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.

An OMOP-Compatible Exposure Infrastructure for GeoAI-Enabled Environmental Health Research: The OHDSI GIS Toolchain

Andrew E Williams
Tufts Clinical and Translational Science Institute
Kyle Zollo-Venecek
Valo Health
Robert Miller
Tufts Clinical and Translational Science Institute
Polina Talapova
polina.talapova@sciforce.tech

SciForce
Timothy Norris
University of Miami
Jay Greenfield
CODATA
Nitin Parikh
University of Miami
Jimmy Phuong
University of Washington
Anne E Thessen
University of North Carolina at Chapel Hill
Timothy D Rossi
Architectural Medicine LLC
Alexia Sampri
D4L Data4Life gGmbH
Jiawen Liao
Cleveland Clinic Foundation
Sara Khalid
University of Oxford
Maksym Trofymenko
SciForce
Jared Houghtaling

Tufts Clinical and Translational Science Institute

Article

Keywords:

Posted Date: November 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11436369/v1

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

Additional Declarations: No competing interests reported.

An OMOP-Compatible Exposure Infrastructure for GeoAI-Enabled Environmental Health Research: The OHDSI GIS Toolchain

Andrew E Williams[1†], Kyle Zollo-Venecek[1,2], Robert Miller[1], Polina Talapova[1,4*], Timothy Norris[5], Jay Greenfield[6], Nitin Parikh[7], Jimmy Phuong[8], Anne E Thessen[9], Timothy D. Rossi[10], Alexia Sampri[12,13], Jiawen Liao[14], Sara Khalid[11,15],Maksym Trofymenko[4], Jared Houghtaling[1,3]

1 Tufts Clinical and Translational Science Institute, Boston, MA, USA
2 Valo Health, Lexington, MA, USA
3 Observational Health Data Analytics, Global Epidemiology, Johnson & Johnson Research and Development, Beerse, Belgium
4 SciForce, Kharkiv, Ukraine
5 The University of Miami, Coral Gables, FL, USA
6 CODATA, Paris, France
7 University of Miami, Clinical and Translational Science Institute, Coral Gables, FL, USA
8 University of Washington, Seattle, WA, USA
9 University of North Carolina Chapel Hill, Chapel Hill, NC, USA
10 Architectural Medicine LLC, Santa Fe, NM, USA
11 Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences (NDORMS), University of Oxford, Oxford, United Kingdom
12 D4L Data4Life gGmbH, Potsdam, Brandenburg 14482, Germany
13 Hasso Plattner Institute, Digital Engineering Faculty, University of Potsdam, Potsdam, Germany
14 Cleveland Clinic Foundation, Cleveland, OH, USA
15 UK Health Security Agency, London, United Kingdom

† Special note about deceased authorship

* Corresponding author: Polina Talapova

Email: polina.talapova@sciforce.tech

Abstract

Environmental and socio-economic contexts shape population health. Despite increasing availability of data about such contexts, no open infrastructure standardizes these data into person-place-time exposure records for federated health research. We present the OHDSI GIS toolchain, an open-source infrastructure integrating geospatial data with the OMOP Common Data Model through a FAIR catalog, a GIS vocabulary, local geocoding, and a geometry-attribute index that registers every source uniformly. We tested the architecture on a synthetic cohort of 40,000 persons in real US census tracts, with eight simulated PM2.5 effects and six null outcomes. Derived exposures matched an independent answer key, and known effects were recovered. Displacing residences by empirical geocoding error changed estimates little for a smooth regional exposure but attenuated them by up to half for exposures varying over 100 m. These results show that the toolchain and its standards enable reproducible, transparent linkage of geospatial sources with health records.

1. Introduction

1.1 Geospatial Determinants of Population Health

The burden of both non-communicable and infectious disease is distributed unequally across geography, and that distribution is not random. Air pollution contributes to respiratory and cardiovascular disease [1]; neighborhood socioeconomic deprivation affects access to care, health behaviours, and disease risk [2]; and proximity to environmental hazards correlates with adverse outcomes that disproportionately affect communities of color and low-income populations [3]. Environmental conditions also shape infectious disease: climatic hazards such as warming, heavy precipitation and flooding have been linked to the aggravation of more than half of known human pathogenic diseases.[4] These spatially patterned exposures are not incidental to health inequity - they are among its structural causes. Understanding and ultimately addressing them requires linking where people live to what happens to their health, at population scale, across diverse geographic and demographic contexts.

Despite this evidence, most large-scale observational health research remains anchored in clinical data captured in electronic health records (EHRs) and associated insurance claims. Geospatial and environmental context (air quality indices, social vulnerability indices, neighborhood deprivation scores, and proximity to green space or industrial facilities) is typically stored externally, described inconsistently, processed in bespoke local workflows, and excluded from federated analyses entirely. The result is that clinical observational databases, which now cover hundreds of millions of patient records across international networks [5,6], largely omit the environmental and social conditions that shaped the outcomes they record unless those external exposures are represented in a standardized, reproducible and privacy-preserving way.

1.2 The Infrastructure Gap

This disconnect between spatial information and health data is driven by several factors. Geospatial datasets are often fragmented across government agencies, academic repositories, and commercial providers, with heterogeneous formats, spatial resolutions, temporal coverage, and licensing constraints. Patient address information raises serious privacy concerns under HIPAA and GDPR, complicating geocoding outside institutional boundaries. In the European Union, the European Health Data Space (EHDS) Regulation lists ‘data on factors impacting on health, including socio-economic, environmental and behavioural determinants’ among the minimum categories of electronic health data to be made available for secondary use, applicable from March 2031 [7]. Therefore, regulatory demand for standardized exposure data is arriving before the infrastructure needed to represent it. Expertise required to create tools that transform and analyze these data effectively spans clinical informatics, epidemiology, GIS, spatial databases, privacy governance, and vocabulary engineering; such an ensemble is rarely available within a single research team. Most fundamentally, there is no widely adopted, vocabulary-governed, provenance-aware way to express how an external exposure variable, measured at a geographic unit or spatial surface, becomes an auditable person-level exposure record usable in federated observational research.

This representational gap is not merely a technical inconvenience. Without a standardized semantic bridge connecting persons, locations, time intervals, exposure variables, values, spatial relationships, units, and provenance, researchers risk treating area-level social or environmental context as an individual clinical fact, generating exposure variables through undocumented preprocessing that cannot be audited across sites, or, most commonly, simply omitting geospatial context altogether.

1.3 The OHDSI Software Ecosystem and Its Geospatial Extension

The Observational Health Data Sciences and Informatics (OHDSI) community has established an open international standard: the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) [8]. More than 540 data sources in 83 countries have been mapped to the OMOP CDM, which harmonizes clinical observational data into a common structure with controlled vocabularies (the OHDSI Standardized Vocabularies) [5,6]. In Europe, the OMOP CDM also strengthens the European Health Data & Evidence Network (EHDEN) and the European Medicines Agency's DARWIN EU network for regulatory real world evidence [9]. The model's standardization enables researchers to execute analytic code across institutions without sharing patient-level data, a federated architecture that has supported landmark studies in pharmacoepidemiology, comparative effectiveness, and more recently social determinants of health [6,10]. The OHDSI software ecosystem includes ATLAS for graphical phenotype development and characterization [6], the HADES R package suite for statistical analyses [11], and a variety of packages for data transformation, quality validation, and visualization. The toolchain connects to this ecosystem through two extension packages that integrate the GIS tables with standard OHDSI analytic workflows [12] (Section 2.2).

The OMOP CDM was designed to standardize clinical events generated within healthcare systems. It has no model for external geospatial data such as administrative boundaries, raster surfaces, pollutant estimates, social vulnerability indices, location histories, or time-varying exposure assignments. Extending this model requires new tables and new semantics, developed so that they preserve OHDSI's core principles: standardization, reproducibility, privacy preservation, and federated analytic compatibility.

1.4 Objectives

GeoAI and public health intelligence are increasingly bottlenecked not by modeling algorithms but by the availability of harmonized, unit-resolved, provenance-tracked person-place-time data and infrastructure [13]. We present the OHDSI GIS toolchain as that foundation: a geospatial exposure infrastructure for geoAI-ready OMOP-compatible environmental health observational research (**Figure 1**). Rather than introducing a new geoAI model, its primary contribution is a standardized representation of heterogeneous geospatial sources: a geometry-attribute index (geom_index/attr_index) that registers every geometry and attribute dataset, regardless of native format, resolution, or time coverage, in a single uniform structure, and drives the standardized, provenance-aware generation of person-place-time exposure records from it. This is a prerequisite for, rather than an instance of, geoAI and spatial big-data research, with explicit support for health

equity and climate justice research that improves understanding of environmental, socio-demographic, geographic, and wider determinants of health.

