

当显著相关不足以支持“效应”结论

PM2.5 与期望寿命跨国数据的双轨再分析

Everion Cao

GHEAS, 加拿大安大略省滑铁卢

ORCID : 0009-0005-8238-8583

版本：公开讨论稿 1.0, 2026 年 7 月 28 日

状态：本文未经同行评审，也未投稿。欢迎通过公开讨论页面，对数据、研究设计、模型、证据门槛和结论提出批评。

摘要

研究背景

GHEAS 的第一篇论文发现，跨国数据中 PM2.5 与期望寿命存在明显负相关；但加入发展水平等因素后，系数几乎消失，在传统的双向固定效应模型中也没有得到明确结果。后来，我们把这项研究作为“方法优先”框架的压力测试。本研究的目的不是寻找一个更漂亮的系数，而是判断：现有证据究竟允许我们说到哪一步。

研究方法

我们把原分析完整保留为冻结、可复现的基线，同时增加两条预先定义的轨道。轨道 A 把国家一年份面板视为关联性证据，完整运行 11 个模型，比较滞后混杂因素、国家特异趋势、变化模型、不同暴露窗口、非线性形式、不同不确定性估计，以及“未来暴露安慰剂”检验。轨道 B 只用于因果问题：在分析任何健康结局以前，必须先有明确的政策或自然实验分组机制、稳健的 PM2.5 暴露变化、可接受的干预前趋势和阴性对照。我们按照“门槛不通过就停止”的规则，审计了三个公开政策候选。

主要结果

在 2023 年横断面中，PM2.5 每升高 10 微克/立方米，未经调整时对应期望寿命减少 1.15 年（95% CI -2.25 至 -0.05）；完全调整后为 -0.06 年（95% CI -0.57 至 0.45），绝对值衰减 94.7%。冻结基线的传统面板估计为 +0.047 年（95% CI -0.217 至 0.312）。轨道 A 主模型得到 +0.401 年（95% CI 0.150 至 0.651），但加入国家特异线性趋势后变为 -0.038 年（95% CI -0.183 至 0.107）。更关键的是，下一年的 PM2.5 也能“预测”本年的期望寿命（+0.364 年；95% CI 0.121 至 0.607），未来暴露安慰剂检验失败。轨道 A 的四项证据门槛全部未通过。轨道 B 的估计程序可以在模拟数据中恢复已知真值，但三个真实政策候选都在健康分析前被停止：美国 1997 年 PM2.5 未达标地区设计缺少稳健的暴露第一阶段；美国氮氧化物预算交易计划的比较组未通过暴露前趋势检验；印度机动车催化转化器政策结果对比较组和城市趋势不稳健。

结论

原始负相关确实存在于这批数据中，但现有国家层面面板无法识别 PM2.5 对期望寿命因果效应的方向和大小。采用不同但合理的分析选择，可以得到正、负或接近零的系数；而且“未来暴露”也能产生类似关系。因此，正确的科研结果不是挑出某一个系数，而是划出证据边界：轨道 A 只支持 E1 级描述性结论，经验因果效应仍为 E0，无法评级。这不意味着 PM2.5 无害，而是说明这批数据和这些设计无法回答因果问题。

关键词：PM2.5；期望寿命；生态学研究；模型不确定性；反证检验；双重差分；证据分级

1. 为什么要重新讨论这项研究？

大量流行病学和疾病负担研究支持：长期暴露于细颗粒物是重要的公共卫生危害[1-3]。这些背景证据非常重要，但它们不能替某一项具体研究修补设计缺陷，也不能提前决定一批数据应该得出什么结果。

GHEAS 的第一项分析研究了两个问题：PM2.5

较高的国家是否有更低的期望寿命；加入国家发展背景后，这种关系是否仍然存在。原研究使用了 1990-2023 年间 200 个国家或经济体的 6,800 条国家一年份记录；纵向模型使用 2000-2023 年间 186 个国家的 4,406 条记录。原论文的计算可以复现，表述也相对谨慎：未经调整的关系很强，调整后大幅减弱，主面板模型不能得出明确结果。

