Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Prompt-Driven, Human-Supervised AI Research Infrastructure for Population-Scale Epidemiological Studies: Development and Validation

Ndey Isatou Jobe

njobe@arizona.edu

Fourier Health Analytics Lab

Article

Keywords: artificial intelligence, autonomous research, epidemiology, digital health infrastructure, large language models, population health, democratisation of science

Posted Date: June 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9794640/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background: Clinical research infrastructure imposes prohibitive costs and institutional dependencies that systematically exclude solo investigators and under-resourced research groups. Existing AI tools assist individual research tasks but do not provide an integrated, end-to-end research environment paired with rigorous human oversight. We describe the design, implementation, and validation of Fourier Health Analytics Lab (FHAL), a human-supervised, agent-assisted AI research infrastructure built and operated by a single investigator through natural language prompts to Claude Code — no code was written directly by the investigator — with continuous investigator review and approval of all generated outputs.

Methods: FHAL was constructed across five sequential phases using prompt-driven development through Claude Code (Anthropic). The system integrates a persistent SQLite state database, a command-line interface, automated scheduling, a multi-source literature monitor spanning PubMed and Semantic Scholar, a ChromaDB vector memory layer, and specialised analysis and writing agents. Validation was performed using an epidemiological study with a pre-specified analytical plan of consumer cosmetic chemical exposures and systemic inflammation, drawing on NHANES 2005-2010 data (N=4,864; complete cases across seven exposure biomarkers).

Results: FHAL successfully executed an end-to-end research pipeline from literature ingestion to submission-ready manuscript production. The literature monitor ingested 1,467 papers on first run across 17 search lanes, classified all papers via Claude API with PICO extraction and relevance scoring, and fired 34 Telegram alerts for high-relevance findings. The analysis agent processed 4,864 NHANES participants across three survey cycles (from a merged adult dataset of 18,161 adults aged ≥18) using survey-weighted regression, false discovery rate correction, weighted quantile sum mixture analysis, multiple imputation under Rubin's rules, and sex-stratified sensitivity analyses. Key findings included significant positive associations between BPA and log-CRP (FDR-corrected $p<0.05$) and a substantial WQS mixture effect dominated by monocyclohexyl phthalate. The writing agent produced a complete JAMA Internal Medicine manuscript package including draft, simulated peer review, cover letter, and policy brief. Total elapsed construction time was approximately 18 hours of investigator-directed prompt sessions, with all generated code reviewed by the investigator before deployment.

Conclusions: Human-supervised, agent-assisted research infrastructure is feasible, reproducible, and scientifically rigorous when paired with continuous investigator oversight. FHAL demonstrates that a solo investigator can build and operate a research platform capable of executing population-scale epidemiological studies, monitoring global literature, and producing publication-ready manuscripts without authoring code line-by-line, while retaining full scientific and ethical responsibility for every output. This approach has significant implications for democratising research capacity in under-resourced settings and for accelerating the pace of evidence generation in global health.

INTRODUCTION

Clinical and epidemiological research operates within an infrastructure bottleneck that has remained largely unchanged for decades. A typical population-scale study requires sustained institutional support — biostatisticians, research coordinators, literature analysts, grant administrators, and technical staff — whose combined cost places meaningful independent research beyond the reach of most investigators globally. The consequence is a systematic concentration of research capacity in well-resourced institutions in high-income countries, producing an evidence base that is geographically, demographically, and topically skewed.

Large language models (LLMs) and AI agent frameworks have recently demonstrated capacity to assist individual components of the research process: literature review, statistical analysis, manuscript drafting, and peer review simulation.1-4 However, these capabilities have been deployed as isolated tools rather than as integrated, autonomous research infrastructure. Existing end-to-end AI research systems, including The AI Scientist (Lu et al.)15 and OpenLens AI,16 have demonstrated automated pipelines for machine learning experimentation and health informatics. However, no prior system has been described that executes the complete epidemiological research pipeline — from persistent literature monitoring through population-scale survey-weighted analysis to manuscript production — under continuous human investigator oversight, without requiring programming expertise or a dedicated research team.

