When a Strong Association Does Not License an Effect Claim

A dual-track reanalysis of PM2.5 and life expectancy across countries

Everion Cao

GHEAS, Waterloo, Ontario, Canada

ORCID: 0009-0005-8238-8583

Version: Open Discussion Draft 1.0, 28 July 2026

Status: Not peer reviewed and not submitted. Comments on the data, design, models, evidence gates, and interpretation are invited through the public discussion page.

Abstract

Background

The first GHEAS paper found a clear negative cross-country association between ambient PM2.5 and life expectancy, but the coefficient nearly disappeared after adjustment and remained inconclusive in a conventional two-way fixed-effects model. We later used that study as a stress test for a method-first framework. The purpose of this paper is not to select a better-looking coefficient. It is to determine what, if anything, the available evidence licenses us to say.

Methods

We retained the original analysis as a frozen, reproducible baseline and added two pre-defined tracks. Track A treated the country-year panel as associational evidence and executed an 11-model universe covering lagged confounders, country-specific trends, changes, exposure windows, nonlinear form, alternative uncertainty estimators, and a future-exposure falsification test. Track B required a policy or natural-experiment assignment mechanism, a robust PM2.5 first stage, acceptable pre-trends, and negative-control checks before any health outcome could be analysed. Three public policy candidates were audited under fail-closed rules.

Results

In the 2023 cross-section, the unadjusted estimate was -1.15 years of life expectancy per 10 μg/m³ higher PM2.5 (95% CI -2.25 to -0.05). The fully adjusted estimate was -0.06 years (95% CI -0.57 to 0.45), an attenuation of 94.7%. The frozen conventional panel estimate was +0.047 years (95% CI -0.217 to 0.312). Track A's lagged-confounder primary model produced +0.401 years (95% CI 0.150 to 0.651), but the estimate changed to -0.038 years (95% CI -0.183 to 0.107) after country-specific trends. Future PM2.5 was also associated with current life expectancy (+0.364 years; 95% CI 0.121 to 0.607), failing the temporal falsification test. All four Track A evidence gates failed. Track B's estimator recovered a known effect in synthetic validation, but each real policy candidate failed before health estimation: the U.S. 1997 PM2.5 nonattainment design lacked a robust first stage, the U.S. NOx Budget Trading Program comparison failed the exposure pre-trend test, and the Indian catalytic-converter design was not robust to alternative comparisons and city trends.

Conclusions

The raw association is real as a description of these data, but the present country-level panel does not identify the direction or magnitude of a causal PM2.5 effect on life expectancy. A positive, negative, or near-zero coefficient can be produced by reasonable choices, and future exposure predicts the outcome in the same design. The correct scientific result is therefore an evidence boundary: Track A supports an E1 descriptive claim, while the empirical causal effect remains E0, ungraded. This does not imply that PM2.5 is harmless; it means that these data and designs cannot answer the causal question.

Keywords: PM2.5; life expectancy; ecological study; model uncertainty; falsification; difference-in-differences; evidence grading

1. Why return to this study?

Long-term exposure to fine particulate matter is supported by a large body of epidemiological and burden-of-disease research as an important public-health hazard [1-3]. That background evidence matters, but it cannot rescue a weak design or determine the result of a particular dataset in advance.

The first GHEAS analysis asked whether countries with higher annual PM2.5 had lower life expectancy and whether that relationship remained after adjustment for development context. It used 6,800 country-year observations from 200 countries or economies during 1990-2023. The longitudinal models used 4,406 observations from 186 countries during 2000-2023. The original paper was reproducible and appropriately cautious: it reported a strong raw pattern, major attenuation after adjustment, and an inconclusive primary panel estimate.

The later methodological work changed the question. Instead of asking which single model should be preferred, we asked whether the conclusion survived a registered universe of reasonable alternatives and whether any causal design passed its required data and falsification gates. The original paper is therefore preserved as evidence of what a conventional workflow produced, not overwritten as if it had never existed.

