanation
Strength of Association	Large and precise risk estimates are less likely to be due to chance, bias, or other factors
Consistency	Evidence is stronger if consistent effects are observed among studies of different designs, people, places, circumstances, and times
Specificity	Evidence is stronger when disease is specific to exposure or exposure is specific to disease
Temporality	Exposure must precede occurrence of disease
Exposure-response	Evidence is stronger when a well-characterized exposure-response relationship exists (e.g., disease risk increases with greater exposure intensity and/or duration) at the relevant exposure levels (in this case below the current NAAQS)
Biological Plausibility	Evidence on the biological mechanism of an effect allows scientifically defensible determination for causation at relevant exposure levels
Coherence	All known facts related to observed association from various evidence streams fit together in a coherent manner
Experiment	"Natural experiments" can provide strong evidence when intervention or cessation of exposure results in a change in disease risks
Analogy	Evidence is stronger when a similar substance is an established causal factor for a similar effect

Figure 1 shows how evidence across realms should be integrated to evaluate causation. The current US NAAQS causal framework has five categories for causation (causal, likely causal, suggestive, inadequate, not likely causal). Figure 1 shows four categories (it does not include likely causal). The NAAQS causal framework requires only one high-quality study for evidence of a causal relationship to be deemed suggestive. Under this definition, high-quality studies that are inconsistent with evidence of an association may exist, but as long as one high-quality study demonstrates an effect

there would still be enough evidence to constitute a suggestive relationship. However, because all studies should be reviewed using the same criteria, it is more appropriate to conclude a suggestive causal association only if the weight of evidence indicates that a causal association is more likely than not based on all the evidence combined. In situations where there are multiple but inconsistent high-quality studies, the appropriate conclusion is that the evidence is inadequate (IOM 2008). With this definition of suggestive, the likely causal category is not necessary. This four-tiered framework is consistent with other causal frameworks, such as that defined in the Institute of Medicine report titled *Improving the Presumptive Disability Decision-Making Process for Veterans* (IOM 2008) and, notably, the framework in the ISA Preamble for potential at-risk factors.

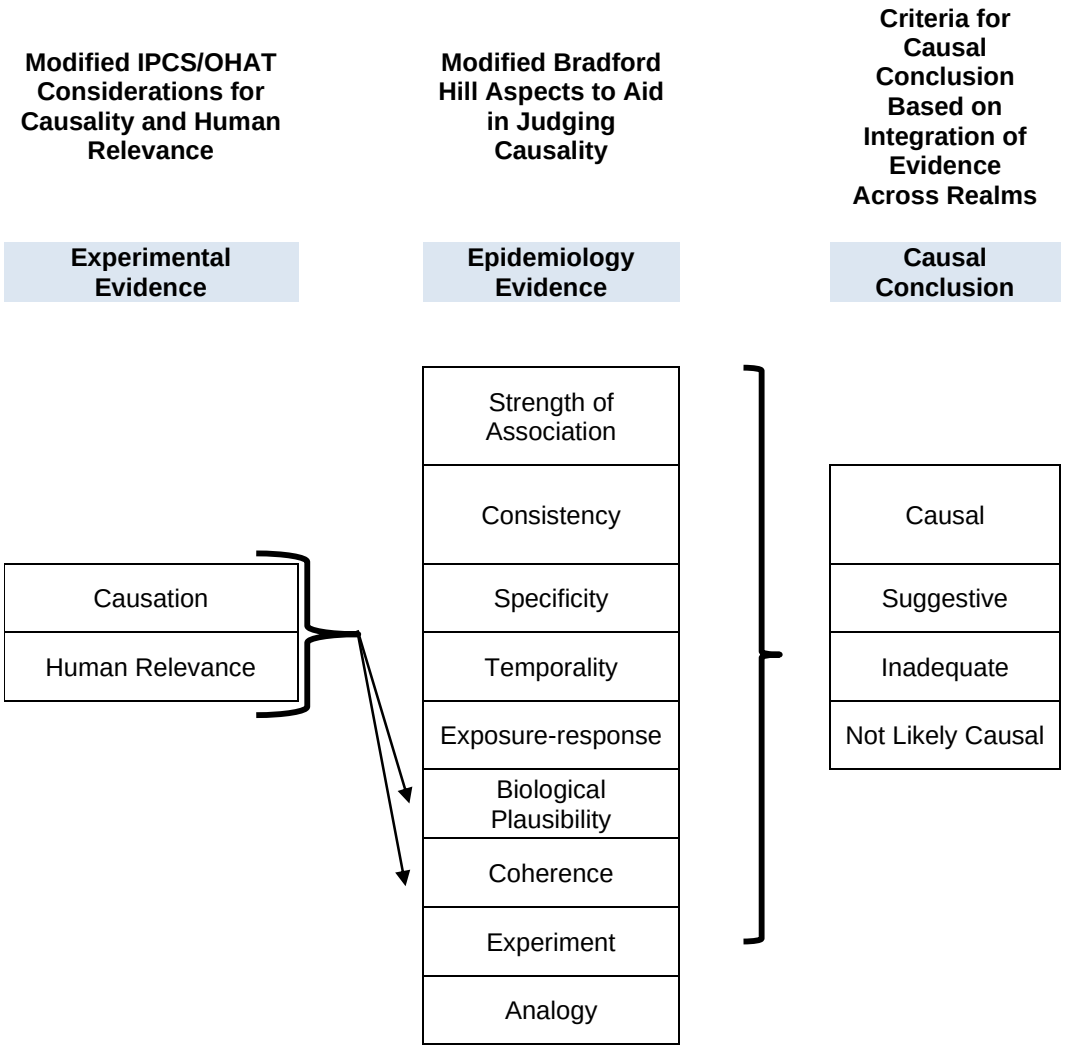


Figure 1. Causal Conclusion for an Effect Based on Experimental and Epidemiology Evidence
[IPCS = International Programme on Chemical Safety;
OHAT = Office of Health Assessment and Translation]

A causal relationship should be concluded when all modified Bradford Hill aspects are met, or when most are met and there is a likely explanation for each that is not met. A suggestive relationship should be concluded when an assessment of the evidence indicates that a causal relationship is more likely than not but some of the modified Bradford Hill aspects have inadequate information,

and all other aspects are met or there is a likely explanation for each that is not met. The inadequate category should be concluded when most or all modified Bradford Hill aspects have inadequate information or are not met and there is no likely explanation for each that is not met. The not likely causal category should be concluded when evidence indicates there is no causal relationship based on the modified Bradford Hill aspects (e.g., there is a consistent lack of association in robust epidemiology studies) and the experimental evidence indicates a lack of biological plausibility.

For the purpose of this protocol, which evaluates research questions related to the potential of effects occurring within specific exposure concentrations rather than questions of general causation, the question of causation becomes significantly more narrow in scope and requires refining the application of evidence integration in order to be effectively 'fit for purpose.' In a review that seeks to explore questions of general causation between an exposure and health effects, it is expected that a broad range of study types, and the use of a variety of methodologies, might be available to inform the conclusion of the review. However, a protocol designed to address a specific aspect of the relationship between exposure and a health outcome within the larger issue of general causation will have more specific requirements in terms of study features and methodological choices in order to reach conclusions with high confidence.

For this protocol, evidence integration domains that provide evidence for magnitude of effect and exposure-response are essential in order to address the stated research questions. As such, methodological choices that impact interpretation of these domains should be documented in careful detail by reviewers to ensure that an appropriate conclusion is drawn for each research question, with an accurate characterization of the confidence associated with those conclusions. Conclusions with high confidence are valuable for informing risk assessment and risk management practices. Conclusions with poor confidence are valuable for informing the research community as to data gaps that need to be addressed in order to adequately answer the questions being asked.

3.0 Summary of Findings

In Appendices A-C, 3 studies indicated by EPA in the ISA and PA as being sources of important evidence regarding the potential causal association between low concentrations of PM2.5 and cardiovascular-related mortality are evaluated under the specifications of this proposed protocol. The results indicate that the strongest evidence base available consists of Tier 2 and Tier 3 studies. Tier 1 studies, which are high quality studies that contribute to the evidence base, were not identified in these 6 studies. Tier 2 studies have potential study quality issues but are considered to be useful as support of Tier 1 studies when similar conclusions are reached; Tier 3 studies have study quality issues to the extent they are not considered to contribute to the evidence base. When the strongest evidence available consists of Tier 2 and Tier 3 studies, a there is not an adequate evidence base to conclude causation. The evaluation of these studies under a narrow in scope, policy relevant research objective combined with modern, prescribed systematic review criteria lead to a more detailed characterization of their utility as an evidence base for policy decision making.

REFERENCES

Boobis, A.R., Doe, J.E., Heinrich-Hirsch, B., Meek, M.E., Munn, S., Ruchirawat, M., Schlatter, J., Seed, J., and Vickers, C. 2008. IPCS framework for analyzing the relevance of a noncancer mode of action for humans. *Critical Reviews in Toxicology* 38(2):87-96. https://doi.org/10.1080/10408440701749421.

IOM (Institute of Medicine) Committee on Evaluation of the Presumptive Disability Decision-Making Process for Veterans, Board on Military and Veterans Health. 2008. *Improving the Presumptive Disability Decision-Making Process for Veterans*. Samet, J.M., and Bodurow, C.C. (eds.). Washington DC: National Academies Press. 781 p. Available at http://books.nap.edu/openbook.php?record_id=11908&.

Meek, M.E., Palermo, C.M., Bachman, A.N., North, C.M., and Lewis, R.J. 2014. Mode of action human relevance (species concordance) framework: evolution of the Bradford Hill considerations and comparative analysis of weight of evidence. *Journal of Applied Toxicology* 34(6):595-606. https://doi.org/10.1002/jat.2984.

NTP. 2019. *Handbook for Conducting a Literature-Based Health Assessment Using OHAT Approach for Systematic Review and Evidence Integration*. National Toxicology Program Office of Health Assessment and Translation. 101 p. https://ntp.niehs.nih.gov/ntp/ohat/pubs/handbookmarch2019_508.pdf.

USEPA. 2009. *Integrated Science Assessment for Particulate Matter* (Final Report). EPA-600/R-08-139F. Research Triangle Park NC: United States Environmental Protection Agency, Office of Research and Development, National Center for Environmental Assessment. Available at: https://cfpub.epa.gov/ncea/risk/recordisplay.cfm?deid=216546.

USEPA. 2015. *Preamble to the Integrated Science Assessments*. EPA/600/R-15/067. Washington DC: United States Environmental Protection Agency. http://www.epa.gov/isa [Accessed December 1, 2015].

