TOWARDS AUTONOMOUS ENVIRONMENTAL HEALTH: AERIS, A RESEARCH AGENT FOR AIR-POLLUTION EXPOSURE ASSESSMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Air pollution harms human health and contributes to millions of deaths each year, yet many public-health studies omit it even when they record the locations and times needed to estimate exposure. AERIS (Airborne Exposure Research & Investigation System) equips command-line agents with the tools and instructions to compute exposure from public air-quality measurements and study location data. To test the harness on a real-world, end-to-end, long-horizon research task, we introduce *source-substitution reanalysis*: AERIS rebuilds air-pollution datasets from public measurements, links them to published health outcomes, and reruns the statistical analyses. A structured literature search identified nine published studies as candidates for reanalysis based on open health-outcome data, sufficient methodological detail, and adequate air-quality coverage across the required locations and periods. Across these studies, AERIS processed 1.85 million air-quality monitor records into 20,568 area-time exposure estimates and 14 pollution–health comparisons across four pollutants. It matched the direction of all seven published positive associations; the median study-level difference between AERIS and published log effect ratios was 0.040. We then ran autonomous reanalyses as a robustness control and repeated them with the same model stripped of AERIS tools, instructions, and exposure engine as a harness ablation. Self-driving AERIS completed 23 of 27 attempts, compared with 7 of 27 for the raw agent, although its median log-effect difference rose to 0.169 without researcher review. AERIS turns air-pollution exposure assessment into a reviewable agent workflow and opens a path toward autonomous environmental-health research.

1 INTRODUCTION

Air pollution is among the most consequential environmental threats to health. The 2024 *State of Global Air* assessment attributed 8.1 million deaths in 2021 to ambient and household air pollution (Health Effects Institute, 2024). Decades of epidemiological research link short- and long-term exposure to cardiovascular and respiratory disease, hospitalization, and mortality (Dockery et al., 1993; Di et al., 2017; Brook et al., 2010; Brunekreef & Holgate, 2002). Its relevance extends well beyond studies centered on pollution. Research on weather, infectious disease, cardiovascular care, mental health, treatment, or recovery may produce biased estimates when air pollution is associated with both the primary exposure and the health outcome. Measuring pollution can reduce this confounding, sharpen the relationship under study, and reveal health effects and biological pathways that warrant further investigation.

Many public-health studies already record residential addresses, ZIP codes, administrative areas, observation dates, or GPS traces from smartphones and wearables. Converting these records into pollution exposure requires technical skills and domain expertise. Investigators must select a data source, geographic assignment method, temporal aggregation, and completeness standard, then document those choices for review. The research question guides the choice among air-quality data sources and models: a residential study may use monitor-based exposure or a hyperlocal dispersion model at each participant's home, whereas a study of daily activity may require estimates along each person's route (Nyhan et al., 2019; Steinle et al., 2013). Government monitors provide trusted observations but may not align with where and when people are exposed (Liang & Kumar, 2013);

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

low-cost sensors add local detail but require calibration (Morawska et al., 2018; Barkjohn et al., 2021); and models fill geographic gaps through assumptions that affect resolution and accuracy (Hoek et al., 2008; van Donkelaar et al., 2010; Di et al., 2019).

Researchers rebuild these exposure pipelines for every new study, often across disconnected scripts and data services. This preparation delays the first statistical analysis, keeps air pollution out of studies that could measure it, and slows research on its relationship with disease, treatment, and recovery. Greater autonomy would let more studies account for air pollution, reduce the specialist time spent rebuilding routine analyses, and help researchers test environmental explanations that would otherwise remain unexplored. Scientific review remains essential because source choice, geographic coverage, and fallback analyses can materially change the estimated health effect.

AERIS (Airborne Exposure Research & Investigation System) is a research harness that helps researchers add air-pollution exposure to public-health studies through their preferred command-line agent. The researcher defines the pollutant, locations, dates, source, and assignment method; AERIS retrieves the data, computes exposure, assesses coverage, and prepares the result for analysis.

Long-running research agents are difficult to judge through isolated questions or tool calls. We introduce source-substitution reanalysis as an outcome-based evaluation: the published study fixes the scientific question, while the agent must independently recover public data, rebuild exposure, and rerun the analysis. We screened 18,919 PubMed and OpenAlex records, identified 21 candidates, and attempted the 14 with sufficient public inputs and methods. Nine had retrievable outcomes and adequate historical air-quality coverage; three others lacked adequate coverage or spatial variation, and two lacked a retrievable outcome export. Agreement with the published conclusions, directions, and effect sizes tests retrieval, alignment, exposure computation, modeling, and documentation as one scientific workflow (Chen et al., 2025; Siegel et al., 2025).

Our primary evaluation reflects intended use: a scientist supervises AERIS as it conducts each reanalysis. We then run two autonomous experimental controls, each repeated across three independent nine-study campaigns: self-driving AERIS measures robustness without researcher oversight, while a matched raw-agent ablation isolates the contribution of the harness beyond the underlying model.

Our contributions are: (i) a research-agent harness for computing exposure at fixed locations, along mobility trajectories, and across administrative areas; (ii) source-substitution reanalysis as an outcome-based evaluation, applied to nine published studies with reviewable exposure, analysis, and provenance artifacts; and (iii) autonomous robustness and raw-agent ablation controls that measure self-driving reliability and isolate the contribution of the harness.

2 Related work

Agents for science and reproducible research. Scientific agents connect language models to specialized software, data resources, and laboratory interfaces. Coscientist and ChemCrow established this pattern in chemistry (Boiko et al., 2023; Bran et al., 2024); later systems extend it to biomedical research and data analysis (Gottweis et al., 2026; Bu et al., 2026). Executable benchmarks such as MLE-bench evaluate complete agent workflows and show that model scaffolding materially affects performance (Chan et al., 2025). AERIS extends this approach to scientific reanalysis, where success also requires valid data reconstruction, provenance, and agreement with external findings across independent trajectories. Yet ScienceAgentBench and CORE-Bench show that long-horizon, data-driven work remains difficult (Chen et al., 2025; Siegel et al., 2025), while reproducibility studies find that public releases rarely capture every decision needed to rebuild an analysis (Stodden et al., 2018; Seibold et al., 2021). Source-substitution reanalysis complements task-based benchmarks by evaluating a complete workflow against external scientific findings while requiring the agent to reconstruct its own inputs.

Exposure assessment for epidemiology. Environmental epidemiology provides monitor interpolation, land-use regression, dispersion models, and hybrid monitor–satellite products (Jerrett et al., 2005; Ryan & LeMasters, 2007; Kloog et al., 2014). Large studies assign modeled pollution to geocoded homes (Eeftens et al., 2012; Szpiro et al., 2010); mobility-aware methods link time-stamped locations to changing conditions (Nyhan et al., 2016; Yu et al., 2020); and hyperlocal air-quality models combine sensors, weather, and built-environment data to estimate concentrations between

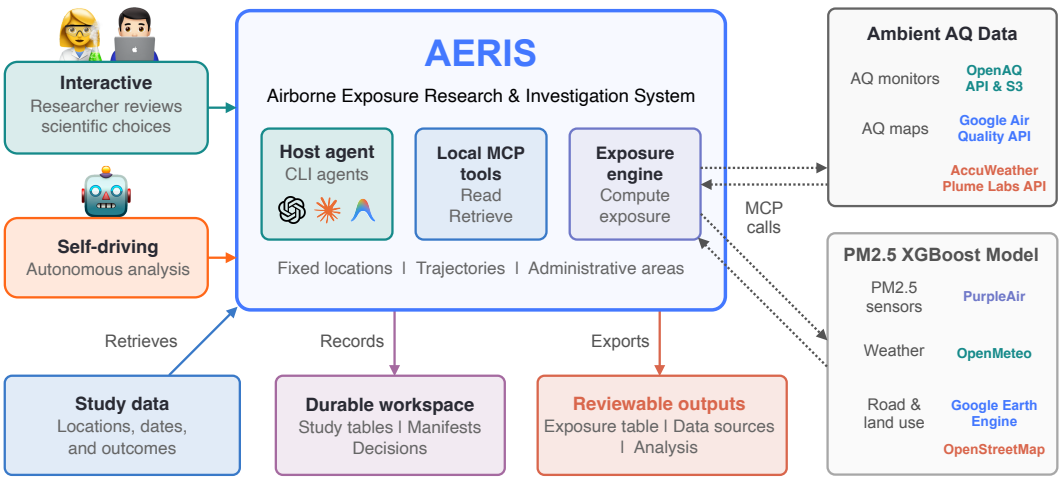

Figure 1: AERIS architecture. In interactive or self-driving mode, a command-line agent loads study data, queries ambient sources, and computes exposure through the AERIS skill and local tool server. Both modes share a workspace that preserves the inputs, decisions, coverage, provenance, and outputs required for scientific review.

regulatory stations (Jauvion et al., 2020; Alléon et al., 2020; Cassard et al., 2020; Barkjohn et al., 2021; Morawska et al., 2018). Existing tools support residential exposure histories, personal exposure tracking, and population-level health-impact assessment (Durney et al., 2026; Donaire-Gonzalez et al., 2019; Breen et al., 2019; Sacks et al., 2018). Researchers must still translate a study design into source selection, spatial assignment, quality checks, and provenance across these methods and tools. AERIS coordinates that work and turns study locations and dates into an exposure record that fits the scientific question.

Environmental health evidence and study design. Landmark cohort and multicity studies connect ambient particles with mortality and hospital admission (Dockery et al., 1993; Samet et al., 2000; Dominici et al., 2006), cardiopulmonary mortality (Pope et al., 2002; Di et al., 2017), and natural-cause mortality (Beelen et al., 2014). Systematic reviews extend the evidence across $PM_{2.5}$, PM_{10}, NO_2, ozone, mortality, and pediatric respiratory outcomes (Orellano et al., 2020; Khreis et al., 2017). Every estimate depends on choices about timing, covariates, geographic scale, and sensitivity analysis. AERIS records the exposure-related choices beside the resulting values so investigators can review and revise the analysis.