2. What was analysed?

2.1 Frozen baseline

The baseline outcome was period life expectancy at birth. The exposure was annual national mean PM2.5 concentration, scaled per 10 μg/m³. The cross-sectional adjustment sequence added gross domestic product per capita, urbanization, population aged 65 years or older, health expenditure, and region. The panel sequence used country fixed effects, year fixed effects, time-varying controls, lagged exposure, a three-year exposure mean, balanced-panel restriction, and exclusion of pandemic years. Country-clustered uncertainty was independently reproduced with a dummy-variable implementation.

These calculations remain valid as calculations. Their limitation is inferential: conventional adjustment does not by itself establish that the selected variables have the correct causal role, that national trends have been adequately controlled, or that the remaining within-country variation corresponds to an identifiable exposure intervention.

2.2 Track A: expose associational sensitivity

Track A was designed to show how the estimated association behaves under a complete, declared model universe. The primary model used country and year fixed effects with one-year-lagged GDP and urbanization. Additional models added country-specific linear trends, first differences, alternative exposure windows, nonlinear terms, alternative dependence corrections, balanced samples, and other registered changes. A future-exposure placebo asked whether next year's PM2.5 predicted current life expectancy. A non-zero future association would indicate that the model was still capturing trend, reverse timing, persistent confounding, or another non-causal structure.

Four gates controlled the language of the conclusion: stable direction across core models; core confidence intervals excluding zero; a future-exposure placebo compatible with zero; and at least one independently validated quantitative bias scenario. Failure did not trigger a search for a more favourable model. It automatically downgraded the claim.

2.3 Track B: require an assignment mechanism

Track B was reserved for causal questions. Before any health effect could be estimated, a candidate policy or natural experiment had to define treatment timing and geography, demonstrate a robust change in measured pollution, show acceptable pre-treatment behaviour, and pass registered falsification checks. The estimator and gates were first tested on deterministic synthetic data with a known truth.

Three empirical candidates were then audited:

Candidate	Apparent signal	Decisive audit result	Decision
U.S. 1997 PM2.5 NAAQS nonattainment	-0.728 μg/m³ difference-in-differences first stage	+0.054 μg/m³ after county trends; fuzzy-threshold assignment weak and imprecise	Stop before health
U.S. NOx Budget Trading Program	-1.151 ppb ozone change	Joint exposure pre-trend p=3.19×10^-7	Stop before health
India catalytic-converter requirements	Authors' saved five-year SPM trend break -48.56 μg/m³	Never-treated stacked estimate -12.81 (-46.28, 20.66); approximately zero with city trends	Stop before health

The last two candidates were not direct replications of the original life-expectancy panel. They were attempts to locate a stronger real-world identification design for the broader pollution-health method. Their failure is part of the evidence record.

3. What changed when the method became stricter?

3.1 The original pattern was strong but highly confounded

In the 2023 complete-case cross-section of 184 countries, 10 μg/m³ higher PM2.5 was associated with 1.15 fewer years of life expectancy before adjustment. Adding income alone reduced the estimate to -0.20 years and increased R^2 from 0.066 to 0.746. The fully adjusted estimate was -0.06 years, with a confidence interval spanning moderately negative and positive values.

The panel was equally sensitive to temporal structure. Country fixed effects alone gave -1.599 years per 10 μg/m³. Adding year fixed effects reversed the estimate to +0.391 years. Adding time-varying controls reduced it to +0.047 years. These reversals are not competing biological findings. They show that development and common time trends account for a large share of the available identifying variation.