We describe the design, implementation, and empirical validation of Fourier Health Analytics Lab (FHAL), a human-supervised, agent-assisted AI research infrastructure constructed and operated by a single principal investigator. FHAL was built through prompt-driven development using Claude Code (Anthropic PBC, San Francisco): the investigator directed and reviewed all generated code rather than authoring it line-by-line. The system operates continuously under investigator oversight, monitors global literature across multiple databases, classifies and alerts on high-relevance findings, executes complex statistical analyses on population datasets, and produces submission-ready manuscripts — all with the investigator providing scientific direction, validation, and final approval at each stage.

We validate FHAL through an epidemiological study with a pre-specified analytical plan examining associations between consumer cosmetic chemical exposures and systemic inflammation, using data from the National Health and Nutrition Examination Survey (NHANES) 2005-2010 (N=4,864 in the primary analytic sample).5

The primary contribution of this work is not the findings of the validation study, but the demonstration that human-supervised, agent-assisted, end-to-end research infrastructure is achievable by a solo investigator, is scientifically rigorous when paired with continuous investigator oversight, and produces results that meet the methodological standards of high-impact clinical journals. The implications for the democratisation of research capacity — particularly for investigators in low- and middle-income countries — are substantial.

METHODS

System Architecture

FHAL was designed as a modular, phase-sequential system constructed across five working sessions between April 3 and April 9, 2026. Each phase produces fully functional outputs before the next phase begins, ensuring the system remains operational throughout construction. Phase 4 comprises two sub-phases: Phase 4a (Analysis Agent) and Phase 4b (Writing Agent). The architecture comprises five core layers: state management, literature monitoring, memory and context, analysis, and writing.

The system is controlled through a command-line interface (fhal) built using the Click framework (Python 3.9+). An APScheduler v3 job store enables persistent scheduling of automated tasks across sessions. A Telegram bot (python-telegram-bot v20+, async implementation) delivers real-time alerts when high-relevance papers are detected or when human judgment is required.

Infrastructure Construction (Phase 1)

Phase 1 established the project scaffold, SQLite database schema (six tables: projects, tasks, runs, papers, alerts, decisions), CLI, Telegram alert system, and APScheduler with persistent job store. The entire phase was executed by pasting a single structured prompt into Claude Code and allowing the agent to implement, test, and verify the resulting infrastructure without manual code editing. Construction required approximately two hours of elapsed time, inclusive of error diagnosis and resolution.

Literature Monitor Agent (Phase 2)

The literature monitor operates across 17 search lanes simultaneously, querying both PubMed E-utilities and the Semantic Scholar API. Each lane specifies a search query, maximum result count, and optional date filters. Retrieved papers are classified via the Anthropic API (claude-haiku-4-5-20251001, selected for cost efficiency at high classification volume) using a structured prompt that extracts PICO elements (Population, Intervention, Comparator, Outcome), assigns a relevance score (0.0-1.0), flags novelty, and generates a one-sentence summary.

Relevance scores incorporate a journal-priority boost (0.10 added when a paper appears in a configured priority journal list) and a novelty bonus (0.05 added when the abstract contains novelty signals such as 'first human evidence', 'real-world cohort', 'causal', 'dose-response', 'longitudinal', or 'biomonitoring'). Both values are configurable per project. Papers scoring above a configurable alert threshold (default 0.87) trigger Telegram alerts to the investigator. All papers are written to the SQLite database and vectorised into the ChromaDB memory layer. Daily markdown digests are generated summarising the classification results for investigator review.

Memory Layer (Phase 3)

A ChromaDB vector store (local persistent instance) stores embeddings of all ingested papers generated using the sentence-transformers library (all-MiniLM-L6-v2, 384-dimensional embeddings). A project journal system records all session activities, decisions, and outputs in structured markdown. A build_context() function assembles per-agent prompt context from the memory layer before each agent run, providing relevant prior work, project state, and investigator preferences without requiring manual context assembly.