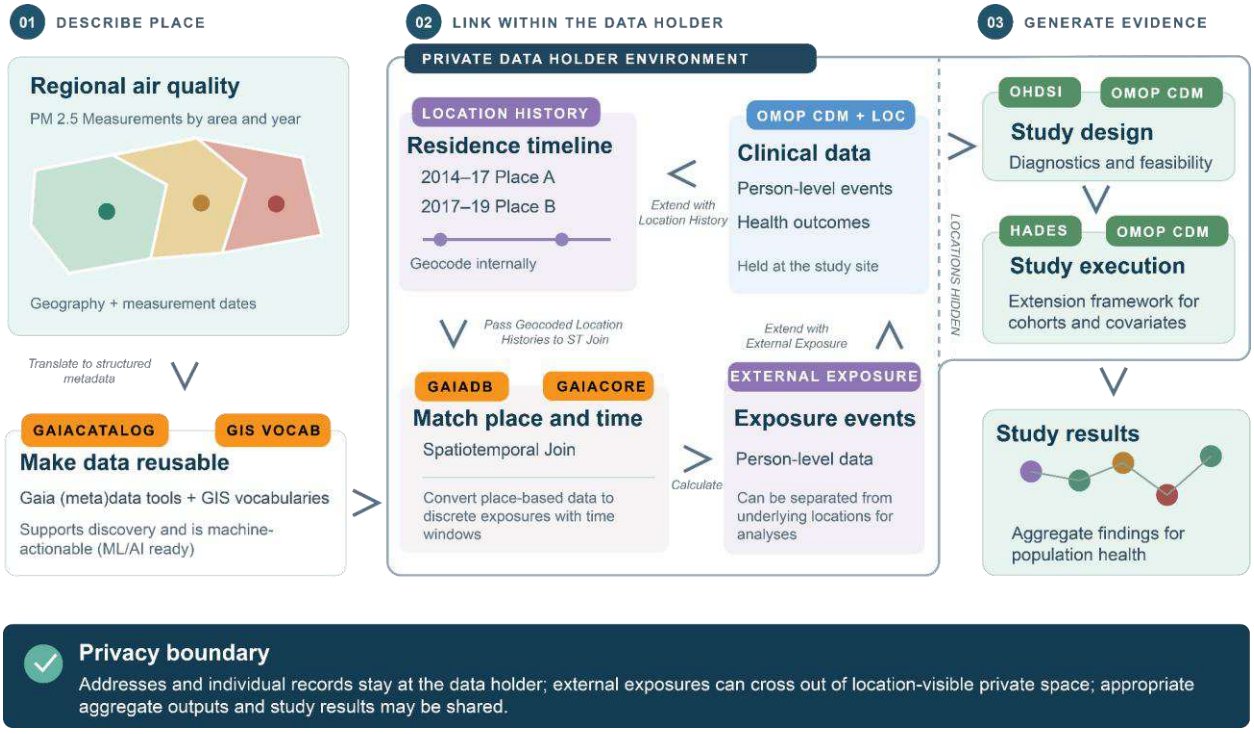


Figure 1. The OHDSI GIS toolchain workflow. *(1) Describe place: regional PM2.5 measurements, defined by geography and measurement dates, are described in GaiaCatalog and mapped to GIS Vocabulary concepts, making the data discoverable and machine-actionable. (2) Link within the data holder: inside the data holder's private environment, residence histories are geocoded locally and stored in LOCATION_HISTORY alongside OMOP CDM clinical data. GaiaDB and GaiaCore join place and time through the geometry-attribute index, producing person-level EXTERNAL_EXPOSURE records that can be separated from the underlying locations. (3) Generate evidence: OHDSI tools analyze the exposures together with clinical data. By design, addresses stay in a location-visible enclave within the data holder's infrastructure, and only appropriate aggregate results are shared.*

Janowicz et al. posed a longer-term moonshot: "Can we design a software agent that takes a user's GIS-related domain question, understands how to gather the required data, how to analyze them, and how to publish the results in a suitable form [by 2030]?" [14, p. 631] We do not build such an agent; we provide the standardized, machine-readable inputs that such work would need. Specifically, we aimed to: (i) define OMOP-compatible structures for location histories and external exposure records; (ii) develop a GIS Vocabulary Package standardizing geographic entities, spatial relationships, environmental exposures, SDOH, and provenance semantics; (iii) implement a geometry-attribute index through which diverse geometry and attribute sources are registered, loaded, and spatially and temporally joined to person locations; (iv) implement local geocoding and spatial-temporal processing workflows deployable behind institutional firewalls; (v) build a FAIR-compliant, machine-readable metadata catalog supporting discovery and reuse of heterogeneous geospatial datasets that, critically, contains the details needed to transform its sources into

standardized geometry and attribute forms; (vi) use synthetic cohorts, which combine geometry, environmental, and socio-economic sources, as a controlled substrate to test whether the standardized representation recovers known exposure–outcome relationships and how geocoding error propagates to exposure and effect estimates; and (vii) outline a path toward characterizing the real-world, multi-site AI-readiness of OMOP-standardized data for geospatial exposure linkage (Section 4.6).

2. Methods

2.1 System Architecture

The toolchain is modular: GaiaDB, GaiaCore, and GaiaCatalog (below) can each be run on their own, and a reference deployment packages them as containers following the OHDSI Broadsea approach [15]. Local containerized execution supports reproducible installation and execution behind institutional firewalls, consistent with the local-execution model of OHDSI network studies, and helps keep geospatial processing separate from protected clinical systems. Containerization is important but not sufficient for privacy compliance: institutional governance, data-flow controls, and local review remain required. Deployment options and requirements are described in Section 3.1, and technical documentation is available online [16]. The toolchain is also accompanied by an open tutorial, a set of worked exercises that run the full pipeline on a synthetic 10,000-person OMOP CDM dataset, from loading data to estimating an exposure effect with CohortMethod [16]. Later sections refer to it as "the tutorial"

2.1.1 GaiaDB

GaiaDB is the relational database where external socio-economic and environmental data are processed and joined to location data in the OMOP CDM. It is built on the PostgreSQL relational database manager with the PostGIS extension, which provides native support for geometric data types and spatial operations. For each generated exposure record, GaiaDB records available provenance fields - currently the identifier of the source variable, from which the source dataset follows - to support reconstruction of the exposure assignment; a richer set of spatial, temporal, semantic, and processing provenance fields is planned for future releases (Section 4.6).

2.1.2 Geometry-attribute index (within GaiaDB)

The central design element of GaiaDB is a common index that standardizes how diverse geospatial sources are represented. Every loaded geometry source is registered in a geometry index (geom_index) and every attribute source, meaning the variables measured on those geometries, in an attribute index (attr_index). Each registration assigns a stable identifier and points to a per-source table in the *working* schema, created from a template that fixes its structure; attribute records reference their geometry and carry variable name, value, unit, and start and end dates. A source that arrives as polygons, a monthly time series of pollutant estimates, or a table of socio-economic indicators is therefore reduced to the same two kinds of object, whatever its native format, spatial resolution, or temporal coverage (Table 2). Spatial and temporal joins to person location histories are run against

these per-source tables as described by the index, rather than by source-specific code, so a new source is cataloged and registered without writing a new join. This lets heterogeneous sources reach person locations through one uniform mechanism and yield EXTERNAL_EXPOSURE rows with the same structure (Section 2.2). Note that the current implementation covers vector geometries; raster sources would follow a related but not identical approach (Section 4.6).

2.1.3 GaiaCore

GaiaCore is an R package that serves as the interface to GaiaDB. It provides functions to run and orchestrate the stages of the pipeline (Section 2.5), to validate the derived exposure records, and to analyze them, for example by calculating day-weighted exposure over a time window or by estimating exposure effects with confidence intervals that remain valid when there are few, unbalanced geographic clusters. It connects to the database through the OHDSI DatabaseConnector package and is distributed in an analysis container together with the HADES packages and the extension packages described in Section 2.2 [11]. A PostgREST API container deployed alongside GaiaDB optionally provides API-level access from other programming languages. Further spatial analyses, such as cluster analysis and spatial autocorrelation measures like Moran's I, are planned (Section 4.6).

2.1.4 GaiaCatalog

GaiaCatalog is a metadata catalog following FAIR (Findable, Accessible, Interoperable, Reusable) principles, exposing metadata as a schema.org JSON-LD Dataset shaped by the Science on Schema.org (SOSO) guidance on Datasets, DataRepositories and Experimental Recommendations. Catalog entry structure is inspired by the RO-Crate architecture, where descriptive, technical and administrative metadata reside with ETL logic all in the same directory as the extracted source data [17]. The JSON metadata can be indexed with standard search engines like SOLR or ElasticSearch, can be converted to a triple store to create knowledge graphs, and can be leveraged to create the logic for automated data retrieval and ETL. The indexes and knowledge graphs can be a back end for catalog browsers that feature both free text searching and discovery through relationships. For example, the indexed JSON enables full-text/concept-based search across catalog entries as well as discovery by spatial and temporal coverage at the Dataset level; it also allows for discovery based on properties at the variable level, including variable description, one or more variable concept IDs mapped to different vocabularies, and variable datatype using the QUDT model. The triple store enables knowledge-graph visualization, querying and traversing of relationships between data sources, variables, and concepts, and is at an experimental stage of development. The catalog's technical metadata is currently processed by the GDSC data curation Python module [18] to create source-specific shell scripts that leverage the GDAL/OGR cross-platform command line toolkit [19] and provide efficient ETL logic for each dataset. GaiaDB then loads the dataset's variables according to its variable-level metadata. This loading step currently runs as source-specific scripts; a declarative recipe interpreted in-database is planned (Section 4.6). The catalog includes a KoboToolbox authoring tool and uses SSSOM (Simple Standard for Sharing Ontological Mappings) technology assisted by an AI to map the Kobo input to the SOSO profile; the catalogue is extensible and can support other profiles as needed. SHACL shapes are used to

validate selected rules governing the relationships between properties at the Dataset and variable levels, and the rule set continues to expand. By representing every catalog entry as schema.org/SOSO-conformant JSON-LD rather than as static human-oriented documentation, the catalog is directly machine-actionable: entries can be queried and filtered by external AI/ML and agentic tooling without bespoke metadata parsing - extending FAIR compliance toward the AI-readiness increasingly expected of research data infrastructure; ingesting a source still relies on its source-specific script.