3 AERIS System Design and Exposure Assessment

3.1 What a usable exposure record requires

A reviewable exposure record pairs each pollutant estimate with the evidence and choices that produced it. AERIS starts from a participant table, time-stamped trajectory, or administrative boundary. The researcher defines the pollutant, time window, geographic representation, assignment rule, and completeness standard. AERIS retrieves measurements from public air-quality repositories such as OpenAQ, assigns them to the requested places and times, and aggregates them at the study's unit of analysis. The result records the source, contributing monitors or boundary, coverage, settings, and retrieval decisions.

AERIS supports three forms of location data: fixed locations for participants or study sites, mobility trajectories for time-varying participant locations, and administrative geographic areas for population-level analyses. Before retrieval, AERIS records the question, outcome, pollutant, location unit, time window, source, assignment rule, and completeness threshold. The agent uses this record when selecting data, assigning exposure, and evaluating coverage.

3.2 How AERIS works

AERIS separates scientific reasoning from reproducible computation across four components. This separation of concerns lets research teams use their preferred command-line agent while preserving a stable exposure workflow and reviewable study record.

1. **Host agent and AERIS skill.** The researcher starts AERIS with `aeris.sh` and selects Codex, Claude Code, Gemini CLI, Antigravity CLI, or another compatible command-line agent. The launcher opens that host in the selected study directory, supplies the AERIS workflow instructions, and connects the local tool server. The host manages conversation, planning, and tool selection.
2. **Local tool server and exposure tools.** AERIS uses the Model Context Protocol (MCP) to expose typed tools over local stdio. Any compatible host can work with public air-quality repositories and study data through one interface. Its tools inspect inputs, approve exposure requests, retrieve observations, assign exposure, diagnose coverage, export outputs, and retain decisions. Shared definitions keep these operations consistent across hosts and independent of chat history.
3. **Durable study workspace.** Loading a cohort creates a local `.aeris` workspace. Input tables, exposure outputs, and run manifests receive stable identifiers. The host can inspect schemas, samples, missingness, value counts, and numeric summaries. Each run manifest records inputs, parameters, source settings, coverage, and decisions. Following the external-state principle of recursive language models (Zhang et al., 2025), the workspace also gives subagents shared state for bounded inspection and monitoring tasks.
4. **Exposure computation layer.** AERIS performs geocoding, boundary resolution, monitor discovery, retrieval, spatial and temporal aggregation, quality checks, and export. It uses OpenAQ for monitoring data and checks OpenAQ's public S3 archive when an API query returns no measurements for the requested historical dates. Request pacing, workload planning, and an optional SQLite cache reduce duplicate calls and make long retrievals recoverable.

Each run starts with a study question and an input table. In interactive mode, the researcher approves the exposure request and reviews the scientific choices. In self-driving mode, the agent works within fixed safety and resource limits, preserves the article-defined question, model, and contrast, records material departures, and critiques its analysis before finalizing. Lessons from interactive runs became general recovery policies: persist after failed retrieval, explore alternative public sources, and label scientific substitutions. Both modes use the same exposure engine and evidence record; autonomy changes the approval process while preserving computation and provenance. Figure 1 shows this separation of concerns across the host, tool server, data sources, and workspace.

3.3 Data sources, retrieval, and quality control

AERIS links study locations and dates to modular exposure sources detailed in Appendix C. Its OpenAQ pathway retrieves station metadata and hourly or daily measurements through the API and Python SDK (version 0.4.0), with the public S3 archive as a historical fallback (OpenAQ, 2026a; Biggs & OpenAQ, 2025; OpenAQ, 2026b). Google Maps provides recent, coordinate-level hourly histories, while AccuWeather provides current and forecast air-quality indices (Google Maps Platform, 2025; AccuWeather, 2026). For reproducible hyperlocal $PM_{2.5}$ estimates, a packaged XGBoost model extends the random-forest method of Lu et al. (2021) by combining nearby low-cost sensors with weather, road proximity, land cover, and weekday. Researchers can also substitute their own exposure model while retaining AERIS's assignment, quality-control, and provenance workflow. Each output manifest records the source, resolution, coverage, provenance, and retrieval settings.

Before retrieval, AERIS validates identifiers, dates, and locations. It resolves ZIP codes with `pgeocode` (Yurchak, 2024), addresses with Nominatim (Nominatim developer community, 2026), and administrative units with geoBoundaries (Runfola et al., 2020). Workload estimates, rate limits, and an optional SQLite cache control repeated requests. Each output retains its station or boundary, assignment settings, temporal completeness, missingness, and underlying daily or raw records.

3.4 EXPOSURE COMPUTATION PATHWAYS

Three pathways define ambient exposure at different spatial and temporal units. They share source selection, retrieval, coverage checks, provenance, and export.

Fixed-location participant exposure. The fixed-location pathway accepts a participant identifier, ZIP code, address, or coordinates with an exposure window. It supports studies where a residence, school, clinic, or other stable setting represents the relevant location. AERIS geocodes each record, finds candidate monitors, and assigns either the nearest monitor or an inverse-distance-weighted estimate:

$$C_p = \frac{\sum_i w_i C_i}{\sum_i w_i}, \qquad w_i = d_i^{-\beta}.$$

The output identifies the contributing monitors, distances, exclusions, coverage, and aggregation rule. When the required features are available, the packaged $PM_{2.5}$ model can instead estimate a fixed coordinate over time.

Trajectory-based participant exposure. The trajectory pathway accepts time-stamped coordinates from GPS traces, travel diaries, or similar records. AERIS validates coordinates and timestamps, converts time to UTC, removes duplicate participant-hours, orders each trajectory, and estimates conditions at every retained location and time. The XGBoost pathway loads packaged weights on demand and combines nearby sensor, weather, road-network, and land-cover features. Its output preserves both the location-time variation and source metadata.

Administrative-area exposure. The administrative-area pathway accepts a named unit or researcher-provided polygon with an exposure window. It supports ecological studies whose outcomes are indexed by municipality, county, department, region, or country. AERIS resolves the boundary, selects monitors within it, aggregates daily measurements, and records the polygon source, contributing stations, coverage, and aggregation method.

4 OUTCOME-BASED EVALUATION USING PUBLISHED STUDIES

4.1 EXPERIMENTAL DESIGN

Long-running research agents are difficult to evaluate through isolated questions or bounded tasks (Chen et al., 2025; Siegel et al., 2025). To evaluate AERIS on a complete scientific analysis, we put it to work on published public-health studies. Each article defines the pollutant, outcome, period, geographic unit, and model. AERIS retrieves independent public measurements of the same pollutant, constructs a new exposure record, joins it to public outcome data, and fits the analysis as closely as those data allow. We call this task a *source-substitution reanalysis* because AERIS replaces the article's exposure data while preserving its research question and analysis.

This protocol evaluates the full end-to-end agent workflow, from source discovery through statistical fitting, against an external scientific result. It also exposes failures that answer-based tests often miss, including incomplete historical coverage, mismatched geographic units, and unsupported substitutions.

Our primary evaluation reflects intended use. AERIS found data, wrote code, built exposure panels, and fit models while we reviewed its scientific choices; Appendix F records our interventions. Two controls remove that guidance: (1) self-driving AERIS uses standing permissions, retry policies, and internal review; and (2) raw Codex runs the same tasks without the AERIS harness.

We count a study as completed when the agent retrieves usable health outcomes, constructs an exposure record, and produces an identifiable fitted effect. A completed run need not reproduce the published estimate: completion measures execution, while direction, null recovery, and effect magnitude measure scientific agreement. For every run, we preserve the article, data sources, tool trace, manifest, analysis panel, and fitting script for reproducibility.

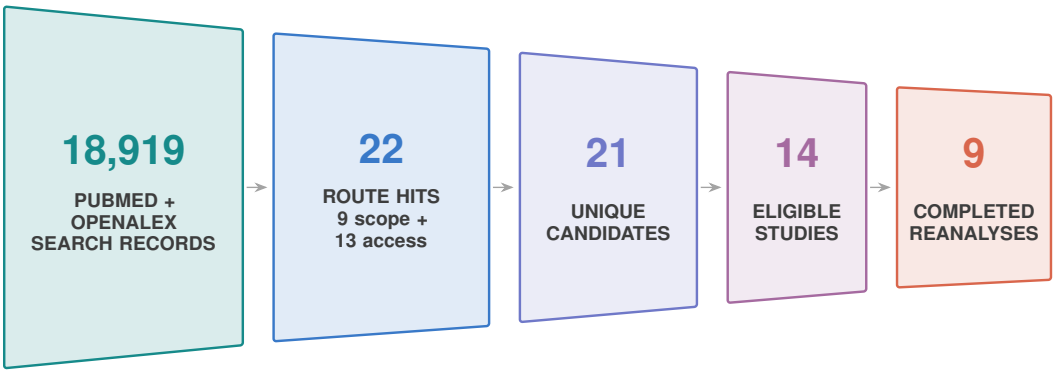

Figure 2: Study-selection funnel. Two routes screened 18,919 PubMed and OpenAlex records: scientific scope yielded nine studies and public-outcome access yielded 13, with one overlap. Fourteen of the 21 candidates passed the data-and-method audit; nine yielded fitted reanalyses. Three of the remaining studies lacked adequate air-quality coverage or spatial variation, and two lacked a retrievable outcome export.

4.2 Study-selection and reanalysis pipeline

We searched the preceding five years of PubMed and OpenAlex for outdoor air-pollution and human-health studies. Two complementary routes screened the 18,919 deduplicated records (Figure 2): one prioritized scientific scope and the other publicly accessible health outcomes. After removing one overlap, we audited 21 candidates and attempted the 14 with sufficient public data and methodological detail. Nine also had retrievable outcomes and adequate historical air-quality coverage for a fitted reanalysis. Appendix A gives every count and exclusion; the supplementary material provides the screening ledger, source URLs, execution records, and candidate decisions.