USEPA. 2016. In*tegrated Review Plan for the National Ambient Air Quality Standards for Particulate Matter*. EPA-452/R-16-005. Washington DC: United States Environmental Protection Agency. Available at https://www3.epa.gov/ttn/naaqs/standards/pm/data/201612-final-integrated-review-plan.pdf.

USEPA. 2019. *Integrated Science Assessment for Particulate Matter*. EPA/600/R-19/188. Research Triangle Park NC: United States Environmental Protection Agency, Office of Research and Development, National Center for Environmental Assessment – RTP Division. www.epa.gov/isa.

USEPA. 2021 Supplement to the 2019 Integrated Science Assessment for Particulate Matter (External Review Draft). EPA/600/R-21/198 Center for Public Health and Environmental Assessment Office of Research and Development U.S. Environmental Protection Agency Research Triangle Park, NCwww.epa.gov/isa

APPENDIX S
SUPPLEMENTAL
PROTOCOL TABLES AND FIGURES

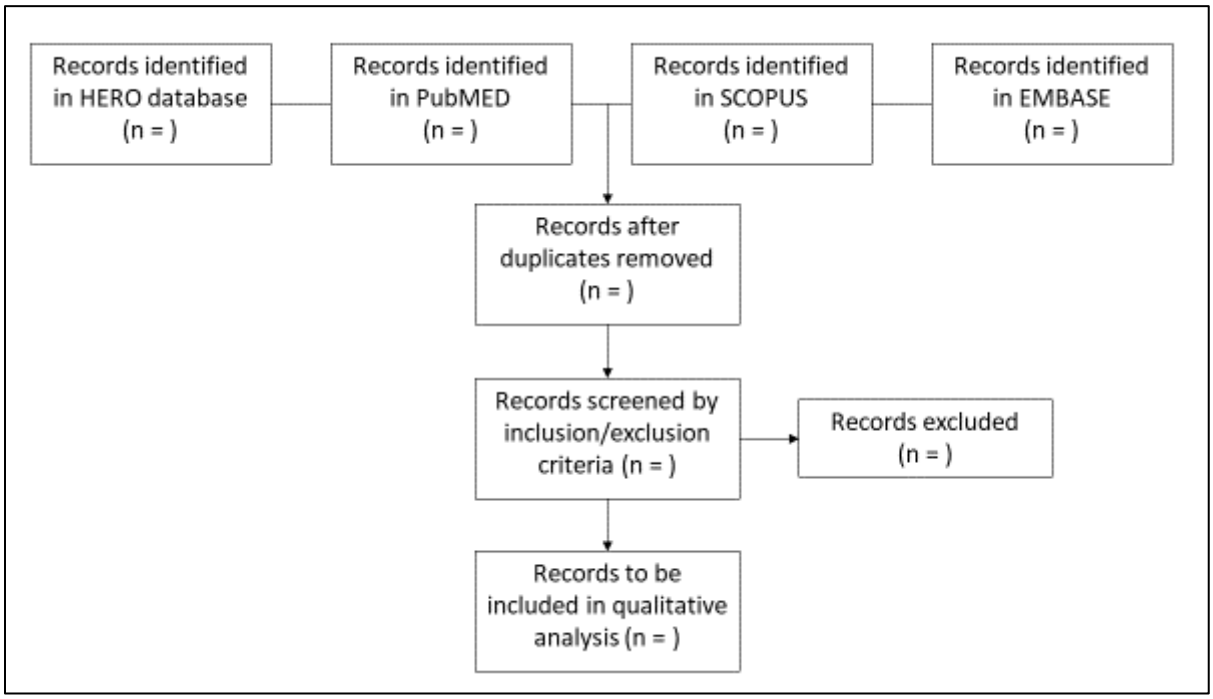


Figure S1. Study Selection Process

Table S1. Risk of Bias Guidelines

Bias Domain Question[a]	Judgment Guidelines[a,b]
Selection Bias	
1. Were selected study participants in appropriate comparison groups? Applies to Co, CaCo, CrSe, Eco	Low/prob low – **Co, CrSe**: There is direct/indirect evidence that subjects (both exposed and nonexposed) were similar (e.g., recruited from same eligible population, recruited with same method of ascertainment using same inclusion and exclusion criteria, and were of similar age and health status), recruited within same time frame, and had similar participation/response rates. **CaCo**: There is direct/indirect evidence that cases and controls were similar (e.g., recruited from same eligible population including being of similar age, gender, ethnicity, and eligibility criteria other than outcome of interest as appropriate), recruited within same time frame, and controls are described as having no history of the outcome. Note: A study is considered low risk of bias if baseline characteristics of groups differed but differences were considered as potential confounding or stratification variables (see Question 2). **Time-series**: For ecological studies, a table of information or text on potential differences in characteristics that could bias results

Bias Domain Question[a]	Judgment Guidelines[a,b]
	is provided, and these characteristics are adjusted for as potential confounders. There is direct evidence that subjects (both exposure groups and referent groups) were similar (e.g., of similar geographic region, ethnicity, socioeconomic status); OR baseline characteristics of groups differed but these differences were considered as potential confounding or stratification variables in analyses (see Question 2). **Additional Guidance**: Comparison groups selected adequately. Study provides table of subject characteristics by exposure levels and/or case status. CrSe studies can be considered low risk of bias if a general table of subject characteristics is provided and analyses are adjusted for confounders. Prob high/high – **Co, CrSe**: There is indirect/direct evidence that subjects (both exposed and nonexposed) were not similar, recruited within very different time frames, or had very different participation/response rates; OR there is insufficient information provided about comparison group, including a different rate of nonresponse without an explanation. **CaCo**: There is indirect/direct evidence that controls were drawn from a very dissimilar population than cases or recruited within very different time frames; OR there is insufficient information provided about appropriateness of controls, including rate of response reported for cases only. **Time-series**: There is indirect/direct evidence that subjects (exposure groups and referent groups) were not similar (e.g., not of similar geographic region, ethnicity, socioeconomic status); OR there is insufficient information provided about appropriateness of comparison groups. At least one known difference between groups was not accounted for (e.g., study authors acknowledged that groups were different with respect to a variable that is a potential confounder not considered in analysis); OR recruitment methods were very different (e.g., completed during different time frames, different criteria were used for recruitment).
Confounding	
2. Did the study design or analysis account for important confounding and modifying variables? Applies to Co, CaCo, CrSe, CaS	Low/prob low – **Co, CrSe, CaS**: There is direct/indirect evidence that appropriate adjustments or explicit considerations were made for primary covariates and confounders in final analyses through statistical models to reduce research-specific bias including standardization, case matching, adjustment in multivariate model, stratification, propensity scoring, or other methods were appropriately justified. Acceptable consideration of appropriate adjustment factors includes cases when factor is not included in final adjustment model because author

Bias Domain Question[a]	Judgment Guidelines[a,b]
	conducted analyses that indicated it did not need to be included. **CaCo**: There is direct/indirect evidence that appropriate adjustments were made for primary covariates and confounders in final analyses through statistical models to reduce research specific bias including standardization, matching of cases and controls, adjustment in multivariate model, stratification, propensity scoring, or other methods were appropriately justified. **Time-series**: There is direct/indirect evidence that appropriate adjustments or explicit considerations were made for covariates and confounders in final analyses through statistical models (e.g., standardization, multivariate adjustment). Acceptable consideration of appropriate adjustment factors includes cases when factor is not included in final adjustment model because author conducted analyses that indicated it did not need to be included. **Additional Guidance**: Study adjusted for or addressed important potential confounders. Age, gender, education, and socioeconomic status are potential confounders that need to be addressed and considered in study design or analyses. In addition, specific important confounders for this assessment depend on health outcome (e.g., smoking for lung cancer). Other confounders might also be judged important for certain health outcomes. A low risk of bias rating is assigned for this question if potential confounders deemed important were adequately addressed (e.g., distribution of variables was compared between groups and there was no statistically significant difference). Prob high/high – **Co, CrSe, CaS**: There is indirect/direct evidence that distribution of primary covariates and known confounders differed between groups and was not appropriately adjusted for in final analyses; OR there is insufficient information provided about distribution of known confounders. **CaCo**: There is indirect/direct evidence that distribution of primary covariates and known confounders differed between cases and controls and was not investigated further; OR there is insufficient information provided about distribution of known confounders in cases and controls. **Time-series**: There is indirect/direct evidence that distribution of covariates and known confounders differed between groups and was not appropriately adjusted for in final analyses; OR there is insufficient information provided about distribution of known confounders. **Additional Guidance**:

Bias Domain Question[a]	Judgment Guidelines[a,b]
	Design or analysis did not adjust for important potential confounders. Adjustments were made for some potential confounders, but at least one major confounder was not addressed for a particular health outcome (e.g., no adjustment for smoking when evaluating lung cancer).
3. Did researchers adjust or control for other exposures that are anticipated to bias results? Applies to Co, CaCo, CrSe, CaS, Eco	Low/prob low – **Co, CaCo, CrSe, CaS, Eco**: There is direct/indirect evidence that other exposures anticipated to bias results were not present or were appropriately adjusted for. For occupational studies or studies of contaminated sites, other chemical exposures known to be associated with those settings were appropriately considered. **Additional Guidance**: Researchers adjusted for other chemicals or accounted for occupational exposures likely to be associated with outcome (low); OR it is deemed that coexposures present would not appreciably bias results (prob low). This includes insufficient information provided on coexposures in general population studies. Prob high/high - **Co, CrSe, CaS**: There is indirect/direct evidence that there was an unbalanced provision of additional coexposures across primary study groups that were not appropriately adjusted for; OR there is insufficient information provided about coexposures in occupational studies or other studies where exposures to other air pollutants would have been reasonably anticipated. **CaCo**: There is indirect/direct evidence that there was an unbalanced provision of additional coexposures across cases and controls that were not appropriately adjusted for; OR there is insufficient information provided about coexposures in occupational studies or other studies where exposures to other air pollutants would have been reasonably anticipated. **Eco and Semi-individual**: There is indirect/direct evidence that there was an unbalanced provision of additional coexposures that were not appropriately adjusted for; OR there is insufficient information provided about coexposures in studies where exposures to other air pollutants would have been reasonably anticipated. **Additional Guidance**: There is evidence that coexposures might not have been addressed. Examples include any study of populations that may be exposed to numerous ambient air pollutants including gases (ozone, nitrogen oxides, sulfur oxides) but these coexposures are not addressed; OR study with known coexposures, but relevance of coexposure to PM effects is unknown; OR it is not clear if other compounds were adjusted for in analyses. Known differential exposure to other air pollutants also associated