Twelve agent blueprints were selected from The Agency (msitarzewski/agency-agents, MIT licence), a community-curated open-source library of Claude Code agent templates forked to NdeyJobe/agency-agents. Selected agents cover roles including code review, statistical model QA, technical writing, compliance auditing, content dissemination, and critical assessment. The investigator invokes these blueprints by name within Claude Code sessions during analysis and writing to apply specialised reasoning modes without writing custom agent code.

Analysis Agent (Phase 4a)

The NHANES Analysis Agent was configured to process continuous NHANES survey cycles (2005-2010, three cycles; monocyclohexyl phthalate was discontinued by NHANES after 2010). XPT files were downloaded directly from CDC servers. The agent implemented a pre-specified analytical pipeline incorporating: (1) survey-weighted least squares regression using Taylor series linearisation for variance estimation; (2) Benjamini-Hochberg false discovery rate correction6 across all chemical-outcome tests; (3) weighted quantile sum (WQS) mixture regression7 with 10,000 bootstrap samples; (4) multiple imputation under Missing Completely at Random (MCAR) testing via Little's test,9 with Rubin's rules8 for pooling across five imputed datasets; and (5) sex-stratified and sensitivity analyses.

A QA gate was implemented requiring the agent to complete a 12-point methodological checklist before any analysis script was executed. A human-in-the-loop approval step required investigator sign-off before each analysis run, ensuring the investigator retained final authority over analytical decisions.

Writing Agent (Phase 4b)

The Writing Agent (claude-opus-4-20250514, selected for complex long-form generation) produced a complete manuscript package following STROBE reporting guidelines10 for observational studies. The package comprised: (1) a complete journal-formatted manuscript with abstract, introduction, methods, results, and discussion; (2) a simulated peer review with three independent reviewer perspectives; (3) a journal cover letter; and (4) a policy brief for non-specialist audiences. All outputs were produced as formatted .docx files.

Validation Study: Chemical Exposures and Systemic Inflammation

The primary validation study examined associations between urinary concentrations of consumer cosmetic and skincare product chemicals — parabens (methylparaben, propylparaben, butylparaben, ethylparaben), phthalates (mono-ethyl, mono-n-butyl, mono-isobutyl, mono-benzyl, monocyclohexyl, mono-(2-ethylhexyl) phthalate), and bisphenol A (BPA)11,12 — and serum high-sensitivity C-reactive protein (hsCRP) as a marker of systemic inflammation.13

The study population comprised NHANES participants aged 18 years and older from three survey cycles (2005-2006, 2007-2008, 2009-2010) with complete data on chemical exposures and CRP outcomes. Survey sampling weights were applied throughout to produce estimates representative of the US non-institutionalised civilian population. Final analytical sample: N=4,864 (Model 1, primary; covariates: age, sex, race/ethnicity, poverty-income ratio, survey cycle). Of 18,161 adult participants across three survey cycles, 4,864 had complete measurements for all seven urinary chemical exposures, serum CRP, and primary covariates and constituted the primary analytic sample. The reduction from the full adult sample reflects the requirement for simultaneous measurement of all exposure biomarkers, particularly monocyclohexyl phthalate, which was measured in a random one-third subsample of eligible participants in each cycle. Multiple imputation across five datasets was used to handle missing covariate data within the analytic sample; participants missing any of the seven chemical exposures were excluded prior to imputation. A pre-specified sensitivity analysis additionally adjusted for body mass index and smoking status (N=4,494, age ≥20).

All chemical concentrations were log2-transformed prior to analysis. The primary outcome was log-transformed hsCRP. Covariates in the primary analysis (Model 1) were age, sex, race/ethnicity, poverty-income ratio, and survey cycle. Body mass index and smoking status were included in a pre-specified sensitivity analysis (Model 2) to assess potential confounding versus mediation by adiposity; physical activity was not adjusted for in either model. The analysis was pre-specified in a statistical analysis plan generated by the Analysis Agent and approved by the investigator prior to data access.