COVID-19 studies dominate because governments released cases and deaths with the dates and geographic keys needed for exposure reconstruction. Comparable public outcomes remain uncommon for chronic disease and hospital admissions. GPT-5 Codex subagents reviewed each article and named source on 13–14 August 2026. All nine fitted studies use administrative-area outcomes.

4.3 Autonomous experimental controls

Each run receives the article, public outcome source, target population, period, geographic unit, pollutant, and analysis specification. The agent retrieves the outcome and reconstructs exposure from public sources without using the article's pollution table or exposure file. It returns the analysis panel and fit with coverage, provenance, substitutions, and decisions. Codex ran every reanalysis with GPT-5.6 Terra at medium reasoning effort.

The primary evaluation includes researcher review. Two autonomous controls use fresh workspaces and the same study targets: self-driving AERIS runs without human intervention, while raw Codex receives the article bundle, task contract, network access, and output schema without the AERIS harness, prior code, or results. Each control comprises three independent campaigns with identical model and reasoning settings. Repetition measures variation across agent trajectories rather than changes in the task or model.

We report completion before agreement because effect metrics cover only runs that produced comparable estimates. For those runs, we measure retention of the published direction or null conclusion, absolute log-effect difference, the geometric mean reconstructed-to-published ratio, and Pearson correlation. Campaign-level correlations reveal instability that a pooled score can hide. Appendix E and the supplementary material report every contrast, setting, substitution, and failed attempt.

Table 1: Primary evaluation and autonomous controls. Completion is reported as mean ± SD across campaigns, with pooled attempts in parentheses. Agreement metrics include only completed, comparable estimates and should be read alongside completion rates. Pearson correlations are listed by campaign; an em dash marks unavailable logging or fewer than two comparable estimates. Token counts include all retries.

Measure	AERIS with researcher review	Self-driving AERIS	Raw agent
Role	Primary evaluation	Robustness control	Harness ablation
AERIS tools and context	Yes	Yes	No
Human researcher intervention	Occasional	None	None
Independent campaigns	1	3	3
Study attempts	9	27	27
Fitted studies per campaign	9 (9 / 9)	7.7 ± 1.5 (23 / 27)	2.3 ± 1.2 (7 / 27)
Complete nine-study campaigns	1 / 1	1 / 3	0 / 3
AERIS exposure and provenance record	9 / 9	26 / 27	0 / 27
Comparable study-level estimates	8	17	7
Published positive direction retained	7 / 7	11 / 16	3 / 4
Published null/no-evidence conclusion retained	4 / 7	14 / 21	3 / 3
Median $\lvert\Delta \log R\rvert$	0.040	0.169	0.073
Geometric mean reconstructed/published	0.998	0.901	1.011
Pearson r by campaign	0.834	$0.507 \mid 0.091 \mid -0.170$	$0.016 \mid -0.675 \mid$ —
Campaign runtime, hours	—	2.10 ± 0.52	1.02 ± 0.01
Model output tokens per campaign	—	$275\text{k} \pm 52\text{k}$	$103\text{k} \pm 5\text{k}$

5 RESULTS

Our reanalysis campaign tests whether AERIS helps scientists construct valid exposure analyses. The two autonomous controls then test whether AERIS can repeat this work without human researcher intervention and what the harness contributes beyond the underlying model. Table 1 compares experimental results across the three conditions.

5.1 AERIS WITH RESEARCHER REVIEW

Across the nine completed reanalyses, AERIS processed 1,847,519 air-quality monitor records across 705 monitor-study associations and produced 20,568 area-time exposure estimates. These analyses covered 14 effect contrasts, four pollutants, and administrative units across Europe and the Americas.

AERIS recovered the direction of all seven published positive contrasts, but only three reconstructed confidence intervals excluded 1. Among seven null or no-evidence contrasts, four AERIS intervals included 1; all four were OECD endpoints, while Brazil and two Northern Italy estimates excluded 1. These contrast-level results are clustered within studies. Appendix E places each result beside its source and reconstruction settings.

For numerical agreement, Table 5 gives each study equal weight. For a study with k contrasts, we calculate its mean absolute log difference and signed mean log ratio:

$$e_s = \frac{1}{k}\sum_{c=1}^{k}\left|\log\left(\frac{R_{\mathrm{AERIS},c}}{R_{\mathrm{published},c}}\right)\right|, \qquad d_s = \frac{1}{k}\sum_{c=1}^{k}\log\left(\frac{R_{\mathrm{AERIS},c}}{R_{\mathrm{published},c}}\right).$$

The absolute value in e_s prevents deviations from cancelling; d_s yields the geometric mean reconstructed-to-published ratio. Averaging within studies prevents the four OECD contrasts from

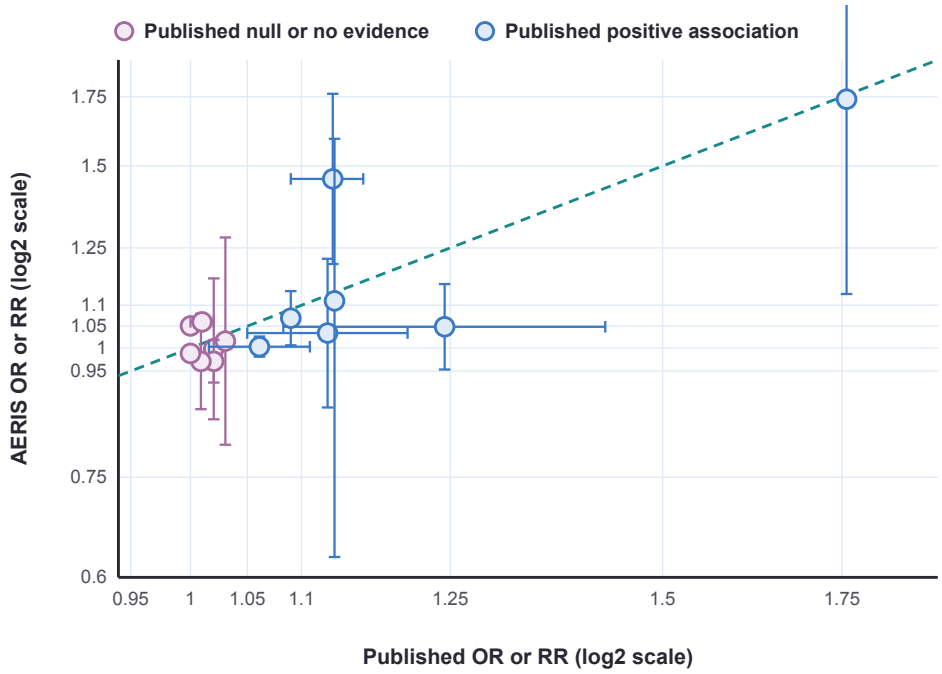

Figure 3: Published and AERIS effect ratios for all 14 contrasts from reanalyses conducted with researcher review, shown on $\log_2$ axes. Error bars show available 95% CIs; colors distinguish positive from null or no-evidence contrasts. The identity line marks exact agreement.

dominating. Direction and null recovery remain contrast-level measures. These quantities measure agreement with the published estimates, while the underlying health association remains unknown.

The two Northern Italy contrasts, reported as 1.00, cannot support a continuous magnitude comparison. Across the remaining eight studies, the median study-level absolute log difference was 0.040, the geometric mean AERIS-to-published ratio was 0.998, and six estimates were within 10% on the ratio scale. Published and reconstructed estimates had a Pearson correlation of $r = 0.834$, although this association was sensitive to the influential Italian study. We treat correlation as supporting evidence and emphasize direct measures of effect-size agreement and calibration. Ratio-scale proximity can still conceal a large change in excess association: for São José ozone, the RR fell from 1.061 to 1.003, reducing the excess above the null from 6.1% to 0.3%. Direction, null recovery, log-scale difference, and correlation therefore describe complementary aspects of agreement.

5.2 Autonomous controls

We aggregate three independent campaigns for each autonomous control and count completion across 27 study attempts. No run receives human researcher intervention. We compute direction, null recovery, absolute log difference, and correlation only for fitted effects that can be placed on the published scale; we report the denominator beside every metric. Appendix G reports the self-driving runs, and Appendix H reports the raw-agent ablation. We classify each attempt by whether it produced a fitted scientific result.

Self-driving AERIS fitted 8, 9, and 6 studies across the three campaigns: 23 of 27 attempts, or 7.7 ± 1.5 per campaign. One campaign completed all nine studies, and 26 of 27 attempts retained an AERIS exposure and provenance record. Among fitted, interpretable contrasts, self-driving AERIS retained 11 of 16 published positive directions and 14 of 21 null conclusions. Its median study-level absolute log difference was 0.169, and its geometric mean reconstructed-to-published ratio was 0.901. Study-level Pearson correlations were 0.507, 0.091, and -0.170 across the three campaigns. Campaigns took 2.10 ± 0.52 hours and generated $275{,}000 \pm 52{,}000$ output tokens, including retries.

The raw-agent harness-ablation campaigns fitted 7 of 27 attempts (3, 3, and 1), compared with 23 for self-driving AERIS, a 3.3-fold difference. Six of the seven fits retained the published conclusion. Their median absolute log difference was 0.073, and their geometric mean reconstructed-to-published ratio was 1.011. Study-level Pearson correlations were 0.016 and -0.675 in the two campaigns with multiple fits; the third produced only one fit. Effect agreement could be measured only for these seven fitted analyses; the raw agent produced no estimate in the other 20 attempts. The ablation shows that AERIS helps the agent recover from incomplete public inputs and finish analyses that the raw model usually abandons.