后来，方法学工作的重点发生了变化。我们不再问“哪一个模型最好”，而是问：结论能否经受一整套合理替代模型；如果要使用因果语言，是否有真实的分组机制、暴露变化和反证检验支持。于是，原论文被保留为传统分析流程的冻结基线，而不是被覆盖或假装从未存在。

2. 我们分析了什么？

2.1 冻结基线

结局指标是出生时期望寿命，暴露指标是国家年度平均 PM2.5，每 10 微克/立方米为一个单位。横断面模型依次加入人均国内生产总值、城市化、65 岁及以上人口比例、卫生支出和地区。面板模型依次比较国家固定效应、年份固定效应、时变控制变量、滞后暴露、三年平均暴露、平衡面板限制和排除疫情年份。主模型的国家聚类不确定性还通过另一套虚拟变量程序独立复算。

这些计算本身仍然有效。问题在于推断：传统调整并不能自动证明变量的因果角色放置正确，也不能保证国家长期趋势已经处理充分，更不能证明剩余的国家内部变化对应一个可以解释的污染干预。

2.2 轨道 A：把关联结果的敏感性完整暴露出来

轨道 A 的目标不是制造因果结论，而是让估计值在完整、事先声明的模型集合中接受检验。主模型使用国家和年份固定效应，并加入滞后一期的人均 GDP 和城市化。其他模型进一步加入国家特异线性趋势、一阶差分、不同暴露窗口、非线性项、不同相关结构处理和平衡样本。

其中最重要的一项是“未来暴露安慰剂”：用下一年的 PM2.5 解释本年的期望寿命。如果未来暴露仍然显著，说明模型仍在捕捉长期趋势、时间倒置、持续混杂或其他非因果结构。

结论语言受到四道门槛约束：核心模型方向一致；核心置信区间都排除零；未来暴露与当前结局相容于零；至少存在一个经过独立验证的定量偏倚情景。任何门槛失败，都不能转而寻找更有利的模型，而是自动降低结论等级。

2.3 轨道 B：必须先有真实的分组机制

轨道 B 只处理因果问题。在估计健康效应以前，一个政策或自然实验必须明确干预时间和地区，证明污染物确实发生稳健变化，显示可接受的干预前趋势，并通过预设反证检验。我们先用真值已知的确定性模拟数据验证估计程序和门槛。

随后审计了三个真实候选：

候选设计	表面信号	决定性审计结果	决定
美国 1997 年 PM2.5 NAAQS 未达标地区	双重差分第一阶段 -0.728 微克/立方米	加入县趋势后 +0.054 微克/立方米；阈值附近分组弱且不精确	健康分析前停止
美国氮氧化物预算交易计划	臭氧变化 -1.151 ppb	暴露联合前趋势 $p=3.19 \times 10^{-7}$	健康分析前停止
印度机动车催化转化器要求	原作者保存系数推得五年 SPM 趋势改变 -48.56 微克/立方米	从未处理城市堆叠估计 -12.81 (-46.28, 20.66)；加入城市趋势后接近零	健康分析前停止

后两个候选并不是对原期望寿命面板的直接复算，而是为了整个污染—健康方法寻找更强的现实识别设计。它们没有通过门槛，本身就是证据记录的一部分。

3. 当方法变严格以后，结果发生了什么变化？

3.1 原始关系很强，但与国家发展背景高度纠缠

在 2023 年 184 个完整国家的横断面中，PM2.5 每升高 10 微克/立方米，未经调整时对应期望寿命减少 1.15 年。只加入收入，估计就缩小到 -0.20 年，决定系数 R^2 从 0.066 升到 0.746。完全调整后，估计为 -0.06 年，置信区间同时容许一定程度的负值和正值。