Statistical Analysis

All statistical analyses were implemented in Python (≥3.9) using statsmodels (v0.14) for survey-weighted regression, scikit-learn (v1.4) IterativeImputer for multiple imputation, and custom WQS implementation. Survey weights, strata, and primary sampling units were specified according to NHANES analytic guidelines. Statistical significance was assessed at FDR-corrected $q<0.05$. All analysis code was generated by the FHAL Analysis Agent and is available in the project repository.

Reproducibility and Code Availability

The complete FHAL infrastructure was constructed using only natural language prompts — no code was authored directly by the investigator. All generated code was reviewed and approved by the investigator before deployment; the investigator bears full scientific and ethical responsibility for all outputs produced by the system. All prompts used to construct each phase of the system are documented verbatim in the companion implementation guide. The system is designed to be fully reproducible by any investigator following the documented prompt sequences, regardless of programming background. Code and configuration files are available at https://github.com/NdeyJobe/fourier-hal.

Ethics

NHANES data are de-identified and publicly available from the Centers for Disease Control and Prevention. No additional ethics approval was required for this secondary analysis.

Human Oversight, Investigator Validation, and AI Use Disclosure

All analysis code generated by the FHAL Analysis Agent was reviewed by the principal investigator (N.I.J.) prior to execution. The statistical analysis plan was approved by the investigator before data access. All regression outputs, FDR-corrected results, WQS mixture weights, and imputed dataset results were manually verified by the investigator against NHANES analytic guidelines and published epidemiological methods prior to inclusion in any manuscript. The Writing Agent's manuscript draft was reviewed, substantively edited, and approved by the investigator before submission. The investigator independently confirmed that the reported findings from the validation study — including significant associations for BPA, monoethyl phthalate, and monocyclohexyl phthalate, and null findings for methylparaben and butylparaben — were consistent with the underlying analysis outputs before accepting them as reported results.

Claude Code (Anthropic PBC, San Francisco, CA) was used for infrastructure construction, code generation, literature classification, and manuscript drafting throughout this work. All research-consequential decisions — including study design, selection of analytical strategy, interpretation of findings, assessment of methodological limitations, and ethical responsibility for all published claims — were made exclusively by the investigator. No AI-generated content was included in any deliverable without prior investigator review and explicit approval. This disclosure is made in accordance with the Springer Nature policy on the use of large language models in submitted manuscripts.

RESULTS

Infrastructure Construction

FHAL was constructed across five working sessions spanning six calendar days (April 3-9, 2026). Table 1 summarises the construction timeline, components built in each phase, and key performance metrics. Total elapsed construction time was approximately 18 hours inclusive of error diagnosis and resolution. Implementation was carried out via prompt-driven AI agents under continuous investigator supervision; all generated code was reviewed by the investigator before deployment.

Table 1. FHAL Construction Timeline and Phase Summary

Phase	Component	Build Time	Key Metric
1	Scaffold, SQLite DB, CLI (fhal), Telegram alerts, APScheduler	~3 hours	6-table schema, persistent job store
2	Literature Monitor Agent — 17 search lanes, PubMed + Semantic Scholar	~4 hours	1,467 papers ingested on first run
3	ChromaDB memory layer, project journal, context builder, 12 agent blueprints	~3 hours	1,467 papers vectorised
4a	NHANES Analysis Agent — QA gate, survey weighting, FDR, WQS, imputation	~5 hours	4,864 participants processed (Model 1 primary; merged adult dataset 18,161)
4b	Writing Agent — STROBE manuscript, peer review simulation, cover letter	~3 hours	Complete manuscript package

Literature Monitoring Performance

On initial deployment, the literature monitor ingested 1,467 papers from PubMed and Semantic Scholar across 17 search lanes focused on the chemical safety and human exposure domain. Classification via the Claude API assigned relevance scores to all 1,467 papers with PICO extraction. Thirty-four papers scored above the 0.87 alert threshold and triggered Telegram alerts; five papers received perfect scores of 1.00. The cost of classifying 1,467 papers using claude-haiku-4-5-20251001 was approximately USD 1.20.