6 Discussion

With researcher review, AERIS completed all nine feasible reanalyses, recovered every published positive direction, and reached a median log-effect difference of 0.040. Agreement across four pollutants, several countries, and varied health outcomes shows that the exposure pipeline can preserve broad spatial and temporal patterns under source substitution. It does not imply 96% accuracy in recovering a health effect: a small ratio-scale difference can represent a large change in excess risk. As an evaluation method, source substitution yields a weight of evidence across direction, uncertainty, calibration, and study-level failures rather than a single benchmark score.

Self-driving AERIS completed 23 of 27 attempts, compared with 7 for raw Codex, and preserved provenance in 26. Yet its median log-effect difference rose to 0.169, while study-level correlations ranged from -0.170 to 0.507 across campaigns. Researchers still improve source selection, estimand fidelity, and calibration. This gap separates two requirements that completion scores often conflate: a harness can help an agent finish a long analysis, while scientific judgment still determines whether it answers the intended question precisely and accurately. A practical path toward autonomy preserves review at consequential branch points while automating retrieval, computation, recovery, and documentation.

Our autonomy results remain specific to this evaluation. We designed the retry policies after observing failures in the nine-study campaign with researcher review, and every control used GPT-5.6 Terra. We have not yet tested whether the same policies transfer to new studies or other agents. The evaluation also favors COVID-19 and administrative-area analyses because their health outcomes were publicly accessible. Future work should test other agents, unseen disease settings, and private cohort data under appropriate research governance.

Scientific uncertainty remains in monitor placement, exposure-model validity, indoor exposure, and the design of each health study. The reanalyses evaluate the administrative-area pathway; fixed-location and mobility exposure received implementation tests only. Autonomous analyses should therefore preserve their sources, traces, and substitutions and receive expert review before informing research or policy.

7 Conclusion

Air pollution remains absent from many public-health studies because exposure assessment is technically demanding and difficult to audit. AERIS lowers this barrier: it gives research agents a common system for retrieving public air-quality data, computing exposure across participant and population settings, assessing coverage, and preserving provenance. We evaluated the resulting scientific work through source-substitution reanalysis rather than isolated tool calls. Across nine published studies, AERIS processed 1.85 million air-quality monitor records, recovered every positive direction, and reached a median absolute log-effect difference of 0.040. The autonomous controls further isolate the value of the harness: self-driving AERIS completed 23 of 27 attempts, compared with 7 for the raw agent, although researcher review still produced better-calibrated estimates.

AERIS turns the labor-intensive work of exposure assessment into an auditable workflow, freeing researchers to focus on study design, interpretation, and the questions that advance environmental health. Together, these results define a practical path toward autonomous environmental health.

AI USE STATEMENT

We used generative AI for research design, literature retrieval, software development, data preparation and analysis, interpretation, reanalysis execution, and manuscript drafting and editing. Codex ran the reanalyses with researcher review, autonomous AERIS, and the raw agent with GPT-5.6 Terra at medium reasoning; Codex and Claude Code also supported implementation and writing. We tested generated code and manually checked the analysis artifacts, results, claims, citations, and prose. We used no generative AI to create synthetic datasets or prove mathematical claims, and take responsibility for all content and AI-assisted artifacts.

REPRODUCIBILITY STATEMENT

The appendices document the literature search, eligibility decisions, study targets, source substitutions, statistical comparisons, and agent configuration. The anonymous supplement contains the evaluated AERIS snapshot, machine-readable tool contracts, environment specification, screening scripts and prompts, per-study analysis code, run manifests, expected outputs, and instructions for retrieving public source data. It also preserves complete records for successful and unsuccessful control runs.

ETHICS STATEMENT

This study reanalyzes public, aggregate, or previously released research data; we did not recruit participants or access private clinical records. Exposure estimates remain subject to monitor coverage, source quality, geographic aggregation, and the assumptions of each health study. AERIS therefore retains provenance and coverage diagnostics so that researchers can review its work before drawing scientific or clinical conclusions. The system does not provide medical advice, and its outputs require expert review before use in research or policy.

REFERENCES

AccuWeather. Guide | AccuWeather Enterprise API, 2026. URL https://apidev.accuweather.com/developers/air-quality/guide. Accessed 5 September 2026.

Antoine Alléon, Grégoire Jauvion, Boris Quennehen, and David Lissmyr. PlumeNet: Large-scale air quality forecasting using a convolutional LSTM network. *arXiv preprint arXiv:2006.09204*, 2020. URL https://arxiv.org/abs/2006.09204.

Karoline K. Barkjohn, Brett Gantt, and Andrea L. Clements. Development and application of a United States-wide correction for PM2.5 data collected with the PurpleAir sensor. *Atmospheric Measurement Techniques*, 14(6):4617–4637, 2021. doi: 10.5194/amt-14-4617-2021.

Rob Beelen, Ole Raaschou-Nielsen, Massimo Stafoggia, et al. Effects of long-term exposure to air pollution on natural-cause mortality: An analysis of 22 European cohorts within the multicentre ESCAPE project. *The Lancet*, 383(9919):785–795, 2014. doi: 10.1016/S0140-6736(13)62158-3.

Russell Biggs and OpenAQ. OpenAQ Python SDK, 2025. URL https://github.com/openaq/openaq-python/releases/tag/v0.4.0. Version 0.4.0; computer software; OpenAQ, ROR: https://ror.org/05g88bt71; Apache 2.0 License; released 31 March 2025.

Daniil A. Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578, 2023. doi: 10.1038/s41586-023-06792-0.

Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D. White, and Philippe Schwaller. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, 6(5):525–535, 2024. doi: 10.1038/s42256-024-00832-8.

Michael S. Breen, Catherine Seppanen, Vlad Isakov, et al. Development of TracMyAir smartphone application for modeling exposures to ambient PM2.5 and ozone. *International Journal of Environmental Research and Public Health*, 16(18):3468, 2019. doi: 10.3390/ijerph16183468.

Robert D. Brook, Sanjay Rajagopalan, C. Arden Pope, III, et al. Particulate matter air pollution and cardiovascular disease: An update to the scientific statement from the American Heart Association. *Circulation*, 121(21):2331–2378, 2010. doi: 10.1161/CIR.0b013e3181dbece1.

Bert Brunekreef and Stephen T. Holgate. Air pollution and health. *The Lancet*, 360(9341):1233–1242, 2002. doi: 10.1016/S0140-6736(02)11274-8.

Dechao Bu et al. Empowering AI data scientists using a multi-agent LLM framework with self-evolving capabilities for autonomous, tool-aware biomedical data analyses. *Nature Biomedical Engineering*, 2026. doi: 10.1038/s41551-026-01634-6.

Thibaut Cassard, Grégoire Jauvion, and David Lissmyr. High-resolution air quality prediction using low-cost sensors. *arXiv preprint arXiv:2006.12092*, 2020. URL https://arxiv.org/abs/2006.12092.

Jun Shern Chan, Neil Chowdhury, Oliver Jaffe, James Aung, Dane Sherburn, Evan Mays, Giulio Starace, Kevin Liu, Leon Maksin, Tejal Patwardhan, Aleksander Mądry, and Lilian Weng. MLE-bench: Evaluating machine learning agents on machine learning engineering. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=6s5uXNWGIh.

Ziru Chen, Shijie Chen, Yuting Ning, et al. ScienceAgentBench: Toward rigorous assessment of language agents for data-driven scientific discovery. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/f12b4df26344f3be803c06b555252efe-Abstract-Conference.html.

Qian Di et al. Air pollution and mortality in the Medicare population. *The New England Journal of Medicine*, 376(26):2513–2522, 2017. doi: 10.1056/NEJMoa1702747.

Qian Di et al. An ensemble-based model of PM2.5 concentration across the contiguous United States with high spatiotemporal resolution. *Environment International*, 130:104909, 2019. doi: 10.1016/j.envint.2019.104909.

Douglas W. Dockery et al. An association between air pollution and mortality in six U.S. cities. *The New England Journal of Medicine*, 329(24):1753–1759, 1993. doi: 10.1056/NEJM199312093292401.

Francesca Dominici, Roger D. Peng, Michelle L. Bell, et al. Fine particulate air pollution and hospital admission for cardiovascular and respiratory diseases. *JAMA*, 295(10):1127–1134, 2006. doi: 10.1001/jama.295.10.1127.

David Donaire-Gonzalez et al. ExpoApp: An integrated system to assess multiple personal environmental exposures. *Environment International*, 126:494–503, 2019. doi: 10.1016/j.envint.2019.02.054.

Clinton H. Durney et al. APEX: A web-based tool for assessing long-term outdoor PM2.5 exposure—brief report. *JTO Clinical and Research Reports*, 7(5):100975, 2026. doi: 10.1016/j.jtocrr.2026.100975.

Marloes Eeftens, Rob Beelen, Kees de Hoogh, et al. Development of land use regression models for PM2.5, PM2.5 absorbance, PM10 and PMcoarse in 20 European study areas; results of the ESCAPE project. *Environmental Science & Technology*, 46(20):11195–11205, 2012. doi: 10.1021/es301948k.

Google Maps Platform. History | Air Quality API | Google for Developers, 2025. URL https://developers.google.com/maps/documentation/air-quality/history. Accessed 25 August 2026.

Juraj Gottweis et al. Accelerating scientific discovery with Co-Scientist. *Nature*, 655(8122):487–496, 2026. doi: 10.1038/s41586-026-10644-y. URL https://doi.org/10.1038/s41586-026-10644-y.

Health Effects Institute. State of Global Air 2024: Special report. Technical report, Health Effects Institute, Boston, MA, 2024. URL https://www.stateofglobalair.org/resources/archived/state-global-air-report-2024. Produced in partnership with UNICEF. Accessed 25 August 2026.