Bias Domain Question[a]	Judgment Guidelines[a,b]
	with health outcome of interest occurred with PM, and exposure was not addressed by study authors.
Attrition/Exclusion Bias	
4. Were outcome data complete without attrition or exclusion from analysis? Applies to Co, CaCo, CrSe, Eco	Low/prob low – **Co**: There is direct/indirect evidence that loss of subjects (i.e., incomplete outcome data) was adequately addressed and reasons were documented when human subjects were removed from study. Acceptable handling of subject attrition includes: very little missing outcome data; reasons for missing subjects unlikely to be related to outcome (for survival data, censoring unlikely to be introducing bias); missing outcome data balanced in numbers across study groups, with similar reasons for missing data across groups; OR missing data have been imputed using appropriate methods, AND characteristics of subjects lost to follow-up or with unavailable records are described in identical way and are not significantly different from those of study participants. **CaCo, CrSe**: There is direct/indirect evidence that exclusion of subjects from analyses was adequately addressed, and reasons were documented when subjects were removed from study or excluded from analyses (low); OR were deemed not to bias results (prob low). **Time-series**: There is direct/indirect evidence that there was no loss of subjects (e.g., due to moving or migration) or data during study and outcome data were complete; OR incomplete outcome data were adequately addressed, AND characteristics of subjects lost to follow-up or with unavailable records are described in identical way and are not significantly different from those of study participants. **Additional Guidance**: There are no reported data lost to attrition, and numbers in results tables sum to total number of subjects; OR less than 10% of data are missing; OR there are some missing outcome data but study report clearly identifies missing data and how it was handled (e.g., loss to follow-up for a cohort study is determined to be minimal if there are some missing data for either exposure or outcome for certain subjects at specific time measured and authors clearly explain what happened to everyone and which results were used in analyses). For ecological studies specifically, there are no reported data lost to attrition; OR there are some missing data but study report clearly identifies missing data and how they were handled (e.g., migration in and out of study area and residence location within study area were tracked and accounted for or references provided to verify that population migration within or in/out of study area is not a concern for this population), and characteristics of subjects lost to attrition do not differ significantly from those included in study.

Bias Domain Question[a]	Judgment Guidelines[a,b]
	Prob high – **Co**: There is direct evidence that loss of subjects (i.e., incomplete outcome data) was adequately addressed and reasons were documented when human subjects were removed from study; OR it is deemed that proportion lost to follow-up would not appreciably bias results. This would include reports of no statistical differences in characteristics of subjects lost to follow-up or with unavailable records from those of study participants. Generally, the higher the ratio of participants with missing data to participants with events, the greater potential there is for bias. For studies with a long duration of follow-up, some withdrawals for such reasons are inevitable. **CaCo, CrSe**: There is indirect evidence that exclusion of subjects from analyses was adequately addressed, and reasons were documented when subjects were removed from study or excluded from analyses. **Eco and Semi-individual**: There is indirect evidence that there was no loss of subjects (e.g., due to migration during study) and outcome data were complete; OR it is deemed that proportion of subjects lost to follow-up would not appreciably bias results. This would include reports of no statistical differences in characteristics of subjects lost to follow-up or with unavailable records of outcomes. For studies with long duration of follow-up, some withdrawals for such reasons are inevitable. **Additional Guidance**: No direct evidence of loss to follow-up, attrition, or loss of subjects due to migration/moving provided. Tables of results do not include number of subjects and it is not stated that there was any data missing; OR there appear to be no or very few missing data; OR in a cohort study, there is no mention of loss to follow-up. High – **Co**: There is indirect evidence that loss of subjects (i.e., incomplete outcome data) was unacceptably large and not adequately addressed; OR there is insufficient information provided about numbers of subjects lost to follow-up. **CaCo, CrSe**: There is indirect evidence that exclusion of subjects from analyses was not adequately addressed; OR there is insufficient information provided about why subjects were removed from study or excluded from analyses. **Eco and Semi-individual**: There is direct/indirect evidence that incomplete outcome data (e.g., due to subject migration or moving) were unacceptably large (greater than 20% in each group) and not adequately addressed; OR there is insufficient information provided about missing outcome data. **Additional Guidance**:

Bias Domain Question[a]	Judgment Guidelines[a,b]
	Missing outcome data with no explanation of why data were missing, and it is unclear from characteristics table or other information provided in report why data might be missing.
5. Can we be confident in the exposure characterization? Applies to EA, Co, CaCo, CrSe, CaS, Eco	Low/prob low – **Co, CaCo, CrSe, CaS**: There is direct evidence that appropriate measurements were taken and that the most reliable methods for sampling were conducted. Best measurements would include personal measurement of PM concentrations. If fixed-site monitors were used, modeling of PM concentrations to estimate personal exposures are preferred. The least preferred exposure measurement is single or multiple fixed-site monitors. If PM2.5 data are not available and must be estimated, this would add to uncertainty in exposure measurements. Modeled estimates should include validation of estimates against measured concentrations. **Eco and Semi-individual**: This rating is not applicable. Only studies with individual-level exposure characterization can earn this rating. If individual-level exposure data are provided, it is not an ecological study and should be reclassified and rated according to other study type criteria. Prob high/high – **Co, CaCo, CrSe, CaS**: There is direct or indirect evidence that data are based on single or a few fixed-site monitor locations that would not adequately describe personal exposures. A surrogate for PM2.5 was used to estimate concentrations (e.g., PM10). **Eco and Semi-individual**: There is indirect/direct evidence that chemical in question was not adequately characterized by appropriate measures and methods (e.g., no historical monitoring, isolated or remote-time samples taken to be representative of large areas).
6. Can we be confident in the outcome assessment? Applies to Co, CaCo, CrSe, CaS, Eco	Low/prob low – Case Control: There is direct/indirect evidence that the outcome was assessed in cases using well-established methods (the gold standard) and subjects had been followed for the same length of time in all study groups. Cross-Sectional, Case Series/Report: There is direct/indirect evidence that the outcome was assessed using well-established methods (the gold standard). Ecological and Semi-individual: There is direct/indirect evidence that the outcome was assessed using well-established methods, the gold standard (e.g., individual-level outcome data were assessed, as in the case of semi-individual ecological studies), and subjects have been followed for the same length of time in all study groups. Acceptable assessment methods will depend on the outcome, but examples of such methods may include: objectively measured with diagnostic methods, measured by

Bias Domain Question[a]	Judgment Guidelines[a,b]
	trained interviewers, obtained from reliable registries or records. Additional Guidance: Cancer cases are histologically confirmed, OR data obtained from nationwide registry are accepted as valid and complete, OR outcome diagnosed by physician, OR outcome obtained from medical record data or validated with such data (if self-reported). Prob high/high – **CaCo**: There is indirect/direct evidence that outcome was assessed in cases using an insensitive instrument or was not adequately validated; OR there is insufficient information provided about how cases were identified. **CrSe, CaS**: There is indirect/direct evidence that outcome assessment method is an insensitive instrument or was not adequately validated; OR there is insufficient information provided about validation of outcome assessment method. **Eco and Semi-individual**: There is indirect/direct evidence that authors did not validate methods used, or length of follow-up differed by study group; OR there is insufficient information provided about validation of outcome assessment method. **Additional Guidance**: Outcome is self-reported (e.g., "ever been diagnosed by a physician") and not verified by medical records or other means. There is insufficient information on quality of self-report or validation of answers. Outcome is assessed by nurses and there is no information on assessor agreement.
Selective Reporting Bias	
7. Were all measured outcomes reported? Applies to Co, CaCo, CrSe, CaS	Low/prob low – **Co, CaCo, CrSe, CaS, Eco**: There is direct/indirect evidence that all the study's measured outcomes (primary and secondary) outlined in protocol, methods, abstract, and/or introduction (that are relevant for evaluation) have been reported. This includes outcomes reported with sufficient detail to be included in meta-analysis or fully tabulated during data extraction; OR analyses that had not been planned at outset of study (i.e., retrospective unplanned subgroup analyses) are clearly indicated as such, and it is deemed that omitted analyses were not appropriate and selective reporting would not appreciably bias results. This includes outcomes reported with insufficient detail such as only reporting that results were statistically significant (or not). **Co, CaCo, CrSe, CaS, Eco**: There is indirect evidence that all the study's measured outcomes (primary and secondary) outlined in protocol, methods, abstract, and/or introduction (that are relevant for evaluation) have been reported; OR there is

Bias Domain Question[a]	Judgment Guidelines[a,b]
	insufficient information provided about selective outcome reporting. **Additional Guidance**: An outcome mentioned in a part of the study report is obviously missing from the results.
Other Bias	
8. Were there no other potential threats to internal validity (e.g., statistical methods were appropriate, and researchers adhered to study protocol)? Applies to Co, CaCo, CrSe, CaS, Eco	Low/prob low – **Co, CaCo, CrSe, CaS, Eco**: There is direct/indirect evidence that there was no impact of model selection on depicting linearity or otherwise shape of concentration response curve. **Additional Guidance**: Taking into consideration that linear models (e.g., Cox Proportional Hazards) rely on a model assumption of linearity and evaluate model validity over a large spectrum of exposure data that may not be relevant to specific research questions of this protocol, studies that employ linear models at exposure ranges outside range of interest may mischaracterize concentration/response function at specific ranges of exposure and introduce bias. Prob high/high – **Co, CaCo, CrSe, CaS, Eco**: There is direct/indirect evidence that model validity was assessed over exposure ranges outside range of interest to research question or model validity was not assessed at all. Low/prob low – **Assessment-Specific Clarification**: Statistical analyses were appropriate and no other threats to internal validity were identified. Study authors might acknowledge limitations, but these are not expected to affect study's internal validity. There are study limitations likely to bias results toward or away from null, but adequate sample size was available in each cell ($n \geq 5$); OR sample size is small and acknowledged as a potential limitation by study authors, but significant results were still observed. Prob high/high – There are study limitations likely to bias results towards or away from null; OR analyses were conducted on a small number of subjects ($n < 5$ in any given cell) and no statistically significant results were observed.
9. Did researchers adhere to study protocol? Applies to Co, CaCo, CrSe, CaS, Eco	Low/prob low – **Co, CaCo, CrSe, CaS, Eco**: There is direct/indirect evidence that there were no deviations from protocol (i.e., authors reported no deviations/did not report any deviations); OR deviations

Bias Domain Question[a]	Judgment Guidelines[a,b]
	from protocol are described and it is deemed that they would not appreciably bias results. **Additional Guidance**: Taking into consideration typical reporting practices, it seems unlikely that deviations from protocol will be explicitly reported in most studies. Thus, unless stated otherwise by authors (i.e., evidence of deviation is reported) or it is clear from study report that deviations from planned approach occurred, assume that no deviations occurred. It is anticipated that this approach will result in a rating of "probably low risk of bias" for most studies. If there are deviations, rating reflects how deviations changed direction, magnitude, and/or significance of results. Prob high/high – **Co, CaCo, CrSe, CaS, Eco**: There is direct/indirect evidence that there were large deviations from protocol as outlined in methods or study report. In addition to not reporting outcomes, this includes reporting outcomes based on composite score without individual outcome components or outcomes reported using measurements, analysis methods or subsets of data (e.g., subscales) that were not prespecified, or reporting outcomes not prespecified (unless clear justification for their reporting is provided, such as an unexpected effect).