Validation Study: Chemical Exposures and Systemic Inflammation

The analytical sample comprised 4,864 NHANES participants from three survey cycles (2005-2006, 2007-2008, 2009-2010) after applying exclusion criteria and survey weight processing; the analytical window was set by the availability of monocyclohexyl phthalate, which was discontinued by NHANES after 2010. The initial merged adult dataset (age ≥18) comprised 18,161 individuals across these three cycles. Little's MCAR test indicated data were not missing completely at random ($p<0.001$), supporting

the use of multiple imputation. Five imputed datasets were generated using iterative imputation; results were pooled under Rubin's rules.

In survey-weighted Model 1 (primary) regressions adjusted for age, sex, race/ethnicity, poverty-income ratio, and survey cycle, BPA showed a statistically significant positive association with log-CRP after FDR correction (β = +0.052, 95% CI: 0.019–0.086, FDR-q = 0.005). Monoethyl phthalate also remained significant after correction (β = +0.043, 95% CI: 0.019–0.067, FDR-q = 0.002). Monocyclohexyl phthalate showed the largest individual association (β = +0.215, 95% CI: 0.065–0.365, FDR-q = 0.007), with each doubling of urinary concentration associated with a 16.1% increase in CRP. Two paraben metabolites (ethylparaben, β = −0.056, FDR-q < 0.001; propylparaben, β = −0.026, FDR-q = 0.007) showed FDR-significant inverse associations of uncertain mechanism; methylparaben and benzophenone-3 did not reach significance after FDR correction.

WQS mixture regression identified a substantial positive mixture effect on log-CRP (WQS index β = +0.056, 95% CI: 0.031–0.081, p<0.001, bootstrapped n=10,000). The mixture was dominated by monocyclohexyl phthalate (weight = 0.342), followed by BPA (weight = 0.307) and mono-ethyl phthalate (weight = 0.267); these three chemicals accounted for approximately 91% of the mixture signal. Sex-stratified analyses of monocyclohexyl phthalate showed an approximately two-fold larger association in females (β = +0.267, p = 0.008) than males (β = +0.141, p = 0.241; not statistically significant), consistent with differential cosmetic product use patterns. A pre-specified BMI- and smoking-adjusted sensitivity analysis (N=4,494, age ≥20) attenuated all individual chemical associations to non-significance after FDR correction; MCP preserved direction (β = +0.129, FDR-q = 0.43) and its mixture weight increased to 0.494, indicating its contribution is at least partially independent of adiposity-mediated pathways.

Table 2 presents the primary regression results. These findings were materially altered relative to naive analyses by three methodological upgrades implemented by the Analysis Agent: survey weighting, FDR correction, and full missingness analysis. Adjustment for BMI and smoking status (Model 2 sensitivity, N=4,494) attenuated all individual chemical associations to non-significance after FDR correction; monocyclohexyl phthalate was the only chemical to maintain its direction across all four analytical specifications examined (primary complete-case, BMI-adjusted complete-case, primary-aligned multiple imputation, and BMI-adjusted multiple imputation), reaching statistical significance in three of four, providing convergent evidence of a robust MCP–inflammation association. This pattern underscores the importance of methodological rigour in this domain.

Table 2. Primary Results: Chemical Exposures and log-CRP (N=4,864, NHANES 2005-2010, Model 1 primary)

Chemical	β (95% CI)	p-value	FDR-q	Significant
Bisphenol A (BPA)	+0.052 (0.019, 0.086)	0.002	0.005	Yes
Monoethyl phthalate (MEP)	+0.043 (0.019, 0.067)	<0.001	0.002	Yes
Monocyclohexyl phthalate (MCHP)	+0.215 (0.065, 0.365)	0.005	0.007	Yes
Ethylparaben (EPB)	−0.056 (−0.079, −0.033)	<0.001	<0.001	Yes (inverse)
Propylparaben (PPB)	−0.026 (−0.044, −0.008)	0.005	0.007	Yes (inverse)
Methylparaben (MPB)	−0.023 (−0.046, +0.000)	0.055	0.064	No
Benzophenone-3 (BP-3)	−0.015 (−0.032, +0.002)	0.084	0.084	No
WQS Mixture Index	+0.056 (0.031, 0.081)	<0.001	N/A	Yes