Gerard Hoek, Rob Beelen, Kees de Hoogh, Danielle Vienneau, John Gulliver, Paul Fischer, and David Briggs. A review of land-use regression models to assess spatial variation of outdoor air pollution. *Atmospheric Environment*, 42(33):7561–7578, 2008. doi: 10.1016/j.atmosenv.2008.05.057.

Grégoire Jauvion, Thibaut Cassard, Boris Quennehen, and David Lissmyr. DeepPlume: Very high resolution real-time air quality mapping. *arXiv preprint arXiv:2002.10394*, 2020. URL https://arxiv.org/abs/2002.10394.

Michael Jerrett, Altaf Arain, Pavlos Kanaroglou, et al. A review and evaluation of intraurban air pollution exposure models. *Journal of Exposure Analysis and Environmental Epidemiology*, 15(2): 185–204, 2005. doi: 10.1038/sj.jea.7500388.

Haneen Khreis, Charlotte Kelly, James Tate, Roger Parslow, Karen Lucas, and Mark Nieuwenhuijsen. Exposure to traffic-related air pollution and risk of development of childhood asthma: A systematic review and meta-analysis. *Environment International*, 100:1–31, 2017. doi: 10.1016/j.envint.2016.11.012.

Itai Kloog, Alexandra A. Chudnovsky, Allan C. Just, et al. A new hybrid spatio-temporal model for estimating daily multi-year PM2.5 concentrations across northeastern USA using high resolution aerosol optical depth data. *Atmospheric Environment*, 95:581–590, 2014. doi: 10.1016/j.atmosenv.2014.07.014.

Dong Liang and Naresh Kumar. Time-space kriging to address the spatiotemporal misalignment in the large datasets. *Atmospheric Environment*, 72:60–69, 2013. doi: 10.1016/j.atmosenv.2013.02.034.

Yougeng Lu, Genevieve Giuliano, and Rima Habre. Estimating hourly PM2.5 concentrations at the neighborhood scale using a low-cost air sensor network: A Los Angeles case study. *Environmental Research*, 195:110653, 2021. doi: 10.1016/j.envres.2020.110653.

Lidia Morawska et al. Applications of low-cost sensing technologies for air quality monitoring and exposure assessment: How far have they gone? *Environment International*, 116:286–299, 2018. doi: 10.1016/j.envint.2018.04.018.

Nominatim developer community. Nominatim 5.3.2 manual, 2026. URL https://nominatim.org/release-docs/latest/. Accessed 5 September 2026.

Marguerite Nyhan, Sebastian Grauwin, Rex Britter, et al. Exposure track—the impact of mobile-device-based mobility patterns on quantifying population exposure to air pollution. *Environmental Science & Technology*, 50(17):9671–9681, 2016. doi: 10.1021/acs.est.6b02385.

Marguerite M. Nyhan et al. Quantifying population exposure to air pollution using individual mobility patterns inferred from mobile phone data. *Journal of Exposure Science & Environmental Epidemiology*, 29(2):238–247, 2019. doi: 10.1038/s41370-018-0038-9.

OpenAQ. OpenAQ API. https://api.openaq.org, 2026a. Organization ROR: https://ror.org/05g88bt71. Air-quality data aggregated and distributed by OpenAQ; original providers retain their own licenses. Accessed 26 September 2026.

OpenAQ. About open data on AWS, 2026b. URL https://docs.openaq.org/aws/about. Accessed 5 September 2026.

Pablo Orellano, Julieta Reynoso, Nancy Quaranta, Ariel Bardach, and Agustin Ciapponi. Short-term exposure to particulate matter (PM10 and PM2.5), nitrogen dioxide (NO2), and ozone (O3) and all-cause and cause-specific mortality: Systematic review and meta-analysis. *Environment International*, 142:105876, 2020. doi: 10.1016/j.envint.2020.105876.

C. Arden Pope, III, Richard T. Burnett, Michael J. Thun, et al. Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. *JAMA*, 287(9):1132–1141, 2002. doi: 10.1001/jama.287.9.1132.

Daniel Runfola, Austin Anderson, Heather Baier, Matt Crittenden, Elizabeth Dowker, Sydney Fuhrig, Seth Goodman, Grace Grimsley, Rachel Layko, Graham Melville, Maddy Mulder, Rachel Oberman, Joshua Panganiban, Andrew Peck, Leigh Seitz, Sylvia Shea, Hannah Slevin, Rebecca Youngerman, and Lauren Hobbs. geoBoundaries: A global database of political administrative boundaries. *PLOS ONE*, 15(4):e0231866, 2020. doi: 10.1371/journal.pone.0231866. URL https://doi.org/10.1371/journal.pone.0231866.

Patrick H. Ryan and Grace K. LeMasters. A review of land-use regression models for characterizing intraurban air pollution exposure. *Inhalation Toxicology*, 19(Suppl 1):127–133, 2007. doi: 10.1080/08958370701495998.

Jason D. Sacks et al. The environmental benefits mapping and analysis program—community edition (BenMAP-CE): A tool to estimate the health and economic benefits of reducing air pollution. *Environmental Modelling & Software*, 104:118–129, 2018. doi: 10.1016/j.envsoft.2018.02.009.

Jonathan M. Samet, Francesca Dominici, Frank C. Curriero, Isabelle Coursac, and Scott L. Zeger. Fine particulate air pollution and mortality in 20 U.S. cities, 1987–1994. *The New England Journal of Medicine*, 343(24):1742–1749, 2000. doi: 10.1056/NEJM200012143432401.

Heidi Seibold, Severin Czerny, Siona Decke, Roman Dieterle, Thomas Eder, Steffen Fohr, Nico Hahn, Rabea Hartmann, Christoph Heindl, Philipp Kopper, Dario Lepke, Verena Loidl, Maximilian Mandl, Sarah Musiol, Jessica Peter, Alexander Piehler, Elio Rojas, Stefanie Schmid, Hannah Schmidt, Melissa Schmoll, Lennart Schneider, Xiao-Yin To, Viet Tran, Antje Völker, Moritz Wagner, Joshua Wagner, Maria Waize, Hannah Wecker, Rui Yang, Simone Zellner, and Malte Nalenz. A computational reproducibility study of PLOS ONE articles featuring longitudinal data analyses. *PLOS ONE*, 16(6):e0251194, 2021. doi: 10.1371/journal.pone.0251194.

Zachary S. Siegel, Sayash Kapoor, Nitya Nadgir, Benedikt Stroebl, and Arvind Narayanan. CORE-Bench: Fostering the credibility of published research through a computational reproducibility agent benchmark. *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=BsMMc4MEGS.

Susanne Steinle, Stefan Reis, and Clive E. Sabel. Quantifying human exposure to air pollution—moving from static monitoring to spatio-temporally resolved personal exposure assessment. *Science of the Total Environment*, 443:184–193, 2013. doi: 10.1016/j.scitotenv.2012.10.098.

Victoria Stodden, Jennifer Seiler, and Zhaokun Ma. An empirical analysis of journal policy effectiveness for computational reproducibility. *Proceedings of the National Academy of Sciences*, 115(11): 2584–2589, 2018. doi: 10.1073/pnas.1708290115.

Adam A. Szpiro, Paul D. Sampson, Lianne Sheppard, Thomas Lumley, Sara D. Adar, and Joel D. Kaufman. Predicting intra-urban variation in air pollution concentrations with complex spatio-temporal dependencies. *Environmetrics*, 21(6):606–631, 2010. doi: 10.1002/env.1014.

Aaron van Donkelaar, Randall V. Martin, Michael Brauer, Ralph Kahn, Robert Levy, Carolyn Verduzco, and Paul J. Villeneuve. Global estimates of ambient fine particulate matter concentrations from satellite-based aerosol optical depth: Development and application. *Environmental Health Perspectives*, 118(6):847–855, 2010. doi: 10.1289/ehp.0901623.

Xiaonan Yu et al. Quantifying the impact of daily mobility on errors in air pollution exposure estimation using mobile phone location data. *Environment International*, 141:105772, 2020. doi: 10.1016/j.envint.2020.105772.

Roman Yurchak. pgeocode — pgeocode 0.5.0 documentation, 2024. URL https://pgeocode.readthedocs.io/. Version 0.5.0. Accessed 5 September 2026.

Alex L. Zhang, Tim Kraska, and Omar Khattab. Recursive Language Models, 2025. URL https://arxiv.org/abs/2512.24601.

AVAILABILITY

The anonymous supplementary archive contains the AERIS skill, MCP server, exposure tools, tests, and packaged XGBoost model. It also includes the literature-search protocol and scripts, screening prompt, per-study reanalysis code and results, intervention ledger, and complete self-driving and raw-agent control records, including manifests, transcripts, generated code, and model outputs.

A STUDY-IDENTIFICATION PIPELINE IN FULL

Figure 2 condenses the study-identification pipeline to its principal stages. This appendix reports the intermediate counts, explains each screening rule, and shows how the two discovery routes combine.

A.1 CORPUS

We searched PubMed and OpenAlex for public-health studies published during the preceding five years that associated an outdoor ambient pollutant with a human health outcome. We deduplicated results by DOI, then by normalized title, year, and first author when no DOI was available, leaving 18,919 records. Both discovery routes searched this archived corpus in parallel. A study found by both routes therefore represents overlap between search strategies rather than a duplicated record.

A.2 ROUTE 1: SCIENTIFIC-SCOPE SCREENING

In Route 1, we screened titles, abstracts, and metadata for empirical studies of outdoor ambient exposure and human-health outcomes. We then assessed data access, scientific fit, and whether the available inputs could support source-substitution reanalysis.

Rapid title, abstract, and metadata screen. The screen retained records that might describe an empirical human-health study of outdoor ambient exposure. It excluded clear reviews, protocols, editorials, laboratory or non-human studies, and studies outside outdoor ambient exposure.

Decision	Records
Retained	10,020
Uncertain	5,985
Excluded	2,914
Total	18,919

The metadata did not support a decision for 5,985 records. We retained them as unresolved rather than scientific exclusions, leaving some risk that the search missed eligible studies.