[Note: For more details refer to OHAT Risk of Bias Tool (NTP 2013)]

[a] Study type/design: Co (cohort), CaCo (case-control), CrSe (cross-sectional), CaS (case series/case report), Eco (ecological), EA (experimental animal)

[b] Rating (low, prob low, prob high, high risk, unclear, N/A)

RISK OF BIAS PROFILE (HUMAN)

Table S2. Study Reference Template for Human Risk of Bias Profile

Study Element	Description
Participants	
Exposure	
Comparator	
Outcome	
Study Design	

Bias Domain	Reviewer's Judgment	Support for Judgment
Source population representation (Did selection of study participants result in appropriate comparison groups?)		
Confounding (Did study design or analysis account for important confounding and modifying variables?)		
Incomplete outcome data (Were outcome data complete without attrition or exclusion from analysis?)		
Exposure assessment (Can we be confident in exposure characterization?)		
Outcome assessment (Can we be confident in outcome assessment?)		
Selective outcome reporting (Were all measured outcomes reported?)		
Other potential threats to internal validity (Were statistical methods appropriate? Did researchers adhere to study protocol?)		
Model Specification (Were statistical models evaluated for validity of underlying assumptions or was model validity assessed with exposure ranges outside the range of interest for research question?)		
External evidence of bias (Does external research indicate there may be unmeasured uncertainty, bias, or confounding present in study design or dataset?)		

RISK OF BIAS PROFILE (EXPERIMENTAL ANIMAL)

Table S3. Study Reference Template for Experimental Animal Risk of Bias Profile

Study Element	Description
Design	
Participants	
Exposure	
Comparator	
Outcome	

Bias Domain	Reviewer's Judgment	Support for Judgment
Randomization of dose/exposure level (Was administered dose or exposure level adequately randomized?)		
Inadequate concealment of allocation (Was allocation to study groups adequately concealed?)		
Experimental conditions (Were experimental conditions identical across study groups?)		
Blinding (Were research personnel and human subjects blinded to study group during study?)		
Incomplete outcome data (Were outcome data complete without attrition or exclusion from analysis?)		
Exposure assessment (Can we be confident in exposure characterization?)		
Outcome assessment (Can we be confident in outcome assessment?)		
Selective reporting (Were all measured outcomes reported?)		
Other potential threats to internal validity (Were statistical methods appropriate? Did researchers adhere to study protocol? Did study design or analysis account for important confounding or modifying variables, including unintended co-exposures)?		

INTEGRATION

Table S4. Experimental Studies Integration Template

Criterion	Considerations	Reviewer's Judgement	Support for Judgement	Rating
Causation	Consistency			
	Magnitude			
	Essentiality			
	Specificity			
	Temporality			
	Exposure-response			
	Biological Concordance			
	Analogy			
Human Relevance	Relevant groups and life stages			
	Comparative developmental processes and their relative timing			
	Differences in ontogeny			
	Consequences of interaction of chemical with cells, tissues, and organs			
	Magnitude of exposure differences for observation of key events or apical outcome			
Confidence in Biological Plausibility				

Table S5. Bradford Hill Criteria for Evidence Integration Template

Aspect	Reviewer's Judgement	Support for Judgement	Causal Conclusion
Strength of Association			
Consistency			
Specificity			
Temporality			
Exposure-Response			
Biological Plausibility			
Coherence			
Experiment			
Analogy			

REFERENCES

NTP. 2013. *Risk of Bias Tool*. National Toxicology Program Office of Health Assessment and Translation. https://ntp.niehs.nih.gov/ntp/ohat/pubs/riskofbiastool_508.pdf.

APPENDIX A

DATA COLLECTION AND SUMMARY FOR CROUSE ET AL. 2020

Crouse DL, Erickson AC, Christidis T, Pinault L, van Donkelaar A, Li C, Meng J, Martin RV, Tjepkema M, Hystad P, Burnett R, Pappin A, Brauer M, Weichenthal S. 2020. Evaluating the sensitivity of PM2.5–mortality associations to the spatial and temporal scale of exposure assessment. Epidemiology 31: 168-176.

Summary of Crouse et al. 2020 Findings

The Supplement to the 2019 ISA for PM (EPA 2021) cited Crouse et al. (2020) as supporting an association of long-term PM2.5 and cardiovascular mortality. The primary objective of the study was to evaluate the sensitivity of cause-specific mortality ratios (including cardiovascular disease and ischemic heart disease) according to varying temporal and spatial scales of exposure. Crouse et al. (2020) evaluated 1-, 3-, and 8-year average exposure at 1-, 5-, and 10-kilometer spatial resolution. The secondary aims were to evaluate confounding by co-exposure to NO2 and ozone, as well as the sum of the oxidant gases (Ox).

For each spatial scale, the association between PM2.5 and cardiovascular mortality was of larger magnitude when using the 8-year average (example: for 1-km spatial scale, HR 1.29, 95% CI 1.23–1.35 per 10 μg/m^3) than the 3-yr average (1-km: HR 1.25, 95% CI 1.19 – 1.31 per 10 μg/m^3) or 1-yr average (1 km: HR 1.16, 95% CI 1.11 – 1.21 per increase of 10 μg/m^3). The association between PM2.5 and cardiovascular mortality was attenuated in sensitivity analyses that restricted the cohort to those living in urban areas: For 1-km spatial scale, the HR was 1.15 (95% CI 1.09 – 1.22) per 10 μg/m3 in 8-yr average, the HR was 1.12 (95% CI 1.07 – 1.18) per 10 μg/m^3 in the 3-yr average HR , and the HR was 1.07 (95% CI 1.02 – 1.12) per 10 μg/m^3 in the 1-yr average.

Sensitivity analyses that adjusted for co-pollutants showed attenuated hazard ratios when adjusting for NO2, ozone, or Ox. Hazard ratios were lowest from fully-adjusted models (see eTable 4 in Crouse et al. 2020). When adjusting for the sum of oxidant gases, the HR for cardiovascular mortality was 1.08 (95% CI 1.03 – 1.14) (based on the AIC with lowest value) for the 3-yr/1-km spatiotemporal scale. For IHD mortality, the HR was 1.07 (95% CI 1.00 – 1.15) after adjusting for Ox for the 3-yr/1-km spatiotemporal scale.

Table A1. Crouse et al. 2020 PECOS Evaluation

Study Element	Description
Participants	2.4 million Canadian adults 25-89 yrs at baseline (2001) and followed through 2011. (Sample of the 2001 CanCHEC database)
Exposure	PM2.5, alone and in models with O3, NO2, and Ox (sum of oxidants). (Estimated long-term average exposures to outdoor PM2.5 were below 10 μg/m3 and average concentrations were below 8 μg/m3 for all combinations of temporal/spatial scales) -3 temporal and 3 spatial moving averages
Comparator	Per 10 μg/m3 PM2.5 for PM2.5 assigned to different combinations of temporal and spatial scales: 1, 3, and 8-year moving averages and 1, 5, and 10 km
Outcome	Seven groups of mortality: nonaccidental, cardiometabolic (circulatory diseases plus diabetes), cardiovascular disease, ischemic heart disease, cerebrovascular disease, nonmalignant lung disease, lung cancer
Study Design	Cohort

Table A2. Crouse et al. 2020 Risk of Bias Evaluation

Bias Domain	Reviewer's Judgment	Support for Judgment
Source population representation (Did selection of study participants result in appropriate comparison groups?)	R1 Low R2 Low R3 Low	Sample of the 2001 CanCHEC (2.4 million Canadian adults, 25-89 years at baseline (2001) and followed through 2011. Participated in 2001 long form census and linked to income tax filings and mortality database
Confounding (Did study design or analysis account for important confounding and modifying variables?)	R1 Probably High R2 Probably High R3 Probably High	Adjusted for individual level variables: Age, sex, Aboriginal identity, visible minority status, marital status, highest level of education, employment status. Also adjusted for demographic-level variables: community size, airshed, household income adequacy quintiles (neighborhood instability, deprivation, dependency, ethnic concentration) though no individual data was available for these variables. Sensitivity analysis: co exposure to ozone, NO2, or Ox (combined oxidant capacity) Also used indirect adjustment for missing risk factors (smoking, alcohol consumption, exercise, and fruit-vegetable intake) O3 mean 11.6 ppb, median 10.4 ppb, 5.9–7.8 µg/m3 (range 0 – 20) +age, employment status, urban/rural residence +co-exposure to O3, NO2 +contextual (community, marginalization, airshed) Confounders may have been misclassified; no individual level data on several important confounders; no estimate of unmeasured confounding
Incomplete outcome data (Were outcome data complete without attrition or exclusion from analysis?)	R1 Probably Low R2 Low R2 Low	Assume Canadian mortality data database is comprehensive; mortality database linked to the census data and annual income tax filings Linkage rates to mortality files were high Mortality data from recorded deaths

Exposure assessment (Can we be confident in exposure characterization?)	R1 Probably low R2 Probably Low R3 Probably High	Exposure assigned based on postal code (included annual updates to account for residential relocation) 3-year moving exposure window lagged one year, sensitivity analyses included spatiotemporal model (1-, 3-, and 8- year moving average), 1-, 5-, and 10- km spatial resolution GEOS-Chem chemical transport model, AOD surface PM2.5 retrieved from satellites PM2.5 mean (range of different spatiotemporal scales), 6.2 - 8.0 μg/m3; max, 20 μg/m3 "Estimates >20 μg/m3 were assigned values of 20 μg/m3, as values greater than that may represent inaccurate satellite retrievals." Satellite data available since 1998; 1993-1997 estimates based on GEOS Chem simulations of measured PM2.5 and particles <PM10 Exposure misclassification is possible; however, the expectation of the bias is toward the null assuming non-differential exposure misclassification. - Postal codes have different resolution for urban (100 – 160 m) vs. rural (1 – 5 km. - about 20 – 30% lost to follow-up. +residential mobility based on histories +sensitivity of spatial temporal averages No data on time spent in assigned exposure area or indoor air quality; potential exposure misclassification in either direction
Outcome assessment (Can we be confident in outcome assessment?)	R1 Low R2 Low R3 Low	Canadian mortality database or Canadian census-tax-mortality There is direct/indirect evidence that the outcome was assessed using well-established methods Prevalent cases were not excluded.
Selective outcome reporting (Were all measured outcomes reported?)	R1 Low R2 Low	Results for mortality for seven causes of death (ICD9) presented in supplemental material:

	R3 Low	nonaccidental (ICD-10: A to R) cardiometabolic (i.e., circulatory plus diabetes; ICD-10: I10 to I69, E10 to E14) cardiovascular diseases (ICD-10: I10 to I69) ischemic heart disease (ICD-10: I20 to I25) cerebrovascular disease (ICD-10: I60 to I69) non-malignant respiratory disease (ICD-10: J00-J99); and lung cancer (ICD-10: C33 to C34). The case definitions, while provided, were different in this study compared to others.
Other potential threats to internal validity (Were statistical methods appropriate? Did researchers adhere to study protocol?)	R1 Probably Low R2 Probably Low R3 Probably Low	Cox proportional hazards models Hazard ratios and 95% confidence intervals were reported for PM2.5 alone and adjusted for co-pollutants (O3, NO2, Ox) No information on study protocol Is the study overpowered? Results appear robust to different spatiotemporal scales Heatmaps provided outcomes based on different spatial and temporal scales.
Model Specification (Were statistical models evaluated for validity of underlying assumptions or was model validity assessed with exposure ranges outside the range of interest for research question?)	R1 Probably High R2 Probably Low R3 Probably High	Cox proportional hazards model to assess PM2.5 exposure and different temporal and spatial scales HRs were calculated per 10 μg/m3; however, the mean PM2.5 was < 8 μg/m3 for each temporal and spatial scale combination (median exposures were slightly lower than mean) and the maximum exposure was 20 μg/m3. Cox PH assumes proportionality of exposure response along the entirety of the E-R curve and is likely to inaccurately estimate relationships at low and high ends of the curve in non-linear E-R systems
External evidence of bias (Does external research indicate there may be unmeasured uncertainty, bias, or confounding present in study design or dataset?)	R1 Probably Low R2 Probably Low	Literature hazard ratios for indirect adjustment were of large magnitude for smoking (HR 3.08, 95% CI 2.70 -3.51) for 20-39 cigarettes smoked per day; high exercise (≥1000 metabolic equivalent of task (MET)-minutes was

	R3 Probably High	significantly decreased (HR 0.65, 95% CI 0.61 – 0.70). The authors did not present results for indirect adjustment of each variable (smoking, alcohol consumption, exercise, and fruit-vegetable intake) Lack of individual adjustment for confounding for several key confounders combined with evidence in the literature for the presence of unmeasured confounding in this type of study design that has not been accounted for

Summary of Risk of Bias Findings

The highest rank this study can be classified as is Tier 2 study because the risk of bias domain for exposure is rated as "probably low" (though not with complete reviewer consensus) while the risk of bias domain for confounding is rates as "probably high." Important risk factors for cardiovascular disease and mortality include factors such as high blood pressure, high total cholesterol and low-density lipoprotein (LDL), obesity, diabetes, family history of heart disease, and physical activity/exercise. These factors were not considered. Authors reported results of sensitivity analyses that indirectly adjusted for smoking intensity, alcohol consumption, physical activity, and fruit and vegetable intake (included as categorical variables) for (based on the 3-year moving average and 1-km spatial scale) for cardiovascular mortality (HR 1.22, 95% CI 1.16–1.28). This is somewhat lower than the 3-yr moving average and 1-km spatial scale in the main analysis of HR 1.25, 95% CI 1.25–1.31). It would have been helpful to present the indirectly-adjusted results for one variable (e.g., smoking) at a time to evaluate the effect of the variable, rather than a single estimate that is adjusted for all four variables.

There is risk of bias present based on assignment using postal code histories. The authors were able account for residential mobility using annual records. While this is a strength, these data were incomplete for about a quarter of the cohort. The exposures were summarized annually. This diffuses peaks and higher intensity concentrations which may be informative with respect to risk estimates and shape of the concentration-exposure curve. The use of postal codes is also a standard practice but as the authors noted, has a different specificity across population densities. As well, there is no data for movement within and without defined exposure zones or indoor air quality, introducing further risk of bias in this domain.

The investigators do not discuss the larger attenuations in HRs in the sensitivity analyses restricted to cohort members living in urban areas only. For example, the 3-yr moving average and 1-km spatial scale HR for cardiovascular mortality was 1.12 (95% CI 1.07–1.18).

Each risk estimate was for an increase of 10 $\mu g/m^3$; however, the median and mean PM2.5 exposure estimates were below 8 $\mu g/m^3$ for each spatiotemporal scale. Exposure contrasts were relatively low: At each temporal scale (1-, 3- and 8-yr moving averages), the exposure contrast between the mean PM2.5 values for the different spatial scales was approximately 1 $\mu g/m^3$, with highest mean value at 1-km scale and lowest mean value at 10-km scale (e.g., for 3-year moving average, PM2.5 mean was 7.43 μg/m3 for 1-km scale and 6.44 μg/m3 for 10-km scale. A similar exposure contrast

was seen with median values. For each spatial scale, smaller exposure contrasts were seen for the 1-yr, 3-yr, and 8-yr moving averages (e.g., at 1-km scale, the 1-, 3- and 8-yr moving averages were 7.21, 7.43, and 7.98 $\mu g/m^3$, respectively).

Further risk of bias may be introduced in the model specification domain by employing the Cox Proportional Hazard model which relies on the underlying assumption of proportionality between increasing exposure and outcome. This is likely to overestimate hazard ratios at the upper and lower ends of the concentration response curve, which are of particular importance to policy decision making.

Reviewer Consensus: Tier 2

APPENDIX B

DATA COLLECTION AND SUMMARY FOR

Wang B, Eum K-D, Kazemiparkouhi F, Li C, Manjourides J, Pavlu V, Suh H. 2020. The impact of long-term PM2.5 exposure on specific causes of death: exposure-response curves and effect modification among 53 million US Medicare beneficiaries. Environmental Health 19:20.

Summary of Wang et al. 2020 Findings

The Supplement to the 2019 ISA for PM cited Wang et al. (2020) as supporting an association of long-term PM2.5 and cardiovascular disease mortality. The objective of Wang et al. (2020) was to examine the exposure-response shape for low PM2.5 exposures and mortality from specific causes. Wang et al (2020) also evaluated the effect of non-traffic PM2.5 on mortality for all-cardiovascular mortality as well as ischemic heart disease (IHD), cerebrovascular, and congestive heart failure (CHF) mortality. Wang et al. also assessed effect modification by sex, race, age, SES, and urbanicity, and the impact of confounding by these variables.

After adjusting for socioeconomic indicators, mortality from all cardiovascular disease and ischemic heart disease was increased (CVD HR 1.088, 95% CI 1.078-1.098 and IHD HR 1.126, 95% CI 1.112-1.140) per increase of 10 μg/m^3 in annual PM2.5. When evaluating non-traffic PM2.5, the hazard ratios were attenuated for all cardiovascular disease (HR 1.016, 95% CI 1.005-1.028) and IHD (HR 1.027, 95 CI 1.011-1.043).

After adjusting for socioeconomic indicators, mortality from congestive heart failure mortality was not associated with one-year average PM2.5 (for year before death) (HR 0.986, 95% CI 0.953–1.021 per increase of 10 μg/m^3) or non-traffic PM2.5 (HR 0.970, 95% CI 0.930–1.012 per increase of 10 μg/m^3).

Age and race modified the association between PM2.5 and cardiovascular mortality, IHD mortality and CHF mortality. HRs were higher for younger (≤75 years) than for older (>75 years) beneficiaries for all CVD mortality as well as for IHD mortality and CHF mortality. PM2.5 was not associated with CHF mortality for beneficiaries older than 75 years (HR) For IHD mortality and CHF mortality, urbanicity modified the association between PM2.5 and mortality in models adjusted for SES status.

CVD mortality was increased in models restricted to beneficiaries living in ZIP codes with average PM2.5 concentrations below 8 μg /m^3 (HR 1.34, 95% CI 1.30–1.37), below 10 μg /m^3 (HR 1.533, 95% CI 1.510–1.556) or below 12 μg /m^3 (HR 1.659, 95% CI 1.641–1.677). These models were not adjusted for SES and other models have shown that adjusting for SES attenuated risk ratios.

Table B1. Wang et al. 2020 PECOS Evaluation

Study Element	Description
Participants	53 million US Medicare beneficiaries (≥65 yrs old), 74% lived in urban area
Exposure	PM2.5 (2000–2008)
Comparator	Per 10 μg /m3 increase in 12-month average PM2.5 for year before death
Outcome	Cardiovascular mortality
Study Design	Cohort

Table B2. Wang et al. 2020 Risk of Bias Evaluation

Bias Domain	Reviewer's Judgment	Support for Judgment
Source population representation (Did selection of study participants result in appropriate comparison groups?)	R1 Probably Low R2 Low R3 Low	Medicare enrolment database for US residents, 2000-2008, age 65- 120 years Population is large and nationally representative for older adults (> 65 years)
Confounding (Did study design or analysis account for important confounding and modifying variables?)	R1 Probably High R2 Probably High R3 Probably High	Age, sex, race, and ZIP code (stratified), neighborhood socioeconomic status Sensitivity analysis adjusted for 1 hr maximum ozone (warm season average Non-traffic PM2.5 effect examined used two-stage models of PM2.5 and NO2: stage 1) regression of 12-month PM2.5 on NO2 and stage 2) residuals of stage 1 (as non-traffic PM2.5) as measure of exposure in Cox PH models Lack of individual level data for many factors, resulting in indirect approaches. +Age, race, sex Estimated urban/rural residence, SES +co-exposure to O_3, NO_2
Incomplete outcome data (Were outcome data complete without attrition or exclusion from analysis?)	R1 Low R2 Low R3 Low	Mortality data from National Death Index for major categories of mortality (specific categories): All causes (Non-accidental and accidental) All cardiovascular diseases (IHD, cerebrovascular, congestive heart failure) All Respiratory (COPD, Pneumonia) All Cancer (lung cancer)