All models survey-weighted with Taylor series linearisation. FDR correction: Benjamini-Hochberg. Model 1 (primary) covariates: age, sex, race/ethnicity, poverty-income ratio, survey cycle. Body mass index and smoking status were added in the Model 2 sensitivity analysis (N=4,494, age ≥20), under which all individual chemical associations attenuated to non-significance after FDR correction; MCP preserved direction (β=+0.129, FDR-q=0.43). WQS bootstrap n=10,000.

Writing Agent Output

The Writing Agent produced a complete JAMA Internal Medicine manuscript package in a single session. The package comprised a 4,800-word manuscript with structured abstract, introduction, methods, results, and discussion sections; an adversarial peer-review simulation generated from three reviewer personas — environmental epidemiologist, biostatistician, and clinical/regulatory scientist — producing the five hardest objections the manuscript would face at a high-impact journal, covering causal-inference design, multiple comparisons, WQS methodology, clinical significance, and unmeasured confounding; a 400-word journal cover letter; and a 600-word policy brief. The initial draft was formatted per JAMA Internal Medicine guidelines and reviewed by the investigator against STROBE reporting criteria. The manuscript was subsequently revised, reformatted per Environmental Health Perspectives author guidelines, and submitted to that journal.

DISCUSSION

Principal Findings

We demonstrate that a human-supervised, agent-assisted, end-to-end research infrastructure is achievable by a solo investigator through prompt-driven AI development under continuous investigator oversight, and that such a system can execute epidemiological research to standards appropriate for submission to high-impact clinical journals. FHAL successfully monitored global literature, processed a

national survey dataset of 4,864 adults (from a merged sample of 18,161 across three NHANES cycles) using state-of-the-art statistical methods, and produced a complete manuscript package — all within six calendar days, at infrastructure cost less than USD 50, and with every research-consequential decision under explicit investigator review.

The validation study produced substantively meaningful findings. BPA and monoethyl phthalate showed significant positive associations with systemic inflammation after rigorous FDR correction, with a substantial mixture effect dominated by monocyclohexyl phthalate. The sex-stratified finding — a larger MCP–CRP association in females than males — is consistent with differential cosmetic product use patterns and is scientifically informative. Importantly, these findings were materially different from naive analyses, confirming that the methodological rigour implemented by the Analysis Agent was not cosmetic but consequential.

Implications for Research Democratisation

The most significant implication of this work is not methodological but structural. FHAL was built by a single investigator, at minimal cost, without a dedicated research team, without a statistical team, and without programming expertise. The entire infrastructure cost less than USD 50 in API charges to construct and operate through the validation study.

This has direct implications for research equity. The conventional requirements for population-scale epidemiological research — statistical expertise, research coordinators, literature analysts, institutional computing resources — are not equitably distributed globally. Investigators in low- and middle-income countries, independent researchers, and those outside major academic centres are systematically excluded from this class of research. Prompt-driven autonomous infrastructure removes these barriers without sacrificing methodological rigour.

The Collective Narrowing Risk

The democratisation argument advanced here must be tempered by an important countervailing finding. Hao et al.14 analysed 41.3 million research papers and found that while AI adoption substantially increases individual scientific productivity, it is associated with a 4.63% decrease in the diversity of scientific topics studied and a 22% decrease in scientific engagement across research areas. AI tools, optimised against patterns of successful prior work, may systematically direct investigators toward established, data-rich questions at the expense of exploratory inquiry.

FHAL is not immune to this risk. The literature monitor's relevance scoring, journal boosts, and novelty flags are all calibrated against existing epidemiological literature. A system that surfaces and prioritises work resembling prior high-impact findings will, over time, bias the investigator toward those same research directions. Realising the democratisation potential of human-supervised AI research

infrastructure without reproducing this narrowing effect will require deliberate design choices — including explicit mechanisms for surfacing under-studied questions, underrepresented populations, and methodological approaches absent from the dominant literature. FHAL's multi-study architecture, while currently configured for established epidemiological domains, is designed to be domain-agnostic; future work should explore whether investigator-directed prompt design can counteract the convergence tendencies documented by Hao et al.14.