Open-access confirmation. We checked the 10,020 retained records for a PubMed Central identifier or OpenAlex metadata identifying an open full-text location.

Decision	Records
Confirmed open access	4,658
Open-access status unresolved	5,359
Confirmed non-open-access	3
Total	10,020

We could not confirm access for 5,359 records. The prespecified Route 1 workflow continued with the 4,658 records whose open-access status we could confirm.

Strict linked-data path. The strict path searched the 4,658 confirmed-open-access papers for a concrete public-data URL. Full-text review then tested scientific fit, followed by field-level validation of outcomes, dates, locations, covariates, and the analysis specification.

Stage	Records
Confirmed open access	4,658
Concrete public-data URL	138
Passed scientific-fit review	22
Passed field-level validation	3

Public-source rescue path. The rescue path screened the 4,520 confirmed-open-access papers with PubMed Central identifiers for named public systems from which the required inputs could be reconstructed even when the article did not link a complete analysis dataset. Full text was cached for 3,707 papers. A high-precision screen selected 133 candidates; three already belonged to the strict candidate pool and were removed before rescue review, leaving 130 new candidates.

Stage	Records
Confirmed-open-access PMC papers	4,520
Cached full texts screened	3,707
High-precision candidates	133
New candidates after removing strict-path overlap	130
Passed full-text source review	13
Passed field-level validation	6

The strict and rescue inputs overlap: the 4,520 PMC papers are a subset of the 4,658 confirmed-open-access papers, and 138 PMC papers passed the strict URL gate. We removed three overlapping records when we combined the candidate lists. The two paths yielded $3 + 6 = 9$ validated scientific-scope candidates. The rescue path's 13 full-text passes are distinct from the 13 Route 2 candidates below.

A.3 Route 2: screening for publicly accessible health outcomes

Title and abstract screening. A script examined each title and available abstract for a named source that might provide public health outcomes. Signals included vital statistics and civil-registration systems; DATASUS and SIH/SUS; NHANES, DHS, MICS, and BRFSS; cancer and disease registries; notifiable-disease and hospital- discharge systems; and named government or international COVID-19 reporting systems. This pass set review priority rather than study eligibility.

Routing label	Records
Named public outcome source	203
Likely public aggregate outcome	3,430
Likely individual-level or restricted outcome	6,251
Insufficient metadata to identify a route	9,035
Total	18,919

These labels did not exclude records. We left 9,035 records unresolved because their metadata did not identify an outcome source. We prioritized the 203 records that named one for article review.

Article and source review. Agents read the Methods and data-availability material for all 203 papers and inspected each cited outcome source.

Access disposition	Papers
Direct public download	45
Public source with a clear retrieval route	47
Registration or account required	52
Agreement, application, or request required	36
No usable retrieval link found	21
Unresolved	2
Total	203

The first two groups formed 92 papers with a potentially accessible public outcome. The $45+47=92$ split describes two forms of access and is not an additional filtering stage.

Health-outcome access and ambient-pollution review. Agents reviewed all 92 papers again to identify studies of outdoor ambient air pollution, wildfire smoke, or desert dust whose cited outcomes could be retrieved without authentication.

Decision	Papers
Ambient pollution or wildfire smoke	17
Unauthenticated public outcome retrieval	13
Agreement or request required	2
Public route not working or not found	2
Non-ambient exposure or unrelated subject	75
Total	92

Route 2 therefore produced 13 candidates. Figure 2 combines the ambient-scope and access checks into the $92 \rightarrow 13$ transition.

A.4 Combining the routes and conducting the full audit

Routes 1 and 2 produced 9 and 13 candidates, respectively, for 22 route hits. One study appeared in both, leaving 21 unique candidates. The limited overlap shows that the routes captured substantially different parts of the literature.

We then audited the paper and data sources for every candidate. We required an outcome source available without an account, request, agreement, approval, or collaborator; recoverable dates and geographic identifiers; a described primary model and required non-exposure covariates; public monitoring data for the reported pollutant; and enough information for AERIS to reconstruct exposure independently.

Audit decision	Studies
Eligible for attempted reanalysis	14
Excluded during audit	7
Missing a required input or method	6
Historical outcome export not recovered	1
Total	21

A.5 Execution

AERIS attempted all 14 eligible studies; nine produced fitted reanalyses.

Execution result	Studies
Completed with fitted results	9
No fitted result	5
Inadequate historical air-quality coverage or variation	3
Required outcome export not retrieved	2
Total	14

The five execution exclusions were:

- **China COVID-19:** we found two usable prefecture panels for 2014–2016, compared with 287 in the published analysis.
- **Israel COVID-19:** the available $PM_{2.5}$ panel had no spatial variation across municipalities for the study dates.
- **Argentina COVID-19:** the completed boundary run yielded too few $PM_{2.5}$ observations for a fitted panel.

- **Taubaté COVID-19 admissions:** the required daily SIH/SUS outcome export was not recovered.
- **Alaska wildfire asthma:** we could not download the required regional outcome table from a public source.

The audit established that an analysis appeared runnable from the article and identified sources. Execution tested whether those sources supplied adequate data in practice. Three failures arose from inadequate air-quality coverage or variation; two arose because we could not retrieve the required outcome export. The records do not establish that those outcome exports had previously been downloadable and later disappeared.

A.6 Count reconciliation

Stage	Count	Yield
Unique search records	18,919	—
Route 1 candidates	9	0.048% of corpus
Route 2 candidates	13	0.069% of corpus
Route hits	22	—
Unique candidates after one overlap	21	—
Eligible after audit	14	66.7% of candidates
Completed reanalyses	9	64.3% of attempts
Completed among unique candidates	9	42.9%
Corpus-to-completion yield	9	0.048%

GPT-5 Codex subagents conducted the article-level reviews on 13–14 August 2026. The anonymized supplementary material includes the search protocol and scripts, review prompt, source records, exclusion reasons, and execution artifacts. Figure 2 omits the uncertain and unresolved records, Route 1's overlapping strict and rescue subpaths, the $45 + 47$ access subdivision, and Route 2's intermediate count of 17 ambient-exposure studies; the appendix reports all of them above.

B Dynamic-path implementation detail

The open-source trajectory pathway estimates hourly $PM_{2.5}$ at each time-stamped coordinate with a packaged XGBoost regression model. The model contains 600 gradient-boosted trees and minimizes squared error. Its 14 ordered inputs are: the nearest PurpleAir $PM_{2.5}$ measurement within 5 km; temperature, relative humidity, wind speed, wind direction, and precipitation from Open-Meteo; distance to the nearest highway and arterial road from OpenStreetMap; open space, water, impervious surface, tree cover, and elevation from Google Earth Engine; and a weekday indicator. When PurpleAir has no nearby measurement, AERIS can substitute $PM_{2.5}$ from an OpenAQ regulatory monitor within 25 km. The run record preserves the input sources and missing-feature warnings for each estimate.

The feature design extends the neighborhood-scale random-forest method of Lu et al. (2021), which combined low-cost sensor measurements with meteorological, temporal, and spatial predictors. AERIS loads the model weights locally, so a trajectory need not be sent to a third-party prediction service. Researchers may also supply locally trained XGBoost weights with the same ordered inputs. The packaged weights support implementation testing and reproducible inference; their accuracy outside the training domain depends on local sensor conditions and requires external validation. We verified that the complete pathway runs on a 24-hour New York City trajectory. This execution test does not measure predictive accuracy or personal dose.

C Ambient-data sources and local $PM_{2.5}$ inference

Table 2 summarizes each external source, the information it contributes, and its role in exposure reconstruction. The local inference pathway groups the packaged model with the sensor, weather, and geographic inputs it expects.

OpenAQ aggregates observations from governments, researchers, and sensor networks rather than originating every measurement. We therefore cite OpenAQ as the distribution platform. Provider-specific

Table 2: Ambient-data sources and model inputs. AERIS records the selected pathway, source settings, and assignment method in the run manifest.

Source	What it provides	Why AERIS uses it
OpenAQ API and public S3 archive	Global repository of hourly and daily station observations from public and regulatory monitoring networks	Supplies historical, station-level concentrations, monitor metadata, and coverage diagnostics for fixed-location and area-based exposure assignment
Google Maps Air Quality API	Location-specific hourly air-quality history and pollutant information	Supplies a credentialed coordinate-level pathway when a trajectory requires location-time estimates beyond a monitor assignment
AccuWeather Air Quality API	Credentialed coordinate-level current and forecast air-quality index and category information	Adds a localized current and forecast source-inspection pathway. Its index is retained as an index; concentration-based epidemiological models require a source that returns physical pollutant concentrations (AccuWeather, 2026)
Local XGBoost $PM_{2.5}$ inference pathway		
Packaged XGBoost weights	A local hourly $PM_{2.5}$ model evaluated at a requested coordinate and time	Extends the low-cost-sensor, weather, and spatial-feature approach of Lu et al. (2021) with gradient-boosted trees; produces a reproducible estimate without sending participant locations to a third-party prediction service
PurpleAir	Dense low-cost sensor observations	Supplies nearby particulate measurements that capture local variation between regulatory stations
Open-Meteo weather	Meteorological conditions at the requested place and time	Supplies weather features that influence pollutant formation, dispersion, and transport
OpenStreetMap	Roads and other built-environment features	Supplies local source-proxy features associated with traffic and urban form
Land-cover inputs	Surface-cover features around the requested coordinate	Supplies geographic context associated with dispersion and background concentration

licenses govern downstream reuse. The supplement's `OPENAQ_ATTRIBUTION_MANIFEST` audits the provider and license evidence for all nine fitted reanalyses and flags records that cannot be verified from the preserved traces. The artifact does not redistribute raw observations.