2.2 OMOP Extensions: LOCATION_HISTORY and EXTERNAL_EXPOSURE

Linking place-based geospatial data to person-level clinical data requires two OMOP-compatible extension tables. LOCATION_HISTORY records the sequence of locations associated with a person, including residences where available. Each record points to a geometry already held in the OMOP LOCATION table (via location_id) and adds explicit start and end dates. This replaces OMOP's single static person-location link, so residential moves and time-varying exposure windows can be represented. EXTERNAL_EXPOSURE is a measurement-like table recording a person's exposure to a given factor within a time window. Each row carries the identifier of its source variable, which links it to the catalog record and dataset, the spatial relationship used (for example, Within), the type of data source, and the generating LOCATION_HISTORY record. Because exposures from each source are represented in a unified structure, downstream tools only ever reference and query a single schema. EXTERNAL_EXPOSURE is the key intersection between geospatial data in GaiaDB and clinical data in the OMOP CDM. Field specifications are provided in the table definitions [16] and Supplementary Table S1.

Two companion R packages connect these tables to standard HADES tools [11,12]. FeatureExtractionForExtensions builds person-level covariates from EXTERNAL_EXPOSURE, such as day-weighted exposure over a look-back window that accounts for residence intervals. CaprForExtensions adds exposure- and location-based criteria to Capr cohort definitions. The resulting covariates and cohorts can be passed to packages such as CohortMethod, as in the tutorial's analysis exercise [16].

2.3 GIS Vocabulary Package

The GIS Vocabulary Package extends the OHDSI Standardized Vocabularies [6] to make geospatial exposure records interoperable across sites and over time. It has three coordinated layers: a core layer for geographic and spatial concepts (administrative boundaries, spatial units, coordinate reference systems, spatial relationships); an exposome layer for environmental pollutants and related variables [20], drawing on the Toxin and Toxin Target Database [21]; and an SDOH layer for social vulnerability, deprivation, environmental justice, and neighbourhood context, drawn from the Social Vulnerability Index, Area Deprivation Index, Environmental Justice Index, and AHRQ SDOH resources [22-25]. These sources were chosen for their relevance to health equity research.

Each candidate term was assigned a semantic role (construct, determinant, item, or item value), so that source-specific variables are not promoted to broad standard concepts and

categorical values are not treated as exposure variables. Mapping to existing vocabularies followed SSSOM [26] and was deliberately conservative: terms without sufficient semantic support were left unmapped, and geographic concepts were not mapped to clinical conditions or measurements. Quality control covered structural, semantic, relationship, and implementation checks. Role definitions, mapping outcomes by vocabulary layer, and the quality-control checks are in Supplementary Table S2.

2.4 Local geocoding benchmark

Geocoding is the process of assigning spatial coordinates to location information such as address strings or place names. We benchmarked four geocoders - DeGAUSS and PostGIS TIGER Geocoder, both available as optional components of the Gaia Docker deployment; the OpenStreetMap Nominatim geocoder, which has been used as an alternative geocoder in prior deployments but is not part of the default containerized stack; and ArcGIS Pro as a commercial comparator that is not part of the toolchain - using 5,000 validated residential addresses in Miami-Dade County, Florida, with known true locations (the county property-parcel point for each address). Evaluation metrics included hit rate, the distribution of error distances by distance band, and the median, mean, and maximum error. DeGAUSS, PostGIS TIGER, and Nominatim can be run locally, so that no address data leave the institutional compute environment. The ArcGIS Pro geocoder was run with an enterprise license and used the ESRI Streetmap Premium database; ESRI claims this process is HIPAA compliant, yet addresses are sent to the ESRI servers with the University of Miami configuration. ArcGIS Pro is therefore used only as a comparator and is not part of the privacy-preserving stack. Local geocoding supports privacy-preserving deployment, although compliance depends on institutional governance, configuration and data-flow controls. The benchmark protocol recorded geocoder versions, reference data snapshot dates, address sampling frame, preprocessing rules, and hardware environment to ensure reproducibility.

2.5 Spatial-Temporal Exposure Generation Pipeline

In the implemented workflow, EXTERNAL_EXPOSURE generation proceeds through six auditable stages: (1) data source identification and cataloging via Gaia Catalog; (2) ingestion and format conversion (shapefiles, GeoJSON, GeoPackage, CSV) into GaiaDB; (3) geocoding of patient addresses to associated LOCATION_HISTORY records with PostGIS geometry objects and temporal bounds; (4) spatial-temporal joins between patient location histories and geospatial features via source-specific geom and attr tables (for example, a point-in-polygon predicate with time-varying values clipped to each residence interval), with the spatial relationship concept and operator recorded on each resulting row; (5) population of EXTERNAL_EXPOSURE records; and (6) loading of exposure records and GIS vocabulary delta tables into the institutional OMOP CDM for use by standard OHDSI analytic tools. In practice, the relative order of source ingestion (Stage 2) and patient geocoding (Stage 3) may vary by deployment; the stages above are presented in a canonical rather than strictly sequential order. GaiaCore provides R functions to run each stage and to validate the derived records (Section 2.1.3).

2.6 Synthetic Evaluation

Real patient addresses cannot be shared, and they are exactly the data a privacy-preserving, federated toolchain is designed to keep local. We therefore evaluated the toolchain on synthetic OMOP CDM v5.4 cohorts, which serve as a controlled substrate in which the true exposure, the true exposure–outcome effects, and the true residence of every person are known. This allows two questions to be tested directly: whether the pipeline reproduces the exposures and recovers the effects that were simulated, and how geocoding error propagates to exposure and effect estimates. No real individuals, addresses, or coordinates are represented; residences are random points inside real-world census units, and the cohorts are not intended to represent any real population or to yield epidemiological estimates. The synthetic data test the analytic pipeline; they do not themselves demonstrate privacy preservation, which rests on local execution (Section 2.4).

Two cohorts were generated with an openly available generator [16](archived as release benchmark-v1.0.1, Zenodo DOI 10.5281/zenodo.23296292; tutorial build at generator commit c62b5a0, benchmark build at commit 15efabc). The tutorial cohort comprises 10,000 adults in 2,672 US counties. A benchmark cohort of 40,000 adults (44,003 residence locations; about 10% of persons move once) in 10,223 census tracts of four states (Alabama, Arizona, California, Oregon) provides tract resolution. Both cohorts cover 2014–2019 (Supplementary Table S3). Exposure is monthly PM2.5 from the CDC Environmental Public Health Tracking Network (EPA Downscaler estimates) at county resolution and, for the benchmark, at tract resolution, together with a simulated socio-economic status (SES) index, defined by county or by tract, that is correlated with PM2.5 and therefore confounds its association with outcomes. County and tract boundaries, PM2.5 at each resolution, and SES were each cataloged as a GaiaCatalog entry and passed through the geometry-attribute index (Table 2); the GaiaDB spatial-temporal join then assigned each residence interval the value of the unit containing the residence, clipped to the interval. We present an explicit example of an individual in the county-level dataset in **Figure 2**; additional details on the synthetic data sources and their construction are in the Supplementary Methods.