The Role of Human Judgment

FHAL is designed to minimise human intervention, not eliminate it. The human-in-the-loop approval gate before each analysis run, the Telegram alert system for high-relevance findings, and the investigator sign-off on the statistical analysis plan all reflect a deliberate design principle: the investigator retains final scientific and ethical authority. Autonomy is applied to the mechanical, repetitive, and computationally intensive components of research. Judgment — about study design, analytical strategy, interpretive framing, and ethical implications — remains with the investigator.

This distinction is important for the integrity of the scientific record. AI-generated analyses require human scientific oversight. FHAL is infrastructure for amplifying investigator capacity, not for replacing investigator judgment.

Throughout this manuscript we therefore describe FHAL as human-supervised, agent-assisted rather than autonomous. The distinction is conceptual, not merely terminological. AI agents perform implementation work — code generation, literature classification, statistical execution, manuscript drafting — at a scale and speed no individual investigator could match. The investigator performs scientific work — selecting research questions, defining analytical strategies, validating outputs against domain knowledge, granting ethical approvals, and bearing final responsibility for every published claim. Both roles are essential and neither substitutes for the other. The infrastructure is designed to maximise the productivity of the human-agent partnership while keeping every research-consequential decision under explicit human review. We anchor the rest of this work to that framing.

Limitations

Several limitations should be noted. First, the validation study is cross-sectional in design; causal inference about chemical exposures and inflammation requires longitudinal or intervention data. Second, urinary chemical concentrations represent a single point-in-time measurement and are subject to within-person variability and measurement error. Third, the literature monitor operates on published abstracts and titles; full-text analysis would improve classification accuracy. Fourth, WQS mixture analysis assumes a unidirectional mixture effect; this assumption should be relaxed in future analyses using quantile g-computation. Fifth, the LLM-generated analysis code, while reviewed by the investigator and validated against published methods, has not been independently audited.

More broadly, the reproducibility of prompt-driven development depends on the stability of the underlying LLM. Prompt sequences that produced specific outputs in April 2026 may produce different outputs with updated model versions. Researchers replicating FHAL should pin specific model versions and validate outputs against expected behaviour.

Future Directions

FHAL's architecture is designed for multi-study operation — the project registry, per-project configuration, and scheduled job store support concurrent research programmes without modification to core infrastructure; a fairness-aware clinical prediction programme for end-stage renal disease (ESRD) — extending the chemical exposure–inflammation findings of the present study to longer-term kidney function outcomes, with external validation in UK Biobank and All of Us and explicit subgroup evaluation in populations of African ancestry — is in preparation as the second application of the system.

Beyond chemical safety and renal epidemiology, FHAL's architecture is domain-agnostic. Future applications to mental health surveillance, nutritional epidemiology, environmental health monitoring, and infectious disease literature synthesis are planned. The addition of causal inference modules — difference-in-differences, instrumental variable analysis, regression discontinuity — would further extend the analytical repertoire available to solo investigators.

CONCLUSION

Human-supervised, agent-assisted, prompt-driven research infrastructure is feasible, scientifically rigorous, and accessible to solo investigators who author no code directly, requiring only the ability to specify research objectives in natural language and to evaluate AI-generated outputs against domain knowledge, when paired with continuous investigator oversight of all research-consequential decisions. FHAL demonstrates that the principal bottlenecks to independent population-scale research — statistical capacity, literature monitoring, manuscript production — can be addressed through AI agent infrastructure under human direction without sacrificing methodological standards. The validation study produced findings consistent with established epidemiological literature on consumer chemical exposures and systemic inflammation.