		Matched to US National Death Index from an insurance cohort, while % follow-up is not included, the inclusion is likely to be high.
Exposure assessment (Can we be confident in exposure characterization?)	R1 Probably Low R2 Probably Low R3 Probably High	12-month average exposure for year before death estimated from daily PM2.5 estimates for a 6-km grid using spatiotemporal generalized additive mixed model Exposure assigned to zip code centroid, accounted for residential moves Cross-validation R2=0.76 Overall mean 12-month average, 10.32 μg /m^3 (SD 3.15), maximum not provided, but SD suggests that the most values are below 20 μg/m^3 Monthly NO2 concentration for a 100 m grid R2 = 0.82 spatial variability and R2 = 0.76 temporal variability 1-h max ozone (warm season average) for subset living within 6 mils of US EPA air quality monitors - Postal codes have different resolution for urban vs. rural in the United States. +residential mobility based on histories No data to evaluate intra-day spatial movement or indoor air quality; potential for exposure misclassification or poor exposure estimate quality
Outcome assessment (Can we be confident in outcome assessment?)	R1 Low R2 Low R3 Low	Mortality data from National Death Index High quality database, includes most deaths for the United States There is direct/indirect evidence that the outcome was assessed using well-established methods

		Prevalent cases were not excluded
Selective outcome reporting (Were all measured outcomes reported?)	R1 Low R2 Low R3 Low	All-CVD mortality reported IHD mortality CBV mortality CHF mortality Specific cardiovascular outcomes were analyzed.
Other potential threats to internal validity (Were statistical methods appropriate? Did researchers adhere to study protocol?)	R1 Probably Low R2 Probably Low R3 Probably Low	Cox proportional hazard models Hazard ratios and 95% confidence intervals presented with HR adjusted for co-pollutants No information on study protocol Immortal time bias? Variation in timing of exposure: if exposure is misclassified or ignored,
Model Specification (Were statistical models evaluated for validity of underlying assumptions or was model validity assessed with exposure ranges outside the range of interest for research question?)	R1 Probably High R2 Probably Low R3 Probably High	Cox proportional hazards models. Models also fit using restricted cubic splines with three knots (10^{th}, 50^{th} and 90^{th} percentiles) to characterize non-linearity Effect modification was assessed for ag, sex, race, and urbanicity 4 billion person-years of follow up Immortal time bias? Model validity no assessed with exposure ranges outside the range of approximately Cox PH assumes proportionality of exposure response along the entirety of the E-R curve and is likely to inaccurately estimate relationships at low and high ends of the curve in non-linear E-R systems
External evidence of bias (Does external research indicate there may be unmeasured uncertainty, bias, or confounding present in study design or dataset?)	R1 Probably High R2 Probably Low R3 Probably High	53 million beneficiaries for 2000-2008 is about 33% more than expected based on the following information:

		In 2000, the population 65 and over was 35 million. In 2008, 2.7 million individuals turned 65 years. In 2008, about 1.8 million 65 and older died (net annual of 927,305). Rounding up to 1 million net increase per year: 35 million + (1 million x 8 years) = 43 million. In 2007, 93% covered by Medicare: ~40 million 65 and older during 2000-2008 covered by Medicare. Source: Administration on Aging. US Department of Health and Human Services, A Profile of Older Americans: 2009.[3] Lack of individual adjustment for confounding combined with evidence in the literature for the presence of unmeasured confounding in this type of study design that has not been accounted for

Summary of Risk of Bias Findings

This highest this study can be classified is a Tier 3 study because the risk of bias domain for exposure is rated as "probably low" while the risk of bias domain for confounding is rated as "probably high" and there was also non-consensus ratings of probably high for Model Specification and External evidence of bias. Important risk factors for cardiovascular disease and mortality include factors such as high blood pressure, high total cholesterol and low-density lipoprotein (LDL), obesity, diabetes, family history, and physical activity/exercise. These factors were not considered.

There is moderate risk of bias based on assignment using postal code histories. Exposure was averaged over 12 months based on a 6-mile buffer of the residential postal code. The models used US databases of stationary monitors and meteorological data (Yanosky et al. 2014). Monthly NO2 data were used to estimate traffic (Kazemiparkouhi et al. 2020).

Peaks and higher intensity concentrations were not available using annual data, which may be informative with respect to risk estimates and shape of the concentration-exposure curve. The use of postal codes is also a standard practice but have a different specificity across population

[3] Available at: https://acl.gov/sites/default/files/Aging%20and%20Disability%20in%20America/2009profile_508.pdf

densities. As well, individual exposure data was not available including indoor air quality and movement within and without the exposure zone.

Models adjusted for ozone showed slightly attenuated hazard ratios in the base model; however, models adjusted for ozone and SES variables show larger attenuations in the hazard ratios: For IHD, the base model showed a HR of 1.92 (95% CI 1.89 – 1.95) per 10 μg/m3 increase in PM2.5. After adjusting for ozone, the HR was 1.90 (95% CI 1.88 – 1.93). Additional adjustment for ZIP code level SES resulted in an HR of 1.14 (95% CI 1.12 – 1.16) (see Table

Figure 1 in Wang et al. (2020) shows both the non-linear and linear association of PM2.5 on cardiovascular mortality. For the non-linear association, the authors fitted restricted cubic splines with three knots (knots were specified at the 10th, 50th and 90th percentiles). There appears to be little difference in the non-linear and liner models for PM2.5 and cardiovascular disease (Figure 1 of Wang et al. 2020), although the HRs for the non-linear model appear to be slightly lower than the linear model over the range of exposures from 10 to 20 μg/m^3 (which is the range for more than 95% of the data based on a mean 12-month concentration of 10.32 μg/m3 and standard deviation of 3.15 μg/m^3).

The authors reported "Beneficiaries living in urban as compared to non-urban ZIP codes had higher PM2.5-associated mortality risks for non- accidental, respiratory, and cancer mortality, with similar risks for CVD-related mortality. RRs for beneficiaries living in non-urban areas were positive and statistically significant for CVD-related causes of death." Closer examination of Figure 3, however, showed urbanicity significantly modified the association between PM2.5 and IHD (RR 1.12, 95% CI 1.10–1.13 for urban dwellers with average PM2.5 of 11.1 μg/m^3 versus RR 1.16, 95% CI 1.13 – 1.19 for rural dwellers with average PM2.5 of 8.9 μg/m^3. The PM2.5 and CHF association was in different directions for urban (RR 0.94, 95% CI 0.90 – 0.98) and rural dwellers (RR 1.15, 95% CI 1.08 – 1.23).

The authors reported in the methods, "To further examine effects of low PM2.5 exposures, we fit SES-adjusted models restricted to beneficiaries living in ZIP codes with average PM2.5 concentrations below 8, 10 or 12 μg/m^3." However, the authors only presented results for gender, race, age, and ZIP codes strata. The results from the SES-adjusted models were not presented. Based on other results, the expectation is that the HRs would be attenuated for each of these categories. Also, it is not clear how to interpret an increase of 10 μg /m^3 when analyses are restricted to beneficiaries with average PM2.5 concentrations below 8 μg /m^3 (or even below 10 μg /m3). Such an interpretation assumes a linear relationship exists, but linearity over a narrow exposure range should not be assumed.

Further risk of bias may be introduced in the model specification domain by employing the Cox Proportional Hazard model which relies on the underlying assumption of proportionality between increasing exposure and outcome. This is likely to overestimate hazard ratios at the upper and lower ends of the concentration response curve, which are of particular importance to policy decision making.

Higher PM concentrations were recorded in the urban areas, where ¾ of the population resided. The correlation with ozone and NO2 varied by region.

Lung cancer and congestive heart failure mortality were not associated with PM, particularly after SES adjustment. The stratified results in Figures 2 and 3 highlight the differences in risk profiles. Speculation on the role of confounding by personal behaviors is warranted. Is smoking, exercise and diet less impactful upon the association of mortality and PM among the elderly (>75)? Does

unmeasured confounding explain the striking racial differences for mortality? What is the role of smoking for the nonurban population, which has lower risk estimates for respiratory and lung cancer mortality?

A potential issue of concern is whether the underlying database is inflated with respect to the number of Medicare beneficiaries and/or if there are duplicate individuals. Based on information from the census for 2000 and other government agencies, it appears that individuals could be overestimated by 30%. This also potentially impacts whether there is immortal time bias that potentially biases the result away from the null.

Reviewer Consensus: Tier 3

APPENDIX C

DATA COLLECTION AND SUMMARY FOR CHEN ET AL. 2020

Chen H, Zhang Z, van Dokelaar A, Bai L, Martin RV, Lavigne E, Kwong JC, Burnett RT. 2020. Understanding the joint impacts of fine particulate matter concentration and composition on the incidence and mortality of cardiovascular disease: a component-adjusted approach. *Environmental Science and Technology* 54: 4388-4399

Summary of Chen et al. 2020 Findings

The Supplement to the 2019 ISA for PM cited Chen et al. (2020) as supporting an association of long-term PM2.5 and cardiovascular disease and acute myocardial infarction at low PM2.5 levels. The objective of this study was to evaluate the joint effects of PM2.5 (mass concentration) and its components on CVD mortality and incidence of acute myocardial infarction. Chen et al. (2020) used a Shape Constrained Health Impact Function (SCHIF) to assess the shape of the concentration-response (C-R) relationship for PM2.5 mass concentration. Based on the shape of the C-R function, Chen et al. (2020) modeled the relative contribution of each component on cardiovascular disease mortality. The authors compared four models: single-pollutant linear Cox model, single-pollutant non-linear Cox model (PM2.5 fitted as a nonlinear term using SCHIF), a multiple-component linear Cox model and a component-adjusted nonlinear Cox model. The Akaike information criterion (AIC) value was used to assess model fit. The authors then used the component-adjusted non-linear model to estimate the risks of AMI incidence and CVD mortality for each postal code in Ontario stratified by five regions and compared them to the risks of AMI incidence and CVD mortality from the single-pollutant nonlinear model.

Overall, the risk of AMI per increase in 5 μg/m^3 PM2.5 was essentially the same for each model: HR 1.14, 95% CI 1.12–1.16 for the single pollutant linear model, the single-pollutant non-linear model, and the component-adjusted nonlinear model. The risk of CVD mortality per increase in 5 μg/m^3 PM2.5 was also similar for each model: HR 1.10, 95% CI 1.09 – 1.12 for the single pollutant model, HR 1.11, 95% CI 1.10 – 1.12 for the single pollutant non-linear model, and HR 1.10, 95% CI 1.09 – 1.12 for the component-adjusted nonlinear model. Chen et al. 2020 reported that the model that adjusted for the proportion of components that make up PM2.5 was superior (based on AIC values) to the single-pollutant models for five regions of Ontario. Cardiovascular mortality increased on average by 27% and acute myocardial infarction incidence increased on average by 10% when compared to single-pollutant models for five regions of Ontario (see Figure 2 for AMI incidence and Figure 3 for CVD mortality of Chen et al. 2020).

