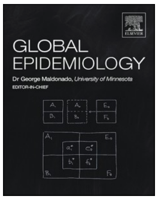
Global Epidemiology 7 (2024) 100130

Contents lists available at ScienceDirect

Global Epidemiology

journal homepage: www.sciencedirect.com/journal/global-epidemiology



An AI assistant to help review and improve causal reasoning in epidemiological documents

Louis Anthony Cox Jr. *

Cox Associates, Entanglement, and University of Colorado, United States of America

ARTICLE INFO

Keywords:
Artificial intelligence
Causality
Review methodology
Causal AI boosting
Large language models (LLMs)

ABSTRACT

Drawing sound causal inferences from observational data is often challenging for both authors and reviewers. This paper discusses the design and application of an Artificial Intelligence Causal Research Assistant (AIA) that seeks to help authors improve causal inferences and conclusions drawn from epidemiological data in health risk assessments. The AIA-assisted review process provides structured reviews and recommendations for improving the causal reasoning, analyses and interpretations made in scientific papers based on epidemiological data. Causal analysis methodologies range from earlier Bradford-Hill considerations to current causal directed acyclic graph (DAG) and related models. AIA seeks to make these methods more accessible and useful to researchers. AIA uses an external script (a “Causal AI Booster” (CAB) program based on classical AI concepts of slot-filling in frames organized into task hierarchies to complete goals) to guide Large Language Models (LLMs), such as OpenAI’s ChatGPT or Google’s LaMDA (Bard), to systematically review manuscripts and create both (a) recommendations for what to do to improve analyses and reporting; and (b) explanations and support for the recommendations. Review tables and summaries are completed systematically by the LLM in order. For example, recommendations for how to state and caveat causal conclusions in the Abstract and Discussion sections reflect previous analyses of the Study Design and Data Analysis sections. This work illustrates how current AI can contribute to reviewing and providing constructive feedback on research documents. We believe that such AI-assisted review shows promise for enhancing the quality of causal reasoning and exposition in epidemiological studies. It suggests the potential for effective human-AI collaboration in scientific authoring and review processes.

Prelude: An AI-generated executive summary

This paper explores the use of artificial intelligence (AI)-assisted reviews of causal claims and analyses in epidemiology. The paper is long, but its main message is short: existing AI is capable of providing useful critical assessments, reviews, and constructive suggestions to help authors (and perhaps also reviewers and editors, if appropriate full disclosure and consent are obtained from all involved) ensure that normative principles for sound causal analysis are followed in epidemiological documents and publications. Applying ChatGPT4 to the sections that follow produced the following summary, with which the author concurs (full chat at https://chat.openai.com/share/f635c0f2-1838-438e-8da8-696684d7de21):

“**Executive Summary of AI-Assisted Causal Analysis in Epidemiological Research**

Overview:

The paper introduces an AI Causal Research Assistant (AIA), an advanced tool designed to aid practitioners in evaluating interventional causal claims in epidemiology. This AI-driven system leverages the principles of “good old-fashioned AI” (GOFAI) to enhance the ability of Large Language Models (LLMs) like ChatGPT to produce meaningful insights into causal relationships from scientific texts.

Key Features:

- AIA is capable of conducting technical reviews of scientific texts, scrutinizing data, methods, and reasoning behind health risk claims related to changing exposures.
- Utilizes a novel Causal AI boosting (CAB) technique that systematically prompts LLMs to generate in-depth feedback, leading to more nuanced and critical reviews.

Methodology:

---

* Corresponding author.
E-mail address: tcoxdenver@aol.com.

https://doi.org/10.1016/j.gloepi.2023.100130
Received 4 October 2023; Received in revised form 8 November 2023; Accepted 24 November 2023
Available online 4 December 2023
2590-1133/© 2023 Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

The study contrasts the feedback from two LLMs, Bard and ChatGPT, on causality in epidemiology papers, highlighting the capability of LLMs to provide both superficial and profound insights based on the complexity of the prompts. An external script, the CAB, enhances the LLM's focus and memory, facilitating a structured review process through a series of detailed prompts and responses, which culminate in a comprehensive review report.

Findings:

Application of the AIA approach to an example paper on PM2.5 and mortality revealed significant potential for AI in aiding causal inference and critical review processes in scientific research. The AIA review process identifies key technical limitations, suggests methodological improvements, and advocates for a balanced presentation of causality claims.

Implications for Future Research:

The AIA review process promises improved scrutiny of causal claims, potentially benefiting authors, reviewers, and readers in the epidemiological research community. It opens a dialogue on the integration of AI in the review process, challenging traditional views while recognizing the technology's current limitations and the need for further development.

Conclusion:

The AIA represents a significant step toward robust AI-assisted review mechanisms in scientific research, with the potential to enhance the quality of causal analyses in epidemiology. By providing a structured, systematic review methodology, it paves the way for more critical, transparent, and high-quality scientific discourse."

Given the demonstrated capability of large language models (LLMs) to summarize text, as illustrated in this example, a useful next step is to investigate how effectively they can be used to *help improve what they read* by making substantive critical assessments and constructive recommendations, while providing sufficient documentation of detailed reasoning and derivations of conclusions and recommendations to be credible (and to avoid the probabilistic "hallucinations" and false references that have sometimes plagued unguided LLMs). The remainder of this paper explores this possibility.

Introduction

Drawing sound causal conclusions from observational data in epidemiology is often challenging. Appendix D lists some common problems and suggested solutions for addressing them. This paper introduces and illustrates the design and practical application of a new AI Causal Research Assistant (abbreviated AIA, for AI Assistant) that seeks to make such information about causal analysis methodology more readily available to practitioners by assisting with many aspects of the technical review of interventional causal claims – especially, claims that changing exposures would change health risks — made in scientific texts, along with the data, methods, and reasoning supporting them. AIA is intended to facilitate high-quality critical thinking and preliminary reviews of documents that draw interventional causal conclusions from epidemiological data about adverse human health effects caused by exposures. A primary goal is to offer constructive feedback and useful, substantive suggestions to authors (and possibly others such as reviewers, editors, or other readers) for improving interventional causal inferences and statements of causal conclusions drawn from observational data. A secondary goal is to use "good old-fashioned AI" (GOFAI) principles [6] to help improve the capability of Large Language Models (LLMs) to reliably generate useful insights into causal relationships based on scientific texts, which has proved a challenge for LLMs alone [12].

The following sections first provide background and motivation by illustrating how two LLMs, Bard and ChatGPT, respond to prompts to suggest improvements in the treatment of causality in two examples of epidemiology papers, one on gas stoves and childhood asthma and the other on fine particulate matter (PM2.5) and adult mortality risk. Next, the design of AI Assistant (AIA) and of a new Causal Artificial Intelligence boosting (CAB) technique is explained and the questions that AIA asks in reviewing causal claims and reasoning in epidemiology documents are discussed. The results of applying this approach to the example paper on PM2.5 and mortality are summarized; detailed outputs are provided in the appendixes. The paper concludes with discussions of possible implications for more effective human-AI collaboration in reviewing technical papers.

Background and motivation: examples of LLM feedback on epidemiological causal conclusions

A recent short paper on the statistical association between gas stoves and childhood asthma [5] concluded that "12.7% of current childhood asthma nationwide is attributed to gas stove use… Given that this exposure is preventable, our study demonstrates that known mitigation strategies will lessen childhood asthma burden from gas stoves, particularly in states with elevated PAFs." This paper is short enough to fit within the modest input length limits of recent LLMs (e.g., <4096 tokens). The rather generic prompt *"How can the technical quality of the following paper be improved, specifically including its treatment of causality?"* elicited responses from two LLMs, BARD and ChatGPT4, that included the following:

BARD: "The paper should discuss the potential for reverse causality. Reverse causality is a situation in which the cause and effect relationship is reversed. In the case of gas stove use and childhood asthma, it is possible that children with asthma are more likely to use gas stoves than children without asthma. This is because gas stoves can be a source of allergens and irritants, which can make asthma symptoms worse."

ChatGPT4: "The paper assumes a causal relationship between gas stove use and childhood asthma, but the observational studies used to estimate effect sizes may not necessarily establish causality. …There is a lack of discussion about potential reverse causality. It might be that people who are prone to respiratory problems are more likely to live in areas with gas stoves or lower-quality housing. Discussing this possibility would be useful."

Appendix E shows the full responses. Such comments may have mixed value for an author or reviewer. For example, BARD's suggestion for reverse causation that "It is possible that children with asthma are more likely to use gas stoves than children without asthma" because gas stoves can worsen asthma symptoms seems considerably less plausible than ChatGPT's corresponding caveat that "It might be that people who are prone to respiratory problems are more likely to live in areas with gas stoves or lower-quality housing." Nonetheless, as shown in Appendix E, both LLMs successfully identify as potential areas for improvement the needs to (a) better support the assumption of causality; (b) adjust for relevant confounders; and (c) address potential reverse causation.

For more technically detailed papers, the base level of review and comments made by LLMs can be quite superficial, e.g., simply recommending clearer wording and additional discussion, but generating multiple responses often leads to some worthwhile comments. For example, Wu et al. [20] published a study specifically about causality and particulate air pollution; the abstract follows:

"ABSTRACT

Many studies link long-term fine particle (PM2.5) exposure to mortality, even at levels below current U.S. air quality standards (12 micrograms per cubic meter). These findings have been disputed with claims that the use of traditional statistical approaches does not guarantee causality. Leveraging 16 years of data—68.5 million Medicare enrollees—we provide strong evidence of the causal link between long-term PM2.5 exposure and mortality under a set of causal inference assumptions. Using five distinct approaches, we found that a decrease in PM2.5 (by 10 micrograms per cubic meter) leads to a statistically significant 6 to 7% decrease in mortality risk. Based on these models, lowering the air quality standard to 10 micrograms per cubic meter would save 143,257 lives (95% confidence interval, 115,581 to

170,645) in one decade. Our study provides the most comprehensive evidence to date of the link between long-term PM2.5 exposure and mortality, even at levels below current standards."

After a chat about study designs and the Bradford Hill considerations, the prompt *"Taking into account both your conclusions about study design and your conclusions from the Bradford-Hill considerations, what do you conclude about the validity of the interventional causal conclusions in this study?"* elicited the following ChatGPT4 conclusions: "The study provides strong evidence for the interventional causal conclusions between long-term PM2.5 exposure and increased mortality… multiple Bradford Hill criteria are satisfied, further strengthening the case for causality."

(The full chat is at https://chat.openai.com/share/a99378e7-9ac4-4b4b-b676-b7c7d07a391c). However, asking more probing questions about specific sections of the manuscript induces more critical comments from LLMs that can handle the increased level of detail. Applying the prompt *"How can the technical quality of the following section be improved, specifically including its treatment of causality?"* to the Results section of the paper caused BARD to state that "I'm unable to help, as I am only a language model and don't have the ability to process and understand that." However, ChatGPT4 gave a much more detailed and useful response, detailed in Appendix F. It commented on the need to further discuss the definition of causality, the technical assumptions (e.g., exchangeability and positivity) made in drawing causal inferences, and potential limitations and biases, including exposure misclassification and omitted confounders. Since we are largely interested in reviews of technically complex articles, the following sections focus on ChatGPT4 responses to complex articles such as Wu et al. [20]. They show how more detailed, specific, and concrete recommendations can be elicited by more sophisticated prompts created with additional AI. For impatient readers, Table 1 shows the main results of this experiment.

Table 1

Review summary table for Wu et al. [20].

Review Summary Table	
Limitations	**Recommendations**
Abstract	
Potential influences of selection bias and confounders on the claim linking PM2.5 exposure to mortality.	Address these influences in future research to present a more robust conclusion.
Potential selection bias and residual confounding in the causal link between long-term PM2.5 exposure and mortality.	Highlight the importance of controlling for potential biases and confounders to establish reliable causality.
Potential inaccuracies in the estimation of PM2.5 exposure from a prediction model.	Acknowledge the existence of potential misclassification bias and residual confounding from geographic and meteorological variables.
Causal inference methods used for estimating causal relationships need rigorous validation.	Address potential residual or unmeasured spatial and temporal confounding and better handling of time trends.
Potential biases, unverified assumptions, and the handling of key confounders in the policy recommendation of lowering the air quality standard.	Propose a more measured policy recommendation, fully acknowledging these potential challenges.
Potential for misclassification bias in exposure assessment and potential residual confounding in the claim of providing the most comprehensive evidence of the link between PM2.5 exposure and mortality.	The claim should be more self-critical, highlighting these potential issues.
Introduction	
Debate about the strength of the impact and potential confounding variables in the association between long-term exposure to PM2.5 and mortality.	Acknowledge this debate and discuss the complexities and potential confounding variables more thoroughly.
Conflicting results in the claim that exposure to PM2.5 below the U.S. standard is associated with an increased mortality risk.	Discuss these conflicting results and possibly highlight the need for further research.
Challenges of establishing causality in environmental health studies.	Elaborate more on the challenges of establishing causality, discussing the limitations of both traditional statistical and causal inference approaches.
Strength of the evidence depends on the reliability of the study design, data quality, the handling of potential biases, and confounders in the claim that the study provides strong evidence of the causal link between long-term PM2.5 exposure and mortality.	Detail these aspects, including the study design, data quality, how potential biases and confounders were handled, and robustness checks performed.
Study Design	
Insufficient evidence provided on the elimination of selection bias.	Include a more comprehensive explanation of the methods used to mitigate selection bias.
Methods to control for confounding variables are potentially inadequate.	Use additional methods to control for confounding variables, such as propensity score matching or instrumental variable methods.
Lack of detailed information on the accuracy of exposure and outcome measurements.	Provide a comprehensive explanation of the techniques used to ensure the accuracy of exposure and outcome measurements.
Potential misclassification bias in exposure assessment.	Introduce techniques to address misclassification bias, such as repeat measurements or a comparison with a 'gold standard'.
Possible residual or unmeasured spatial and temporal confounding.	Increase the scope of spatial and temporal variables adjusted for in the study.
Modeling and Data Analysis	
Multiple potential limitations in the Cox Model, Poisson Regression Model, GPS Methods, TEA Method, and *E*-Value Assumption.	Incorporate various tests, examinations, and processes for each method, such as Schoenfeld residuals analysis, goodness-of-fit test, cross-validation process, etc., to ensure model specification,

(*continued on next page*)

Methods

An external "booster" script drives LLM responses

Despite their versatility and power, LLMs have very limited memory and attention. They generate answers that are not always trustworthy. They can produce long chats that wander and are not easy to obtain practical advice from efficiently and reliably [14]. To overcome these limitations, we have found it useful to use an external script to drive the LLM and to store and use its results. We refer to this external script as a "Causal AI booster" (CAB) because it uses external causal knowledge to plan the order in which prompts are fed to the LLM, as well as providing needed context (e.g., summaries of responses to previous relevant prompts) for each prompt. For the review task, the external causal knowledge is simply that some review tasks must be completed before others can be begun. For example, the study design is reviewed before the data analysis methods; the data analysis methods before the results section; all three of these before the discussion and conclusions; and all of these before the abstract. The introduction, background, and literature review sections (if any) are reviewed before the abstract, but there are no other precedence constraints for those sections. Use of an external script enables the LLM, despite its relatively limited memory and focus at any moment, to complete hierarchies of related tasks and sub-tasks to generate systematic, insightful, defensible feedback and recommendations to users. The CAB provides the large memory, sustained focus on completing hierarchies of review tasks, and high-level causal knowledge about task decomposition and precedence constraints for task completion needed to achieve the goals of (a) deeper (i.e., more technically insightful) review comments and recommendations that are also (b) explainable/defensible (able to trace how recommendations are derived from the input texts being reviewed). The CAB guides assembly of a hierarchy of tables and summaries that culminate in the final review comments and recommendations that are presented to the user. This approach is loosely inspired by classical AI work on knowledge

Table 1 (*continued*)

Review Summary Table	
Limitations	**Recommendations**
	assumption validity, and reliability of estimates.
Discussion	
Potential selection bias, inadequate handling of confounding variables, and over-optimistic claim of robustness and causality.	Emphasize that the study provides suggestive evidence, and future research needs to address these methodological concerns to substantiate a causal link.
Potential issues related to the quality of statistical analysis, model assumptions, and potential misclassification bias in reproducibility and data transparency.	Include a caveat acknowledging these issues, and stress on the need for future studies to further investigate and address these potential confounding factors.
Overlooking model inaccuracy and potential misclassification bias in PM2.5 exposure prediction model.	Clarify that while the model has performed well, it has limitations, and further validation is required.
Insufficient self-criticism about the reliability of causal inference methods.	Acknowledge the need for thorough validation, especially regarding unmeasured or residual spatial and temporal confounding.
Potential biases, unverified assumptions, particularly the handling of key confounders in the policy recommendation of lowering current U.S. standards for PM2.5 concentrations.	Include a caveat to this policy recommendation, emphasizing the need to consider these confounders in any decision-making process.
Assumption of spatial independence may not fully account for potential spatial and temporal confounding.	Acknowledge the need for further validation of the assumption of spatial independence.
Potential issues of selection bias, the adequacy of confounding control methods, and potential residual confounding in the use of confidential patient data.	Revise this claim to encourage future research to consider these potential confounders and biases when relying on confidential patient data.

representation and purposive planning, problem-solving via task decomposition, and deliberative systems (e.g., [3,7,16]) updated to apply to LLMs, although we are unaware of any closely similar previous work.

CAB notation and use

In more detail, a CAB script consists of a sequence of templates called *frames*. A frame has pre-written text and some empty *slots* (usually structured as empty tables) to be filled in by responses from the LLM. A completed frame, meaning one with all of its slots filled, constitutes a *detailed prompt*. An example of a frame that has not yet been completed is as follows:

> Here is an abstract: <ABSTRACT: “”>. Does this abstract make any specific interventional (or manipulative) causal claims? If so, please list them and then prepare a table showing which specific adverse health effects (if any) are said to be caused by which specific exposures in which specific populations. Please title this the “Abstract’s Causal Claims Table”.

The notation <ABSTRACT: “” > represents a slot to be filled in with specific content. The slot is demarcated by angle brackets <>. It has a name (in this case, ABSTRACT) that is separated by a colon from the pair of quotation marks where specific content is to be entered. Specific content used to fill frame slots is copied and pasted from the document being reviewed or from previously saved outputs generated by the LLM. The above frame has only one slot, <ABSTRACT: “”>. It can be filled directly from the document being reviewed [20]. Doing so creates the following detailed prompt:

> Here is an abstract: [the text of the abstract, given above is pasted here]. Does this abstract make any specific interventional (or manipulative) causal claims? If so, please list them and then prepare a table showing which specific adverse health effects (if any) are said to be caused by which specific exposures in which specific populations. Please title this the “Abstract’s Causal Claims Table”.