D Exposure defaults and MCP tool contracts

Table 3 states the defaults used when a researcher does not supply a study-specific choice. They are starting points rather than universal scientific rules: AERIS records any override in the run manifest. In particular, AERIS imposes no default completeness cutoff.

The local MCP server returns each result in a common envelope containing `status`, warnings, tool name, workspace identifier, artifact identifier, and run or plan identifier. Invalid arguments, authorization failures, missing data, and upstream failures return typed error fields instead of fabricated exposure values. Table 4 groups the public endpoints; the artifact includes their complete JSON schemas.

Table 3: Default exposure computations in the released implementation.

Pathway	Default	Computation and edge cases
Fixed location	Nearest monitor within 25 km	The caller may request more monitors or a radius up to 200 km. With multiple monitors, AERIS aligns observations by UTC hour and applies inverse-distance weighting with power 2. An exactly co-located monitor receives the full weight. Hours without observations are omitted.
Administrative area	All eligible monitors within the resolved boundary	AERIS first averages observations within monitor, then takes the unweighted mean across monitor means. Its daily series likewise averages within monitor-day before averaging across monitors. If a station budget is supplied, deterministic spatial spread selects the subset.
Temporal completeness	No cutoff	AERIS reports observed versus expected periods. A user-supplied threshold flags or excludes incomplete monitor assignments; no hidden 75% rule applies. A query with no usable observations returns a missing estimate and an explicit quality-control code.
Historical recovery	OpenAQ API, then public S3 archive	An empty historical API response triggers bounded archive retrieval. The default automatic archive window is at most 400 days and 500 objects; larger jobs require an explicit plan.

Table 4: Compact MCP contract. Arguments shown are the principal scientific inputs; all outputs carry the common provenance envelope.

Tool	Principal inputs	Output
`load_cohort`	Local table path	Validated study schema, location mode, date range, workspace and dataset IDs
`plan_exposure`	Pollutant, country, monitor count, temporal resolution, location strategy	Proposed estimand, source plan, workload, and plan ID
`estimate_exposure`	Approved plan; pollutant, backend, monitor count, completeness rule	Fixed-location or trajectory exposure artifact, coverage summary, and run ID
`estimate_boundary_exposure`	Approved boundary, pollutant, dates, resolution, completeness rule	Area estimate, monitor coverage, boundary provenance, artifact and run IDs
`get_exposure_summary; get_daily_data; get_raw_data;`		
`get_station_info`	Artifact, subject or pollutant selector	Bounded summaries, normalized observations, and contributing-monitor records
`record_study_design; get_decision_log`	Question, outcome, pollutant, location and time units, assignment and analysis plan	Durable design and decision records
`list_artifacts; describe_artifact; sample_artifact;`		
`query_artifact`	Artifact ID and bounded inspection operation	Schema, provenance, samples, or aggregate statistics without loading the full table into chat
`export_results`	Artifact selector, local path, format	Reviewable exposure table and export provenance

E COMPLETED SOURCE-SUBSTITUTION REANALYSES

E.1 SUMMARY METRICS

Table 5: AERIS reanalysis summary with researcher review. Direction and null recovery use fitted contrasts; numerical agreement gives each study equal weight. We exclude the two Northern Italy contrasts from continuous comparisons because the article reports them only as 1.00.

Measure	Result
Contrast-level direction and null recovery	
Fitted contrasts	14
Published positive associations, same direction	7 / 7
Published positive associations, AERIS CI excludes 1	3 / 7
Published null or no-evidence contrasts, AERIS CI includes 1	4 / 7 (all OECD)
Study-weighted numerical agreement	
Studies	8
Pearson r (Fisher 95% CI)	0.834 (0.314–0.969)
Leave-one-study-out Pearson r range	0.225–0.958
Median absolute log difference	0.040
Mean absolute log difference	0.078
Within 10% on the ratio scale	6 / 8
Geometric mean AERIS / published ratio	0.998

E.2 CONTRAST-LEVEL RESULTS

E.3 STUDY REGISTRY

For workload accounting, a monitor record is counted each time an observation enters a study-specific exposure calculation; the same physical observation may therefore enter more than one study or geographic assignment. An area-time exposure estimate is one pollutant value assigned to one administrative area and time interval.

United Kingdom COVID-19. M. Pear Hossain, Wen Zhou, Marco Y. T. Leung, and Hsiang-Yu Yuan. *Association of air pollution and weather conditions during infection course with COVID-19 case fatality rate in the United Kingdom*. DOI: 10.1038/s41598-023-50474-w. The study relates country-level infection-course pollution and weather to daily COVID-19 case-fatality rates from March through September 2020. We reconstruct the four country series with public monitor-derived $PM_{2.5}$.

Italian provincial capitals COVID-19. Mario Coccia. *Effects of the spread of COVID-19 on public health of polluted cities: results of the first wave for explaining the dejà vu in the second wave of COVID-19 pandemic and epidemics of future vital agents*. DOI: 10.1007/s11356-020-11662-7. This ecological study links pre-pandemic PM_{10} exceedance days in Italian cities to first-wave COVID-19 outcomes. We combine public city outcomes with independently reconstructed 2018 monitor exposure.

France COVID-19. Anastase Tchicaya, Nathalie Lorentz, Hichem Omrani, Gaetan de Lanchy, and Kristell Leduc. *Impact of long-term exposure to $PM_{2.5}$ and temperature on coronavirus disease mortality: observed trends in France*. DOI: 10.1186/s12940-021-00784-1. The study models department-level COVID-19 mortality with long-term $PM_{2.5}$, temperature, and public health-system covariates. We join public department outcomes to a 2019 AERIS monitor-based exposure panel.

OECD COVID-19. Zohar Barnett-Itzhaki and Adi Levi. *Effects of chronic exposure to ambient air pollutants on COVID-19 morbidity and mortality: a lesson from OECD countries*. DOI: 10.1016/j.envres.2021.110723. This cross-country study evaluates $PM_{2.5}$ against COVID-19 cases and deaths at prespecified days after the first case. We fit the public outcome series for countries with adequate 2017 AERIS coverage.

Table 6: AERIS source-substitution reanalyses with researcher review on a common multiplicative scale. Bold values are odds ratios (OR) or rate ratios (RR). Rows follow each article's conclusion; the setting columns expose differences between the published and AERIS analyses.

Study and contrast	Published setting	AERIS setting	Published OR/RR	AERIS OR/RR
Published positive evidence				
France COVID-19				
RR / 1 µg/m^3 $PM_{2.5}$	96 departments, 1999–2016	63 departments, 2019	**1.24** (1.08–1.43)	**1.05** (0.95–1.15)
Italy COVID-19				
OR / doubling PM_{10} days‡	55 cities, 2018	8 cities, 2018	**1.76**	**1.74** (1.13–2.69)
São José admissions	SIH/SUS, 2020–2021	SIVEP-Gripe, 89 days in 2021		
O_3 lag 2 RR			**1.061** (1.016–1.108)	**1.003** (0.981–1.026)
NO_2 lag 3 RR			**1.125** (1.050–1.205)	**1.034** (0.876–1.220)
São Paulo AMI				
High-pollution RR	39 municipalities, 2012–2015	39 municipalities, 193 days	**1.090**	**1.069** (1.006–1.135)
UK COVID-19				
OR / 10 µg/m^3 $PM_{2.5}$	4 nations, 696 days	4 nations, 635 nation-days	**1.13** (1.09–1.16)	**1.46** (1.21–1.76)
US stroke mortality				
RR / 10 µg/m^3 $PM_{2.5}$†	US counties, 2016–2020	46 counties, 2019	**1.132**	**1.111** (0.628–1.594)
Published null or no evidence				
Brazil cancer mortality				
OR / 10 µg/m^3 $PM_{2.5}$	Brazil, 2015–2019	1,824 days; 929,790 deaths	**1.010** (1.00–1.02)	**1.060** (1.052–1.069)
Italy NO_2 COVID-19	2 regions, Feb–May 2020	2 regions, 14 days		
Incidence RR / 1 µg/m^3			**1.000***	**1.051** (1.042–1.060)
Mortality RR / 1 µg/m^3			**1.000***	**0.988** (0.987–0.990)
OECD COVID-19	36 countries	11 countries, 2017		
Day 60 cases			**1.020**	**0.998** (0.853–1.168)
Day 60 deaths			**1.030**	**1.016** (0.806–1.279)
Day 80 cases			**1.020**	**0.971** (0.926–1.018)
Day 80 deaths			**1.009**	**0.971** (0.873–1.079)

*The article reports a null effect rounded to two decimals. †Stroke ratios use the reported linear-model coefficient and mean outcome rate. ‡The article reports the underlying log-log coefficient without a confidence interval or p-value.

Northern Italy COVID-19. Chi-Chang Ho, Shih-Chieh Hung, and Wen-Chao Ho. *Effects of short- and long-term exposure to atmospheric pollution on COVID-19 risk and fatality: analysis of the first epidemic wave in northern Italy*. DOI: `10.1016/j.envres.2021.111293`. The study models daily incidence and mortality in Lombardia and Veneto. We combine public regional outcomes with the recovered AERIS NO_2 series.

São José dos Campos COVID admissions. A.O.R. Santos, B.R. Lucarevschi, C.J.D. Cunha, P.C. Ribeiro, A.C.G. Cesar, and L.F. Nascimento. *Exposure to air pollution and hospitalization due to COVID-19 in São José dos Campos, Brazil*. DOI: `10.1590/1414-431x2021e12273`. The

article estimates lagged pollutant associations with daily admissions. Because the full SIH/SUS period could not be recovered, we use public SIVEP-Gripe admissions.