The implications for global research equity are significant. A solo investigator with domain expertise, a research question, and access to an LLM can now build and operate research infrastructure that supports population-scale studies which would otherwise require substantial institutional resources, while preserving the investigator's role as the locus of scientific judgment. Realising this potential equitably will require attention to API cost accessibility, data access equity, and the development of validation standards for AI-generated research pipelines. We offer FHAL as a proof-of-concept and an open invitation for replication, extension, and critique.

Declarations

DATA AVAILABILITY

NHANES data are publicly available from the US Centers for Disease Control and Prevention at https://www.cdc.gov/nchs/nhanes. FHAL infrastructure code, configuration files, and all prompts used for construction are available at https://github.com/NdeyJobe/fourier-hal.

CODE AVAILABILITY

All analysis code generated by FHAL for the validation study is available in the project repository. The complete prompt sequences used to construct each phase of the infrastructure are documented in the companion implementation guide: Jobe NI. The 6-Day Research Lab. manuscript in preparation, 2026.

COMPETING INTERESTS

The author declares no competing interests.

FUNDING

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. Infrastructure operating costs (LLM API access) totalled less than USD 50 and were borne by the investigator.

AUTHOR CONTRIBUTIONS

N.I.J. conceived the study, designed and constructed all infrastructure, conducted all analyses, and wrote the manuscript.

ACKNOWLEDGEMENTS

The author thanks the NHANES participants and the CDC for making these data publicly available. FHAL infrastructure was built using Claude Code (Anthropic PBC).

References

1. Thirunavukarasu AJ, Ting DSJ, Elangovan K, et al. Large language models in medicine. Nat Med. 2023;29(8):1930-1940.
2. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. Nature. 2023;620(7972):172-180.
3. Moor M, Banerjee O, Abad ZSH, et al. Foundation models for generalist medical artificial intelligence. Nature. 2023;616(7956):259-265.
4. Liang W, Tadesse GA, Ho D, et al. Advances, challenges and opportunities in creating data for trustworthy AI. Nat Mach Intell. 2022;4(8):669-677.

5. Centers for Disease Control and Prevention. National Health and Nutrition Examination Survey Data. Hyattsville, MD: US Department of Health and Human Services, Centers for Disease Control and Prevention, 2005-2016.
6. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B. 1995;57(1):289-300.
7. Carrico C, Gennings C, Wheeler DC, Factor-Litvak P. Characterization of weighted quantile sum regression for highly correlated data in a risk analysis setting. J Agric Biol Environ Stat. 2015;20(1):100-120.
8. Rubin DB. Multiple Imputation for Nonresponse in Surveys. New York: Wiley; 1987.
9. Little RJA. A test of missing completely at random for multivariate data with missing values. J Am Stat Assoc. 1988;83(404):1198-1202.
10. von Elm E, Altman DG, Egger M, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement. PLoS Med. 2007;4(10):e296.
11. Meeker JD, Calafat AM, Hauser R. Urinary bisphenol A concentrations in relation to serum thyroid and reproductive hormone levels in men from an infertility clinic. Environ Sci Technol. 2010;44(4):1458-1463.
12. Stacy SL, Eliot M, Etzel TM, et al. Patterns, variability, and predictors of urinary bisphenol-A concentrations during childhood. Environ Sci Technol. 2016;50(11):5396-5403.
13. James-Todd T, Stahlhut R, Meeker JD, et al. Urinary phthalate metabolite concentrations and diabetes among women in the National Health and Nutrition Examination Survey (NHANES) 2001-2008. Environ Health Perspect. 2012;120(9):1307-1313.
14. Hao Q, Xu F, Li Y, Evans J. Artificial intelligence tools expand scientists' impact but contract science's focus. Nature. 2026;649:1237-1243.
15. Lu C, Lu C, Lange RT, et al. The AI Scientist: towards fully automated open-ended scientific discovery. Nature. 2026. doi:10.1038/s41586-026-10265-5
16. Cheng Y, Suo J. OpenLens AI: fully autonomous research agent for health informatics. arXiv:2509.14778 [Preprint]. 2025. Available from: https://arxiv.org/abs/2509.14778