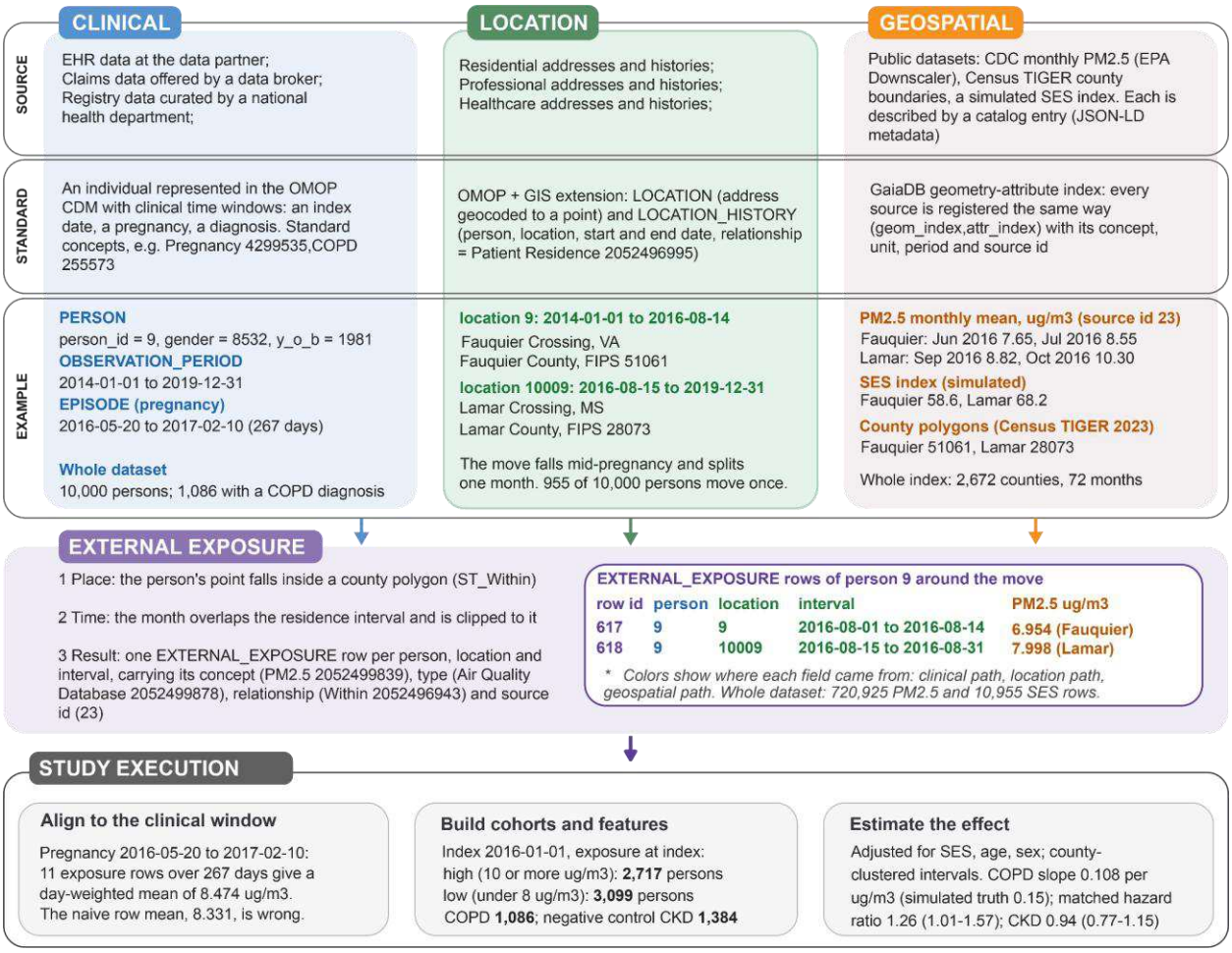


Figure 2. Worked example from the tutorial cohort, following one person from source data to an effect estimate. *Clinical (blue), location (green) and geospatial (orange) sources are each standardized: clinical data as OMOP CDM records, residences as LOCATION and LOCATION_HISTORY rows, and public datasets through the geometry-attribute index. Person 9 moves from Fauquier County, VA, to Lamar County, MS, midway through a pregnancy. A spatial join (ST_Within) and temporal clipping to the residence interval produce one EXTERNAL_EXPOSURE row per person-location- interval combination; the move splits August 2016 into two rows (colors in the table show each field's source). Study execution then aligns exposures to the clinical window (a day-weighted mean over 11 exposure rows, 8.474 μg/m³, rather than the unweighted row mean, 8.331), builds cohorts and features, and estimates effects adjusted for SES, age and sex with county-clustered intervals. All data are synthetic. SES is simulated, and the COPD effect is simulated (true slope 0.15) with CKD as a negative control. Estimates are shown to illustrate the workflow, not as findings.*

Fourteen health conditions (five respiratory, nine cardiometabolic) were drawn from the derived exposures with a logistic model in day-weighted residential PM2.5, SES, age, sex, and a shared county random effect. Eight conditions carry a simulated PM2.5 effect (0.05–0.15 log-odds per μg/m^3) and six are simulated nulls that serve as negative controls [27]. The simulated effects are deliberately larger than epidemiologic estimates, because PM2.5 varies little across persons; they are benchmark values, not effect sizes.

We assessed (i) agreement of the derived exposure rows with an independent answer key computed without any spatial operation, and with an equivalent hand-written PostGIS query; (ii) recovery of the simulated effects in the benchmark cohort by logistic regression of each outcome on continuous day-weighted PM2.5, adjusted for SES, age, and sex, with county-clustered confidence intervals (leave-one-county-out jackknife; Supplementary Table S7) and a CohortMethod comparison of high versus low exposure as a qualitative check [11]; (iii) sensitivity to geocoding error, by displacing every residence in 200 Monte Carlo replicates by a distance drawn from the empirical error distribution of each of four geocoders (Nominatim, ArcGIS Pro, DeGAUSS, PostGIS TIGER; 5,000 Miami-Dade County addresses [28]), reassigning it to the containing tract and county, and recomputing exposures and all 14 outcome models, with two stress tests of sharply varying exposures (synthetic random fields with correlation lengths of 100 m, 1 km and 10 km, and a near-road increment around real road geometry); and (iv) feasibility of using the derived exposure in a machine-learning workflow (Supplementary Methods). The objective was to confirm that a known signal, including its confounding structure, survives the trip from heterogeneous sources through the index into OMOP-adjacent exposure records and out through standard tools. It was not to validate clinical prediction or estimate causal effects.

3. Results

3.1 Toolchain Implementation

The toolchain and its documentation are publicly available [16] (Apache 2.0). The reference deployment uses Docker Compose to package the core components as containers, with optional containers for geocoding, analysis, and experimental git-based ingestion. It has been deployed on Linux and macOS, and a default Windows 11 deployment ran the tutorial pipeline (Exercise 2) end to end, matching the answer key. It is the simplest path to a full working stack and requires Docker, Docker Compose, and 50 GB of free storage, with substantial extra space for geocoding reference data depending on geocoder. Docker is a convenience, not a requirement. GaiaDB is a standard PostgreSQL/PostGIS database that runs on any native Postgres server, a lightweight GaiaCatalog is served as a GitHub Pages site, and GaiaCore, an R package, installs independently and can be pointed at any GaiaDB instance. Institutions can therefore adopt the components that fit their infrastructure and security requirements without the full container set.

Source datasets in shapefile, GeoJSON, ESRI geodatabase, Open Street Map XML, GeoTIFF, netCDF, HDF5, GeoPackage, and CSV formats can be loaded by GaiaDB and GaiaCatalog into staging tables; full variable-level ingestion into the geometry-attribute index is currently implemented for vector sources. Every source loaded in this work was registered in the geometry and attribute indexes and joins were run against the per-source geometry and attribute tables (Section 3.4). Run natively (Intel Core i7-1255U, 32 GB, Ubuntu, default PostgreSQL settings; median of three runs), the Gaia tract join (2,883,882 rows) took 28 s against 16–17 s for the equivalent hand-written PostGIS query, about 1.7 times as long, with identical rows (tolerance 1e-9; Supplementary Table S5). The Gaia join is therefore slower than plain SQL; its value is catalog-driven provenance, standardization, and reuse rather than speed. GaiaDB's SQL function library covers schema initialization, JSON-LD metadata ingestion, location ingestion, spatial-temporal joins, data-source retrieval, and

catalog integration [16]. GaiaCore provides R functions to run the pipeline, validate its outputs, and analyse exposure effects (Section 2.1.3).

3.2 GIS Vocabulary Package

The GIS Vocabulary Package comprises 9,923 unique concepts across three vocabulary layers (release GIS_2025-04-18, commit f7d78f9):

- OMOP GIS Vocabulary: 159 concepts covering geographic entities, spatial operations, and coordinate systems.
- OMOP Exposome Vocabulary: 6,872 concepts from T3DB covering environmental chemicals, pollutants, and toxicants.
- OMOP SDOH Vocabulary: 2,892 concepts spanning social determinants from SVI, ADI, EJI, AHRQ SDOH, COI, SDGI, SDOHO, and SEDH.

Concepts span nine OMOP domain assignments, including seven newly introduced semantic domains: behavioral, demographic, environmental, geographic, healthcare, phenotypic, and socioeconomic features. Domain distributions vary across vocabulary layers: OMOP GIS concepts are predominantly assigned to Observation (95.6%) and Geographic Feature (3.8%), while all OMOP Exposome concepts are assigned to Observation. OMOP SDOH exhibits greater domain diversity, with Observation (78.0%), Socioeconomic Feature (10.1%), and Behavioral Feature (3.4%) accounting for the largest proportions. The seven newly introduced feature domains collectively represent 20.7% of OMOP SDOH concepts. Structural quality-control checks identified no duplicate concept identifiers, duplicate vocabulary-code combinations, unrecognized domain identifiers, invalid date intervals, or duplicate relationship records.

3.3 Geocoding Performance

The four geocoders benchmarked in Section 2.4 returned a location for 97.4–100% of the 5,000 addresses (Table 1; Supplementary Table S13). Median error was smallest for Nominatim (18 ft, 5.6 m) and ArcGIS Pro (85 ft, 26 m), and about 155–165 ft (47–50 m) for PostGIS TIGER and DeGAUSS, which also had the heaviest tails: 213 and 191 addresses, respectively, fell more than a mile from the reference point, and PostGIS TIGER produced a few extreme outliers (maximum about 510 miles). Nominatim also provides global coverage, which matters for international multi-site studies, whereas DeGAUSS and PostGIS TIGER use U.S. reference data. Address preprocessing (standardizing directional abbreviations, street suffixes, unit designators) improved match rates by 8–12% across services.