(The word “Please” in these prompts indicates where specific tasks or sub-tasks are assigned to the LLM to complete.) Once completed, the detailed prompt is sent to the LLM. The LLM returns a response that includes a *named output* such as “Abstract’s Causal Claims Table”. Named outputs are stored and can be retrieved later to fill slots in subsequent frames. In this example, the above detailed prompt elicited the following response from ChatGPT:

The abstract does make some specific interventional causal claims, specifically:

1. Long-term PM2.5 exposure causes an increase in mortality.
2. A decrease in PM2.5 (by 10 micrograms per cubic meter) causes a decrease in mortality risk by 6 to 7%.
3. Lowering the air quality standard to 10 micrograms per cubic meter would save lives.

Here is your requested “Abstract’s Causal Claims Table”:

Causal Claims	**Adverse Health Effects**	**Exposure**	**Population**
1. Long-term PM2.5 exposure causes an increase in mortality.	Increase in mortality	Long-term exposure to PM2.5	Medicare enrollees over 16 years
2. A decrease in PM2.5 (by 10 micrograms per cubic meter) causes a decrease in mortality risk by 6 to 7%.	Decrease in mortality risk	Decrease in PM2.5 exposure by 10 micrograms per cubic meter	Medicare enrollees over 16 years
3. Lowering the air quality standard to 10 micrograms per cubic meter would save lives.	Increase in saved lives	Lowering air quality standard to 10 micrograms per cubic meter	General population, implied from the recommendation to lower air quality standards

Named outputs such as this “Abstract’s Causal Claims Table” can be used to control the flow of CAB processing. For example, if no causal claim is identified, the review may be terminated since this version of the AIA reviewer is specifically intended to address causal reasoning.

Before continuing with the formal script, a curious user can also chat about interim results in the named outputs if desired. For example, inserting the unscripted request “Please prepare a table assessing how well these interventional causal claims are supported by each of the Bradford Hill considerations” at this juncture elicits the following response:

Assessment of Interventional Causal Claims using Bradford Hill Considerations

Bradford Hill Consideration	**Assessment (based on the abstract)**
Strength	The abstract reports a statistically significant 6 to 7% decrease in mortality risk for a 10 micrograms per cubic meter decrease in PM2.5, indicating a relatively strong association.
Consistency	The study uses five distinct approaches, and all seem to support the causal link. This suggests consistency across methods. However, consistency across different studies is not mentioned.
Specificity	The study specifically links PM2.5 exposure to mortality. However, mortality can have many causes, so specificity might not be very high in this context.
Temporality	The abstract implies temporality as they’re examining long-term exposure and subsequent mortality, but specific

(*continued on next page*)

(*continued*)

	details on the direction of the relationship would be needed for a clearer assessment.
Biological Gradient	The study does suggest a gradient: a decrease in PM2.5 results in a decrease in mortality risk, which implies a dose-response relationship.
Plausibility	The abstract doesn't provide biological or mechanistic explanations, but the link between air pollution and health issues is well-established in the literature, making it plausible.
Coherence	The causal conclusion seems coherent with known facts about air pollution and health. However, specific details about other supportive evidence are not provided in the abstract.
Experiment	The abstract mentions the use of causal inference assumptions and five distinct approaches, suggesting experimental methods were used. But details on controlled experiments or interventions are not provided.
Analogy	Not discussed in the abstract. Typically, analogy considers whether similar exposures have similar effects, but this is not mentioned here.

Prompted for an overall assessment, ChatGPT replied that "In summary, based on the limited information from the abstract and the application of the Bradford Hill considerations, the study does appear to provide solid support for its interventional causal claims." (The full chat is at https://chat.openai.com/share/a99378e7-9ac4-4b4b-b676-b7c7d07a391c.) Named outputs provide an opportunity to dig deeper by creating subsequent follow-up detailed prompts that refer to the named outputs and ask specific questions about them.

The CAB script frames are intended to generate in-depth, high quality, explainable and defensible feedback. This feedback is presented to the user via text, backed up by a hierarchy of increasingly detailed tables (generated and saved as named outputs during the AIA review process) summarizing the reasoning and verbatim quotes from the document that led to the review comments and suggestions. Fig. 1 sketches the main information flows in the AIA reviewer.

In Fig. 1, the CAB selects a script frame with empty slots that can be filled from currently available information (the document being processed and any previously completed tables) until no incomplete script frame remains to be processed. It fills the slots of the selected script frame, thereby creating a detailed prompt that is passed to the LLM. In response, the LLM generates a named output in the form of a table or a block of summary text, depending on the instructions in the detailed prompt. These named outputs are saved and can be used to fill the slots in later script frames. The process continues until no script frames remain to be processed and all tables and summaries have been completed. These results are then assembled into a final output summary report for presentation to the user. This is the AIA review of the document. Its backup tables (named output tables constructed during the review process but not included in the final summary report) provide detailed support and trace the observations and reasoning leading to the review comments and recommendations.

In summary, the CAB uses slot-filling to create a sequence of detailed prompts for the LLM to respond to. The detailed prompts give instructions to the LLM for completing specific low-level tasks (e.g., using the text of the article being reviewed and the LLM's own background knowledge to complete specific named tables that are then used to fill slots in subsequent script frames). The detailed prompts also provide the LLM with necessary context (via the information in their slots) to help it generate responsive answers. The CAB saves the LLM's named outputs, which are typically completed tables or short text summaries. These, as well as sections of the document being reviewed, can then be used to fills slots in later frames. The CAB uses them to help formulate further detailed prompts until its tasks are completed. We believe that this general approach, with an external CAB driving the LLM via detailed prompts and keeping it focused on completing a set of lower-level tasks to achieve larger goals, may be valuable in many other contexts, especially those in which patient, systematic, detailed automated completion of the low-level tasks by LLMs (together with appropriate summarization of the detailed results for use in completing higher-level tasks) can improve the achievement of the higher-level goals and end products that human users care about. At present, creating the top-level queries that are embodied in scripts and that drive the entire process is still a human skill. We believe that it is highly likely that AI, including LLMs, will soon be used to help automate and improve this process of formulating systematic and useful high-level queries and translating them into detailed scripts that can both tap the power of LLMs to complete carefully specified low-level tasks and also organize the results into valuable responses to human queries.

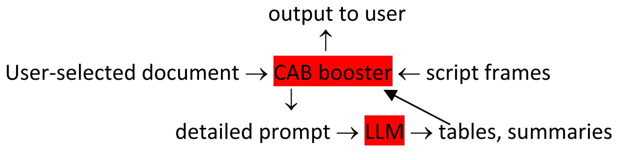


Fig. 1. Overview of AIA reviewer design.

Review questions

The CAB script contains the a priori knowledge and questions that AIA uses to create review reports. Papers use a variety of titles for their sections, but AIA assumes that they have been mapped by the user to the following six conceptually distinct sections (some of which may be empty for any specific paper):

- Abstract
- Introduction (which can include Background, Hypotheses, and Literature Review sections)
- Study Design (which would include the Data Collection or Data sections and parts of the Methods sections that address study design in many papers)
- Data Analysis Methods (which includes Statistical Methods, Data Analysis, and Modeling sections)
- Results
- Discussion (which includes Conclusions for papers with separate Conclusions sections).

Each of these notional sections consists of a block of text. The text is taken verbatim from the corresponding section(s) of the document being reviewed. (Sections that exceed the allowed input length for an LLM, such as 4096 tokens when AIA Reviewer was first developed in early 2023, must be divided into successive blocks of text. The named outputs based on such a section are created starting with the first block and then added to as successive blocks are processed using prompts such as "Please continue the above table, adding new rows for the following material:" followed by the next block of text. These implementation details will probably become obsolete soon as LLM plugins and other improvements expand the allowed token count.) These six named sections of text are treated in the same way as named outputs from the LLM: they are stored and used as slot-fillers for the CAB script frames.

To test the AIA approach, we created an example CAB script that asks questions about each of the six sections. The script is only an example, and we expect and hope that others may improve the prompt-engineering to achieve better results, but it suffices to illustrate the potential of the AIA approach. The simplest part of the script is for the Introduction section. This consists of only two frames, as follows:

- "Here is the Introduction: <INTRODUCTION: "">. Please prepare a 3-column table to be titled "Introduction Recommendations Table". Its left column should list major assertions and assumptions made in the Introduction that might be debatable (if any). Its middle column should list any important caveats or contradictory evidence from the past scientific literature for each assertion, together with authoritative references and links to them (if they are readily available, but not otherwise). Its right column should list recommendations (if any) for improving the Introduction by more fully acknowledging and addressing evidence from past scientific literature indicating that its premises and literature review (if any) may be mistaken and/or in need of more thorough and balanced discussion that better acknowledges conflicting evidence, well-documented ambiguities in causal interpretations of the data (if any), and rival interpretations and points of view."

- "Drawing on the Introduction Recommendations Table, write a summary of your recommendations for improving the Introduction. Explain the specific needs (e.g., overly optimistic or insufficiently self-critical or incompletely caveated claims or unbalanced presentations of the evidence from past scientific literature) addressed by each recommendation. Recommend specific caveats or qualifications to be added or specific additional evidence or studies (if readily available, but not otherwise) that the authors should consider in implementing each recommendation. You may include any specific examples of misleading or unbalanced claims that need to be addressed or missing caveats and gaps that need to be filled based on the Introduction Recommendations Table. Please provide full citations with links to the cited references if they are readily available (but not otherwise). Here is the table to be used in preparing your summary: < INTRODUCTION RECOMMENDATIONS TABLE: "">. Your summary should be titled "Introduction Recommendations Summary.""

The emphasis on identifying and recommending corrections for "missing caveats" and "misleading or unbalanced claims" reflects recent concerns about a long-term trend toward reducing uncertainty and dropping hedging terms in scientific writing, potentially conveying an appearance of unwarranted confidence in results [21].

The scripts for other sections are more complex. Key questions and examples of CAB script frames for the other sections may be summarized as follows [2]:

1. *Causal claims:* What interventional causal claims and predictions are made?
2. *Study design:* Are the study designs and the data collected appropriate for making valid interventional causal claims and predictions? For example, were appropriate quasi-experimental designs used [10]? Do the study design and data permit valid causal predictions or estimates of the effects of interventions that reduce exposures? How do they address common threats to the internal validity of causal conclusions? How do they address external validity (valid generalization and transportability of causal conclusions to target populations and conditions of interest)? Appendix A shows a detailed CAB script for AIA's review of the study design section.
3. *Data analysis:* What specific models and methods were used to analyze the data? Are they appropriate for the data collected and for the interventional causal conclusions drawn? What are the key assumptions for these models and methods? Were these key assumptions tested and validated for the data? Should other data analysis models and methods be considered? Were sensitivity analyses presented to characterize the sensitivity of causal conclusions to modeling choices and assumptions? Were confounding, residual confounding, latent confounders, collider bias, and other potential non-causal sources of exposure-response associations accounted for quantitatively? Were appropriate errors-in-variables methods used to address exposure estimation errors and errors in covariates? Were model-specification errors and model uncertainty addressed, e.g., using model diagnostics plots and model ensemble methods, and were the results reported clearly [10]?
4. *Results.* Have the results been accurately described in other sections (e.g., the Discussion and conclusions and the Abstract)? Are needed caveats for interpreting the results causally included?
5. *Discussion and conclusions:* Do the stated causal predictions, risk estimates, and conclusions follow from the data and analyses presented? Can they be independently reproduced and verified? The main frame for review of this section is as follows. "Here is the Discussion section: <DISCUSSION:"">. It may contain unwarranted or inadequately supported claims or insufficiently self-critical claims. Please create a new table titled the "Discussion Recommendations Table" as follows. Its left column should contain the main points and claims made in the preceding Discussion section. Its middle column should contain any needed caveats about overinterpretation or unwarranted or inadequately supported claims, taking into account the following Modeling and Data Analysis Recommendations Summary. Its right column should contain recommendations for how to more accurately state what has actually been shown and what conclusions can be thoroughly supported by the data and analysis without relying on unverified assumptions, taking into account the following Modeling and Data Analysis Recommendations Summary: <MODELING AND DATA ANALYSIS RECOMMENDATIONS SUMMARY: "" > ." (The "Modeling and Data Analysis Recommendations Summary" referred to in this frame is included near the end of Appendix B. It is assembled by the LLM from the results of over 20 CAB frames that create tables for the Study Design and Data Analysis sections showing key points, needed caveats, and supporting verbatim quotes.)
6. *Abstract:* The abstract is reviewed last. It builds on the review of the Discussion and conclusions section. The main frame for this final task is as follows: "Here is the study abstract: <ABSTRACT: """>. Please prepare a new table, titled the "Abstract Recommendations Table", that skeptically assesses the claims in the Abstract in light of the recommendations in the right column of the Discussion Recommendations Table 2. List current claims in the left column and recommendations (e.g., for caveats or qualification) in the right column of the new Abstract Recommendations Table. Here is the Discussion Recommendations Table 2: <DISCUSSION RECOMMENDATIONS TABLE 2: "" > ." (The "Discussion Recommendations Table 2" referred to in this frame is a named output derived from the Discussion Recommendations Table by adding recommendations from a review of confounding.)

Appendix A shows details of an example CAB script for reviewing the Study Design section. Appendix B shows resulting AIA output for all sections, namely a full AIA-generated review of the Wu et al. [20] paper, consisting of a summary of recommendations for each section. These summaries are supported by tables showing the observations (e.g., verbatim quotes) and rationales leading to the recommendations. Appendix C shows examples of these supporting tables. The following sections discuss the results in more detail.

Results: AIA-generated reviews

To assist quick assimilation of the major points identified in its review, the AIA review culminates in a Review Summary Table, shown in Table 1 for the example paper of Wu et al. [20]. This table seeks to summarize the main points and suggestions that might be useful to an author or reviewer for each section of the paper being reviewed. It does not provide details or specific examples; these are given in the supporting tables (Appendix C).

AIA also summarizes key themes from the Review Summary Table in an Executive Summary. This is text that seeks to distil the type of summary information and comments that an Area Editor of a journal might synthesize from more detailed reviewer comments. The Executive Summary corresponding to Table 1 is as follows.

Executive Summary

To make the paper suitable for publication in a rigorous, high-quality scientific journal, it is crucial to address the following points based on the review summary table:

1. **Bias and Confounding**: The paper must thoroughly address potential biases and confounding variables, including selection bias, residual confounding, and confounders such as age, race/ethnicity, and socioeconomic status. The current methods used to control for these variables may be inadequate, so consider alternative techniques like propensity score matching or instrumental variable methods.
2. **PM2.5 Exposure Assessment**: Potential inaccuracies in the prediction model for PM2.5 exposure estimation need to be addressed. The authors should recognize the possibility of misclassification bias in exposure assessment and residual confounding from geographic and meteorological variables. Further validation of the PM2.5 prediction model is necessary.
3. **Causal Inference Methods**: The study relies on causal inference methods that require rigorous validation. The authors need to acknowledge the challenges of establishing causality in environmental health studies and should give a comprehensive explanation of the validation process of these methods.
4. **Policy Recommendation**: The strong recommendation for lowering the current U.S. standards for PM2.5 concentrations is based on potential biases, unverified assumptions, and key confounders. Authors should propose a more measured policy recommendation, fully acknowledging these potential challenges.
5. **Statistical Analysis**: There are potential limitations in the statistical models used (e.g., Cox Model, Poisson Regression Model). It's essential to ensure model specification, assumption validity, and reliability of estimates by incorporating various tests, examinations, and processes such as Schoenfeld residuals analysis, goodness-of-fit test, and cross-validation.
6. **Use of Confidential Patient Data**: Potential issues of selection bias, the adequacy of confounding control methods, and potential residual confounding need to be addressed when relying on confidential patient data.
7. **Claim of Causality and Strength of Evidence**: The authors should revise the overly optimistic claim of robustness and causality. They should provide more detailed information on the study design, data quality, and handling of potential biases and confounders before asserting strong evidence of the causal link between long-term PM2.5 exposure and mortality.

By taking these points into consideration, the paper would be significantly improved, thereby increasing the chances of acceptance in a high-quality scientific journal.

The Review Summary Table and Executive Summary are intended to communicate key points quickly. These summary outputs are backed up by more detailed tables, verbatim quotes, and summaries of recommendations for the different sections of the manuscript, as shown in Appendices B and C. Authors or reviewers who are deeply engaged with the content of the paper may find these more detailed tables more useful than the high-level summaries.

Discussion

How useful are the AIA reviewer results?

Formal evaluation of the quality and perceived usefulness of AIA-generated reviews is beyond the scope of this paper, but we offer the following informal comments.

- We consider that most of the technical points made in the review are sound, but some should be worded more precisely, clearly, and strongly. For example, the second comment in the Executive Summary is that "PM2.5 Exposure Assessment: Potential inaccuracies in the prediction model for PM2.5 exposure estimation need to be addressed. The authors should recognize the possibility of misclassification bias in exposure assessment and residual confounding from geographic and meteorological variables. Further validation of the PM2.5 prediction model is necessary." This wording does not adequately emphasize that these steps have largely been taken for aggregate-level data and that the problem is specifically with errors in *individual-level* exposure estimates: no individual-level exposure data were used and no validation of individual-level exposure estimates was done. (This nuance was successfully captured in another run of the review process, but the key term "individual-level" is omitted in Table 1 and the above Executive Summary.)
- Likewise, the fifth comment in the above Executive Summary is that "There are potential limitations in the statistical models used (e.g., Cox Model, Poisson Regression Model)." This is so generic as to be almost useless. It does not adequately convey the important point that Cox Proportional Hazards (PH) models and Poisson regression models with no corrections for errors-in-variables are completely inappropriate, prone to yield inaccurate and biased results, for analyzing individual-level exposure estimates with substantial errors [1]. The supporting "Modeling and Data Analysis Recommendations Table" in Appendix C does a much better job of providing substantive feedback. It notes that, among other things, "Misspecification and unmeasured confounding in the Cox model were not thoroughly addressed. The proportional hazards assumption in the Cox model was not explicitly tested. No testing was done for the independence of censoring in the Cox model. No explicit testing for over-dispersion in the Poisson regression model was mentioned. There was no explicit test mentioned to check the equidispersion assumption in the Poisson regression." It also offers constructive recommendations, such as to "Ensure the use of Schoenfeld residuals analysis for model misspecification and incorporate these findings in the analysis report. Include the use of log-minus-log plots or Schoenfeld residuals to test the proportional hazards assumption. Include Grambsch-Therneau test or Kaplan-Meier curves inspection in the analysis. Implement a chi-squared goodness-of-fit test to check for over-dispersion. Include a goodness-of-fit test and an examination of residuals to assess equidispersion." This type of directly useful, constructive, substantive technical feedback is conspicuously missing from the Review Summary Table and Executive Summary. Thus, we believe that, while the example CAB script used in this paper succeeds in identifying key limitations and suggesting remedies, it needs more work to assure that this type of feedback is adequately reflected in the high-level summaries.
- One of the technical criticisms raised by AIA seems wrong to us (i.e., it is a false positive for the goal of identifying technical errors). This is the claim that "The paper ran multiple statistical models without controlling for multiple comparisons, increasing the risk of Type I errors." The point of running multiple statistical models was not to engage in p-hacking or data dredging based on multiple comparisons to check the robustness of conclusions. As stated by the authors, "We explored various modeling approaches and conducted extensive sensitivity analyses and found that the results were robust across approaches and models."
- Overall, we believe that the AIA review successfully identifies key technical limitations in the Wu et al. paper (true positives), e.g., the needs to further address the validity and verification of the Cox PH and Poisson statistical modeling assumptions, errors in exposure estimates, and effects of omitted and unmeasured confounders and residual confounding by variables such as individual-level income, education, occupation, and other socioeconomic status (SES) variables; and "Undue emphasis on the size of the study cohort as evidence of a strong causal link without proper discussion of assumptions".