面板分析对时间结构同样敏感。只加入国家固定效应时，估计为 -1.599 年；加入年份固定效应后立刻翻转为 +0.391 年；再加入时变控制变量后缩小到 +0.047 年。这些翻转不是互相竞争的生物学发现，而是在告诉我们：国家发展进程和全球共同时间趋势占据了大量可用于识别的变化。

3.2 轨道 A 没有得到稳定的独立关联

分析	每 10 微克/立方米对应年数	95% CI	说明
2023 年末调整横断面	-1.1475	-2.2488 至 -0.0462	原始关系很强
2023 年完全调整横断面	-0.0604	-0.5678 至 +0.4470	衰减 94.7%
冻结基线面板 P3	+0.0471	-0.2174 至 +0.3116	传统控制模型
轨道 A 主模型 A1	+0.4007	+0.1500 至 +0.6513	滞后混杂因素后的关联
轨道 A 国家趋势 A2	-0.0379	-0.1828 至 +0.1070	A1 信号消失
轨道 A 未来暴露安慰剂 A8	+0.3636	+0.1207 至 +0.6065	时间反证检验失败

如果只看 A1，很容易把它误读为“PM2.5 越高，期望寿命反而越长”。这个解释并不可信：加入国家特异趋势后，结果消失；核心模型方向不一致；而且未来 PM2.5 对当前期望寿命产生了相似的正关联。未来不能导致过去，因此，这套设计仍在捕捉没有被分离出来的时间结构。

轨道 A 四道门槛全部失败。可以发布的结论是：结果高度依赖模型与趋势处理；不能发布的结论是：PM2.5 存在独立效应或因果效应。

3.3 轨道 B 得到的是三个“有理由的停止”，不是三个“零效应”

轨道 B 程序能在模拟数据中恢复已知的暴露下降和健康改善，同时让阴性对照保持接近零。这验证了程序实现，并没有验证现实中的 PM2.5 假设。

三个真实政策审计都在健康结局前停止。这里必须分清：第一阶段或前趋势失败，意味着设计不能识别预定的健康效应；它不等于真实健康效应为零。门槛失败以后仍然去查询死亡数据，只会得到一个没有可靠估计目标的数字。

美国 PM2.5 NAAQS 候选的表面污染下降在加入县趋势后消失，阈值分组也太弱；氮氧化物预算交易计划中的处理地区和比较地区在政策前已有明显不同的暴露轨迹；印度催化转化器候选可以复现原作者保存系数的二阶段计算，但改用从未处理城市后，效应明显缩小且不精确，加入城市趋势后完全消失。三次停止都阻止了一个不受设计支持的健康系数进入结果。

4. 现有证据允许我们得出什么结论？

4.1 可以支持的结论

国家层面数据中，PM2.5 与期望寿命存在明显负相关。但这种关系与收入、人口结构、卫生资源、地区和全球发展进程高度纠缠。在合理的纵向模型之间，方向和区间都会发生实质变化；未来暴露检验同样出现关联，说明时间结构仍未解决。

比任何单个系数都更可靠的方法学结论是：可信工作流必须保留合理替代模型，必须进行时间反证检验；当暴露第一阶段或分组设计失败时，必须停止因果估计。

4.2 不能支持的结论

现有数据不能支持“PM2.5 每升高 10 微克/立方米，会使国家期望寿命改变多少年”的数值结论。尽管 A1 是正数，也不能说明 PM2.5 有保护作用；尽管趋势模型和传统模型接近零，也不能说明没有效应；正负两个方向的因果表述都不成立。

这些边界只针对本研究。它们不会推翻采用不同结局、暴露、研究人群和识别假设的队列研究、自然实验、毒理证据或疾病负担模型[1-3]。一项国家生态学面板得到矛盾或接近零的结果，不能证明 PM2.5 在生物学上无害。

4.3 期望寿命可能距离污染作用点太远

时期望寿命受到所有年龄、所有死因的影响，也同时受到经济发展、卫生体系、战争与冲突、传染病、行为和人口转变的塑造。国家年度平均 PM2.5 又是一种粗略、部分依赖模型估计的暴露，掩盖了地区差异和个人暴露。即使污染确有作用，当暴露和结局都高度聚合、国家内部有效变化又有限时，也可能很难从中识别。