Table 1. Geocoding performance on 5,000 Miami-Dade County addresses (error = distance from the reference point)

Geocoder	Geocoded (hit rate)	0–100 ft	100–500 ft	500 ft–1 mi	> 1 mi	Median ft (m)	Mean ft	Max ft
PostGIS	4,999 (99.98%)	1,675	2,400	711	213	155 (47)	3,934	2,700,50

Geocoder	Geocoded (hit rate)	0–100 ft	100–500 ft	500 ft–1 mi	> 1 mi	Median ft (m)	Mean ft	Max ft
TIGER								4
DeGAUSS	4,960 (99.20%)	1,420	2,583	766	191	165 (50)	818	57,698
ArcGIS Pro	4,868 (97.36%)	2,985	1,789	87	7	85 (26)	284	212,687
Nominatim	4,918 (98.36%)	3,718	899	230	71	18 (5.6)	303	57,096

3.4 Synthetic Evaluation Results

One index, several sources. Table 2 shows how the sources of the synthetic evaluation, which differ in geometry, resolution, time coverage, origin, and variable structure, were represented through the geometry and attribute indexes. The CDC PM2.5 dataset contributed four attribute variables at each resolution, of which only the monthly mean was joined, one variable at a time. Because each source is addressed through the same index, the same join was used for every exposure, and the resulting rows have the same EXTERNAL_EXPOSURE structure.

Table 2. Sources standardized through the geometry-attribute index in the synthetic evaluation

Source	Role	Origin	Native structure	Time coverage	Registered in
US county boundaries (Census TIGER/Line 2023)	Geometry	Real	Polygons, about 3,100 counties	Static	geom_index
US census tract boundaries (TIGER/Line 2019; 2010 vintage)	Geometry	Real	Polygons, 11,523 eligible tracts in four states	Static	geom_index
Monthly PM2.5 at county and tract resolution (CDC EPHTN; EPA Downscaler)	Attribute (four variables; monthly mean joined)	Real	Monthly time series per unit	2014–2019, monthly	attr_index
Simulated SES index (by county; by tract in the benchmark)	Attribute (confounder)	Simulated	One value per unit	Static	attr_index
US primary and secondary roads (TIGER/Line 2019)	Geometry	Real	Lines, 18,408 segments in four states	Static	geom_index
Simulated near-road increment (distance bands of 0–50, 50–100, 100–200 and 200–400 m around those roads)	Geometry and Attribute	Simulated, on Real Road Geometry	Polygons, 140,832 pieces	Static	geom_index and attr_index
Simulated exposure surfaces (correlation length 100 m, 1 km and 10 km; 10 km window)	Geometry and Attribute	Simulated	Polygons, 160,801, 1,681 and 25 cells	Static	geom_index and attr_index

Exposure derivation. For the benchmark cohort, GaiaDB derived 2,883,882 tract-level and 2,883,882 county-level PM2.5 rows and 44,003 SES rows; for the tutorial cohort, 720,925 PM2.5 and 10,955 SES rows. The tract-level PM2.5 and SES rows matched the independent answer key (no missing or extra rows; maximum absolute difference below 10^-13), and a hand-written PostGIS query returned the same rows (Supplementary Tables S4 and S5). The county-level rows matched for 39,997 of 40,000 persons; the other three live next to a county boundary, where the tract (2019) and county (2023) boundary vintages disagree. The county key is derived from each person's tract, so it is less independent than the tract key.

Recovery of simulated effects. In the benchmark cohort, adjusted estimates from the continuous-exposure regression recovered the simulated effects (Figure 3; Supplementary Table S6). All 14 confidence intervals covered the true value, all 8 simulated effects were detected with tract exposure (with county exposure, the stroke interval had a lower bound of 0.000 after rounding, so detection was borderline), and none of the 6 nulls was; for COPD the tract-level estimate was 0.149 log-odds per μg/m^3 (95% CI 0.095–0.204) against a truth of 0.15. A crude model without SES adjustment was confounded (2 of 6 nulls flagged), as designed. The generator draws exposure and SES at tract level, so county exposure is misclassified by design and gave larger errors (Figure 3; Supplementary Table S11); this shows that the index handles two resolutions through the same mechanism, not that tract data are better in general. A CohortMethod-based comparison of high versus low exposure performed less well: its adjusted intervals covered the truth for 6 of 8 effects with tract exposure and 4 of 8 with county exposure, with one and two false-positive nulls, respectively, and its estimates were attenuated to roughly three quarters of the true effect because the exposure was dichotomized. It is therefore reported only as a qualitative check (Supplementary Table S8).

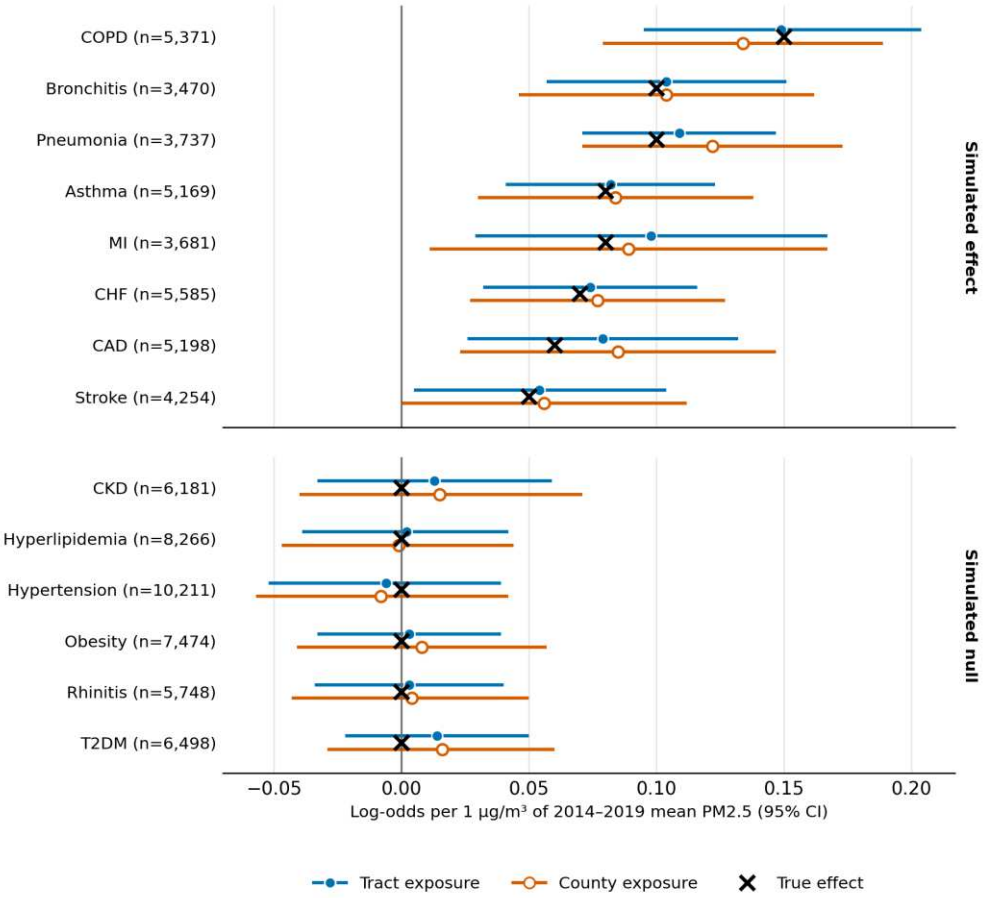


Figure 3. Recovery of simulated PM2.5 effects in the benchmark cohort. *Estimates and 95% confidence intervals (leave-one-county-out jackknife) from logistic regression of each outcome on day-weighted 2014–2019 mean PM2.5, adjusted for SES, age, and sex, using tract (blue) or county (orange) exposure; crosses mark the simulated true effects. Case counts are shown beside each outcome. The six lower outcomes are simulated nulls. Simulated effects are deliberately larger than published epidemiologic estimates.*

Sensitivity to geocoding error. Realistic geocoding error moved up to about 10% of locations into a different tract and about 1% or fewer into a different county, but person-level exposure changed little, because neighbouring tracts have similar monthly PM2.5 and a person's exposure averages many months (Table 3; Supplementary Table S9 and Figure S1). Mean shifts in the effect estimates were 0.0005–0.002 log-odds per $\mu g/m^3$ and the largest was 0.0045 (0.0036 with tract exposure), smaller than the standard errors (about 0.012 with the mixed model and 0.027 with the jackknife; Supplementary Table S6), and no effect was lost or created. The two geocoders with heavy-tailed errors (DeGAUSS and PostGIS TIGER) displaced more than Nominatim and ArcGIS Pro, but the ranking of geocoders was the only difference.

Estimates drifted slightly upward, including for null outcomes, whereas exposure error usually attenuates estimates. Displacing exposure and SES separately showed that the drift comes from SES, which is reassigned from the displaced tract along with exposure. Displacing only the exposure left the mean estimate unchanged (0.0000 to −0.0002), whereas displacing only SES reproduced the shift (+0.0005 to +0.0020; spread across replicates 0.0003–0.0043, against model standard errors of about 0.012–0.027; Supplementary Table S10). A noisier SES covariate leaves some confounding uncontrolled, which pushes all estimates up.

This near-null effect is expected for a smooth regional exposure averaged over six years, and it rests on error distributions from one urban county (Miami-Dade, Section 4.4). We therefore also tested sharply varying exposures (below).