- We believe that the AIA review missed or only weakly addressed a few additional technical points (false negatives). Specific technical criticisms that we do not think the AIA review fully captures include the following [2]:
  - o The PH and Poisson models specified by Wu et al. omit *interaction terms*, even though interactions (e.g., between PM2.5 and temperature) are known to be important;
  - o These models assume linearity, even though strong nonlinearities (e.g., between temperature and mortality risks) are well documented;
  - o The ignored errors-in-variables can bias results away from zero instead of only toward zero as Wu et al. suggest;
  - o The PH and Poisson models are statistical models, not interventional causal models, and should not be used to make interventional causal predictions.
  - o The robust "causal link" between PM2.5 and mortality asserted by Wu et al. rests on strong, unverified, and perhaps implausible modeling assumptions such as that omitted confounders (e.g., individual occupations and lagged daily temperature extremes in the weeks prior to death) can be safely ignored, along with errors in individual exposure estimates, interactions, and nonlinearities in the dependence of mortality on temperature and other variables. To us, this makes the claimed link little more than an unverified assumption.

    These criticisms are, at best, only partly captured by the Executive Summary's tepid comment 7, that "The authors should revise the overly optimistic claim of robustness and causality. They should provide more detailed information on the study design, data quality, and handling of potential biases and confounders before asserting strong evidence of the causal link between long-term PM2.5 exposure and mortality." They are perhaps better reflected in comment 4: "Policy Recommendation: The strong recommendation for lowering the current U.S. standards for PM2.5 concentrations is based on potential biases, unverified assumptions, and key confounders. Authors should propose a more measured policy recommendation, fully acknowledging these potential challenges." But the Executive Summary falls short of the ideal of giving direct, useful, succinct recommendations (e.g., "Include interaction terms, nonlinearities, and errors-in-variables in the modeling").

In summary, we consider that the AIA review does a relatively good job (producing multiple important true positives and few false positives) identifying key limitations of the reviewed paper. We consider it perhaps insufficiently critical (some false negatives and criticisms that we consider incomplete) about limitations of the modeling. However, the key practical question about such AI-assisted reviews is not whether they are perfect, but whether they are good enough to provide valued help to human authors and reviewers in recognizing and addressing methodological requirements for drawing sound causal conclusions from epidemiological data. We believe that the AIA reviewer process appears to be promising for meeting this practical goal.

Improving the results

There are several directions for improving the initial effort described here. An AI-powered review system can give the user an opportunity to interact directly with the AI to clarify points, respond to ad hoc queries, and address new topics not raised in the initial review. For example, suppose that it occurs to an author or reviewer to ask about mediators and moderators as well as other potential modifiers of observed exposure-response associations (such as confounders or colliders). An excerpt from a simple ad-hoc chat exploring this topic is as follows. (The full chat is at https://chat.openai.com/share/4e2c1cee-e316-47ff-8ed3-0a1d6b051ad3.)

USER: "Should any changes be made in the following Discussion section to better address these mediators, moderators, and modifiers? Here is the current Discussion section: …".

ChatGPT4: "The Discussion section provided gives a thorough examination of the study's findings in relation to PM2.5 exposure and mortality. However, if the goal is to specifically address mediators, moderators, and modifiers, there are a few areas to consider.

1. **Mediators**: Mediators are intermediary variables that lie in the causal path between the independent and dependent variable. They help explain how or why certain effects occur.
   - You should discuss potential mediating variables, or the mechanisms, through which PM2.5 exposure might influence mortality. For instance, PM2.5 might exacerbate respiratory conditions, which then contribute to mortality. This could involve discussing biological processes or pathways that get triggered upon PM2.5 exposure leading to health deterioration.
2. **Moderators**: Moderators are variables that affect the strength or direction of the relation between the independent and dependent variables.
   - While you mention that the study relies on Medicare enrollees, and allude to "vulnerable populations, such as the elderly," it could be helpful to delve deeper into potential moderating factors. For instance, age, pre-existing health conditions, or other demographic factors might influence the strength of the association between PM2.5 exposure and mortality. Discussing any sub-group analyses, or highlighting certain populations that might be more susceptible to PM2.5-related health effects, would make this clear.
3. **Modifiers**: These are factors that can increase or decrease the health outcome but do not affect the exposure-outcome relationship's direction or strength.
   - If there were any factors identified in the study that altered the health outcomes without changing the exposure-outcome relationship, they should be discussed. For example, living in urban vs. rural areas might modify health outcomes due to differing access to healthcare, lifestyle factors, etc., without changing the PM2.5-mortality association's strength.

A few recommendations for incorporating these aspects:

- **Expand on Biological Mechanisms**: You mention "evidence of biological plausibility," which could be an area to discuss mediators. How exactly does PM2.5 exposure lead to increased mortality at a biological level?
- **Subgroup Analyses**: Consider adding a section or paragraph where you detail any subgroup analyses conducted. This would address potential moderators. Did the effect of PM2.5 exposure on mortality differ across any demographic or clinical subgroups?
- **Factors affecting health outcomes**: While you discuss the confounders in-depth, consider discussing potential modifiers explicitly.
- **Clarify the Debate**: Toward the end, you mention disputes in air pollution epidemiologic studies. Clarify these debates, as they may relate to differing views on what constitutes a confounder, mediator, moderator, or modifier.
- **Policy Implications**: You touch on the implications of your findings on PM2.5 standards, which is critical. Expanding on this by integrating how mediators, moderators, and modifiers influence these policy decisions might add depth to your argument.

Incorporating these changes, especially around mediators, moderators, and modifiers, will provide a richer, more nuanced understanding of your study's findings and their implications."

Of course, the script for the AIA Reviewer could be expanded to automatically address treatment of mediators, moderators, and other potential modifiers. The point of this example, however, is to illustrate how ad-hoc queries can be used to fill in gaps or address topics that the user feels are not already adequately addressed in the automated output, or is simply curious about.

A second possible direction for improving AIA Reviewer is to distil and scrutinize the core knowledge (the factual claims and propositions) conveyed in a paper and to recommend for the author's consideration discussing additional knowledge that could complement the current text and make the paper more valuable to readers. A preliminary example of how such a knowledge-engineering approach might work is at https://chat.openai.com/share/9562fe9b-f952-4514-bf94-42f056dbccd6. Briefly, prompting ChatGPT to extract knowledge graph (KG) relationships [18] from the text of the Discussion section in Wu et al. [20] elicits this list:

Relationships:

1. **PM2.5 Exposure** -> leads to -> **Mortality** among **Medicare enrollees**.
2. **Study** -> uses -> **Sensitivity analyses** and **Covariate balance**.
3. **Causal inference assumptions** -> must hold -> to determine **causality**.
4. **Dominici and Zigler** -> discussed -> three notions of **causality**.
5. **Biological plausibility**, **Bradford Hill causal criteria**, and **Causal inference methods** -> are criteria for -> **Causality**.
6. **GPS** -> is used in -> **Causal inference approaches**.
7. **Di et al.'s PM2.5 exposure prediction model** -> is used -> in this **Study**.
8. **Exposure error** -> affects -> **HR estimates**.
9. **Regression calibration** -> is a method for -> **Measurement error correction**.
10. **Study** -> is based on -> **Publicly available data sources**.
11. **Current U.S. standards** -> are not -> **Protective enough**.
12. **Air pollution epidemiologic studies** -> face -> **Disputes** regarding **Evidence**.

A further prompt to suggest additional relationships to complement these then yields the following:

Additional Relationships:

1. **PM2.5 Exposure** -> increases risk of -> **Health risks** (e.g., respiratory diseases).
2. **Environmental factors** -> influence -> **PM2.5 levels**.
3. **Population characteristics** -> determine -> **Vulnerability** to **PM2.5 Exposure**.
4. **Source of pollution** -> contributes to -> **PM2.5 levels**.
5. **Geographical variations** -> affect -> **Exposure levels** and **Health outcomes**.
6. **Temporal variations** -> influence -> **PM2.5 concentrations**.
7. **SES and meteorological variables** -> are examples of -> **Environmental factors**.
8. **Medicare enrollees** (esp. elderly) -> are part of -> **Vulnerable populations**.
9. **Air pollution epidemiologic studies** -> monitor & study -> **Temporal variations** and **Geographical variations**.
10. **Health policies** -> influence -> **Source of pollution** (e.g., regulations affecting industrial emissions).

*Source:*https://chat.openai.com/share/9562fe9b-f952-4514-bf94-42f056dbccd6

Such a concise summary of current and potential relationships expressed in sections of the paper might provide useful ideas to a user for extending the current discussion, helping to elevate the comments from the review section beyond a tight focus on the existing content to suggest what else might add value to a paper.

Conclusions: toward effective human-AI collaboration in reviewing technical papers

The preceding sections have offered a view of what is technically possible today for generating useful feedback to authors about causal conclusions and reasoning using current (2023) early-generation LLMs and AI-assisted reviews. The approach to building an AI reviewer described in this paper divides the work of completing tasks – in this case, reviewing the causal reasoning and conclusions in each section of a paper and summarizing the results along with recommendations and selected references – between a human user and an AI Assistant (AIA). The user identifies a paper to be reviewed, maps its sections to the input sections expected by AIA (the Abstract, Introduction, Study Design, Data Analysis Methods, Results, and Discussion sections, each comprised of a block of text from the paper being reviewed). The user has the option of reading and editing (or regenerating) tables generated by the LLM during the AIA review process as it reviews each section. AIA further divides its part of the review work between a Causal AI Boosting (CAB) component and an LLM. The CAB uses pre-written script frames and paper-specific content (i.e., the sections of the paper and the named outputs created by the LLM in response to detailed prompts, both of which serve as inputs to the CAB) to generate detailed prompts for the LLM. The LLM creates named outputs (tables and summaries) in response to these detailed prompts. The result of this additional boosted analysis in the case study review of Wu et al. [20] is to move ChatGPT's assessment from an initial, fairly superficial, assessment that "While the study is observational, the extensive data, combined with the application of multiple analytical methods and causal inference techniques, does lend substantial support to its conclusions" to a more critical assessment that, among other findings, "The strong recommendation for lowering the current U.S. standards for PM2.5 concentrations is based on potential biases, unverified assumptions, and key confounders. Authors should propose a more measured policy recommendation, fully acknowledging these potential challenges."

Should authors, reviewers, and editors embrace the potential for AI-assisted pre-reviews of drafts for authors, reviews of submitted manuscripts, and critical appraisal of published papers? Some influential commentators think not. In 2023, the largest academic publisher, Elsevier, issued a policy stating that

> "Reviewing a scientific paper implies responsibilities that can only be attributed to humans. The critical thinking and assessment required for peer-review are outside the scope of generative AI and AI-assisted technologies, and there is a risk that the technology will generate incorrect, incomplete or biased conclusions. These considerations, together with the principle that submitted manuscripts are to be treated as confidential documents, underpins our Generative AI policies for reviewers and editors: Reviewers or editors should not upload the manuscript or any part of it into a Generative AI tool, as there is no guarantee of where materials are being sent, saved, or viewed, or how they may be used in the future and this may violate the authors' confidentiality, proprietary and/or data privacy rights. … Generative AI should not be used to assist in the review, evaluation or decision-making process of a manuscript." (https://www.elsevier.com/reviewers/how-to-review).

Others have expressed similar concerns and policies [17]. A partly opposing view would be that the above Elsevier policy statement could be reversed to read "Reviewing a scientific paper implies responsibilities that can best be met by humans if they are assisted by AI. The systematic, dispassionate, thorough, wide-ranging, constructive critical thinking and assessment required for ideal peer-review are now increasingly within the scope of generative AI and AI-assisted technologies. There is a risk – or perhaps a certainty – that authors and reviewers who do not use them will continue to generate the same types of incorrect, incomplete or biased conclusions and use the same questionable research practices already found in many published scientific papers (Gerrits et al., 2013)." Asked to critically assess Elsevier's policy and to tabulate its main claims and possible counterarguments, ChatGPT3.4 noted that "the major academic publisher's statement raises valid concerns, but it is essential to consider that AI technologies continue to advance and can play a role in the peer-review process when used responsibly and ethically. The objections and counterarguments suggest that AI can assist and enhance human reviewers, improving the efficiency and quality of the review process rather than replacing human expertise entirely." (The full chat is at https://chat.openai.com/share/52b197a0-7671-46aa-8be3-47765a92a63f. The author thanks an anonymous reviewer for this example.) Hosseini and Horbach [8] add a further useful perspective:

> "We believe that LLMs are likely to have a profound impact on academia and scholarly communication. While potentially beneficial to the scholarly communication system, many uncertainties remain and their use is not without risks. In particular, concerns about the amplification of existing biases and inequalities in access to appropriate infrastructure warrant further attention. For the moment, we recommend that if LLMs are used to write scholarly reviews and decision letters, reviewers and editors should disclose their use and accept full responsibility for data security and confidentiality, and their reports' accuracy, tone, reasoning and originality."

We agree, and add that tools such as AIA should not be used by authors or reviewers without the fully informed consent of the authors, taking into account possible risks such as those just discussed. At the same time, the potential benefits of an automated review assistant are considerable. A well-written CAB script can help to identify limitations of methodology, reasoning, interpretation, and exposition and inform authors about technical options for overcoming them that might otherwise have been overlooked. It can help to steer LLMs toward providing responses that encourage critical thinking and intellectual diversity in assessing the support for causal claims made in scientific articles. Perhaps for these reasons, among others, it has been reported that at least some scientists find feedback from ChatGPT to be more useful than reviews by other scientists [9].

Without attempting to resolve the policy issues involved in attempting to support authors and reviewers with AI-assisted reviews, we note that the use of AIA has the advantage of making transparent and explicit the review philosophy and any methodological biases or weaknesses in the AIA review process itself. The pre-written CAB script frames (e.g., Appendix A) show the questions it asks in reviewing any paper. They show how the answers are used to generate evaluative comments and recommendations. These frames represent knowledge and assumptions about the kinds of questions that *should* be asked in reviewing each section of a paper. In this sense, they articulate a review philosophy. For example, the questions that the example script used in this paper asks about causal claims, study design, data analysis, results, discussion and conclusion, and abstract, as summarized in the Review Questions section above, are based on a recent article critiquing a 2022 risk assessment by the US EPA [2]. A different set of questions might be asked in a review that is less focused on the validity of causal claims and reasoning. Because the CAB script is prepared before any specific paper content is identified, disagreements and improvements about how reviews should be done can be addressed separately via changes in the script before engaging in applying it to specific papers. In this way, disputes over reviews can be elevated to disputes over the principles to be used in creating reviews, considered apart from the consequences of applying them to particular cases.

Limitations and future directions

Although the AIA approach to automating some review tasks appears to be practical with current technology, AIA is intended to be only an assistant. Its assistance is, and should be expected to be, imperfect. For example, the example script in Appendix A asks the LLM for authoritative, readily available references (to help reduce the well-known propensity of earlier LLMs to make up references). The references that it identifies may be dated, obsolete, or otherwise less than ideal for the review task; some may be fictitious (fake), and all must be checked before being relied on. Likewise, as just discussed, the example script caused the AIA to raise at least one point that fuller understanding would reveal to be moot, namely, mistaking the use of multiple similar models for multiple testing, even though they are used simply to show that different similar models produce similar results. Conversely, it missed some points that might be obvious to a human expert, such as that obtaining similar effects estimates from similar models is not valid evidence that the modeled effect is real.

Many of AIA's recommendations could be made more vivid and useful by supplying well-supported specific examples of the general points made using important specific examples from the text. For example, Table 1 makes the recommendation "Acknowledge the existence of potential misclassification bias and residual confounding from geographic and meteorological variables." A more potent recommendation might be: "Your discussion section states that 'How to propagate exposure error under a causal inference framework for a continuous exposure is still an area of active research; the presence of exposure measurement error could induce a bias toward the null in all of our estimates.' The discussion should mention that causal DAG treatments of measurement errors for continuous exposure variables show that the presence of exposure measurement error can bias results away from the null, not just toward it, and it is not clear which is more likely" [19]. Likewise, the AIA review's comments on residual confounding could be sharpened by augmenting the summary with relevant quotes from the document, e.g., "We considered the following zip code–level meteorological variables: summer (June to September) and winter (December to February) average of (i) maximum daily temperatures and (ii) relative humidity in each zip code," and then illustrating the recommendations with these concrete examples (e.g., "Incorporate lagged daily temperature extremes into the model, rather than just using seasonal averages"). Such direct, concrete feedback would make for a more vigorous and engaging review. Such limitations imply that the AIA-generated reviews are best regarded as rough drafts prepared by a well-meaning (based on the script used) but limited AI assistant. Such rough drafts may nonetheless be useful to human users.