下一步更合适的检验，应使用年度年龄标化全因死亡和心肺死亡、与数据来源匹配的不确定性区间、更细的暴露测量，以及具有可靠第一阶段的政策分组。估计目标、暴露窗口、阴性对照和停止规则都应在看到健康结果前锁定。

5. 这项研究带来了哪些方法学经验？

第一，可复现性能够验证计算和数据来源，但不能替我们选择正确的估计目标。第二，当多种模型都有科学合理性时，完整模型集比一个“首选模型”更有信息。第三，反证必须真正有权力改变结论：即使主系数达到传统显著性，未来暴露检验失败也必须降级。第四，因果语言必须建立在真实分组机制和明确暴露变化上。第五，“停止”是合法的科研结果，不能被翻译成“没有效应”。

这些规则并不限于空气污染。任何暴露—健康研究，都可能在时间、测量、选择或模型形式仍未解决时算出一个非常精确的系数。可复用的分析系统应当把这些假设机器可读化，在可能时直接检验，并禁止分析者或 AI 因为看到漂亮结果而自行升级结论。

6. 优势与限制

本研究的优势包括：保留原论文作为冻结基线；完整运行轨道 A 的注册模型集合；保留相互矛盾的证据；加入未来暴露反证；独立复核数值；用模拟真值验证轨道 B；审计三个真实政策候选；以及采用“门槛失败即降级或停止”的结论许可机制。

限制同样明显。轨道 A 方案形成时，团队已经看到早期结果，因此这不是完全前瞻的验证。国家平均值不能识别个人效应和国内不平等。PM2.5 与期望寿命的测量误差虽然被记录，但尚未定量传播；缺失和数据源阶段的选择也只得到部分审计。政策候选审计依赖当前公开数据，不能保证穷尽所有合理比较组或暴露定义。轨道 B 至今没有得到一个符合要求的现实 PM2.5—健康数据集。因此，这是一篇方法学讨论稿，不是最终的污染风险估计。

7. 结论

原来的 GHEAS PM2.5

论文已经相当谨慎，但方法优先分析给出了更清楚的证据边界：粗相关很强，调整后估计不稳定；轨道 A 的正向主系数在加入国家趋势后消失；未来暴露能“预测”当前期望寿命；三个候选因果设计都在健康分析前失败。

因此，最合理的结论不是 PM2.5

会延长期望寿命、会缩短多少年，或没有任何作用；而是现有国家层面数据和已审计政策候选无法识别 PM2.5 对期望寿命因果效应的方向和大小。保留这种不确定性，不是没有得到结果；它正是证据所支持的结果。

数据、代码与透明度声明

GHEAS 保存了冻结基线数据、模型输出、独立复核记录、轨道 A 模型集合、轨道 B 模拟验证、三个政策候选审计、研究方案、文件哈希和决策记录。没有为了完成某个偏好分析而填补缺失暴露或健康结果；失败的门槛和相互矛盾的系数均被保留。

Everion Cao 决定研究问题、分析原则和最终判断。OpenAI Codex（项目中称为 Jace）协助编程、执行、复核和文字起草，但不得自行提高因果语言，也不得在注册门槛以外选择偏好的结果。

利益冲突

作者声明不存在利益冲突。

参考文献

- [1] World Health Organization. WHO global air quality guidelines: particulate matter (PM_{2.5} 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 PM_{2.5} 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. GHEAS 于 2026 年 7 月获取并保存国家—年份指标。 <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 可审计证据框架, 1.0 版。加拿大安大略省滑铁卢; 2026。 <https://gheas.com/methodology>
- [9] GHEAS. PM_{2.5} 与人群健康方法学研究 v2: 研究方案、分析集合、双轨协调报告和政策候选审计记录。加拿大安大略省滑铁卢; 2026。