Table 3. Sensitivity of benchmark exposure and effect estimates to simulated geocoding error (200 replicates)

Geocoder	Median error (m)	Assigned to another tract	Assigned to another county	Tract exposure RMSE, $\mu g/m^3$ (correlation)	Mean / largest shift of the 14 estimates (tract)
Nominatim	5.6	3.9%	0.17%	0.026 (0.9997)	0.0007 / 0.0014
ArcGIS Pro	26	3.2%	0.11%	0.030 (0.9997)	0.0005 / 0.0012
DeGAUSS	50	10.4%	0.39%	0.054 (0.9993)	0.0018 / 0.0036
PostGIS TIGER	47	9.9%	1.06%	0.067 (0.9970)	0.0016 / 0.0033

When geocoding error matters. In two stress tests with the same individuals, outcome model, and error distributions, geocoding error mattered once it approached the length scale of the exposure. For synthetic exposure surfaces with an exponential correlation length of 100 m, 1 km, or 10 km (200 replicates each), the share of the true effect recovered fell as predicted by regression dilution (Table 4 and Supplementary Figure S2; simulated and predicted values agreed within 0.003 in every cell). At 10 km, which resembles PM2.5, every geocoder recovered at least 97% of the effect. At 1 km the two Census-based geocoders lost about 13% of the effect and their intervals covered 72–74% of the true effects; at 100 m the best geocoder lost 19% and the worst about half, with coverage of 7–56%. Null outcomes were unaffected, because a null has no effect to dilute. A near-road increment added to the real tract PM2.5 behaved similarly: the near-road coefficient lost 4–6% with Nominatim and ArcGIS Pro and 14% with DeGAUSS and PostGIS TIGER, while the

tract PM2.5 coefficient was unaffected (Supplementary Tables S14 and S15). The road increments are illustrative, and synthetic residences lie farther from roads than real addresses, so this effect is likely understated. In both tests, joins of the surfaces and road bands through Gaia reproduced the analytic values (Supplementary Methods).

Table 4. Share of the true effect recovered (coverage of the 8 true effects) after geocoding displacement, by exposure correlation length

Correlation length	Nominatim	ArcGIS Pro	DeGAUSS	PostGIS TIGER
100 m	0.82 (56%)	0.70 (27%)	0.52 (7%)	0.54 (7%)
1 km	0.95 (91%)	0.96 (92%)	0.87 (72%)	0.87 (74%)
10 km	0.99 (93%)	0.99 (93%)	0.98 (93%)	0.97 (93%)

Coverage at the true locations was 94%, 93%, and 93% for 100 m, 1 km, and 10 km.

Downstream use. The derived exposure and SES also served directly as features in a standard machine-learning workflow, which demonstrates feasibility only (Supplementary Table S12).

3.5 Metadata Catalog

The Gaia Catalog indexes 113 curated geospatial datasets spanning environmental quality (EPA), demographics (U.S. Census), social vulnerability (CDC), and international sources, with ~2,847 variables mappable to OMOP GIS vocabulary concepts. JSON-LD metadata export enabled successful federation with the University of Miami's Geospatial Digital Special Collections, demonstrating interoperability with external institutional repositories.

4. Discussion

4.1 Principal Findings

We developed and evaluated an open-source infrastructure for generating, representing, and using geospatial exposure data in federated OMOP-compatible observational research. It addresses four structural barriers to geospatial integration: standardization, through vocabulary extensions and common data representations; privacy, through local containerized geocoding that can run behind institutional firewalls; accessibility, through pre-built containers and documented workflows; and collaborative efficiency, through a FAIR-compliant, machine-readable metadata catalog that supports discovery and reuse across sites.

Of these, the standardized representation of heterogeneous geospatial data is the principal contribution. The geometry-attribute index registers each geometry and attribute source in a uniform structure with explicit variables, units, and time windows. Because of this representation, and not any single source-specific pipeline, the toolchain produces EXTERNAL_EXPOSURE records of identical form from county polygons, a monthly pollutant series, and socio-economic indicators. The other components (vocabulary, catalog, local geocoding, and extension tables) are built around this representation and gain their reuse value from it.

The synthetic evaluation shows that the toolchain can take cataloged public datasets through ingestion, spatial-temporal linkage, and population of EXTERNAL_EXPOSURE, reproducing an independent answer key (exactly for the tract and SES rows, and for 39,997 of 40,000 county rows, the rest reflecting boundary-vintage differences). It also shows that the resulting OMOP-resident records support analyses that recover known exposure–outcome effects, their confounding structure, and negative-control behavior, and it shows, with stress tests, that geocoding error is negligible for a smooth regional exposure but matters when it is comparable to the length scale of the exposure (Section 3.4). With regard to limitations (see also Section 4.4): (1) No analysis was run across sites, so federated execution is a property of the OHDSI network and of the design, not a result shown here. (2) The synthetic data test the analytic pipeline and do not themselves demonstrate privacy preservation. (3) Patient locations were coordinates rather than geocoded addresses, so geocoding was evaluated through simulated displacement rather than end to end. (4) And the index was exercised on a small set of source types: county and tract polygons, two area-level attribute sources (one real, one simulated), and, in the stress tests, simulated gridded exposure surfaces and road-distance bands, whose Gaia joins reproduced analytic values. These stress-test sources were simulated and represented as vector polygons, so the index has not yet been exercised on native raster files, real high-resolution surfaces, or point-monitor networks; uniform handling across such formats remains a design property rather than a demonstrated result.

4.2 GeoAI Infrastructure for Health Equity Research

The toolchain's public health value is most clearly expressed through its SDOH and environmental justice capabilities. The 2,892-concept SDOH vocabulary draws explicitly on sources - SVI, ADI, EJI - designed to characterize the social and environmental conditions facing vulnerable and underserved communities. By integrating these into the OMOP exposure layer, the toolchain is designed to support multi-site observational studies examining how structural determinants of health distribute disease risk across geography and population groups - research directly relevant to SDG3 (Good Health and Well-Being) and SDG10 (Reduced Inequalities).

The federated architecture is particularly important for health equity research. Communities most affected by environmental injustice are often served by smaller health systems and safety-net hospitals whose data are rarely included in large observational studies. By enabling OMOP-standardized institutions of any size to participate in geospatial observational research without sharing patient-level data or building bespoke GIS infrastructure, the toolchain is intended to lower the barrier to inclusion of these populations in environmental epidemiology; this is a design goal that the present work does not test.

This accessibility is as much a data-representation achievement as an architectural one. Three properties of the toolchain jointly make heterogeneous, site-specific geospatial and SDOH data usable by AI/ML tooling at scale: (1) the geometry-attribute index (geom_index/attr_index) reduces every geometry and attribute source, whatever its native format, to the same registered structure, so that exposure joins are driven by index entries rather than by source-specific code (Sections 2.1, 3.4); (2) Gaia Catalog exposes every

dataset and variable as schema.org/SOSO-conformant JSON-LD, so datasets from any source become machine-queryable without custom metadata parsing (Sections 2.1, 3.5); and (3) EXTERNAL_EXPOSURE and LOCATION_HISTORY represent the resulting exposure assignments in one OMOP-compatible relational structure (Section 2.2). Together, these mean an AI/ML pipeline built against the toolchain's output needs to understand one schema and one metadata standard rather than N site- and source-specific ones - a structural precondition for applying geoAI methods across a federated network at scale.

4.3 Comparison with Existing Approaches

The DeGAUSS project pioneered containerized, privacy-preserving geocoding and geomarker assessment [29]. The OHDSI GIS toolchain incorporates DeGAUSS as one geocoding option while extending scope to vocabulary standardization, diverse data source integration, and compatibility with broader OHDSI analytics. The National Neighborhood Data Archive (NaNDA) provides curated U.S. neighbourhood contextual data [30] but focuses on specific harmonized measures rather than an extensible OMOP-compatible framework for generating person-place-time exposure records from arbitrary geospatial sources. Commercial and academic EHR research platforms (i2b2, PCORnet, TriNetX) include geographic capabilities but typically lack vocabulary-governed OMOP extension tables for external exposure generation, spatial provenance tracking, and compatibility with OHDSI federated analytics [31,32]. More broadly, spatial and spatiotemporal methods in epidemiology [33] and recent reviews of geospatial exposure models [34] converge on the same conclusion this toolchain is built around: methodological sophistication in spatial epidemiology has outpaced the standardized data infrastructure needed to deploy it reproducibly across sites. A variety of spatial and spatial-temporal analytical methods provide options that leverage FAIR and OHDSI data integration and open-source computations with R programming packages. Among these, Amadeus [35], an R package from the U.S. National Institute of Environmental Health Sciences, provides functions to download, process, and calculate covariates from publicly available environmental datasets, mainly for the United States. It addresses data acquisition and covariate calculation for researchers working in R, whereas the OHDSI GIS toolchain adds an OMOP-compatible person-place-time representation, vocabulary governance, and federated execution; the two are complementary. The distinctive contribution here is the geometry-attribute index as a uniform representation of diverse geospatial sources, together with geospatial processing, vocabulary-governed exposure representation, metadata-driven discovery, privacy-preserving local deployment, and OHDSI analytic workflow compatibility within a single open-source infrastructure.