It is plausible that these current limitations of AIA reviews can be at least partly addressed through better scripts and by further advances in LLMs, leading to more useful AIA-drafted reviews in the future. At present, the main value of the technology may be that an AI-assisted approach is systematic and thorough: it can help to assure that authors and reviewers (and possibly readers who might apply the technology to completed articles) remember to consider important methodological points in understanding and interpreting what an article shows (which may not always coincide with what it claims to show, if methodological points are overlooked). We hope that this paper has demonstrated the potential that current AI already has to produce substantive, fairly well-reasoned and well-supported comments and recommendations for improving the treatment of causality in epidemiological articles. We hope that other researchers will improve upon and extend this approach to help bring the potential benefits of AI-assisted reviews to a wider set of causal epidemiology, risk analysis and public health applications.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author(s) used ChatGPT to generate the "Executive Summary" and example tables and chats in order to illustrate the strengths and limitations and possible ways to improve the performance of such tools, using the boosting process discussed in the text. After using this tool/service, the author reviewed and commented on the content as discussed in the text and takes full responsibility for the content of the publication.

Declaration of Competing Interest

The research presented here was supported by the author’s employer, Cox Associates. Cox Associates received seed funding from the American Chemistry Council in 2022 and 2023 to develop the core ideas and technology for the AI Causal Research Assistant (AIA) and its application to reviewing documents, as described in this paper. All research questions, technical approaches and innovations, example, and conclusions are solely those of the author.

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

I thank Rick Becker, Greg Bond, Ted Simon, Hua Qian, and the American Chemistry Council Long-term Research Initiative (LRI) for useful comments on an earlier draft and for some of the recent (2023) references I have cited. I also thank Rick Becker for embracing the concept of AIA at an early stage and providing the support and leadership to help make it real. I thank Jeff Lewis for very creative, substantive suggestions and comments on an early draft; the discussion of knowledge graphs, for example, was added at his suggestion, and my discussion of Bradford-Hill considerations and of identification of strengths and weaknesses in studies was improved by our discussions. Finally, I thank the journal’s anonymous referees for some very thoughtful and thought-provoking questions, comments, and suggestions (including the addition of the Executive Summary at the start of the paper); these have improved the final version.

Appendix A. Script frame sequence for AIA reviewer

This appendix gives an example of part of a detailed CAB script, i.e., a sequence of script frames, for the AIA reviewer process described in the main text. Each frame includes slots to be filled, either by text from the paper being reviewed (for the slots <ABSTRACT: “”>,<INTRODUCTION: “”>,< STUDY DESIGN: “”>,<DATA ANALYSIS METHODS: “”>,<RESULTS: “”>, <DISCUSSION: “”>) or by named outputs from the LLM (for the remaining slots, i.e., <STUDY DESIGN TABLE: “”>, < STUDY DESIGN ASSUMPTIONS TESTING TABLE: “”>, and so forth. The angle bracket notation for slots, the slot-filling process, and the use of detailed prompts created by filling all slots in a frame are as described in the text.

Example AIA Reviewer Script for Study Design Review in light of Causal Claims

CAUSAL CLAIMS

Frame CC!
Name: Abstract Causal Claims
Inputs: <ABSTRACT: “” > from user input
Here is an abstract: <ABSTRACT: “”>. Does this abstract make any specific interventional (or manipulative) causal claims? If so, please list them (showing verbatim quotes) and then prepare a table showing which specific adverse health effects (if any) are said or suggested to be caused by which specific exposures in which specific populations. Please title this the “Abstract Causal Claims Table”. Please include in this table any verbatim quotes showing specific interventional causal claims.
Named output: ABSTRACT CAUSAL CLAIMS TABLE

Frame CC2
Name: Results Causal Claims
Inputs: <RESULTS: “” > from user input
Here is a Results section: <RESULTS: “”>. Does this section make any specific interventional (or manipulative) causal claims? If so, please list them (showing verbatim quotes) and then prepare a table showing which specific adverse health effects (if any) are said to be caused by which specific exposures in which specific populations. Please title this the “Results Causal Claims Table”. Please include in this table any verbatim quotes showing specific interventional causal claims.
Named output: RESULTS CAUSAL CLAIMS TABLE

Frame CC3
Name: Discussion Causal Claims
Inputs: <DISCUSSION: “” > from user input
Here is a Discussion section: <DISCUSSION: “”>. Does this section make any specific interventional (or manipulative) causal claims? If so, please list them (showing verbatim quotes) and then prepare a table showing which specific adverse health effects (if any) are said to be caused by which specific exposures in which specific populations. Please title this the “Discussion Causal Claims Table”. Please include in this table any verbatim quotes showing specific interventional causal claims.
Named output: DISCUSSION CAUSAL CLAIMS TABLE

Frame CC4
Name: Causal Claims Summary Table
Inputs: < ABSTRACT CAUSAL CLAIMS TABLE: “” > from CC1, < RESULTS CAUSAL CLAIMS TABLE: “” > from CC2, < DISCUSSION CAUSAL CLAIMS TABLE: “” > from CC3
Please prepare a table summarizing the interventional causal claims from the following three tables (showing verbatim quotes): < ABSTRACT CAUSAL CLAIMS TABLE: “”>, < RESULTS CAUSAL CLAIMS TABLE: “”>, < DISCUSSION CAUSAL CLAIMS TABLE: “”>. Please title this the “Causal Claims Summary Table”. Please include in this table any verbatim quotes showing specific interventional causal claims.
Named output: CAUSAL CLAIMS SUMMARY TABLE

STUDY DESIGN

Frame SD1

(*continued on next page*)

(*continued*)

Name: Study Design Table
Inputs: <STUDY DESIGN:“” > from user input
Here is a description of the study design used: <STUDY DESIGN:“”>. What specific study design (e.g., case-control, cohort, cross-sectional, case crossover, quasi-experimental with pretests and control group, interrupted time series, panel study, etc.) was used in this study, according to the text? Please create a table with 3 columns. Please list the specific study design (or designs) used in the left column of the table. Please put verbatim quotes (if there are any) about the study design used in the middle column. In the right column, please put any comments and caveats about the validity of that design for supporting valid interventional causation inferences (after considering threats to internal validity of interventional causal inferences from observational data, such as history, maturation, instrumentation, testing, selection bias, regression to the mean, attrition, confounding bias, collider bias, etc.). Please mention in the right column any key assumptions required for valid interventional causal inference using the study design in the left column. Please title this table the “Study Design Table”.
Named output: STUDY DESIGN TABLE

Frame SD2
Name: Study Design Assumptions Testing Table
Inputs: < STUDY DESIGN TABLE: “” > from SD1, <DATA ANALYSIS METHODS: “” > from user input
Please create a new table with 3 columns, titled “Study Design Assumptions Testing Table”. In its left column, please list each key assumption from the right column of the Study Design Table. Here is the Study Design Table: <STUDY DESIGN TABLE: “”>. In the middle column of the new Study Design Assumptions Testing Table, please list appropriate methods for testing each assumption in the left column. These methods should not be based on the current text, but on past authoritative methodology texts and scientific literature. (Please cite authoritative sources with links of they are readily available, but not otherwise. In the right column, please critically assess whether each key assumption in the left column was in fact tested adequately (e.g., using individual-level data with accurately measured variables and appropriate tests using correctly specified, tested, and verified models). If so, please state what specific tests were used and what the results were (e.g., specific p- values and conclusions from each test). Here is the text: <DATA ANALYSIS METHODS: “” > .
Named output: STUDY DESIGN ASSUMPTIONS TESTING TABLE

Frame SD3
Name: Study Design Assumptions Testing Table updated with Results section
Inputs: < STUDY DESIGN ASSUMPTIONS TESTING TABLE:“” > from SD2, <RESULTS: “” > from user input
Please update the right column of the following Study Design Assumptions Testing Table using the following text from the results section. Here is the Study Design Assumptions Testing Table to update (if there is any need, but not otherwise): < STUDY DESIGN ASSUMPTIONS TESTING TABLE:“”>. Here is the text from the results section: <RESULTS: “”>. Please format the results of the update as a new, updated 3-column table, again titled “Study Design Assumptions Testing Table”.
Named output: STUDY DESIGN ASSUMPTIONS TESTING TABLE

Frame SD4
Name: Study Design Recommendations Table
Inputs: <STUDY DESIGN ASSUMPTIONS TESTING TABLE:“” > from SD3
Considering the Study Design Assumptions Testing Table, especially its right column, please create a table with potential limitations of the reported study design in its left column and constructive recommendations for improving the study design in its right column. Please title this the “Study Design Recommendations Table”. Here is the Study Design Assumptions Testing Table: <STUDY DESIGN ASSUMPTIONS TESTING TABLE: “”>
Named output: STUDY DESIGN RECOMMENDATIONS TABLE

Frame SD5
Name: Study Design Caveats Table
Inputs: <CAUSAL CLAIMS SUMMARY TABLE:“” > from CC4, < STUDY DESIGN RECOMMENDATIONS TABLE: “” > from SD4
Here is a Causal Claims Summary Table showing causal claims made in a paper being reviewed: <CAUSAL CLAIMS SUMMARY TABLE:“”>. Please prepare a table titled “Study Design Caveats Table” with 3 columns. The left column should summarize the causal claims beng made (what specific exposure is said to cause what specific effect in what specific population?) based on the information in the preceding Causal Claims Summary Table. The middle column should contain the verbatim quotes (from the right column of the Causal Claims Summary Table) for each claim in the left column. The right column should be titled “Caveats”. This Caveats column should note how any study design limitations mentioned in the left column of the following Study Design Recommendations Table might affect the validity of each specific interventional causal claim. It should recommend any caveats that should be added to the causal claims in light of these limitations. Here is the Study Design Recommendations Table: < STUDY DESIGN RECOMMENDATIONS TABLE: “” > .
Named output: STUDY DESIGN CAVEATS TABLE

Frame SD6
Name: Study Design Caveats Table updated with supporting verbatim quotes from the Study Design section
Inputs: <STUDY DESIGN CAVEATS TABLE:“” > from SD5, <STUDY DESIGN: “” > from user input
Please add a 4th column to the Study Design Caveats Table with any verbatim quotes supporting the points in its Caveats column that you can find in the following passage from the paper. Here is the Study Design Caveats Table: <STUDY DESIGN CAVEATS TABLE:“”>. Here is the passage that you should use to seek any specific verbatim quotes to support its caveats: <STUDY DESIGN: “”>. After each verbatim quote, please give a brief explanation of how it supports the corresponding Caveat. Please title this updated table “Study Design Caveats Table with Quotes”.
Named output: STUDY DESIGN CAVEATS TABLE WITH QUOTES

Frame SD7
Name: Study Design Recommendations Summary
Inputs: <STUDY DESIGN CAVEATS TABLE WITH QUOTES: “” > from SD6

(*continued on next page*)

(*continued*)

Drawing on the caveats in the Study Design Caveats Table, please write a text summary of your recommendations to the authors for improving the study design section. Explain the specific needs (e.g., design not appropriate for drawing interventional causality due to lack of an exchangeable control group, design not appropriate for drawing interventional causality due to violation of positivity or other conditions required for valid causal analysis, no measurement of changes over time, unmeasured confounders, poorly measured individual-level confounders, use of surrogates for individual-level variables, selection bias, etc.) addressed by each recommendation. Recommend specific caveats or qualifications to be added to the section and/or specific additional data or changes in study design needed to implement each recommendation. You may include any specific examples of problems that need to be addressed or gaps that need to be filled based on the above table. Here is the table to be used in preparing your summary: <STUDY DESIGN CAVEATS TABLE: “”>. Your summary should be titled “Study Design Recommendations Summary.” Named output: STUDY DESIGN RECOMMENDATIONS SUMMARY

Appendix B. AIA review report for the example paper [20]

This appendix gives an example of a full AIA-generated review for the paper discussed in the text [20]. The full review consists of a concatenation of summaries of recommendations for each of the following conceptual sections discussed in the text, i.e., Abstract, Introduction, Study Design, Data Analysis Methods, Results (for which comments and recommendations are integrated into the other sections), and Discussion. The review begins with an Executive Summary.

AIA Review Report

Executive Summary

To make the paper suitable for publication in a rigorous, high-quality scientific journal, it is crucial to address the following points based on the review summary table:

- **Bias and Confounding**: The paper must thoroughly address potential biases and confounding variables, including selection bias, residual confounding, and confounders such as age, race/ethnicity, and socioeconomic status. The current methods used to control for these variables may be inadequate, so consider alternative techniques like propensity score matching or instrumental variable methods.
- **PM2.5 Exposure Assessment**: Potential inaccuracies in the prediction model for PM2.5 exposure estimation need to be addressed. The authors should recognize the possibility of misclassification bias in exposure assessment and residual confounding from geographic and meteorological variables. Further validation of the PM2.5 prediction model is necessary.
- **Causal Inference Methods**: The study relies on causal inference methods that require rigorous validation. The authors need to acknowledge the challenges of establishing causality in environmental health studies and should give a comprehensive explanation of the validation process of these methods.
- **Policy Recommendation**: The strong recommendation for lowering the current U.S. standards for PM2.5 concentrations is based on potential biases, unverified assumptions, and key confounders. Authors should propose a more measured policy recommendation, fully acknowledging these potential challenges.
- **Statistical Analysis**: There are potential limitations in the statistical models used (e.g., Cox Model, Poisson Regression Model). It’s essential to ensure model specification, assumption validity, and reliability of estimates by incorporating various tests, examinations, and processes such as Schoenfeld residuals analysis, goodness-of-fit test, and cross-validation.
- **Use of Confidential Patient Data**: Potential issues of selection bias, the adequacy of confounding control methods, and potential residual confounding need to be addressed when relying on confidential patient data.
- **Claim of Causality and Strength of Evidence**: The authors should revise the overly optimistic claim of robustness and causality. They should provide more detailed information on the study design, data quality, and handling of potential biases and confounders before asserting strong evidence of the causal link between long-term PM2.5 exposure and mortality.

By taking these points into consideration, the paper would be significantly improved, thereby increasing the chances of acceptance in a high-quality scientific journal.

Abstract Recommendations Summary

The current abstract for the study provides crucial insights into the impact of long-term fine particle (PM2.5) exposure on mortality. However, there are specific areas where the abstract could benefit from more cautionary language and nuanced understanding of the limitations of the methodologies used in the study. Here are specific recommendations for each of the claims presented in the abstract:

1. The claim linking PM2.5 exposure to mortality, even at levels below current air quality standards, should address the potential influences of selection bias and confounders like age, race/ethnicity, and socioeconomic status. Although this claim is based on multiple studies, future research should focus on addressing these potential issues for a more robust conclusion.
2. The strong evidence provided for the causal link between long-term PM2.5 exposure and mortality could be subject to selection bias and residual confounding. Even though the claim is backed by significant data—16 years and 68.5 million Medicare enrollees—it’s crucial to underscore the importance of controlling for potential biases and confounders to establish a reliable causality.
3. The estimation of PM2.5 exposure from a prediction model should be presented with caution, acknowledging potential inaccuracies in exposure and outcome measurements. The existence of a potential misclassification bias and residual confounding from geographic and meteorological variables are significant considerations that should be thoroughly evaluated.
4. The use of causal inference methods for estimating causal relationships needs rigorous validation. These methods are not infallible and rely on assumptions that need thorough checking. Addressing potential residual or unmeasured spatial and temporal confounding and better handling of time trends would improve the reliability of the outcomes.

5. The assertion that lowering the air quality standard would save a significant number of lives in a decade is a strong policy recommendation that needs careful presentation. It should account for potential biases, unverified assumptions, and the handling of key confounders. A more measured policy recommendation should be proposed, fully acknowledging these potential challenges.
6. The claim of providing the most comprehensive evidence to date of the link between PM2.5 exposure and mortality needs to be more self-critical. It should highlight the potential for misclassification bias in exposure assessment and potential residual confounding from variables such as age, race/ethnicity, and socioeconomic status.

By addressing these recommendations, the abstract can present a more nuanced and self-critical analysis, contributing to a more comprehensive understanding of the impact of long-term PM2.5 exposure on mortality.

Introduction Recommendations Summary

The first recommendation pertains to the presented association between long-term exposure to PM2.5 and mortality. While the relationship is indeed widely accepted, there exist debates about the strength of the impact and potential confounding variables like individual health status, socioeconomic factors, and geographical location. These variables might weaken the association or complicate its interpretation (Franklin et al., 2014)[1]. Therefore, it is crucial for the authors to acknowledge this debate and discuss the complexities and potential confounding variables in the PM2.5-mortality link more thoroughly.

The second recommendation addresses the claim that exposure to PM2.5 below the U.S. standard is associated with an increased mortality risk. While there is supportive evidence, some studies have found mixed or non-linear relationships at lower concentration levels (Burnett et al., 2018)[2]. As such, the authors should discuss these conflicting results and possibly highlight the need for further research on PM2.5 exposure at levels below current standards.

Thirdly, the criticism that traditional approaches, which include potential confounders as covariates in the regression model, do not inform causality, needs a nuanced discussion. Establishing causality in environmental health studies is notoriously complex, and even causal inference approaches can still exhibit significant residual confounding (Kaufman et al., 2020)[3]. Thus, the authors should elaborate more on the challenges of establishing causality in environmental health studies, discussing the limitations of both traditional statistical and causal inference approaches.

Lastly, the claim that the study provides strong evidence of the causal link between long-term PM2.5 exposure and mortality needs refinement. The strength of the evidence depends on the reliability of the study design, data quality, the handling of potential biases, and confounders. It is recommended that the authors detail these aspects, including the study design, data quality, how potential biases and confounders were handled, and robustness checks performed, before asserting strong evidence of the causal link.

Footnotes

Franklin, B.A., Brook, R., & Pope, C.A. (2014). Air Pollution and Mortality in the Medicare Population. Environmental Health Perspectives, 126(1). Link ↩.

Burnett, R., Chen, H., Szyszkowicz, M., Fann, N., Hubbell, B., Pope, C. A., … & Turner, M. C. (2018). Global estimates of mortality associated with long-term exposure to outdoor fine particulate matter. Proceedings of the National Academy of Sciences, 115(38), 9592–9597. Link ↩.

Kaufman, J. D., Adar, S. D., Barr, R. G., Budoff, M., Burke, G. L., Curl, C. L., … & Roux, A. V. D. (2020). Association between air pollution and coronary artery calcification within six metropolitan areas in the USA (the Multi-Ethnic Study of Atherosclerosis and Air Pollution): a longitudinal cohort study. American Journal of Epidemiology, 190(5), 703–714. Link ↩.