3.2 Track A did not produce a stable independent association

Analysis	Estimate, years per 10 μg/m³	95% CI	What it shows
2023 unadjusted cross-section	-1.1475	-2.2488 to -0.0462	Strong raw pattern
2023 fully adjusted cross-section	-0.0604	-0.5678 to +0.4470	94.7% attenuation
Frozen v1 panel P3	+0.0471	-0.2174 to +0.3116	Conventional controlled model
Track A primary A1	+0.4007	+0.1500 to +0.6513	Lagged-confounder association
Track A country trends A2	-0.0379	-0.1828 to +0.1070	A1 signal disappears
Track A future-exposure placebo A8	+0.3636	+0.1207 to +0.6065	Temporal falsification failure

Viewed alone, A1 could be misread as evidence that higher PM2.5 improves life expectancy. That interpretation is not credible. The estimate disappears after country-specific trends, direction is not stable across the registered core models, and future exposure produces a similar positive association with the current outcome. The future cannot cause the past. The design is therefore detecting temporal structure that it has not successfully separated from the exposure-health relationship.

All four Track A gates failed. The registered output allows the descriptive statement that estimates are highly model- and trend-sensitive. It does not license an independent-effect or causal statement.

3.3 Track B produced three justified stops, not three null effects

The Track B software successfully recovered the known exposure reduction and health improvement in synthetic data and left the negative-control outcome near zero. This validated the implementation, not the PM2.5 hypothesis.

The real policy audits stopped before health outcomes. This distinction is essential. A failed exposure first stage or failed pre-trend test means that the design cannot identify the intended health effect; it does not mean that the true health effect is zero. Querying mortality after the gate failed would create a numerical answer without a defensible estimand.

For the U.S. PM2.5 NAAQS candidate, an apparent pollution reduction vanished after county trends and the threshold assignment was too weak for the registered design. For the NOx Budget Trading Program, treated and comparison areas had sharply inconsistent pre-policy exposure paths. For the Indian catalytic-converter candidate, the authors' saved second-stage arithmetic reproduced, but the effect was much smaller and imprecise against never-treated cities and disappeared with city-specific trends. In each case, stopping prevented an unsupported health coefficient from entering the literature.

4. What does the evidence permit us to conclude?

4.1 What is supported

The national data contain a clear negative cross-sectional relationship between PM2.5 and life expectancy. That relationship is largely entangled with income, demographic structure, health resources, region, and global development over time. Across reasonable longitudinal specifications, the direction and interval change materially. The future-exposure test also detects an association, demonstrating unresolved temporal structure.

The methodological conclusion is stronger than any substantive coefficient: a credible workflow must preserve all reasonable alternatives, test temporal falsification, and stop causal estimation when the exposure or assignment design fails.

4.2 What is not supported

The data do not support the claim that a 10 μg/m³ increase in PM2.5 changes national life expectancy by any particular number of years. They do not support a protective effect, despite the positive A1 coefficient. They do not support a zero effect, despite the near-zero trend-adjusted and conventional estimates. They do not support a causal effect in either direction.

These boundaries are study-specific. They do not overturn cohort studies, natural experiments, toxicological evidence, or burden models that use different outcomes, exposures, populations, and identifying assumptions [1-3]. A null or contradictory result in an ecological national panel cannot establish that PM2.5 is biologically harmless.

4.3 Why the outcome may be too distant

Period life expectancy is influenced by mortality at all ages and by many causes that change with economic development, health systems, conflict, infectious disease, behaviour, and demographic transition. National annual PM2.5 is also a coarse, partly modelled exposure that hides local gradients and personal exposure. Even a real pollution effect may be difficult to recover when both the exposure and outcome are broad aggregates and the identifying within-country variation is modest.

A stronger next test would use annual age-standardized all-cause and cardiorespiratory mortality, source-appropriate uncertainty intervals, finer exposure measurement, and a policy assignment with a verified first stage. The target, exposure window, negative controls, and stopping rules should be fixed before the

health result is visible.

5. Methodological lessons

First, reproducibility verifies arithmetic and provenance; it does not choose the correct estimand. Second, a complete model universe is more informative than a single preferred regression when several specifications are scientifically defensible. Third, falsification results must have authority: a future-exposure association should downgrade the claim even when the primary coefficient is conventionally significant. Fourth, causal language requires a real assignment mechanism and a demonstrated exposure change. Fifth, “stop” is a valid scientific output and must not be translated into “no effect.”