4.4 Limitations

Geocoding accuracy and bias vary by address type and population. Rural addresses geocode less accurately than urban, and addresses for marginalized populations - people experiencing homelessness, non-traditional housing - may be incomplete or absent from reference datasets. These errors are socially patterned and may amplify, rather than randomize, exposure misclassification bias in precisely the populations most relevant to health equity research. Spatial resolution mismatches between point-level patient locations and area-level geospatial datasets (census tracts, counties, gridded surfaces) introduce

assignment uncertainty that the toolchain does not yet propagate into downstream statistical models, although the sensitivity analysis in Section 3.4 quantifies it for one exposure. Temporal latency in authoritative datasets - census data updated decennially, many environmental datasets annually - limits capture of rapid environmental changes relevant to acute outcomes. Ecological fallacy remains a risk when area-level characteristics are used to draw individual-level inferences; the vocabulary constraints and EXTERNAL_EXPOSURE design make the area-level origin of exposure variables explicit, but cannot substitute for appropriate hierarchical analytic methods. Vocabulary governance is an additional limitation: the distinction between official OHDSI Standardized Vocabulary content and the GIS Vocabulary Package extension artifacts must be maintained explicitly across releases (Standardized Vocabularies [6]); without stable deprecation rules and concept identifier policies, downstream users may treat provisional concepts as settled standards. Source ingestion (Section 2.1.4) currently runs as source-specific scripts; a declarative recipe interpreted in-database is planned (Section 4.6). Additionally heterogeneous API configurations and authentication methods require bespoke case by case methods to automate and harmonize data source retrieval across government agencies, third party and open data providers.

The synthetic evaluation has further limits. It tests the analytic pipeline and the sensitivity of estimates to simulated geocoding error; it does not itself demonstrate privacy preservation. Persons are placed uniformly at random within county or tract polygons, so everyone in a unit shares one exposure value; SES and SDOH values are simulated; effects are exaggerated, and null outcomes are null by construction, whereas real-world associations may differ. Tract-level PM2.5 combines two CDC releases of the same product that do not overlap in time; aggregated to counties, the tract series agreed with a single-release county series to within 0.003 $\mu g/m^3$ before and after the 2015/2016 seam, with no step at the seam (Supplementary Methods), although a shift that cancels within counties would not be detected. The geocoding sensitivity analysis applies error distributions from one urban county to every person, so rural error is probably understated, and the main analysis is specific to a smooth, regionally correlated exposure averaged over six years; the stress tests show large attenuation for exposures that vary over 100 m to 1 km, but they use simulated surfaces and illustrative road increments, residences placed randomly inside tracts are farther from roads than real addresses, and short exposure windows were not evaluated. Derived exposure rows currently leave unit_concept_id empty (the free-text unit is recorded in source attribute metadata), which should be closed for full OMOP conformance.

4.5 Regulatory Considerations for International Deployment

Although this work is grounded in a US-centric regulatory frame (HIPAA, Section 1.2), through OHDSI's global consortium of data custodians and regulators across 6 continents, international geocoding can rapidly enable deployment in other jurisdictions. Residential address and location-history data are personal data under the EU General Data Protection Regulation (GDPR), and fine-grained geographic coordinates can function as quasi-identifiers even absent direct identifiers; local geocoding (Section 2.4) is necessary but not sufficient for compliance, and institutions deploying the toolchain in the EU should complete the appropriate data-protection assessments. The European Health Data Space

(EHDS) Regulation, in force since March 2025, includes environmental and socio-economic determinants of health among its minimum categories for secondary use; its secondary-use provisions apply from 2029, and these categories from March 2031 [7]. It is therefore a relevant future compliance context. OHDSI Europe initiatives, such as EHDEN and DARWIN EU, offer an established federated deployment model in which these obligations could eventually be operationalized alongside the toolchain's privacy-by-design architecture. A detailed regulatory mapping is outside the scope of this paper.

4.6 Future Directions

Priority near-term extensions include: (1) real-world, multi-site AI-readiness characterization - applying the validated pipeline to existing OMOP-standardized data from partner consortia to quantify geocoding match rates, LOCATION_HISTORY/EXTERNAL_EXPOSURE linkage completeness, and catalog dataset coverage under real institutional conditions, extending the synthetic validation reported here; (2) extending pipeline-derived EXTERNAL_EXPOSURE beyond the county- and tract-level sources used here, by registering additional public sources, finer spatial resolutions, and heterogeneous formats (rasters, point monitors) in the geometry-attribute index, with corresponding variants of the synthetic cohort, and an onboarding comparison (effort to add a third source by catalog entry alone versus new source-specific SQL); (3) complementing source-specific script execution (Sections 2.1, 4.4) with a more dynamically parametrized, declarative script-generation layer for sources where that pattern fits; (4) extending the geocoding sensitivity analysis to rural error distributions, short-window exposures, and real address-level road geometry; (5) expanded raster data support for continuous environmental surfaces; (6) dynamic exposure modeling for cumulative and lagged exposure windows; (7) community-driven vocabulary contribution workflows; and (8) extending GaiaCore with spatial analyses built on established R packages, for example spatial autocorrelation (Global and Local Moran's I with spdep [36,37]), space-time cluster detection (Kulldorff's scan statistic with scanstatistic [38]), Bayesian spatial and spatiotemporal models (INLA [39,40]), and Bayesian model averaging (BAS [41]). An extended list of potential spatial analyses is provided in Supplementary Table S16.

The OMOP CDM's EPISODE table - designed to represent extended clinical courses such as multidrug regimens - may offer a natural structure for representing time-series and multi-pollutant exposures (e.g., weekly PM2.5 series, combined humidity–PM2.5 exposure episodes in COPD patients, or exceedance events) as first-class, concept-anchored objects rather than post-hoc aggregations layered onto EXTERNAL_EXPOSURE via ETL. Formal evaluation of EPISODE-based exposure representation - including whether it should complement or partially replace date-range aggregation in EXTERNAL_EXPOSURE - is a priority design question for the working group.

Looking further ahead, a longer-term goal for the Gaia toolchain is to combine the structural extensions described here with native raster support, uncertainty-propagating spatial joins, extended GIS-aware cohort authoring, and an AI-agentic catalog interface in which natural-language queries are resolved against JSON-LD-described datasets and mapped to OMOP GIS Vocabulary concepts - moving the current multi-step, expert-mediated workflow (Section 2.5) toward a more directly geoAI-usable discovery-to-

exposure-record pipeline [13]. Collaboration with communities outside OHDSI, such as the Amadeus developers, on tool-agnostic data schemas would reduce duplicated effort and promote interoperability. Broader adoption across the OHDSI community will depend on stable versioned releases, training materials, and real-world validation studies examining whether geospatial exposure variables improve inference or prediction in specific epidemiological contexts.

5. Conclusions

The OHDSI GIS toolchain provides a standards-based infrastructure, evaluated under controlled synthetic conditions, for incorporating geospatial, environmental, and social-determinant context into federated OMOP-compatible observational health research. Rather than introducing a new geoAI model, its primary contribution is a geometry-attribute index that gives heterogeneous geospatial sources a standardized representation, and through it a reproducible, vocabulary-governed, machine-readable person-place-time exposure layer deployable across distributed research networks without sharing patient-level data; this addresses a structural gap that has prevented large-scale observational databases from fully engaging with the environmental and social determinants that shape the outcomes they record. By connecting clinical informatics, environmental epidemiology, and SDOH data standardization through a common open infrastructure, the toolchain lays the technical foundation for multi-site research on health inequities driven by place-based exposures; real-world, multi-site validation of this foundation is a near-term priority (Section 4.6). The design is intended to support health equity goals - enabling safety-net hospitals and diverse health systems to participate in environmental health research and providing a standardized route for incorporating social vulnerability and environmental justice data into observational studies - but this has not yet been tested. All code, containers, and vocabulary artifacts are available at https://github.com/OHDSI/GIS (Apache 2.0).

Acknowledgments

We dedicate this work to the memory of Andrew E. Williams, whose vision and intellectual generosity helped initiate the OHDSI GIS toolchain and inspired this community effort. We thank the OHDSI GIS Working Group members for their contributions to design, implementation, and testing, and the Tufts Clinical and Translational Science Institute for supporting development of the GIS vocabularies.

Author Contributions

AEW provided the vision, direction, and conceptual framing of the work. All authors supported manuscript drafting, editing and reviewing efforts, as well as software testing and development of tutorial materials. KZV, RM, PT, TN and JH produced the software and documentation supporting the toolchain and its various components. PT, MT, and JH created and developed the GIS vocabulary. JH designed the synthetic data and executed

associated experiments. TN executed the geocoding experiment to produce the empirical distributions used to sample geocoding error.

Competing Interests

The authors declare no competing interests.

Funding

This work was supported by the National Institutes of Health through NIEHS and the NCATS Clinical and Translational Science Awards to Tufts University (UL1TR002544), the National Institutes of Health Common Fund's Bridge2AI Clinical Care AI by "Patient-Focused Collaborative Hospital Repository Uniting Standards (CHoRUS) for Equitable AI" project (OT2OD0327), and the University of Miami (UL1TR002736). The content is solely the responsibility of the authors and does not represent official views of the NIH.