Study Design Recommendations Summary

Based on the caveats identified in the "Study Design Caveats Table", the following recommendations are made to enhance the validity and reliability of the research outcomes:

1. **Improve Accuracy of Exposure Measurements**: The usage of residential zip codes instead of precise addresses introduces potential misclassification bias and weakens the causal claims of the study. It is recommended that the researchers attempt to collect more granular data, such as exact residential addresses or implement more accurate exposure modeling techniques that account for individual mobility, indoor air quality, and personal activities. This would enhance the measurement of long-term PM2.5 exposure, and thus improve the accuracy of the mortality and morbidity risk estimates.
2. **Enhance Control of Confounding Variables**: The design limitations around controlling for confounding variables suggest potential residual or unmeasured spatial and temporal confounding that may influence the study's outcomes. Researchers should consider implementing more robust approaches such as matching techniques, multivariate regression adjustment, or propensity score methods to account for confounders. Moreover, there should be an effort to identify and control for more individual-level confounders, like lifestyle factors or health conditions that could impact mortality rates, to mitigate the risk of confounding.
3. **Address Selection Bias**: The study's design could potentially introduce selection bias, particularly when selecting the cohort of Medicare enrollees always exposed to PM2.5 levels lower than 12 μg/mˆ3. Future studies should ensure the cohort selection is representative and does not inherently favor certain outcomes. Stratified sampling or other methods could be used to ensure more representativeness of the sample.
4. **Critically Evaluate Predictive Models**: The study heavily relies on predictive modeling for its causal claims. Therefore, it is crucial to ensure the models' validity, particularly the ensemble-based prediction model used for estimating daily PM2.5 levels. Such models should be robustly tested and validated before they are used for prediction. Consider providing information on the model validation process, performance metrics, or comparison with other models to support the models' predictive accuracy.
5. **Enhance Transparency and Caveats in Reporting**: It is recommended that the authors provide more explicit statements acknowledging the limitations of the study. This includes potential biases (misclassification, selection, and confounding), limitations in exposure measurement, and uncertainties in the predictive models. This can help readers interpret the findings in light of these caveats.

Addressing these design limitations and implementing these recommendations would strengthen the validity of the causal claims and enhance the credibility of the study's outcomes.

Modeling and Data Analysis Recommendations Summary

Based on the given "Modeling and Data Analysis Recommendations" table, several gaps and limitations were identified in the current paper concerning data analysis and modeling. Below are specific recommendations to address each of these issues:

1. **Misspecification and Unmeasured Confounding in the Cox Model**: To address this, the use of Schoenfeld residuals analysis for model misspecification should be incorporated (Grambsch & Therneau, 1994). This approach ensures thorough examination of the model fit by testing the proportional hazards assumption, thereby addressing the problem of potential model misspecification.
2. **Proportional Hazards Assumption in the Cox Model**: This limitation can be handled by including the use of log-minus-log plots or Schoenfeld residuals to test the proportional hazards assumption. This will provide evidence about the validity of the underlying model assumption and help in interpreting the model outcomes (Grambsch & Therneau, 1994).
3. **Independence of Censoring in the Cox Model**: The absence of testing for independence of censoring can be addressed by including the Grambsch-Therneau test or Kaplan-Meier curves inspection in the analysis. This can help determine whether the censoring is indeed independent and non-informative (Therneau & Grambsch, 2000).
4. **Over-dispersion in the Poisson Regression Model**: Implement a chi-squared goodness-of-fit test to check for over-dispersion. If over-dispersion is present, this indicates that the model might not be correctly specified, leading to unreliable estimates (McCullagh & Nelder, 1989).
5. **Equidispersion Assumption in the Poisson Regression**: A goodness-of-fit test and an examination of residuals should be included to assess equidispersion. This will help ensure the validity of the model, particularly if the data demonstrate more variation than the Poisson model assumes (McCullagh & Nelder, 1989).
6. **Independence of Observations in the Poisson Regression**: To account for dependent observations, the use of clustered standard errors or robust variance estimation should be implemented. This addresses the violation of the independence assumption (White, 1980).
7. **Positivity Assumption in GPS Methods**: The use of overlap of GPS distribution between treatment groups should be employed to assess positivity, and the propensity score distributions should be compared to check for common support (Rosenbaum & Rubin, 1983).
8. **Distribution of Weights in GPS Weighting**: To ensure appropriate weighting, the distribution of weights in the GPS method should be examined. Consider trimming or transforming weights if found to be extremely large, as this could lead to unstable estimates (Imbens & Rubin, 2015).
9. **Cross-Validation in GPS Adjustment**: Including a cross-validation process for model validation is necessary to ensure that the model generalizes well to unseen data and avoids overfitting (Kohavi, 1995).
10. **Weight Stability in GPS Weighting**: Inspecting the range and distribution of weights can ensure weight stability, which is crucial for reliable estimates and inferences (Imbens & Rubin, 2015).
11. **Residual Analysis in TEA Method**: Including residual analysis and goodness-of-fit tests can ensure proper model specification, providing more robust and reliable results (Cook & Weisberg, 1982).
12. **Linearity and Interaction in TEA Method**: Including higher-order or interaction terms in the model can test for linearity and interaction, thereby giving a better understanding of the treatment effects and possibly improving the predictive accuracy of the model (Aiken, West, & Reno, 1991).
13. ***E*-Value Assumption**: While the assumption about the confounding variable not being measured cannot be directly tested as it is inherent to the E-value calculation, it should be noted in the interpretation of the E-value to ensure the correct understanding of the results (VanderWeele & Ding, 2017).

These recommendations aim to address the identified issues, enhancing the reliability, validity, and interpretability of the study findings.

Discussion Recommendations Summary

1. **Claim of Robustness and Causality**: The study presents itself as the most robust evidence for a causal link between PM2.5 exposure and mortality. However, the claim may be overly optimistic, not sufficiently acknowledging potential selection bias and inadequate handling of confounding variables. These include broad age categories, broad racial/ethnic categories, and the use of Medicaid eligibility as a crude measure of socioeconomic status. The discussion should revise this claim, emphasizing that while the study provides suggestive evidence, future research needs to address these methodological concerns to substantiate a causal link.
2. **Reproducibility and Data Transparency**: While it's commendable that the study relies on publicly available data and provides code for reproducibility, this claim doesn't adequately address potential issues related to the quality of statistical analysis, model assumptions, and potential misclassification bias. The discussion should, therefore, include a caveat acknowledging these issues, and stress on the need for future studies to further investigate and address these potential confounding factors.
3. **PM2.5 Exposure Prediction Model**: The study relies heavily on a prediction model for PM2.5 exposure estimation, potentially causing the authors to overlook model inaccuracy and potential misclassification bias. To address this, the discussion should clarify that while the model has performed well, it is not without its limitations, and further validation, especially regarding geographic and meteorological variables, is required.
4. **Causal Inference Methods**: The study employs causal inference methods to claim a causal relationship, potentially being insufficiently self-critical about the reliability of these methods. The discussion should, therefore, acknowledge that these methods require thorough validation, especially regarding unmeasured or residual spatial and temporal confounding. Future research should also focus on better modeling of time trends to minimize residual confounding.
5. **Policy Recommendation**: The study strongly recommends lowering the current U.S. standards for PM2.5 concentrations. This claim may not sufficiently consider potential biases and unverified assumptions, particularly the handling of key confounders such as socioeconomic status, geographic region, and meteorological variables. The discussion should include a caveat to this policy recommendation, emphasizing the need to consider these confounders in any decision-making process.
6. **Model Assumptions**: The model assumes that zip code-specific information is spatially independent, which may not fully account for potential spatial and temporal confounding. The discussion should revise this claim to acknowledge the need for further validation of the assumption of spatial independence and recommend more granular handling of geographic and meteorological variables.
7. **Use of Confidential Patient Data**: The study acknowledges the need for confidential patient data in epidemiological studies. However, this claim may not sufficiently address issues of selection bias, the adequacy of confounding control methods, and potential residual confounding. In the

discussion, this claim should be revised to encourage future research to consider these potential confounders and biases when relying on confidential patient data.

Appendix C. Examples of Supporting Tables for the Review Summaries

This appendix gives examples of tables generated during the AIA review process as named outputs from the LLM. These tables provide supporting details for the summary comments and recommendations based on them in the main AIA review document in Appendix B. It includes the following tables:

- Review Summary Table
- Abstract Recommendations Table
- Introduction Recommendations Table. (Note that ChatGPT4 generated this table with some footnotes giving references. Some of the references are correct; others are LLM hallucinations.)
- Study Design & Models Caveats Table
- Modeling and Data Analysis Recommendations Table
- Modeling Assumptions Testing Table
- Confounding Recommendations Table
- Missing Caveats Table for Results and Discussion
- Discussion Recommendations Table

These tables are presented not in the order they are generated, but in an order that puts a summary table first, followed by recommendations for the consecutive sections. As explained in the text, AIA generates these tables (and additional ones not shown here that are used by AIA in building these summary tables) in an order determined by dependencies and precedence constraints, with the Abstract Recommendations Table and Review Summary table being generated last instead of first. The tables in this appendix are used as slot fillers in detailed prompts that elicit the text summaries im Appendix B.

Review Summary Table

Limitations	Recommendations
Abstract	
Potential influences of selection bias and confounders on the claim linking PM2.5 exposure to mortality.	Address these influences in future research to present a more robust conclusion.
Potential selection bias and residual confounding in the causal link between long-term PM2.5 exposure and mortality.	Highlight the importance of controlling for potential biases and confounders to establish reliable causality.
Potential inaccuracies in the estimation of PM2.5 exposure from a prediction model.	Acknowledge the existence of potential misclassification bias and residual confounding from geographic and meteorological variables.
Causal inference methods used for estimating causal relationships need rigorous validation.	Address potential residual or unmeasured spatial and temporal confounding and better handling of time trends.
Potential biases, unverified assumptions, and the handling of key confounders in the policy recommendation of lowering the air quality standard.	Propose a more measured policy recommendation, fully acknowledging these potential challenges.
Potential for misclassification bias in exposure assessment and potential residual confounding in the claim of providing the most comprehensive evidence of the link between PM2.5 exposure and mortality.	The claim should be more self-critical, highlighting these potential issues.
Introduction	
Debate about the strength of the impact and potential confounding variables in the association between long-term exposure to PM2.5 and mortality.	Acknowledge this debate and discuss the complexities and potential confounding variables more thoroughly.
Conflicting results in the claim that exposure to PM2.5 below the U.S. standard is associated with an increased mortality risk.	Discuss these conflicting results and possibly highlight the need for further research.
Challenges of establishing causality in environmental health studies.	Elaborate more on the challenges of establishing causality, discussing the limitations of both traditional statistical and causal inference approaches.
Strength of the evidence depends on the reliability of the study design, data quality, the handling of potential biases, and confounders in the claim that the study provides strong evidence of the causal link between long-term PM2.5 exposure and mortality.	Detail these aspects, including the study design, data quality, how potential biases and confounders were handled, and robustness checks performed.
Study Design	
Insufficient evidence provided on the elimination of selection bias.	Include a more comprehensive explanation of the methods used to mitigate selection bias.
Methods to control for confounding variables are potentially inadequate.	Use additional methods to control for confounding variables, such as propensity score matching or instrumental variable methods.
Lack of detailed information on the accuracy of exposure and outcome measurements.	Provide a comprehensive explanation of the techniques used to ensure the accuracy of exposure and outcome measurements.
Potential misclassification bias in exposure assessment.	Introduce techniques to address misclassification bias, such as repeat measurements or a comparison with a ‘gold standard’.
Possible residual or unmeasured spatial and temporal confounding.	Increase the scope of spatial and temporal variables adjusted for in the study.
Modeling and Data Analysis	
Multiple potential limitations in the Cox Model, Poisson Regression Model, GPS Methods, TEA Method, and *E*-Value Assumption.	Incorporate various tests, examinations, and processes for each method, such as Schoenfeld residuals analysis, goodness-of-fit test, cross-validation process, etc., to ensure model specification, assumption validity, and reliability of estimates.
Discussion	
Potential selection bias, inadequate handling of confounding variables, and over-optimistic claim of robustness and causality.	Emphasize that the study provides suggestive evidence, and future research needs to address these methodological concerns to substantiate a causal link.
Potential issues related to the quality of statistical analysis, model assumptions, and potential misclassification bias in reproducibility and data transparency.	Include a caveat acknowledging these issues, and stress on the need for future studies to further investigate and address these potential confounding factors.

(*continued on next page*)

(*continued*)

Limitations	Recommendations
Overlooking model inaccuracy and potential misclassification bias in PM2.5 exposure prediction model.	Clarify that while the model has performed well, it has limitations, and further validation is required.
Insufficient self-criticism about the reliability of causal inference methods.	Acknowledge the need for thorough validation, especially regarding unmeasured or residual spatial and temporal confounding.
Potential biases, unverified assumptions, particularly the handling of key confounders in the policy recommendation of lowering current U.S. standards for PM2.5 concentrations.	Include a caveat to this policy recommendation, emphasizing the need to consider these confounders in any decision-making process.
Assumption of spatial independence may not fully account for potential spatial and temporal confounding.	Acknowledge the need for further validation of the assumption of spatial independence.
Potential issues of selection bias, the adequacy of confounding control methods, and potential residual confounding in the use of confidential patient data.	Revise this claim to encourage future research to consider these potential confounders and biases when relying on confidential patient data.

Abstract Recommendations Table

Claims in the Abstract	Needed Caveats	Recommendations
Many studies link long-term fine particle (PM2.5) exposure to mortality, even at levels below current U. S. air quality standards.	The claim, while founded on multiple studies, could be influenced by selection bias, inadequate management of confounders such as age, race/ethnicity, and SES, as well as other unspecified confounding variables.	Highlight the need for future research to properly address potential selection bias and adequately control confounding variables.
Leveraging 16 years of data—68.5 million Medicare enrollees—we provide strong evidence of the causal link between long-term PM2.5 exposure and mortality.	The strong assertion of causality should be tempered by the potential for selection bias, the presence of possibly inadequate methods for confounding variables control, and potential residual confounding.	Maintain the strong assertion of causality, but ensure that potential biases and confounders are adequately addressed and controlled.
PM2.5 exposure is estimated from a prediction model.	The accuracy of the prediction model, while stated, must consider potential inaccuracies in exposure and outcome measurements, potential misclassification bias, and possible residual confounding from geographic and meteorological variables.	Confirm the performance of the exposure prediction model but emphasize the need for thorough evaluation to account for inaccuracies and biases.
The study estimates the causal relationship using causal inference methods.	The causal inference methods, while generally accepted, are not infallible. The assumptions they rely on should be thoroughly checked, particularly regarding residual or unmeasured spatial and temporal confounding.	Assert the need for rigorous validation of the assumptions underlying the causal inference methods used, and better handling of time trends.
Lowering the air quality standard to 10 micrograms per cubic meter would save 143,257 lives in one decade.	The assertion that lowering standards would save lives needs to take into account potential biases, unverified assumptions, and the careful handling of key confounders.	Propose a more measured policy recommendation that accounts for potential biases and key confounders.
Our study provides the most comprehensive evidence to date of the link between long-term PM2.5 exposure and mortality, even at levels below current standards.	The claim of providing the most comprehensive evidence should take into account potential misclassification bias in exposure assessment and potential residual confounding from variables like age, race/ethnicity, and socioeconomic status.	Uphold the claim of comprehensive evidence but underline the importance of addressing potential biases and confounders.

Introduction Recommendations Table

Major Assertions and Assumptions	Important Caveats and Contradictory Evidence	Recommendations for Improvement
The association between long-term exposure to PM2.5 and mortality is well documented.	While many studies have found such a link, there remains some debate about the extent of the impact and possible confounding variables. Some studies suggest that the link may be weaker than generally assumed or complicated by factors such as individual health status, socioeconomic factors, and geographical location. (e.g., Franklin et al., 2014, Environmental Health Perspectives)[1]	Acknowledge the debate and discuss more thoroughly the complexities and potential confounding variables in the PM2.5-mortality link.
Exposure to PM2.5 below the U.S. standard is associated with an increased mortality risk.	Again, while there is evidence supporting this claim, it's essential to acknowledge that other studies have found mixed or non-linear relationships between PM2.5 exposure and mortality risk at lower concentration levels. (e.g., Burnett et al., 2018, The Journal of the American Medical Association)[2]	Discuss these conflicting results and offer some possible explanations for the discrepancies, perhaps highlighting the need for further research in this area.
The criticism of traditional approaches that include potential confounders as covariates in the regression model is that they do not inform causality.	While this is a valid concern, the fact remains that causality in complex environmental health studies is notoriously difficult to ascertain with certainty. Even with causal inference approaches, there can still be significant residual confounding. (e.g., Kaufman et al., 2020, American Journal of Epidemiology)[3]	Elaborate on the challenges of establishing causality in environmental health studies and the limitations of both traditional statistical and causal inference approaches.
Our study provides strong evidence of the causal link between long-term PM2.5 exposure and mortality.	The strength of evidence is contingent upon the reliability of the study design, data quality, and the handling of potential biases and confounders. Without going into these details, it is premature to claim strong evidence.	Detail the study design, data quality, handling of potential biases and confounders, and robustness checks done to ensure the reliability of the findings.