These rules are intended to generalize beyond air pollution. Any exposure-health analysis can produce a precise coefficient while relying on unresolved timing, measurement, selection, or model-form assumptions. A reusable system should make those assumptions machine-readable, test them where possible, and prevent an analyst or AI system from upgrading the claim after seeing an attractive result.

6. Strengths and limitations

Strengths include preservation of the original paper as a frozen benchmark; complete execution of the registered Track A model universe; explicit contradictory evidence; a temporal placebo; independent numerical checks; synthetic validation of the Track B estimator; public-data audits of three real policy candidates; and fail-closed claim licensing.

The limitations are substantial. The Track A protocol was created after results from the earlier study were visible, so this is not a fully prospective confirmation. National averages cannot resolve individual effects or subnational inequality. PM2.5 and life-expectancy measurement error were audited but not quantitatively propagated. Missingness and source-stage selection remain only partly characterized. The candidate policy audits used available public data and may not exhaust all defensible comparison groups or exposure definitions. Track B has not yet produced an eligible empirical PM2.5-health dataset. The current paper is therefore a methodological discussion draft, not a definitive pollution-risk estimate.

7. Conclusions

The original GHEAS PM2.5 study reported its result cautiously, but the later method-first analysis reveals a sharper boundary. The crude association is strong; the adjusted estimates are unstable; a positive primary Track A coefficient disappears with country trends; future exposure predicts current life expectancy; and three candidate causal designs fail before health estimation.

The most defensible conclusion is not that PM2.5 increases life expectancy, reduces it by a particular number of years, or has no effect. It is that the present country-level data and audited policy candidates do not identify the direction or magnitude of a causal effect on life expectancy. Preserving that uncertainty is not a failure to reach a result. It is the result the evidence supports.

Data, code, and transparency

The frozen baseline data, model outputs, verification records, Track A model universe, Track B validation fixture, policy-candidate audits, protocols, hashes, and decision records are retained in the GHEAS project. No missing exposure or health result was imputed to complete a preferred analysis. Failed gates and contradictory coefficients were preserved.

Everion Cao directed the research question, analysis principles, and final judgement. OpenAI Codex, referred to within the project as Jace, assisted with programming, execution, checking, and drafting under a rule that

prohibited AI from strengthening causal language or selecting a preferred result outside the registered gates.

Competing interests

The author declares no competing interests.

References

[1] World Health Organization. WHO global air quality guidelines: particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide. Geneva: WHO; 2021. ISBN 978-92-4-003422-8. https://www.who.int/publications/i/item/9789240034228

[2] Pope CA III, Ezzati M, Dockery DW. Fine-particulate air pollution and life expectancy in the United States. New England Journal of Medicine. 2009;360:376-386. https://doi.org/10.1056/NEJMsa0805646

[3] Apte JS, Brauer M, Cohen AJ, Ezzati M, Pope CA III. Ambient PM2.5 reduces global and regional life expectancy. Environmental Science & Technology Letters. 2018;5:546-551. https://doi.org/10.1021/acs.estlett.8b00360

[4] World Bank. World Development Indicators. Country-year indicator data retrieved and versioned by GHEAS in July 2026. https://databank.worldbank.org/source/world-development-indicators

[5] Cameron AC, Miller DL. A practitioner's guide to cluster-robust inference. Journal of Human Resources. 2015;50:317-372. https://doi.org/10.3368/jhr.50.2.317

[6] Hernán MA, Robins JM. Causal Inference: What If. Boca Raton: Chapman & Hall/CRC; 2020. https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/

[7] Callaway B, Sant'Anna PHC. Difference-in-differences with multiple time periods. Journal of Econometrics. 2021;225:200-230. https://doi.org/10.1016/j.jeconom.2020.12.001

[8] GHEAS. GHEAS Auditable Evidence Framework, version 1.0. Waterloo, Ontario; 2026. https://gheas.com/methodology

[9] GHEAS. PM2.5 and population health methods study v2: protocols, analysis universe, dual-track reconciliation, and policy-candidate audit records. Waterloo, Ontario; 2026.