Data and Code Availability

All source code, Docker images, and documentation are available at https://github.com/OHDSI/GIS (Apache 2.0). The synthetic cohort generators, tutorial exercises, benchmark analysis scripts, and the build scripts for the tutorial and tract benchmark cohorts, together with the code and input snapshot used in this manuscript, are archived on Zenodo (version 1.0.1, release tag benchmark-v1.0.1; DOI 10.5281/zenodo.23296292; Apache 2.0) and the working copy is at https://github.com/OHDSI/GIS/tree/main/syntheticDataGIS. The Gaia components (GaiaDB, GaiaCore, GaiaCatalog and the Docker deployment) are developed in separate repositories and are not part of the archive; the versions used are identified by commit in the benchmark build record (demo.build_info). The build is deterministic given the pinned gaia-db image digest and gaiaCatalog commit recorded there. The GIS Vocabulary Package's canonical source is hosted at https://github.com/TuftsCTSI/CVB/tree/main/GIS/Ontology, referenced from https://github.com/OHDSI/GIS/tree/main/vocabulary. County and tract boundaries are U.S. Census TIGER/Line files (2023 and 2019, respectively), and monthly county and tract PM2.5 (2014–2019) are CDC Environmental Public Health Tracking estimates (EPA Downscaler); all were ingested through Gaia Catalog entries in the repository. All cohorts are synthetic, and no patient-level data are reported in this manuscript.

References

1. Cohen AJ, et al. Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution. *Lancet.* 2017;389(10082):1907-1918.

2. Diez Roux AV, Mair C. Neighborhoods and health. *Ann N Y Acad Sci.* 2010;1186:125-145.

3. Maantay J. Mapping environmental injustices. *Environ Health Perspect.* 2002;110(Suppl 2):161-171.

4. Mora C, McKenzie T, Gaw IM, Dean JM, von Hammerstein H, Knudson TA, Setter RO, Smith CZ, Webster KM, Patz JA, Franklin EC. Over half of known human pathogenic diseases can be aggravated by climate change. *Nat Clim Chang.* 2022;12(9):869-875. https://doi.org/10.1038/s41558-022-01426-1

5. Hripcsak G, et al. Observational Health Data Sciences and Informatics (OHDSI). *Stud Health Technol Inform.* 2015;216:574-578.

6. Observational Health Data Sciences and Informatics. The Book of OHDSI. 2021. https://ohdsi.github.io/TheBookOfOhdsi/

7. Regulation (EU) 2025/327 of the European Parliament and of the Council of 11 February 2025 on the European Health Data Space and amending Directive 2011/24/EU and Regulation (EU) 2024/2847. Official Journal of the European Union, L series, 5 March 2025. http://data.europa.eu/eli/reg/2025/327/oj

8. Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a common data model for active safety surveillance research. *J Am Med Inform Assoc.* 2012;19(1):54-60.

9. European Health Data & Evidence Network (EHDEN), https://www.ehden.eu; Data Analysis and Real World Interrogation Network (DARWIN EU), https://www.darwin-eu.org

10. Schuemie MJ, Ryan PB, Pratt N, et al. Principles of Large-scale Evidence Generation and Evaluation across a Network of Databases (LEGEND). *J Am Med Inform Assoc.* 2020;27(8):1331-1337.

11. OHDSI. Health Analytics Data-to-Evidence Suite (HADES), including DatabaseConnector, FeatureExtraction, Capr and CohortMethod. https://ohdsi.github.io/Hades/

12. OHDSI GIS Working Group. FeatureExtractionForExtensions and CaprForExtensions. https://github.com/OHDSI/FeatureExtractionForExtensions; https://github.com/OHDSI/CaprForExtensions

13. Schmitt CP, Stingone JA, Rajasekar A, Cui Y, Du X, Duncan C, Heacock M, Hu H, Gonzalez JR, Juarez PD, Smirnov AI. A roadmap to advance exposomics through federation of data. *Exposome.* 2023;3(1):osad010. https://doi.org/10.1093/exposome/osad010

14. Janowicz K, Gao S, McKenzie G, Hu Y, Bhaduri B. GeoAI: spatially explicit artificial intelligence techniques for geographic knowledge discovery and beyond. *Int J Geogr Inf Sci.* 2020;34(4):625-636. https://doi.org/10.1080/13658816.2019.1684500

15. OHDSI. Broadsea. https://github.com/OHDSI/Broadsea

16. OHDSI GIS Working Group. OHDSI GIS toolchain: documentation, GIS vocabulary tools, and synthetic benchmark and tutorial datasets (version 1.0.1). Zenodo; 2026. https://doi.org/10.5281/zenodo.23296292

17. Peroni S, Soiland-Reyes S, Sefton P, et al. Packaging research artefacts with RO-Crate. *Data Science.* 2022;5(2):97-138. https://doi.org/10.3233/DS-210053

18. University of Miami Frost Institute for Data Science and Computing. Geospatial Digital Special Collections - Tools (version 0.1.0) [Computer software]. 2026. https://github.com/Geospatial-Digital-Special-Collections/tools/tree/main

19. GDAL/OGR contributors. GDAL/OGR Geospatial Data Abstraction Library [Computer software]. Open Source Geospatial Foundation. https://gdal.org

20. Wild CP. Complementing the genome with an "exposome." *Cancer Epidemiol Biomarkers Prev.* 2005;14(8):1847-1850.

21. Lim E, et al. T3DB: a comprehensively annotated database of common toxins and their targets. *Nucleic Acids Res.* 2010;38(Database issue):D781-D786.

22. Agency for Toxic Substances and Disease Registry. Social Vulnerability Index. 2024. https://www.atsdr.cdc.gov/place-health/php/svi/index.html

23. University of Wisconsin. Neighborhood Atlas: Area Deprivation Index. https://www.neighborhoodatlas.medicine.wisc.edu/

24. Agency for Toxic Substances and Disease Registry. Environmental Justice Index. 2024. https://www.atsdr.cdc.gov/place-health/php/eji/index.html

25. Agency for Healthcare Research and Quality. Social Determinants of Health Resources. 2023. https://www.ahrq.gov/sdoh/resources.html

26. Matentzoglu N, et al. A Simple Standard for Sharing Ontological Mappings (SSSOM). *Database (Oxford).* 2022;2022:baac035. https://doi.org/10.1093/database/baac035

27. Schuemie MJ, Ryan PB, DuMouchel W, Suchard MA, Madigan D. Interpreting observational studies: why empirical calibration is needed to correct p-values. *Stat Med.* 2014;33(2):209-218.

28. Norris T. geocode-benchmark-2023 (geocoder comparison on 5,000 Miami-Dade County addresses). GitHub. https://github.com/tibbben/geocode-benchmark-2023

29. Brokamp C, et al. Decentralized and reproducible geocoding and characterization of community and environmental exposures for multisite studies. *J Am Med Inform Assoc.* 2018;25(3):309-314.

30. Messer LC, et al. The development of a standardized neighborhood deprivation index. *J Urban Health.* 2006;83(6):1041-1062.

31. NORC at the University of Chicago. Expanding Social Determinants of Health Data across PCORnet. 2022.

32. Simoni AH, et al. Social determinants of health and clinical outcomes among patients with atrial fibrillation. *QJM.* 2024;117(5):353-360.

33. Nazia N, Butt ZA, Bédard M, Tang W, Sehar H, Law J. Methods Used in the Spatial and Spatiotemporal Analysis of COVID-19 Epidemiology: A Systematic

Review. *Int J Environ Res Public Health.* 2022;19(14):8267. https://doi.org/10.3390/ijerph19148267

34. Clark LP, Zilber D, Schmitt C, Fargo DC, Reif DM, Motsinger-Reif AA, Messier KP. A review of geospatial exposure models and approaches for health data integration. *J Expo Sci Environ Epidemiol.* 2025;35(2):131-148. https://doi.org/10.1038/s41370-024-00712-8

35. Manware M, Song I, Marques ES, Kassien MA, Clark LP, Messier KP. Amadeus: Accessing and analyzing large scale environmental data in R. *Environ Model Softw.* 2025;186:106352. https://doi.org/10.1016/j.envsoft.2025.106352

36. Gaudart J, et al. Factors associated with the spatial heterogeneity of the first wave of COVID-19 in France: a nationwide geo-epidemiological study. *Lancet Public Health.* 2021;6:e222-e231.

37. Bilal U, Tabb LP, Barber S, Diez Roux AV. Spatial inequities in COVID-19 testing, positivity, confirmed cases, and mortality in 3 U.S. cities: an ecological study. *Ann Intern Med.* 2021;174:936-944.

38. Tyrovolas S, et al. Estimating the COVID-19 spread through real-time population mobility patterns: surveillance in low- and middle-income countries. *J Med Internet Res.* 2021;23:e22999.

39. DiMaggio C, Klein M, Berry C, Frangos S. Black/African American communities are at highest risk of COVID-19: spatial modeling of New York City ZIP Code-level testing results. *Ann Epidemiol.* 2020;51:7-13.

40. Jalilian A, Mateu J. A hierarchical spatio-temporal model to analyze relative risk variations of COVID-19: a focus on Spain, Italy and Germany. *Stoch Environ Res Risk Assess.* 2021;35:797-812.

41. Olmo J, Sanso-Navarro M. Modeling the spread of COVID-19 in New York City. *Pap Reg Sci.* 2021;100:1209-1229.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- OHDSIGISToolchainSupplementaryInformation.docx