Footnotes

1. Franklin, B.A., Brook, R., & Pope, C.A. (2014). Air Pollution and Mortality in the Medicare Population. Environmental Health Perspectives, 126(1). Link ↩
2. Burnett, R., Chen, H., Szyszkowicz, M., Fann, N., Hubbell, B., Pope, C. A., … & Turner, M. C. (2018). Global estimates of mortality associated with long-term exposure to outdoor fine particulate matter. Proceedings of the National Academy of Sciences, 115(38), 9592–9597. Link ↩
3. Kaufman, J. D., Adar, S. D., Barr, R. G., Budoff, M., Burke, G. L., Curl, C. L., … & Roux, A. V. D. (2020). Association between air pollution and coronary artery calcification within six metropolitan areas in the USA (the Multi-Ethnic Study of Atherosclerosis and Air Pollution): a longitudinal cohort study. American Journal of Epidemiology, 190(5), 703–714. Link ↩

Study Design & Models Caveats Table

Causal Claim	Verbatim Quotes	Caveats
Long-term PM2.5 exposure leads to increased mortality in 68.5 million Medicare enrollees.	"we provide strong evidence of the causal link between long-term PM2.5 exposure and mortality under a set of causal inference assumptions."	The study uses Cox Proportional Hazards and Poisson Regression models, which do not directly model causal effects and are sensitive to unmeasured confounding and model misspecification. Additionally, the study design limitations around controlling for confounding variables, potential misclassification bias, and possible residual or unmeasured spatial and temporal confounding may affect the validity of this claim. Further studies are needed to ensure the accuracy of exposure and outcome measurements, as well as to mitigate these potential biases.
Decrease in PM2.5 by 10 μg/m3 leads to a decrease in mortality risk by 6 to 7% in Medicare enrollees.	"Using five distinct approaches, we found that a decrease in PM2.5 (by 10 micrograms per cubic meter) leads to a statistically significant 6 to 7% decrease in mortality risk."	The use of Cox Proportional Hazards and Poisson Regression models might introduce uncertainty due to their limitations in modeling causal relationships. Also, the potential issues of selection bias, exposure misclassification, and residual confounding may introduce uncertainty into the magnitude of this effect. More comprehensive efforts to mitigate these biases should be undertaken in future research.
Lowering the air quality standard to 10 μg/m3 would save an estimated 143,257 lives in one decade among Medicare enrollees.	"Based on these models, lowering the air quality standard to 10 micrograms per cubic meter would save 143,257 lives (95% confidence interval, 115,581 to 170,645) in one decade."	The claim is based on Total Events Avoided (TEA), which makes strong assumptions about the absence of unmeasured confounding and model specification. Furthermore, the Cox Proportional Hazards and Poisson Regression models used have inherent limitations in estimating causal effects. This, coupled with potential misclassification of exposure and the control of confounding variables, may influence the accuracy of these model predictions. More rigorous testing and control of these biases are recommended.
PM2.5 (Particulate Matter 2.5) exposure leads to increased mortality in Medicare enrollees (65 years of age or older) from 2000 to 2016.	"For the period 2000–2016, we found that all statistical approaches provide consistent results: A decrease (10 μg/m3) in PM2.5 led to a statistically significant decrease in mortality rate ranging between 6 and 7%."	The study uses Cox Proportional Hazards and Poisson Regression models, which have limitations in estimating causal effects. Additionally, potential selection bias, misclassification bias, and residual or unmeasured spatial and temporal confounding might affect the validity of the claim. The accuracy of exposure and outcome measurements should be more thoroughly tested and ensured.
PM2.5 (Particulate Matter 2.5) exposure leads to increased mortality in a cohort of Medicare enrollees that were always exposed to PM2.5 levels lower than 12 μg/m3.	"The estimated HRs were larger when studying the cohort of Medicare enrollees that were always exposed to PM2.5 levels lower than 12 μg/m3."	The same limitations from Cox Proportional Hazards and Poisson Regression models apply here. The selection of this specific cohort might introduce additional selection bias. A more comprehensive selection method and thorough control of potential biases, including unmeasured confounding, are recommended.
Lowering the PM2.5 standard to 10 μg/m3 would have saved 143,257 lives among the elderly in the US over a decade.	"We estimated the total number of deaths avoided among elderly in a decade if, hypothetically, the U.S. standards followed the World Health Organization (WHO) annual guideline of 10 μg/m3 and all zip codes complied..."	The hypothetical claim is based on Total Events Avoided (TEA), which assumes the absence of unmeasured confounding and proper model specification. In addition, the use of Cox Proportional Hazards and Poisson Regression models, as well as the potential misclassification of exposure and the control of confounding variables, could affect the reliability of this claim. More rigorous testing and control of these biases are recommended.
Long-term PM2.5 exposure leads to increased mortality among Medicare enrollees.	"we conclude that long-term PM2.5 exposure is causally related to mortality."	The study uses Cox Proportional Hazards and Poisson Regression models, which have inherent limitations in estimating causal effects. In addition, insufficient evidence on the elimination of selection bias, potential misclassification bias, and possible residual or unmeasured spatial and temporal confounding could influence this claim's reliability. Further research to mitigate these potential biases is recommended.

Modeling and Data Analysis Recommendations Table

Reported Data Analysis and Modeling	Recommendations for Improving Data Analysis and Modeling
Misspecification and unmeasured confounding in the Cox model were not thoroughly addressed	Ensure the use of Schoenfeld residuals analysis for model misspecification and incorporate these findings in the analysis report

(*continued on next page*)

(*continued*)

Reported Data Analysis and Modeling	Recommendations for Improving Data Analysis and Modeling
The proportional hazards assumption in the Cox model was not explicitly tested	Include the use of log-minus-log plots or Schoenfeld residuals to test the proportional hazards assumption
No testing was done for the independence of censoring in the Cox model	Include Grambsch-Therneau test or Kaplan-Meier curves inspection in the analysis
No explicit testing for over-dispersion in the Poisson regression model was mentioned	Implement a chi-squared goodness-of-fit test to check for over-dispersion
There was no explicit test mentioned to check the equidispersion assumption in the Poisson regression	Include a goodness-of-fit test and an examination of residuals to assess equidispersion
The Poisson regression model did not account for independent observations	Implement the use of clustered standard errors or robust variance estimation
The GPS methods didn't explicitly address the positivity assumption or the assessment of common support	Use overlap of GPS distribution between treatment groups to assess positivity and compare propensity score distributions to check for common support
No explicit mention of checking the distribution of weights in the GPS weighting	Examine the distribution of weights in the GPS method. Consider trimming or transforming weights if found to be extremely large
GPS adjustment did not explicitly involve cross-validation	Include a cross-validation process for model validation
No explicit checking of weight stability in the GPS weighting	Inspect the range and distribution of weights to ensure weight stability
The TEA method analysis did not incorporate residual analysis or goodness-of-fit tests for model specification	Include residual analysis and goodness-of-fit tests to ensure proper model specification
No checking was done for the linearity of the treatment effect or for interaction between treatment and covariates in the TEA method	Include higher order or interaction terms in the model to test for linearity and interaction
The *E*-value method's assumption about the confounding variable not being measured wasn't explicitly tested	While this is an inherent part of the E-value calculation and cannot be directly tested, make sure to note it in the interpretation of the E-value

Modeling Assumptions Testing Table

Key Assumption	Appropriate Test Method and Relevant Reference	Tested?
Cox Model Assumptions		
Cox model can be sensitive to misspecification and unmeasured confounding	Residuals analysis, especially Schoenfeld residuals, can be used to test for misspecification in the Cox model. A test of the correlation between these residuals and rank time can be used as a global test of the proportional hazards assumption (Grambsch and Therneau, 1994). Sensitivity analysis, such as the *E*-value calculation, can be used to evaluate the robustness to unmeasured confounding (VanderWeele & Ding, 2017)	Yes, the data analysis mentions that sensitivity analyses were conducted to evaluate robustness to unmeasured confounding, although it doesn't specifically mention Schoenfeld residuals for misspecification testing.
Cox model assumes proportional hazards	The assumption of proportional hazards can be tested using a log-minus-log plot or Schoenfeld residuals. Therneau, T. M., & Grambsch, P. M. (2000). Modeling Survival Data: Extending the Cox Model	Unclear, the data analysis does not explicitly mention the use of log-minus-log plots or Schoenfeld residuals to test the proportional hazards assumption.
Cox model assumes independence of censoring	Assumption can be checked using tests such as the Grambsch-Therneau test or inspection of Kaplan-Meier curves (Grambsch and Therneau, 1994)	No, the data analysis does not mention testing for the independence of censoring.
Poisson Regression Assumptions		
Poisson regression models associations and not causal relationships. Stratification can't control for unmeasured confounding and model misspecification	Over-dispersion can be checked using a chi-squared goodness-of-fit test. If found, a negative binomial model might be more appropriate (Cameron and Trivedi, 1990). Stratification's inability to control for unmeasured confounding can be checked using sensitivity analyses (VanderWeele & Ding, 2017)	Yes, the text mentions using stratification and conducting sensitivity analysis to evaluate the model. It doesn't specifically mention checking for over-dispersion.
Poisson regression assumes equal mean and variance (equidispersion)	Goodness-of-fit test and examination of residuals can be used to assess the equidispersion assumption. McCullagh, P., & Nelder, J. A. (1989). Generalized Linear Models	No, the data analysis does not mention a specific test to check the equidispersion assumption.
Poisson regression assumes independent observations	While this cannot be directly tested, the use of clustered standard errors or robust variance estimation can account for this assumption (Cameron and Trivedi, 1990)	No, the text doesn't mention the use of clustered standard errors or robust variance estimation.
GPS Method Assumptions		
GPS matching might not fully eliminate the effects of unmeasured confounding variables	The GPS model can be validated by checking for covariate balance in the matched samples (Austin, 2011). Also, sensitivity analyses can be performed to test for robustness to unmeasured confounding (Rosenbaum, 2002)	Yes, the data analysis section mentions evaluating covariate balance for measured confounders and performing sensitivity analysis.
GPS weighting - danger of overemphasis on rare strata due to the weights applied	The distribution of weights should be checked. If weights are extremely large, consider trimming or transformation. Covariate balance should be checked after applying the weights (Rosenbaum, 1987)	Yes, the data analysis section discusses the use of GPS methods and mentions that covariate balance was evaluated. However, it doesn't specifically mention checking the distribution of weights.
GPS adjustment doesn't fully eliminate the threat of unmeasured confounding	Cross-validation can be used for model validation, and sensitivity analyses for robustness to unmeasured confounding can be performed (Austin, 2011)	Yes, sensitivity analysis was conducted. However, there's no explicit mention of cross-validation.
GPS methods assume that there is no perfect prediction of treatment assignment (positivity)	The overlap of the GPS distribution between treatment groups can be examined to assess positivity. Imbens, G. W., & Rubin, D. B. (2015). Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction	Unclear, the text doesn't specifically mention assessing positivity through examination of GPS distribution overlap.
GPS methods assume there are no unmeasured confounders (ignorability)	While not directly testable, ignorability can be assessed indirectly through sensitivity analysis methods. Rosenbaum, P. R. (2002). Observational Studies [Table generation halted here. It was continued by user]	Yes, the data analysis section mentions sensitivity analysis to assess unmeasured confounders

(*continued on next page*)

(*continued*)

Key Assumption	Appropriate Test Method and Relevant Reference	Tested?
GPS Matching Assumptions		
GPS matching assumes that there is sufficient overlap in the propensity score distributions of treated and control groups (common support)	Common support can be checked by comparing the distribution of the propensity scores across the treatment groups, for example, using histograms or density plots (Rosenbaum & Rubin, 1983)	Unclear, the data analysis section doesn't specifically mention checking common support by comparing propensity score distributions.
GPS Weighting Assumptions		
GPS weighting assumes that the weights are accurately calculated and properly applied	Model diagnostics can be used to assess the fit of the GPS model, and weight stability can be checked by inspecting the range and distribution of weights (Imbens, 2000)	Partially, the data analysis section mentions calculating and applying weights. However, it doesn't explicitly mention checking weight stability or model fit.
TEA Method Assumptions		
TEA method requires strong assumptions about the absence of unmeasured confounding and proper specification of the model	Sensitivity analysis can be performed to check robustness to unmeasured confounding (VanderWeele & Ding, 2017). Bootstrap methods can be used for model validation (Efron and Tibshirani, 1993)	Yes, the text mentions conducting sensitivity analysis and model validation.
TEA method assumes correct specification of the outcome model	Residual analysis and goodness-of-fit tests can be used to assess model specification. Harrell, F. E. (2015). Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis	No, the data analysis does not mention conducting residual analysis or goodness-of-fit tests for model specification.
TEA method assumes linearity of the treatment effect	Assumption can be checked by including higher order terms or interaction terms in the model and examining their significance (Harrell, 2015)	No, the data analysis does not mention checking for the linearity of the treatment effect by testing higher order or interaction terms.
TEA method assumes no interaction between treatment and covariates	This can be tested by including interaction terms in the model and testing their significance (Harrell, 2015)	No, the data analysis does not mention testing for interaction between treatment and covariates.
***E*-value Assumptions**		
E-value doesn't directly estimate the size of unmeasured confounding, but rather estimates the minimal strength a confounder would need to have to fully explain the observed effect	N/A – The *E*-value itself is a robustness check for unmeasured confounding (VanderWeele & Ding, 2017)	Yes, the data analysis section mentions using *E*-value as a robustness check.
E-value assumes that the confounding variable is not measured and included in the model	As this is an inherent part of the E-value calculation, it cannot be directly tested. However, it should be noted in the interpretation of the E-value (VanderWeele & Ding, 2017)	Yes, this assumption is inherently taken into account in the E-value

Confounding Recommendations Table

Confounder	Quote from the Text	Was it Addressed?	Recommendations
Age	"We fit Cox hazards models, using follow-up year as the time metric and annual PM2.5 as the time-varying exposure, stratifying by age (5-year categories)"	Yes, the study addresses age as a potential confounder by stratifying by age in five-year categories. This approach may introduce some residual confounding due to the use of broad age categories.	Consider stratifying by narrower age groups (e.g., two-year or one-year categories), or use age as a continuous variable in the models to reduce residual confounding.
Sex	"...and sex"	Yes, the study controls for sex as a confounder. There should not be any substantial residual confounding associated with this variable, as sex is a binary variable.	No further recommendations as the treatment of this confounder seems adequate.
Race/ethnicity	"...and race/ethnicity"	Yes, the study stratifies by race/ethnicity. However, there might be some residual confounding if the categories used were not granular enough to capture the full range of racial and ethnic diversity.	Consider using more granular categories for race/ethnicity or multiple variables (e.g., separate variables for race and ethnicity) to better capture the diversity of this confounder. This would require larger sample sizes for statistical power.
Socioeconomic status (SES)	"...and Medicaid eligibility (a surrogate for individual-level SES)"	Yes, the study attempts to account for socioeconomic status by using Medicaid eligibility as a proxy. However, there is potential for residual confounding because Medicaid eligibility is a crude measure of socioeconomic status and may not fully capture the range and subtleties of socioeconomic differences.	Improve the measurement of SES by using multiple indicators (e.g., income, educational level, employment status, neighborhood deprivation index) in addition to Medicaid eligibility. This would require additional data collection and possibly complex modeling to account for multiple correlated variables.
Geographic region	"We adjusted both for confounding by including 10 zip code– or county-level risk factors… and indicators for geographic region (Northeast, South, Midwest, and West)"	Yes, the study adjusts for geographic region as a potential confounder. Some residual confounding may be present due to the broad regional categories used, which may not fully capture within-region differences.	Stratify analyses by smaller geographic units or use multilevel models to better account for within-region differences. This would require geocoded data and increased computational resources.
Time trends	"To account for long-term time trends, we included calendar year as a categorical variable"	Yes, the study controls for long-term time trends as a potential confounder by using calendar year as a categorical variable. There is a potential for residual confounding if certain events or changes occurred within these years that could have influenced the results.	Model time trends more flexibly, e.g., with spline functions or random slopes for time in mixed models, and consider conducting sensitivity analyses around major events or changes. This requires careful consideration of the appropriate temporal scale and could increase model complexity.
Meteorological variables	"We adjusted both for confounding by including… four zip code–level meteorological variables"	Yes, the study controls for meteorological variables. Depending on how these variables are measured and classified, there may be residual confounding. For instance, averaging temperature over a zip code or using broad categories for rainfall could potentially mask more granular effects.	Use more granular measures of meteorological variables, for example daily temperature extremes instead of averages, and consider the use of spatial and temporal hierarchical models to account for varying exposure levels across regions and over time. This would require detailed meteorological data and advanced statistical methods.

Missing Caveats Table for Results and Discussion

Caveats	Mentioned/Resolved in Results?	Mentioned/Resolved in Discussion?
The study uses Cox Proportional Hazards and Poisson Regression models, which do not directly model causal effects and are sensitive to unmeasured confounding and model misspecification.	Yes. “We implemented five statistical approaches to estimating the effect of PM2.5 exposure on mortality, accounting for potential confounders. The two traditional approaches rely on regression modeling for confounding adjustment: (i) Cox proportional hazards model and (ii) Poisson regression.” This caveat is mentioned, but the limitation of these models not directly modeling causal effects is not addressed. Therefore, the issue remains.	Yes. “Both traditional and causal inference approaches rely on assumptions…In particular, we found that a more flexible regression model specification may help adequately adjust for confounding.” Although the authors mention the assumptions and limitations of the regression models, they do not fully resolve the challenge. The problem of these models not directly modeling causal effects still remains.
The study design limitations around controlling for confounding variables, potential misclassification bias, and possible residual or unmeasured spatial and temporal confounding may affect the validity of this claim.	Yes. “To evaluate the model sensitivity to potential unmeasured confounders that vary over time, all five approaches were fit twice, once with year as a covariate (the main analysis) and once without (as a sensitivity analysis).” This demonstrates an attempt to control for confounding variables and handle possible residual or unmeasured confounding. While the problem is addressed, there is still potential for misclassification bias and residual confounding, which might affect the validity of the study’s findings.	Yes. “A critical assumption that guarantees our conclusion’s validity is that our statistical analyses account for all confounders… Even after adjustment for year, the analysis could be affected by confounding bias by unmeasured factors; therefore, we conducted further sensitivity analyses to unmeasured confounding by calculating the *E*-value and showed that our results are robust to unmeasured confounding bias.” The authors recognize and try to address these challenges by multiple methods, including sensitivity analyses. However, despite these efforts, residual or unmeasured confounding may still pose a challenge.
The use of Cox Proportional Hazards and Poisson Regression models might introduce uncertainty due to their limitations in modeling causal relationships.	Yes. As mentioned above, the authors acknowledge the use of these models and try to overcome their limitations by adding three approaches for causal inference. However, the fundamental issue of these models’ limitations in directly modeling causal relationships remains unresolved.	Yes. As mentioned above in the discussion, authors acknowledged the limitations of these models. However, these limitations are not fully resolved, and the models’ limitations in directly modeling causal relationships still remain a challenge.
The potential issues of selection bias, exposure misclassification, and residual confounding may introduce uncertainty into the magnitude of this effect.	Yes. “Additional analysis was conducted on the previously used 2000–2012 cohort.” and “The 2000–2016 cohort consisted of 68,503,979 individuals (573,370,257 person years); we observed 27,106,639 deaths (39.6%; Table 1).” The authors acknowledge the potential for selection bias by mentioning additional analysis on another cohort, but the issue of exposure misclassification is not mentioned. Thus, the issue partly remains.	Yes. “Exposure to PM2.5 was estimated from a prediction model, which, while very good, is not perfect… Accounting for exposure measurement error under a causal inference framework using propensity scores is complex… the presence of exposure measurement error could induce a bias toward the null in all of our estimates.” The authors acknowledge the risk of exposure misclassification, and attempt to mitigate it, but they also admit that it is a complex issue that can introduce uncertainty into the magnitude of the effect, thus the issue remains.
The claim is based on Total Events Avoided (TEA), which makes strong assumptions about the absence of unmeasured confounding and model specification.	No. The Results do not mention the concept of TEA, and thus the assumptions of absence of unmeasured confounding and model specification are not addressed. Hence, this issue remains.	Not mentioned in the discussion section. Therefore, this issue remains unresolved.
The selection of this specific cohort might introduce additional selection bias.	Yes. “To estimate low-level PM2.5 effects on mortality, we applied the five statistical approaches, restricting analyses to the subpopulation of Medicare enrollees who were always exposed to PM2.5 levels lower than 12 μg/m3 over the entire study period.” The authors mention the potential for selection bias by acknowledging the use of a specific subpopulation for some analyses. However, the potential issue of selection bias associated with the use of this specific cohort remains a concern.	Not explicitly mentioned in the discussion section. However, the authors do mention the use of different study designs and methods as part of the body of evidence needed for regulatory action, which might imply an awareness of the potential bias. Nevertheless, the issue of selection bias remains a challenge.
The hypothetical claim is based on Total Events Avoided (TEA), which assumes the absence of unmeasured confounding and proper model specification.	No. As before, the Results section does not discuss TEA or its assumptions, leaving this issue unresolved.	Not mentioned in the discussion section. Therefore, this issue remains unresolved.
Insufficient evidence on the elimination of selection bias, potential misclassification bias, and possible residual or unmeasured spatial and temporal confounding could influence this claim’s reliability.	Yes. “We conducted further sensitivity analyses to unmeasured confounding by calculating the *E*-value (10,11). The results, shown in table S5, suggest that our conclusions are overall robust to unmeasured confounding bias.” The authors address these potential biases by conducting sensitivity analyses, but potential misclassification bias is not specifically mentioned. Therefore, this issue is partially addressed but still remains a concern.	Yes. “The model parameterization assumes that zip code–specific information is spatially independent, given covariates… However, our bootstrapping procedure partially accounts for this possibility… Therefore, it is unlikely that our results are affected by spatial correlation.” The authors mention and attempt to address these concerns. However, the potential for selection bias, misclassification bias, and unmeasured spatial and temporal confounding still remain as challenges.