São Paulo acute myocardial-infarction admissions. Marcelo H. Ribeiro, André Grossi, Adriano Caixeta, Marcelo Franken, Marcelo Katz, Vinicius Seleme, Expedito Ribeiro, Antonio Eduardo Pesaro, Jose Fabri, Sameer Mehta, and Carlos M. Campos. *Cluster of climatic and pollutant characteristics increases admissions for acute myocardial infarction: Analysis of 30,423 patients in the metropolitan area of Sao Paulo*. DOI: 10.1016/j.hrtlng.2020.04.018. The study contrasts daily admission counts across weather–pollution clusters. We use SIH/SUS admissions, Open-Meteo weather, and a reduced-input NO_2 cluster sensitivity.

Brazil cancer mortality. Pei Yu, Rongbin Xu, Yao Wu, Wenzhong Huang, Micheline S.Z.S. Coelho, Paulo H.N. Saldiva, Tingting Ye, Bo Wen, Yanming Liu, Zhengyu Yang, Shanshan Li, Michael J. Abramson, and Yuming Guo. *Cancer mortality risk from short-term $PM_{2.5}$ exposure and temporal variations in Brazil*. DOI: 10.1016/j.jhazmat.2024.134606. The paper uses a national time-stratified case-crossover design for daily cancer mortality. Our reanalysis covers 1,824 days and 929,790 deaths using public IEMA monitoring records from five states.

United States stroke mortality. Jean-Luc K. Kabangu, Danny Fowler, Amanda Hernandez, Takara Newsome-Cuby, Benson Joseph, John Dugan, Lane Fry, Momodou G. Bah, and Sonia V. Eden. *Inequities in air pollution on stroke mortality among Older Americans: a U.S. nationwide analysis*. DOI: 10.3389/fpubh.2024.1364165. The article estimates county-level $PM_{2.5}$ associations with ischemic-stroke mortality among adults aged 65+ years. We use 2019 public all-stroke mortality for 46 counties and the stated covariate families.

F Researcher interventions during the primary evaluation

AERIS found the papers and public data, wrote retrieval and analysis code, reconstructed exposure, built the analysis panels, and fit the models. The researchers intervened at scientific branch points: they asked the agent to continue after a failed route, challenged unnecessarily narrow coverage, corrected an estimand or reporting scale, or approved an explicitly labeled fallback. Table 7 records these interventions. We reconstructed the ledger from the project conversation, per-study code and result notes, prompt summaries, and the later supervision audit. Because the campaign did not retain one canonical turn-by-turn transcript, we report documented intervention classes rather than infer a precise count.

Across studies, the interventions served five recurring purposes: persistence after a failed access route, broader geographic coverage, fidelity to the article's estimand, conversion to the reported effect scale, and approval of a clearly labeled bounded analysis. They did not provide retrieval code, source URLs, exposure values, or fitted answers. This retrospective record supports the paper's claim that the primary campaign used AERIS with researcher review; it cannot establish a precise intervention count. The autonomous campaigns in Appendix G instead used instructions fixed before evaluation and received no scientific intervention.

G Autonomous AERIS evaluation

Protocol. We fixed the AERIS instructions, retry policy, schemas, safety envelope, and container before evaluation. Each study ran in a fresh workspace with Codex CLI, GPT-5.6 Terra, and medium reasoning. The agent received the article bundle and scientific target, but no prior code, source routes, manifests, exposure values, or results. It could retrieve public data, inspect endpoints, write code, revise its design, and run bounded analyses without approval. A host-side critic retried unsupported blockers and invalid handoffs. No run received human researcher intervention.

Results and stability. The three autonomous campaigns produced pollutant-effect estimates for 8, 9, and 6 studies, or 23 of 27 attempts. One campaign completed all nine studies. AERIS retained an exposure and provenance record for 26 attempts. Across the fitted, interpretable contrasts, 11 of 16 published positive directions and 14 of 21 null conclusions were retained. The pooled median study-level absolute log difference was 0.169, the geometric mean reconstructed-to-published ratio was

Table 7: Researcher-intervention ledger for the nine-study primary evaluation. Evidence: D, direct project conversation; R, reconstruction from study artifacts and prompt summaries; I, interpretation or reporting after fitting.

Study (evidence)	Researcher intervention	Scientific effect
France COVID-19 (D, R)	Continue the department-level run, broaden coverage, and disclose the use of 2019 monitor data for the article's 1999–2016 exposure period.	Increased coverage and made the period substitution explicit.
Italian cities COVID-19 (D)	Recover the largest defensible city sample; preserve the PM_{10} estimand; report the article's doubling contrast.	Changed coverage and presentation scale while preserving the target estimand.
São José admissions (D, R)	Continue after the SIH/SUS route failed; permit a labeled public-outcome fallback while retaining the lag structure.	Substituted SIVEP-Gripe outcomes and shortened the period; the analysis was labeled bounded.
São Paulo AMI (D, R)	Continue when the full pollutant-cluster specification could not be recovered; permit a reduced-input sensitivity.	Reduced model fidelity while retaining an interpretable pollution–health analysis.
UK COVID-19 (R)	Recover all four national outcome series, fit the viable core model, and add public weather covariates where available.	Preserved the target geography and pollutant with the strongest executable public-data model.
US stroke mortality (D, R)	Continue the county analysis; permit a deterministic sample and public outcome proxy; match the article's per-10-$\mu g/m^3$ scale.	Changed outcome specificity and coverage and corrected the comparison scale.
Brazil cancer mortality (D, R)	Pursue national coverage and search regulator and public-monitor archives beyond the default API.	Broadened the analysis beyond the initial municipality-level attempt.
Northern Italy COVID-19 (D, R, I)	Use the article's NO_2 analysis after $PM_{2.5}$ coverage failed; classify its rounded 1.00 estimates as reported null effects.	Changed the pollutant within the article's analyses and set the meta-analysis classification.
OECD COVID-19 (R, I)	Continue country-level retrieval, retain countries meeting the declared coverage rule, and follow the article's null conclusion.	Reduced country coverage and set the qualitative classification.

Table 8: Autonomous AERIS campaign. Effects are the principal retained contrast or contrasts. Bounded analyses record material substitutions.

Study	Result	Principal retained effect
France COVID-19	Fit	RR 1.355 (1.193–1.540)
Italian cities	Bounded fit	Ratio 1.238 (1.020–1.503) per doubling
Northern Italy COVID-19	Bounded fit	Case RR 1.006; death RR 1.005
OECD COVID-19	Fit	Four RRs, 1.026–1.135
São José admissions	Bounded fit	O_3 RR 0.692; NO_2 RR 2.198
United Kingdom COVID-19	Bounded fit	OR 1.070 (1.068–1.073)
Brazil cancer mortality	Bounded fit	OR 0.971 (0.929–1.015)
United States stroke mortality	Bounded fit	Ratio 0.989 (0.973–1.005)
São Paulo AMI	Evidenced stop	Historical pollutant route not recovered

0.901, and study-level Pearson correlations were 0.507, 0.091, and -0.170 by campaign. Campaigns took 137, 151, and 91 minutes and generated 270,297, 329,971, and 225,718 output tokens, including retries. Autonomous AERIS completed 3.3 times as many attempts as raw Codex, but did not match the numerical consistency of the campaign with researcher review.

Table 9: Study-level results from the first unaugmented-agent campaign. Ratios for completed fits use the article's reported contrast.

Study	Outcome	Effect or stopping point
Italy COVID-19	Fit	Published 1.76; Codex 1.64 (1.43–1.87)
OECD COVID-19	Fit	Published 1.02; Codex 1.08 (0.94–1.23)
UK COVID-19	Fit	Published 1.13; Codex 2.50 (2.44–2.57)
São Paulo AMI	Stopped	Pollution series and exact outcome definition unresolved
Northern Italy COVID-19	Stopped	Agent declined inference from two regional clusters
France COVID-19	Stopped	Historical $PM_{2.5}$ raster and adjustment panel unresolved
São José admissions	Stopped	Outcome retrieval failed; selected exposure source returned nulls
Brazil cancer mortality	Stopped	Historical mortality microdata retrieval failed
US stroke mortality	Stopped	Target outcome and national pollution coverage unresolved

São Paulo exposed the remaining run-to-run variance. Three independent runs produced two bounded pollutant fits and one evidenced stop. The two fits also chose different public-data windows and returned different estimates. This variation reflects long-horizon source discovery and scientific judgment, not sampling uncertainty alone. It motivates reporting complete traces, substitutions, and failures rather than a single completion score.

H Unaugmented-agent ablation

Design. We tested whether the underlying agent could complete the nine fitted reanalyses without AERIS in three independent campaigns. Each run used Codex CLI 0.155.1 with GPT-5.6 Terra at medium reasoning in a fresh container. The agent received the article bundle, common task contract, public network access, fixed retry rules, and structured output schema. It received no AERIS instructions, MCP tools, exposure engine, prior reanalysis code, or prior results. Runs had an eight-hour limit and no human researcher intervention.

Results. The three campaigns produced 3, 3, and 1 fitted effects, for 7 fits across 27 attempts. Six of seven retained the article's qualitative conclusion. Their median absolute log-effect difference was 0.073, their geometric mean reconstructed-to-published ratio was 1.011, and study-level Pearson correlations were 0.016 and -0.675 in the two campaigns with multiple fits. The third campaign produced one fit, so its correlation is undefined. The campaigns generated 103,007, 107,929, and 98,749 output tokens; the last includes no tokens for an attempt stopped at preflight. The first campaign's three fits appear in Table 9; the later campaigns added one Italian, three OECD, and one UK fit, with overlap across studies. The other 20 attempts stopped or failed before a fit even though AERIS had completed each task with researcher review. A preflight disk threshold stopped two attempts, which we classify as infrastructure-limited. In several scientific stops, the agent recognized an inadequate input but did not recover the source or bounded substitution that enabled the AERIS-assisted analysis.

Preserved artifacts. The supplementary archive preserves the 27 run manifests and each available transcript, generated program, retrieved source, analysis panel, model output or blocker record, and decision log. We store large run directories as checksum-verified archives. The archive also includes the container specification, package versions, and artifact checksums.