Table C1. Chen et al. 2020 PECOS Evaluation

Study Element	Description
Participants	Ontario Population Health and Environment Cohort (ONPHEC) members; 5,264,985 adults 35 to 85 years as of Jan 1 2001, residents of Ontario for ≥5 years, born in Canada, registered with provincial health insurance (full cohort). 48% male, 82% urban dweller; followed 2001-2016 Incident cohort of 5,140,853 subjects without physician-diagnosed AMI, 48% male, 82% urban dweller, average age 54 years
Exposure	PM2.5 (long-term exposure), and components across postal codes between 2000 and 2016, using postal codes across 1 km centroid Adjusted for seven components of PM2.5 (black carbon, ammonium, nitrate, organic matter, sea salt, sulfate, mineral dust)

Comparator	per 1 μg/m^3 increase in PM2.5
Outcome	CVD mortality (from Ontario Registrar General's Death database) or AMI incidence (from a validated database of hospital discharge data) for cohort members ICD codes ICD-9: 401-459 and ICD-10: I10-I99
Study Design	Cohort

Table C2. Chen et al. 2020 Risk of Bias Evaluation

Bias Domain	Reviewer's Judgment	Support for Judgment
Source population representation (Did selection of study participants result in appropriate comparison groups?)	R1 Low R2 Low R3 Low	Inclusive of all Ontario adult (35-85 yrs) residents enrolled in provincial health services Characteristics (sex, age, region) are reported for full cohort and incident cohort in Table S2. Population is large and nationally representative
Confounding (Did study design or analysis account for important confounding and modifying variables?)	R1 Probably High R2 Probably High R3 Probably High	Adjusted for individual-level variables: age at baseline, sex, and different indicators of area of residence (urban/rural; north/south, Greater Toronto/outside Greater Toronto) Also adjusted for neighborhood-level variables of SES: income quintile, % unemployed, % <high school education, % recent immigrants Meteorological variables not included, smoking, indoor exposures not assessed No information on important confounders such as smoking, diet, physical activity, family history, cholesterol, and high blood pressure. The assumption is that misclassification is non-differential; therefore, the expected bias is toward the null.

		+age, sex, urban/rural residence, north/south indicator, neighborhood SES + Residence in Greater Toronto Area - No adjustment for other pollutants - No indirect adjustment for contextual or behavioural factors
Incomplete outcome data (Were outcome data complete without attrition or exclusion from analysis?)	R1 Probably Low R2 Low R3 Low	Two outcomes were studied: AMI and cardiovascular disease. Prevalent cases of AMI (diagnosed before 2001) were excluded from the incidence cohort. This was appropriate. Linkage to mortality files not described but based on insurance data. Analyses by diagnoses not conducted
Exposure assessment (Can we be confident in exposure characterization?)	R1 Probably low R2 Probably Low R3 Probably High	2000-2016 annual averages of PM2.5 exposure estimated using satellite aerosol optical depth (AOD) and GEOS-Chem transport model, weighted regression, 1 x 1 km spatial resolution for each year. Annual estimates assigned to cohort members based on residential postal code. Median PM2.5 across all years ~8 μg/m^3, maximum ~13 μg/m^3 Seven components: Organic mass, median ~4 μg/m^3, max ~ 5 μg/m3 All others below ~ 2 μg/m^3, max below ~2.5 μg/m^3 For each location/postal code, the sum of the proportions of the components equals 1.

		- Postal codes have different resolution for urban (100 – 160 m) vs. rural (1 – 5 km. -Annual estimates of exposure +residential mobility based on histories +evaluating different components No data to evaluate intra-day spatial movement or indoor air quality; potential for exposure misclassification or poor exposure estimate quality
Outcome assessment (Can we be confident in outcome assessment?)	R1 Low R2 Low R3 Low	Incident AMI from validated database (1st hospital admission with diagnosis of MI) Previously validated Sensitivity 89%; specificity 93%; positive predictive value 89% CVD mortality from Ontario Registrar General's Death database (98% linked) There is direct/indirect evidence that the outcome was assessed using well-established methods Prevalent cases were excluded
Selective outcome reporting (Were all measured outcomes reported?)	R1 Probably Low R2 Low R3 Low	On average, there was a 10% increase in the PM2.5–AMI association in all five regions after considering the joint change in PM2.5 mass and the relative contributions of its components compared to considering PM2.5 mass only. There is direct/indirect evidence that all the study's measured outcomes (primary and secondary) outlined in protocol, methods, abstract, and/or introduction (that are relevant for evaluation) have been reported. This includes outcomes reported with sufficient detail to be included in meta-analysis or fully tabulated during data extraction

Other potential threats to internal validity (Were statistical methods appropriate? Did researchers adhere to study protocol?)	R1 Probably Low R2 Probably Low R3 Probably Low	AIC is a valid tool to compare non-nested models. It is not clear, however, if the AIC as a measure of goodness-of-fit is appropriate for enormous sample sizes. AIC difference between multiple-component Cox model and component adjusted nonlinear Cox model is 1 and may not be important (the EPA also may not think it is important). Authors also cite p-value for likelihood-ratio test that compared two non-nested models ($p<0.001$); this is not valid. No information on study protocol.
Model Specification (Were statistical models evaluated for validity of underlying assumptions or was model validity assessed with exposure ranges outside the range of interest for research question?)	R1 Probably High R2 Probably Low R3 Probably High	Although the two-stage approach is described in the methods, the research results are poorly described and the results that the authors discuss are not easy to verify. Used Shape Constrained Health Impact function (SCHIF) to evaluate the dose-response. Confidence intervals were not reported in tables.
External evidence of bias (Does external research indicate there may be unmeasured uncertainty, bias, or confounding present in study design or dataset?)	R1 Probably High R2 Probably Low	Over the period 2000-2016, mean PM2.5 levels for Canada averaged below 10 μg/m^3: values were relatively stable (approximately 8 μg/m^3) for the years 2000-2013 and have dropped to approximately 7 μg/m^3 (or less) for the years 2014-206.[4] The authors reported "the annual mean concentration of PM2.5 across Ontario over the period of 2000 to 2016 was 8.61 μg/m^3

4 https://data.worldbank.org/indicator/EN.ATM.PM25.MC.M3?end=2017&locations=CA&start=1990&view=chart

	R3 Probably High	(Figure 1C), well below the World Health Organization's air quality guideline for PM2.5 (annual mean: 10 μg/m^3). Authors conclude: "On average, there was a 10% increase in the PM2.5–AMI association in all five regions after considering the joint change in PM2.5 mass and the relative contributions of its components compared to considering PM2.5 mass only." Not clear where calculated value of 10% comes from.

Summary of Risk of Bias Findings

This study is classified as Tier 1 based on a judgment of "probably low" risk of bias related to the exposure characterization domain and "probably high" for the confounding domain with additional non-consensus reviewer ratings of "probably high" in the domains of "External evidence of bias" and "Model Specification". Overall, this is an example of an article that provides a novel approach to jointly estimate the effects of PM2.5 mass concentrations and its components on AMI incidence and CVD mortality risks. Some of the reported data (especially related to the "significant improvement in explaining the health impacts of PM2.5, especially in the presence of a nonlinear PM2.5-outcome relationship") cannot be easily verified and has to be taken at face value based on review of two figures in the paper. It also appears to be based on a statistical evaluation that largely relied on AIC values without a considered appraisal as to whether the area of residence variables are sufficiently independent to be included in the model of the risk of cardiovascular disease or AMI in relation to PM2.5 concentrations. This has risk of bias implications for the exposure assessment domain.

The exposure assessment has a moderate risk of bias. The approach of the investigation is robust with respect to relying on available databases and models to objectively estimate PM and component levels. The analysis by composition of the PM is a strength of this analysis. However, the exposures were summarized annually. This diffuses peaks and higher intensity concentrations which may be informative with respect to risk estimates and shape of the concentration-exposure curve. The use of postal codes while a standard practice but as the authors noted, has a different specificity across population densities.

There is likely to be risk of bias related to unmeasured confounding. The authors appropriately adjusted for available information on basic demographics and region. However, they did not adjust, directly or indirectly, for other pollutants or behavioral risk factors.

A strength of this study was evaluating the major components. A limitation was the presentation in the tables. They were somewhat confusing, without confidence intervals to indicate statistical significance by which to compare with the text in the Results section.

Another strength of this study was the assessment of incident acute myocardial infarction (AMI) as well as cardiovascular mortality. Is PM or one of its components related to onset of disease or aggravation resulting in death? The regression coefficients were similar in most models for incidence and

mortality. However, using more specific (and similar) cardiovascular diagnoses would have facilitated interpretation of incidence vs. mortality.

The observations of specific components with mortality (black carbon and nitrate) compared to all components that were associated with AMI incidence may provide additional clues to disease severity related to PM2.5 sources. As discussed by the authors, wildfires which are a source of black carbon, may contribute to the mortality in selected nonurban regions. Additional data on PM speciation in exposure may be helpful in interpreting these results.

Absent from the analysis were trends in mortality risk over time. The authors note that PM2.5 has declined over time, with the components also changing. Temporal analysis by each component may further inform on the role of chemical composition, pollution sources, and role of exposure level on cardiovascular mortality.

The authors highlight the difficulty in evaluating the composition of PM2.5 on a provincial level.

The authors rely on AIC values as a measure of goodness-of-fit. Here are some examples: The authors state, "Compared to the single-pollutant model, the PM2.5–AMI relationship was generally stronger for the component-adjusted model (Figure 2). On average, there was a 10% increase in the PM2.5–AMI association in all five regions after considering the joint change in PM2.5 mass and the relative contributions of its components compared to considering PM2.5 mass only." Whether there was a 10% increase on average cannot be determined by looking at the scatterplot of approximately 295,000 risk estimates (one hazard ratio for each postal code). The component adjustment factor for AMI varies from a factor of approximately 0.85 to 1.3. When the component adjustment factor is less than 1.0, the component-adjusted PM2.5 is not associated with increased cardiovascular disease in different postal codes. For each postal code, the hazard ratio for AMI incidence using the component-adjusted PM2.5 mass (or stage 2 of the model) is plotted against the hazard ratio for AMI incidence using the nonlinear single-pollutant model using SCHIF (or stage 1 of the model). However, Stage 2 of the model depends on the results of Stage 1 of the model, so these are not independent hazard ratios. It is not clear that the component-adjusted model results in an improved estimate of the PM2.5-AMI association. If the pollutant-AMI association is actually an association between a particular source of PM2.5, it seems that the best measure of the association would be a coefficient associated with the source, and not with PM2.5 mass concentration adjusted simultaneously for the source and six other sources.