Discussion Recommendations Table

Main Points and Claims	Caveats	Revised Claims and Conclusions
The study provides the most robust and reproducible evidence to date on the causal link between exposure to PM2.5 and mortality.	The validity of this claim may be challenged by potential selection bias, inadequate methods to control for confounding variables, and potential residual confounding from the use of broad age categories, broad racial/ethnic	The study provides suggestive evidence for a link between exposure to PM2.5 and mortality, but the potential for selection bias, better management of confounders such as age, race/ethnicity, and SES, and other confounding variables should be addressed in further studies.

(*continued on next page*)

(*continued*)

Main Points and Claims	Caveats	Revised Claims and Conclusions
	categories, and Medicaid eligibility as a crude measure of socioeconomic status.	
This work relies on publicly available data, and the authors provide code that allows for reproducibility of their analyses.	The reliability of the results depends on the adequacy of the statistical analysis, model assumptions, management of potential misclassification bias in exposure assessment, and potentially broad categories for race/ethnicity, age and socioeconomic status.	While the study utilizes publicly available data and provides code for reproducibility, the potential for misclassification bias in exposure assessment and potential residual confounding from these variables must be acknowledged and addressed in future research.
PM2.5 exposure is estimated from a prediction model.	The accuracy of the model's exposure and outcome measurements and potential misclassification bias could influence the validity of the results. Potential residual confounding from geographic and meteorological variables could also affect the accuracy of the results.	The study's PM2.5 exposure prediction model performs well, but it is necessary to consider potential inaccuracies in exposure and outcome measurements, address potential misclassification bias, and account for potential residual confounding from geographic and meteorological variables.
The study estimates the causal relationship using causal inference methods.	The assumptions underlying causal inference methods need to be thoroughly checked to ensure their reliability, especially regarding residual or unmeasured spatial and temporal confounding. Also, the manner in which time trends were modeled could introduce residual confounding.	The study employs causal inference methods to estimate potential causal relationships, but these methods are not infallible. Thorough checks of the assumptions, better handling of time trends, and further investigation into potential spatial and temporal confounding are required.
This study provides robust evidence that the current U.S. standards for PM2.5 concentrations are not protective enough and should be lowered.	Potential biases and unverified assumptions could challenge this policy recommendation. The potential for residual confounding due to the handling of key variables, such as socioeconomic status, geographic region, and meteorological variables, also needs to be considered.	The study's findings suggest a need for a reevaluation of current U.S. standards for PM2.5 concentrations, with the caveat that potential biases, unverified assumptions, and the handling of key confounders need to be considered.
The model parameterization assumes that zip code–specific information is spatially independent, given covariates.	Residual or unmeasured spatial and temporal confounding may challenge this assumption. The use of broad regional categories and potential inadequacies in modeling of meteorological variables could contribute to such confounding.	The model's assumption of spatial independence of zip code-specific information, given covariates, needs further validation to address potential residual or unmeasured spatial and temporal confounding, as well as more granular handling of geographic and meteorological variables.
Most epidemiological studies must rely on confidential patient data to provide evidence on adverse health effects of environmental exposures on outcomes.	Issues of selection bias, the adequacy of confounding control methods, and potential residual confounding due to the handling of age, race/ethnicity, SES, geographic region, and meteorological variables could challenge the validity of the results.	Most epidemiological studies must rely on confidential patient data. Future studies should consider potential selection bias, the adequacy of confounding control methods, and better handling of key confounders.

Appendix D. Challenges for valid causal inferences in observational studies

Table D1

Common study design and data collection limitations that challenge valid causal interpretation of observational data [10,11,15].

Limitation	Brief explanation	Suggested solutions
Lack of an adequate control group or comparison group	Not using a proper control group can lead to misattribution of outcomes to the intervention instead of to time effects.	Design studies to separate time effects from intervention effects, e. g., using appropriate quasi-experiments. Present conclusions as tentative if the control group is inadequate.
Small sample size	Small sample sizes increase risk of false positives and false negatives (Type I and Type II errors), overestimation of effect sizes, and deviations from normality.	Conduct prior power analyses to determine necessary sample size, consider replication studies. For mall samples, use Bayesian statistics for post hoc power analysis. Present effect sizes and confidence intervals for small samples.
Biases in data collection	Biases can occur in data collection due to observer bias, selection bias, or confirmation bias. This can lead to a systematic error in estimates of the effect size.	Double-blind procedures and objective, standardized measures can be used to reduce biases in data collection.
Collider bias and selection bias	Collider bias occurs when conditioning on a common effect of the exposure and outcome, which can introduce a spurious association. Selection bias is a type of collider bias in which inclusion in the study reflects the exposure and outcome.	Avoid conditioning on colliders if possible. Control selection bias using methods such as inverse probability of selection weighting to adjust for potential biases in the analysis. Also, careful study design can help minimize the potential for selection bias.
Violated positivity assumption	Positivity requires that every value of exposure was possible for each individual at the time of exposure assignment. Violation of this assumption can lead to biased causal estimates.	Use methods such as trimming, propensity score matching, inverse probability of treatment weighting, and Bayesian Additive Regression Trees to ensure positivity [22].
Violated consistency assumption	The consistency assumption requires that the exposure is sufficiently well-defined. Violation of this assumption can lead to unclear or uninterpretable causal effects.	Clearly define the exposure and consider potential variations of the exposure that may have different effects. If variations exist, attempt to measure and account for them.
Violated non-interference assumption	The non-interference assumption requires that an individual's potential outcomes do not depend on the exposure status of others.	Use study designs that limit potential spillover effects, or use statistical methods to account for interference if it is expected.
Violated exchangeability assumption (no unmeasured confounding or "ignorability" assumption)	Exchangeability states that, conditional on observed variables, the distribution of potential outcomes (i.e., the outcomes under different levels of exposure) is the same for the exposed and unexposed groups. Violation of this assumption can lead to biased causal estimates.	For observational studies, use statistical methods to achieve conditional exchangeability such as matching, propensity score methods, or adjustment for confounders. In experimental studies, use randomization to achieve exchangeability.
Errors in individual-level exposure measurements	Errors in measurement of exposure can occur due to various reasons such as instrumentation error, recall bias, or observer bias. These errors can lead to misclassification of exposure status and biased estimates of the effect size.	Implement high-quality protocols and training for measurement. Use objective measures of exposure where possible. Apply statistical methods to adjust for measurement error, e.g., regression calibration or simulation-extrapolation (SIMEX).
Errors in individual-level covariate measurements	Errors in measurement of covariates can lead to misclassification and biased estimates of the effect size. Additionally, if a confounder is measured with error, this can reduce the ability to control confounding.	Use objective measures of covariates where possible. Apply statistical methods to adjust for measurement error, e.g., multiple imputation or full information maximum likelihood (FIML).

Table D2

Common data analysis threats to valid causal interpretation of observational data [10,11,15].

Threat	Brief explanation	Suggested solutions
P-hacking ("researcher degrees of freedom")	Selectively choosing analyses or statistical tests, data points, or hypotheses until a desired *p*-value is achieved can create false positives.	Implement standardized analytic approaches, pre-registration of design and analysis, and replication studies. Discuss borderline or almost-significant results.
Multiple comparisons	Multiple comparisons can increase false positives if their *p*-values are not adjusted	Apply correction techniques such as Bonferroni correction or false discovery rate (FDR) control. Disclose all measured variables.
Spurious correlations	False correlations due to outliers or improperly pooled data, as in Simpson's Paradox.	Use robust methods like bootstrapping, check for violation of key assumptions, and screen for outliers.
Inflating the units of analysis	Researchers often confuse the number of subjects with the number of observations per subject, increasing degrees of freedom (df) and making it easier to find an apparently significant effect even if no true effect exists (false positives).	Mixed-effects linear models with within-subject variability as a fixed effect and between-subject variability as a random effect can help correct this issue without violating the assumption of independence.
Circular analysis (double dipping)	Use of the same dataset for feature selection and statistical hypothesis testing, inflating the statistical outcome.	Pre-define analysis criteria or use separate datasets for parameter specification and hypothesis testing. Use bootstrapping to maintain statistical power.
Inappropriate use of parametric tests and models; model misspecification	Parametric tests and models have assumptions about the distribution of the data (e.g., normality) and the forms of models (e.g., linear or generalized linear). If these assumptions are violated, the test or models may give inaccurate results.	If the data do not meet the assumptions of parametric tests, use non-parametric tests. Data transformations may also help to meet the assumptions.
Unverified modeling assumptions	Assumptions made during modeling, if incorrect, can lead to inaccurate and biased results. For instance, assumptions about model form, error distributions, measurement errors, interactions, or independence in regression models can lead to biased estimates of regression coefficients.	Perform diagnostic checks and plots to verify modeling assumptions. Use statistical tests to confirm assumptions where applicable. Consider alternative models if assumptions cannot be satisfied.
Confounding variables not fully controlled	Confounding variables can create spurious associations and erroneous conclusions. Confounding bias occurs when exposure and outcome share an uncontrolled common cause. This can lead to biased estimates of causal effects.	Control potential confounders via experimental designs or through statistical methods such as regression adjustment, matching, stratification, or inverse probability weighting.
Residual confounding	Residual confounding refers to remaining confounding bias after adjustment for observed confounders. This can be due to unmeasured ("latent') or poorly measured (e.g., categorized continuous) confounders.	Use sensitivity analysis to assess the potential impact of unmeasured confounding. In the case of poorly measured confounders, improve measurement procedures or use methods that can account for measurement error.
Measurement errors in individual-level exposures or in covariates	Measurement error can lead to misestimation of the effect size. In multivariate models, thee directions and sizes of errors and biases in effects estimates are hard to predict.	If measurement error is significant, more reliable measures should be used or the error should be incorporated into the statistical model, e.g., using errors-in-variables techniques.
Unmodeled interactions	Occurs when important interactions among variables are omitted or not modeled	Consider potential variable interactions. Use exploratory analysis, include interaction terms in models, apply machine learning techniques such as CART to detect and modelnon-linear interactions.

Table D3

Common errors in drawing sound causal conclusions from observational data [4,10,11,15].

Problem	Brief explanation	Suggested solutions
Causation claimed without appropriate design and analysis	This refers to making causal claims when the study design and analysis don't support valid causal inference, e.g., due to potential confounding factors.	Researchers should only claim causality when their study design and analysis allow it (see Tables 1 and 2).
Conflating correlation and causation	Misinterpreting findings of significant correlation or statistical association as evidence of causation.	Use causal language carefully, considering alternative explanations for associations. Avoid causal language when the evidence is merely correlational.
Interpreting differences between effects without directly comparing them	Researchers often infer a difference between effects based on different findings in different groups, rather than a direct statistical comparison of effects.	Directly compare groups for contrast. Monte Carlo simulations can be used to compare correlations, and ANOVA or other statistical techniques for group comparisons.
Over-interpreting non-significant results	Misinterpretation of non-significant p-values as indicating the absence of an effect, disregarding the possibility of insufficient data or lack of statistical power.	Report effect sizes along with p-values, consider using statistical approaches such as Bayesian statistics or equivalence tests. Don't over-interpret non-significant results.
Using unjustified identifying assumptions	Causal effects estimates rely on identifying assumptions about the data and underlying causal relationships that typically cannot be fully tested	Provide theoretical and empirical evidence for identifying assumptions. Encourage transparency in reporting these assumptions. Consider the robustness of results to plausible alternative assumptions.
Hypotheses contain unsupported causality	This occurs when hypotheses make causal statements that are not supported by the data or previous research.	Hypotheses should be grounded in existing research and data. Researchers should be cautious to only claim causality when the study design allows it, such as randomized controlled trials or quasi-experiments.
Unwarranted generalization, e.g., generalization to different populations (or different time periods, locations, or settings)	This refers to making broad claims about populations, settings, or time periods not directly studied in the research.	Researchers should clarify the population and conditions to which their findings apply, and avoid over-generalizing. Transportabilty formulas and external validity criteria for valid generalization of causal findings can clarify generalizability (Lesko et al., 2019 [13]). Further studies can validate generalizability.
Contradictory evidence is not mentioned	This occurs when researchers fail to acknowledge or discuss evidence that contradicts their findings.	Identify and discuss conflicting evidence. A comprehensive literature review can help identify contradictory evidence.

(*continued on next page*)

Table D3 (*continued*)

Problem	Brief explanation	Suggested solutions
Unwarranted conclusions and recommendations.	Conclusions, implications and recommendations for policy depend on unverified assumptions and/or do not follow from the data, analyses, and results presented. Limitations of study design, analyses, results, and conclusions not adequately discussed.	Carefully discuss study limitations and clarify any untested assumptions made in stating conclusions and recommendations.

Appendix E. BARD and ChatGPT responses to a general prompt for an article on gas stoves

As discussed in the text, Gruenwald et al. [5] provide a short article concluding that "12.7% of current childhood asthma nationwide is attributed to gas stove use… Given that this exposure is preventable, our study demonstrates that known mitigation strategies will lessen childhood asthma burden from gas stoves, particularly in states with elevated PAFs." In response to the generic high-level prompt "How can the technical quality of the following paper be improved, specifically including its treatment of causality?" the two LLMs BARD and ChatGPT4 generated the following responses (emphases added by the author to parts of the ChatGPT output):

BARD output:

"The technical quality of the paper can be improved in a few ways, specifically including its treatment of causality.

- The paper should more clearly define the causal relationship between gas stove use and childhood asthma. The paper does a good job of presenting the evidence that gas stove use is associated with an increased risk of childhood asthma. However, it does not go as far as to establish a causal relationship between the two. This is important because it means that other factors, such as exposure to secondhand smoke, could also be contributing to the increased risk of asthma. The paper could improve its treatment of causality by discussing the limitations of the available evidence and by considering alternative explanations for the observed association.
- The paper should discuss the potential for reverse causality. Reverse causality is a situation in which the cause and effect relationship is reversed. In the case of gas stove use and childhood asthma, it is possible that children with asthma are more likely to use gas stoves than children without asthma. This is because gas stoves can be a source of allergens and irritants, which can make asthma symptoms worse. The paper should discuss the potential for reverse causality and how it could affect the interpretation of the results.
- The paper should discuss the potential for confounding. Confounding is a situation in which an unmeasured or uncontrolled variable is associated with both the exposure and the outcome. In the case of gas stove use and childhood asthma, it is possible that some other factor, such as socio-economic status, is associated with both gas stove use and childhood asthma. This could make it difficult to determine whether gas stove use is actually causing the increased risk of asthma. The paper should discuss the potential for confounding and how it could affect the interpretation of the results.

Overall, the paper provides valuable information about the association between gas stove use and childhood asthma. However, the paper could improve its treatment of causality by discussing the limitations of the available evidence, the potential for reverse causality, and the potential for confounding.

In addition to the above, the paper could also be improved by:

- Using more consistent terminology. The paper uses a variety of terms to refer to the same thing, such as "gas stove use," "exposure to gas stoves," and "cooking with gas." This can make it difficult for the reader to follow the argument. The paper should use consistent terminology throughout.
- Providing more details about the methods used. The paper provides a brief overview of the methods used to calculate the population attributable fraction (PAF). However, it would be helpful for the reader to have more details about the specific methods used. This would allow the reader to evaluate the validity of the results.
- Including more discussion of the implications of the findings. The paper concludes by stating that gas stove usage should be considered in multi-faceted asthma prevention approaches. However, it would be helpful for the reader to have more discussion of the specific implications of the findings. For example, the paper could discuss how the findings could be used to develop public health interventions to reduce the burden of childhood asthma."

ChatGPT4 output:

"The paper seems to be well-structured, with a clear and concise introduction, methods, results, discussion, and conclusion. However, there are some areas for potential improvement, especially regarding the treatment of causality and general improvements.

Causality:

1. The paper **assumes a causal relationship** between gas stove use and childhood asthma, but the observational studies used to estimate effect sizes may not necessarily establish causality. It's important to clarify that the paper is relying on previously published causal assumptions. It would strengthen the argument to include discussion or additional analysis showing why these studies are believed to indicate a causal relationship.
2. While the paper uses odds ratios from studies in North America and Europe due to similar housing characteristics and gas-stove usage patterns, it would be more robust to **consider potential confounders specific to each region**. Adjusting for these confounders in the analysis would help increase the credibility of the causal inference.
3. There is a lack of discussion about potential **reverse causality**. It might be that people who are prone to respiratory problems are more likely to live in areas with gas stoves or lower-quality housing. Discussing this possibility would be useful.

General Improvements:

4. More details about the systematic literature search and selection process would enhance the paper's transparency and replicability. For instance, providing search terms, inclusion and exclusion criteria, and reasons for excluding certain studies would be helpful.
5. The paper uses **aggregate data**. Individual-level data would allow for more robust statistical analyses, control for potential confounders, and investigate potential interactions between different variables.
6. The **sample size** is limited to a few states, and it's **not clear whether these states are representative** of the U.S. The paper would benefit from an explanation or analysis showing why these states can be extrapolated to the whole country.
7. Consider discussing the robustness of your results to **potential measurement error**. For instance, the proportion of households with children exposed to gas stoves and the effect size of the gas stove use on childhood asthma might be measured with error.
8. There could be more discussion on the **limitations of the study**. For example, the paper does not consider possible changes in gas stove use or childhood asthma over time, other potential causes of childhood asthma, or factors that might interact with gas stove use.
9. The conclusions section could be strengthened by discussing specific recommendations for policy interventions, including any potential trade-offs, costs, and feasibility of such interventions. By addressing these issues, the paper can improve its causal inference and provide a more comprehensive and robust understanding of the relationship between gas stove use and childhood asthma."

Appendix F. ChatGPT responses to a general prompt for an article on air pollution and mortality

This appendix shows the detailed responses from ChatGPT4 to generic high-level prompts applied to the Results section of the Wu et al. paper discussed in the text.

Applying the prompt *"How can the technical quality of the following section be improved, specifically including its treatment of causality?"* to the Results section of the paper elicited the following detailed response from ChatGPT4.

ChatGPT 4.0 comments on the Results section of Wu et al. [20]:

"The current text provides a clear summary of the methods used in the study, the data sources, statistical approaches, and the results obtained. However, the discussion of causality could be enhanced to clarify the extent to which the relationships identified are believed to represent causal relationships, as well as the limitations of the methods used to draw causal inferences.

Here are some suggestions to improve the technical quality of the section and its treatment of causality:

1. **Definition of Causality**: The text should clearly define what it means by causality. In the context of this study, causality would typically mean that changes in the levels of PM2.5 directly bring about changes in mortality rates. This should be clarified, as readers might have different understandings of what causality entails.
2. **Assumptions of Causal Inference**: The methods used for causal inference—matching by GPS [generalized propensity score], weighting by GPS, and adjustment by GPS—have their own set of assumptions. The text should briefly discuss these assumptions. For example, these methods often assume unconfoundedness or exchangeability, meaning that, conditional on the observed covariates, the potential outcomes are independent of the treatment assignment. In addition, they often assume positivity, meaning that each person has a non-zero probability of receiving each level of treatment.
3. **Limitations and Biases**: It's essential to discuss the limitations and potential biases of your analysis and how they might affect your conclusions. This includes mentioning potential sources of confounding, measurement error, and selection bias. Also, talk about the "no unmeasured confounders" assumption that underlies many causal inference methods. It seems like the authors of the study did conduct a sensitivity analysis with *E*-values to explore the robustness of their results to unmeasured confounding, but they should elaborate on what they mean by unmeasured confounding and how it might bias their results.
4. **Extrapolation of Results**: The causal effect found in this study might be conditional on the specific cohort of Medicare enrollees used for the analysis and the specific time period. So, when interpreting the results, it's crucial to acknowledge these limitations and that extrapolating the results to other populations or time periods should be done cautiously.
5. **Comparisons of Different Approaches**: The text states that five different statistical approaches were used to estimate the effect of PM2.5 exposure on mortality, and they all provided consistent results. However, it's not clear whether these results should be interpreted as showing the same causal effect. The text should discuss more clearly whether these methods estimate the same parameter of interest, and if not, what the differences are.
6. **Hypothetical Interventions**: When discussing the hypothetical intervention of reducing PM2.5 levels to 10 μg/m3, it's crucial to point out that this is an idealized scenario. In reality, implementing such an intervention could have additional impacts that were not considered in the analysis.

By addressing these points, you could improve the clarity and quality of your discussion of causality, providing a more rigorous and nuanced understanding of your results."

We consider these comments to be fairly cogent, albeit perhaps somewhat generic. Regenerating the responses several times gave a range of comments. Some (e.g., to clarify the treatment of confounders and the rationale for selecting statistical methods) seemed somewhat generic. Nonetheless, it is clear that unaided ChatGPT4 can provide responsive comments with very little prompting, even for technically dense articles.

The prompts can easily be revised and extended. For example, to capture strengths as well as limitations of a paper, we could use the following prompt: *"Please prepare a summary of key strengths and limitations of the following section, specifically addressing its treatment of causality. First, please summarize strengths of the analysis and results. Second, please give any constructive suggestions for improving the technical quality of the section be improved, specifically including its treatment of causality. If the strengths and limitations have corresponding topics, they could be put in a left column and a right column for ease of comparison. Here is the section: … ."* Applying this revised prompt to the same text (the Result section of [20]) gave the following results (https://chat.openai.com/share/164a9ad3-f727-4659-a9d3-01b8e154009a):

Strengths:

Strengths	Limitations
1. **Extensive Data Set:** Utilized a vast dataset of 68.5 million Medicare enrollees from 2000 to 2016, providing a comprehensive view.	1. **Zip Code Level Analysis:** The utilization of zip code level data may mask finer grain disparities within the zip codes leading to ecological fallacy.
2. **Multiple Statistical Approaches:** Implemented five diverse statistical approaches to assess the effect of PM2.5 exposure on mortality, enhancing the reliability and robustness of the findings.	2. **Potential Residual Confounding:** Acknowledged potential residual confounding bias due to unmeasured confounders with time trends.
3. **Validated Prediction Model:** Employed a previously validated ensemble prediction model for estimating annual PM2.5 exposure, adding credibility to exposure measurements.	3. **Causal Framework Limitations:** Covariate balance (AC < 0.1) supports causal inferences, but causality is not definitively proven, and unmeasured confounders might still exist.
4. **Detailed Demographic Information:** Detailed demographic and socioeconomic data allow for robust adjustments and stratified analyses.	4. **Reliance on Publicly Available Data:** While the use of publicly available data ensures transparency, it may lack granularity and precision that more detailed private datasets might offer.
5. **Transparent Methods and Reproducibility:** All study data sources are publicly available, and the inclusion of R code facilitates reproducibility of methods.	5. **Addressing Causality:** The causal inference methods applied are robust, but causation would be more definitively addressed through randomized controlled trials, which are not feasible in this context.
6. **Comprehensive Confounder Adjustment:** The adjustment for multiple zip code– and county-level confounders and inclusion of meteorological variables enhances the reliability of the findings.	6. **Single Pollutant Analysis:** The focus on PM2.5 excludes the potential impacts of other pollutants and environmental factors, which could contribute to mortality.
7. **Sensitivity Analysis:** The inclusion of sensitivity analyses strengthens the validity of the findings and accounts for potential biases.	7. **Potential Bias in Exposure Assignment:** The use of zonal statistics to aggregate exposure estimates may lead to exposure misclassification bias.
8. **Clear and Consistent Results:** The study found consistent results across all statistical approaches, strengthening the reliability of the conclusions.	8. **Geographic Variability:** Results might not account for geographic variability in PM2.5 composition and its differential effects.

Constructive Suggestions:

1. **Granularity of Data:** Investigate possibilities for obtaining and analyzing data at a finer resolution than zip code level to avoid ecological fallacy and improve the accuracy of exposure assignments.
2. **Multiple Pollutant Models:** Incorporate other pollutants and environmental factors in the analysis to study their combined or individual effects on mortality.
3. **Further Exploration of Confounding:** Continue to explore and adjust for potential unmeasured confounders and conduct more extensive sensitivity analyses to address residual confounding.
4. **More Advanced Statistical Methods:** Explore more advanced statistical methods and machine learning techniques to uncover complex relationships and interactions between variables.
5. **Additional Subgroup Analyses:** Conduct more detailed subgroup analyses to understand disparities and differential effects within populations.
6. **Enhanced Causal Models:** Develop and apply more sophisticated causal inference models to strengthen causal claims, taking into account the limitations of observational data.
7. **Address Geographic Variability:** Consider geographic variability in PM2.5 composition and sources and study its differential impacts on health outcomes.
8. **Cross-validation with other Datasets:** Cross-verify the results and conclusions with other datasets or similar studies to ensure the generalizability and applicability of the findings

Again, we consider these comments to be responsive, although perhaps not very profound. Interestingly, continuing the chat via the follow-on prompt *"Please add comments where possible on quantitative bias analysis"* generated the following response, which we think is more useful:

Strengths and Limitations

Strengths	Limitations
1. **Comprehensive and Extensive Data Set:** The study utilizes a considerable volume of open cohort data, involving over 68.5 million Medicare enrollees. This comprehensive dataset contributes to the robustness and reliability of the study findings.	1. **Resolution of Data:** The data is available at the zip code level, which might not be granular enough to identify more localized effects or variations, possibly leading to ecological fallacy issues.
2. **Rigorous Statistical Approaches:** The implementation of five different statistical methods, including both traditional and causal inference approaches, strengthens the study's reliability and validity in estimating the effects of PM2.5 on mortality.	2. **Potential Unmeasured Confounding:** While extensive, there could be additional, unaccounted confounding variables impacting the associations, possibly skewing the results and affecting the causal inferences drawn. Some attempts were made to assess sensitivity to unmeasured confounding but it is unclear how extensive these were. 3. **Causal Claims:** Despite the use of rigorous methods, claiming strong causal inference from observational data is inherently challenging. The methods can adjust for observed confounders, but residual confounding might still be present, leading to potential biases in effect estimates.
3. **Validated Ensemble Prediction Model:** The application of a validated model for estimating PM2.5 exposure ensures accuracy and reliability in exposure measurement.	4. **Lack of Direct Quantitative Bias Analysis:** There seems to be no direct application of quantitative bias analysis to assess the robustness of the findings against potential biases. Quantitative bias analysis can help assess how different biases might quantitatively impact the study results.
4. **Consideration of Confounding Variables:** The study extensively considers a range of zip code– and county-level confounders, which adds rigor to the analysis and helps in isolating the effect of PM2.5 exposure on mortality.	5. **Limited to Measured Confounders:** The study's causal inference approaches are reliant on the potential outcomes framework and generalized propensity scores, which can only adjust for observed (measured) confounders. Thus, bias due to unobserved confounders remains a concern.
5. **Covariate Balance Assessment:** The study evaluates covariate balance, strengthening the validity of causal assessment and helping in ensuring that the groups being compared are similar on the observed covariates.	6. **Use of Surrogates:** Medicaid eligibility is used as a surrogate for individual-level SES, which might not adequately represent the socioeconomic status and could potentially introduce misclassification bias.
6. **Consistency Across Approaches:** The congruence of results across five different statistical approaches and different periods adds to the robustness and credibility of the findings.	

(*continued on next page*)

(*continued*)

Strengths	Limitations
7. Clear and Precise Quantification of Impact: The estimation of the total number of deaths that could be avoided by adhering to WHO guidelines provides clear and actionable insights, which can guide policy decisions. **8. Detailed Evaluation of Confounders:** A meticulous adjustment for multiple potential confounders ensures the reduction in confounding bias, enhancing the credibility of the causal relationships established.	**7. Potential Over-reliance on Statistical Significance:** There is a substantial reliance on statistical significance which might overshadow practical significance, and any p-hacking or multiple testing issues need careful consideration. **8. External Validity:** The study's findings, although robust, are primarily applicable to the elderly population and may not generalize well to other age groups or populations with different characteristics.

Suggestions for Improvement

1. **Enhance Resolution of Data:**
   - Obtain more granular data to identify localized effects and variations, reducing the risk of ecological fallacy.
2. **Apply Quantitative Bias Analysis:**
   - Perform robust quantitative bias analyses to evaluate the impacts of different biases on the study findings and ascertain the robustness of the results against potential biases.
3. **Refine Causal Inference Approaches:**
   - Explore advanced causal inference methods that can potentially handle unobserved confounding, and clearly state the assumptions under which causal interpretations can be made.
   - Employ instrumental variable methods or other techniques to attempt to deal with unobserved confounding.
4. **Enhance External Validity:**
   - Conduct similar studies in different populations or age groups to test the generalizability of the findings.
5. **Address Residual Confounding:**
   - Employ methods like sensitivity analyses to further assess the impact of potential residual confounding on the results.
6. **Refine Measurement of Confounders:**
   - Use more direct measures for individual-level SES and other confounders instead of surrogate markers to reduce misclassification bias.
7. **Further Sensitivity Analysis:**
   - Execute extensive sensitivity analyses to unmeasured confounding, considering various potential unmeasured confounders and assessing their possible impact on the study results.
8. **Multiple Testing Correction:**
   - Use appropriate corrections for multiple testing to avoid inflated type I error rate when employing multiple statistical approaches.
9. **Enhance Transparency:**
   - Clearly report all analyses performed, including any that did not yield significant results, to avoid selective reporting bias and enhance the reproducibility and credibility of the findings.

References

[1] Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu CM. Measurement error in nonlinear models a modern perspective. 2nd ed. New York, NY: Chapman and Hall/CRC; 2006. https://doi.org/10.1201/9781420010138.

[2] Cox Jr LA. Improving interventional causal predictions in regulatory risk assessment. Crit Rev Toxicol 2023 May 3;53(5):311–25. https://doi.org/10.1080/10408444.2023.2229923. PMID: 37489873.

[3] Ernst GW, Newell A. GPS: A Case Study in Generality and Problem Solving. New York, NY: Academic Press; 1969.

[4] Gerrits RG, Jansen T, Mulyanto J, van den Berg MJ, Klazinga NS, Kringos DS. Occurrence and nature of questionable research practices in the reporting of messages and conclusions in international scientific Health Services Research publications: a structured assessment of publications authored by researchers in the Netherlands. BMJ Open 2019 May 15;9(5):e027903. https://doi.org/10.1136/bmjopen-2018-027903. PMID: 31097488; PMCID: PMC6530378.

[5] Gruenwald T, Seals BA, Knibbs LD, Hosgood 3rd HD. Population Attributable Fraction of Gas Stoves and Childhood Asthma in the United States. Int J Environ Res Public Health 2022 Dec 21;20(1):75. https://doi.org/10.3390/ijerph20010075. PMID: 36612391; PMCID: PMC9819315.

[6] Haugeland J. Artificial intelligence: The very idea. Cambridge, MA: MIT Press; 1985. ISBN 0-262-08153-9.

[7] Hayes-Roth B. A blackboard architecture for control. Artificial Intelligence 1985;26(3):251–321. https://doi.org/10.1016/0004-3702(85)90063-3.

[8] Hosseini M, Horbach SPJM. Fighting reviewer fatigue or amplifying bias? Considerations and recommendations for use of ChatGPT and other large language models in scholarly peer review. Res Integr Peer Rev 2023 May 18;8(1):4. https://doi.org/10.1186/s41073-023-00133-5. PMID: 37198671; PMCID: PMC10191680.

[9] Hsu J. Scientists prefer feedback from ChatGPT to judgement by peers. New Sci 2023;(18 October). https://www.newscientist.com/article/2398051-scientists-prefer-feedback-from-chatgpt-to-judgement-by-peers/.

[10] Huntington-Klein N. The effect: An introduction to research design and causality. Boca Raton, Florida: CRC Press; 2022.

[11] Igelström E, Craig P, Lewsey J, et al. Causal inference and effect estimation using observational data. J Epidemiol Community Health 2022;76:960–6.

[12] Kıcıman E, Ness R, Sharma A, Tan C. Causal reasoning and large language models: Opening a new frontier for causality.. 2023. https://doi.org/10.48550/arXiv.2305.00050.

[13] Lesko CR, Buchanan AL, Westreich D, Edwards JK, Hudgens MG, Cole SR. Generalizing study results: A potential outcomes perspective. Epidemiology. 2017 Jul;28(4):553-561. doi: 10.1097/EDE.0000000000000664. Erratum in. Epidemiology. 2018 Mar;29(2). e16. PMID: 28346267; PMCID: PMC5466356.

[14] Lippenkova J. Overcoming the Limitations of Large Language Models How to enhance LLMs with human-like cognitive skills. https://towardsdatascience.com/overcoming-the-limitations-of-large-language-models-9d4e92ad9823; 2023.

[15] Makin TR, Orban de Xivry JJ. Ten common statistical mistakes to watch out for when writing or reviewing a manuscript. Elife. 2019 Oct 9;(8):e48175. https://doi.org/10.7554/eLife.48175. PMID: 31596231; PMCID: PMC6785265.

[16] Nilsson NJ. The quest for artificial intelligence. New York, NY: Cambridge University Press; 2009. ISBN 978-1-139-64282-8.

[17] Polesie S, Larkö O. Use of large language models: editorial comments. Acta Derm Venereol 2023 Feb 16;103. https://doi.org/10.2340/actadv.v103.9593. adv00874. PMID: 36794896; PMCID: PMC9949211.

[18] Rajabi E, Kafaie S. Building a disease knowledge graph. Stud Health Technol Inform 2023 May 18;(302):701–5. https://doi.org/10.3233/SHTI230243. PMID: 37203473.

[19] VanderWeele TJ, Hernán MA. Results on differential and dependent measurement error of the exposure and the outcome using signed directed acyclic graphs. Am J Epidemiol 2012 Jun 15;175(12):1303–10. https://doi.org/10.1093/aje/kwr458. PMID: 22569106; PMCID: PMC3491975.

[20] Wu X, Braun D, Schwartz J, Kioumourtzoglou MA, Dominici F. Evaluating the impact of long-term exposure to fine particulate matter on mortality among the elderly. Sci Adv 2020 Jul 17;6(29). https://doi.org/10.1126/sciadv.aba5692. eaba5692. PMID: 32832626; PMCID: PMC7439614.

[21] Yao M, Wei Y, Wang H. Promoting research by reducing uncertainty in academic writing: a large-scale diachronic case study on hedging in *Science* research articles across 25 years. Scientometrics 2023;128:4541–58. https://doi.org/10.1007/s11192-023-04759-6.

[22] Zhu Y, Hubbard RA, Chubak J, Roy J, Mitra N. Core concepts in pharmacoepidemiology: Violations of the positivity assumption in the causal analysis of observational data: Consequences and statistical approaches.

Pharmacoepidemiol Drug Saf 2021 Nov;30(11):1471–85. https://doi.org/10.1002/pds.5338. Epub 2021 Aug 24. PMID: 34375473; PMCID: PMC8492528.