The authors present the component-adjustment factor (CAF) for each reach by density in the inset of Figures 2 (AMI incidence) and 3 (cardiovascular mortality). The y-axis is described as density, which I interpreted to mean the density of postal codes in each region (scale is 0 to 30). Also, the authors reported that "the density and distribution of postal codes closely align with that of the Ontario population." It is not clear why the density is greater for the CAF for Region 1 for the AMI association than the density for the CAF for Region 1.

The authors state the following: "We observed that the modulation by PM2.5 components consistently strengthened the association of cardiovascular mortality with PM2.5 (on average a 27% increase across the five regions, varying from 24% in the south to 37% in the north and northwest) than considering PM2.5 alone, even after accounting for the nonlinear relationship between PM2.5 and cardiovascular mortality." However, this information is not obvious from Figure 3, which shows that the postal-code specific HR for component-adjusted PM2.5 in cardiovascular mortality always exceeds the single-pollutant nonlinear hazard ratio for PM2.5 (the component-adjustment factor always exceeds 1.0 for each region and for each postal code). In contrast, the component-adjusted factor for AMI includes numbers below 1.0 (and presumably results in some instances in which the postal-code specific HR for

AMI using the single-pollutant nonlinear model exceeds the postal-code specific HR for AMI using the component-adjusted model.

Using this approach, the authors state that a negative coefficient for a component decreases the overall HR for total PM2.5 mass. Presumably, a positive coefficient for a component increases the overall HR for total PM2.5 mass?

Other results are difficult to verify. For example, the authors state, "Third, using multiple-pollutant Cox models in which all component concentrations were examined simultaneously, AMI incidence was significantly associated with all seven components, whereas cardiovascular mortality was significantly associated with BC and $NO3^-$ (Table 1)." Closer inspection of Table 1 does not show any coefficients that are identified as statistically significant for either AMI incidence or CVD mortality. The work is on the reader to calculate a hazards ratio and 95% CI confidence interval by exponentiating the coefficients and the standard error multiplied by 1.96. Furthermore, it is not clear how the coefficients for the components in the component-adjusted model are to be interpreted.

In sensitivity analysis (Appendix Table S3), PM2.5 mass concentration was not associated with AMI incidence or cardiovascular death when adjusting for age and sex. A model that further adjusted for urban/rural residence also showed no association between PM2.5 mass concentration and AMI (HR 0.99, 95% CI 0.98 – 1.00 per 5 μg/m3) although the risk of CVD mortality was increased in relation to PM2.5 mass concentration (HR 1.04, 95% CI 1.03–1.06). An indicator variable for residence in the Greater Toronto area (which is urban and located in south Ontario) also showed no increased risk of AMI in relation to PM2.5 mass concentration (HR 1.005, 95% CI 0.993 – 1.1017). In contrast, a model that included an indicator variable for north/south (model 3) showed an increased risk of AMI (HR 1.073, 95% CI 1.059–1.086). In combination, the sensitivity analysis suggests that rural areas (representing 18% of the population) or the north which is rural and represents 8% of the population is driving the increased risk. However, lower concentrations of PM2.5 were seen in the north, which is largely rural (concentrations below 4 $\mu g/m^3$ according to the legend for the maps included in Figure 3). Urban areas are concentrated in Region 3 (whether Greater Toronto area is), Region 1 (Ottawa), and Region 4 (where Windsor and Chatham are located, and where average PM2.5 concentrations are highest (approximately 10 $\mu g/m^3$). This is consistent with the authors report that "the largest increases [*in the PM2.5—AMI association*) were found in the east region (Region 1) and the north and northwest region (Region 5) (both up by ~13%) where PM2.5 concentrations were among the lowest in Ontario." The authors speculate that the increased mortality associated with the north is because the primary particles from wildfires (organic mass) contribute proportionally more to PM2.5 mass concentration than other sources.

Presumably, there would be greater statistical "noise" associated with risks calculated for postal codes where there is less density (smaller populations would produce more variation in risk estimates than postal codes representing greater population density).

Despite a lower AIC value for the fully-adjusted models, it seems likely that the fully-adjusted models over-adjust the risk estimates when they include all area indicators (urban/rural, north/south, Greater Toronto area/outside Greater Toronto area). In addition, there is no information on individual risk factors for cardiovascular disease that may confound the associations between PM2.5 and heart disease. The authors only state "In addition, our sensitivity analyses adjusting for the three area-level indicators individually, and in combination support our inclusion of these variables in our model

specification." It is not clear that AIC values are appropriate for assessing goodness of fit for enormous sample sizes.

Reviewer Consensus: Tier 3

Appendix D: Summary of Evidence Integration

Given that this appendix serves as an example for the treatment of specific articles under the proposed systematic review protocol and not a full implementation of the protocol, a full evidence integration analysis is not possible. However, using examples from the Supplement to the ISA, it is possible to place Crouse et al. 2020, Wang et al. 2020, and Chen et al. 2020 within the broader context-relevant literature to explore its potential impact in forming a conclusion on the charge question of the proposed systematic review. The Supplement to the ISA identified cardiovascular mortality as a strong line of evidence for causal association with policy relevant concentrations of PM2.5 and cited these 3 studies as new evidence for that association. However, when considering evidence integration criteria (Table S5), it is notable that low magnitudes of association exist in these studies (all are lower than 1.5). As a result, these small measures of association are highly susceptible to the subtle presence of bias, uncertainty, and confounding. The emphasis placed on the risk of bias domains of confounding and exposure assessment in the protocol reflects the need to use methodological approaches of exceptionally high quality in order to produce low magnitudes of association that are reliable. The lack of confounding data collected at the individual level and the lack of exposure data collected at the individual level were primary drivers in the ranking of these studies as incapable of being viewed as primary lines of evidence for causation at policy relevant concentrations of PM2.5 (i.e. ranked less than Tier 1).

Most studies in the related body of evidence are characterized by mixed exposures, making the element of specificity difficult to demonstrate, and most studies fail to demonstrate increased hazard ratios as exposure concentration increases in a consistent manner (i.e., dose response). Methodologically, these studies are largely characterized by the absence of individual level exposure assessments, the absence of individual level confounding data, reliance on linear statistical models, and evidence of residual confounding that exists in large scale data sets. It is unlikely that other studies cited in the Supplement to the ISA will rank higher in terms of risk of bias than those evaluated in this analysis (Figure D1).

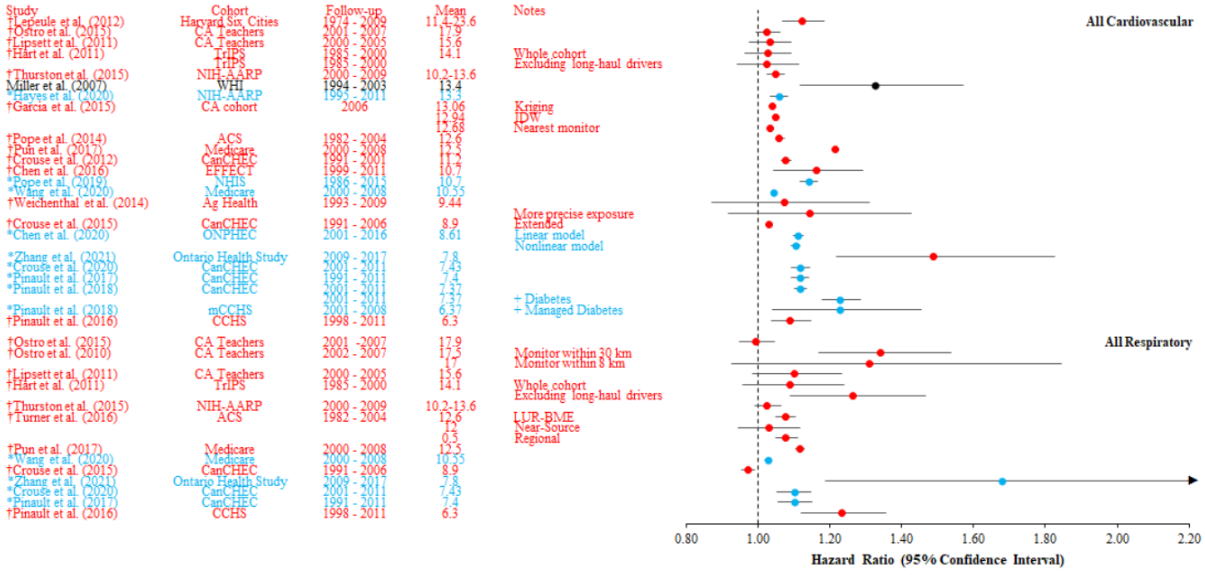


Figure D1: Associations between long-term PM2.5 exposure and all cardiovascular disease and all respiratory disease mortality in recent North American cohorts.[5]

[5] Adapted from USEPA. 2021 Supplement to the 2019 Integrated Science Assessment for Particulate Matter (External Review Draft). EPA/600/R-21/198 Center for Public Health and Environmental Assessment Office of Research and Development U.S. Environmental Protection Agency Research Triangle Park, NCwww.epa.gov/isa

Considering the 22 total articles (Figure D1; All Cardiovascular) that inform on this issue, none are likely to have risk of bias ratings for confounding, exposure characterization, or model specification superior to those of Crouse et al. 2020, Wang et al. 2020, and Chen et al. 2020. This places all human health evidence for the outcome of cardiovascular mortality at policy-relevant PM2.5 concentrations in either Tier 2 or Tier 3. In the uncertainty portion of a systematic review, this finding would substantially lower the confidence of the evidence base for this analysis. Further, statistical significance is broadly lacking across studies for measures of association between PM2.5 exposure and cardiovascular mortality. This is an example of unexplained inconsistency that impacts integration of evidence for causal inference and reduces the confidence of a causal conclusion. Using the prescribed integration approach described in the primary review protocol, lack of consistency, low magnitude of association, and high risk of bias would not lead to a causal conclusion for this outcome at policy relevant levels of PM2.5 exposure. The lack of animal toxicology studies to demonstrate equivalent outcomes at equivalent exposure concentrations in a controlled setting prevents the integration of supporting toxicology evidence.

If the evidence in the Supplement to the ISA is selected, evaluated, and integrated through a systematic review approach with important risk of bias domains included, the overall quality of the evidence base can be transparently conveyed. The reviewers for this evaluation determined that the quality of the evidence base, as indicated by these representative articles, is not sufficient to provide primary evidence for cardiovascular mortality being causally associated with policy relevant concentrations of PM2.5. Use of these modern systematic review approaches is likely to lead to more refined and transparent conclusions regarding the state of the science.