

Agentic AI-Assisted Harmonization of CHoRUS/PostGIS Exposure Variables for OHDSI GIS Observational Research

Polina Talapova^{1,2}, Jared Houghtaling², Maksym Trofymenko¹, Jay Greenfield³, Timothy Norris⁴, Robert Miller², Jimmy Phuong⁵, Anne Thessen⁶, Timothy D. Rossi⁷

¹SciForce, Ukraine/Estonia

²Tufts Clinical and Translational Science Institute, USA

³CODATA, France

⁴The University of Miami, USA

⁵University of Washington, USA

⁶University of North Carolina Chapel Hill, USA

⁷Architectural Medicine LLC, USA

Background

The OHDSI GIS Workgroup has defined a reproducible path for bringing external geospatial, environmental, and social-context data into OMOP-based observational research. Prior work on the GIS Vocabulary Package¹, the GAIA toolchain², and the OMOP GIS extension, including EXTERNAL_EXPOSURE and LOCATION_HISTORY^{3,4} established essential components of this pathway: semantic representation of geography-linked exposure variables, PostGIS-based storage and transformation of external datasets, metadata-driven spatial joins, and provenance-aware linkage of exposure records to persons, locations, and time. The remaining bottleneck is not conceptual but operational. High-value exposure resources, including social vulnerability indices, environmental justice indicators, deprivation indices, and SDoH inventories, are commonly distributed as heterogeneous workbooks, GIS metadata files, and local data dictionaries. These artifacts encode variables, geography levels, values, score definitions, units, methodological notes, and provenance, but they are not immediately usable as OMOP-compatible vocabulary and mapping assets. Manual harmonization is possible, but it is slow, inconsistently documented, difficult to audit, and poorly suited to repeated use across network studies. This gap is increasingly urgent. Agentic AI is moving from experimental assistance toward routine scientific infrastructure. For OHDSI GIS activities, delayed adoption risks preserving manual bottlenecks precisely where the field needs scalable, transparent, and reviewable data preparation. The goal is not to replace domain expertise, but to reassign repetitive cognitive labor to bounded computational agents while keeping semantic judgment, scientific accountability, and implementation governance with subject-matter experts. We implemented Step 1 of an agentic AI pipeline for OHDSI GIS observational research: harmonizing CHoRUS⁵/PostGIS exposure source documentation into a reviewer-ready vocabulary and mapping layer that can support downstream GAIA, EXTERNAL_EXPOSURE, HADES, and ATLAS-facing workflows.

Methods

We implemented the workflow in OpenAI Codex as a repository-bound agentic development environment, using GPT-5.5 with high reasoning and on-request approval. Codex was configured over controlled project files, prompts, scripts, and source artifacts, with generated outputs maintained under version control. Rather than functioning as a single generative assistant, Codex served as an orchestration layer: a main orchestrator coordinated specialized agents for parsing, GIS vocabulary mapping, concept identifier assignment, semantic classification, QA, synthetic-data generation, medical/GIS review preparation, and documentation. The implementation was organized as a version-controlled repository with modular agent code (`src/`), explicit pipeline and mapping configuration (`config/`), vocabulary evidence (`Mappings/` and `Ontology/`), generated outputs, documentation, and automated tests. The reproducible execution path was `python -m src.orchestrator --config config/pipeline_config.yaml`, followed by `pytest`. Bounded prompts defined allowed inputs, required schemas, conservative mapping rules, provenance requirements, uncertainty handling, and SME-escalation criteria. The parser agent normalized CHORUS/PostGIS workbook tabs, including SVI, EJI, ADI, SDI, AHRQ, and SDoH Datasets in PostGIS Exposure, into a unified source dictionary. The mapping agent linked exact local source-code matches to existing GIS Vocabulary Package concepts, while ambiguous or unsupported terms were assigned `noMatch` with reviewer comments. The mapping policy preferred exact local source-code matches (`skos:exactMatch`; confidence 0.92). When no exact concept or existing local mapping evidence was available, cases were retained as `noMatch` (confidence 0.0), with the target concept left unresolved rather than assigned speculatively. Existing local mapping records were reused as reviewer evidence when no exact concept-code match took precedence. The concept ID agent assigned deterministic identifiers from the reserved CVB OHDSI GIS range. The QA agent detected malformed rows, duplicates, mapping gaps, variable-value conflation, geography-to-clinical-concept errors, and unsupported assumptions. Additional executable QA checked required output columns and selected source-field completeness, duplicate source codes within source sets, geography values and expected metadata presence, mapping schema and confidence ranges, required target-concept metadata when a target ID was present, rows flagged for reviewer decisions, person and source-concept links, exposure-date consistency, and synthetic-data provenance through a required notice file and tests of record-level metadata. A synthetic-data agent generated synthetic persons and `EXTERNAL_EXPOSURE`-style records with explicit synthetic provenance. The pipeline is configured to generate 100 synthetic patient records and synthetic exposure records for demonstrating table representation and workflow structure, not clinical or epidemiologic validity. The repository's synthetic-data notice designates both output files as synthetic and states that they contain no real patient data or PHI. Each exposure record is also labeled as synthetic, although patient rows are not individually tagged. All outputs were treated as reviewer-ready drafts requiring SME validation. Unresolved mappings were not automatically adjudicated: `noMatch`, `Needs review`, and lower-confidence rows were explicitly routed to GIS/medical review, while source-concept-ID reconciliation, `EXTERNAL_EXPOSURE` modeling approval, and production CVB/OHDSI release remained human-governed decisions outside autonomous agent authority.

Results

The pipeline generated a unified source dictionary, reviewer-ready GIS vocabulary mappings, deterministic source concept IDs, mapping and structural QA artifacts, critical-risk reports, and synthetic EXTERNAL_EXPOSURE test fixtures. The produced results are shown in Figure 1:

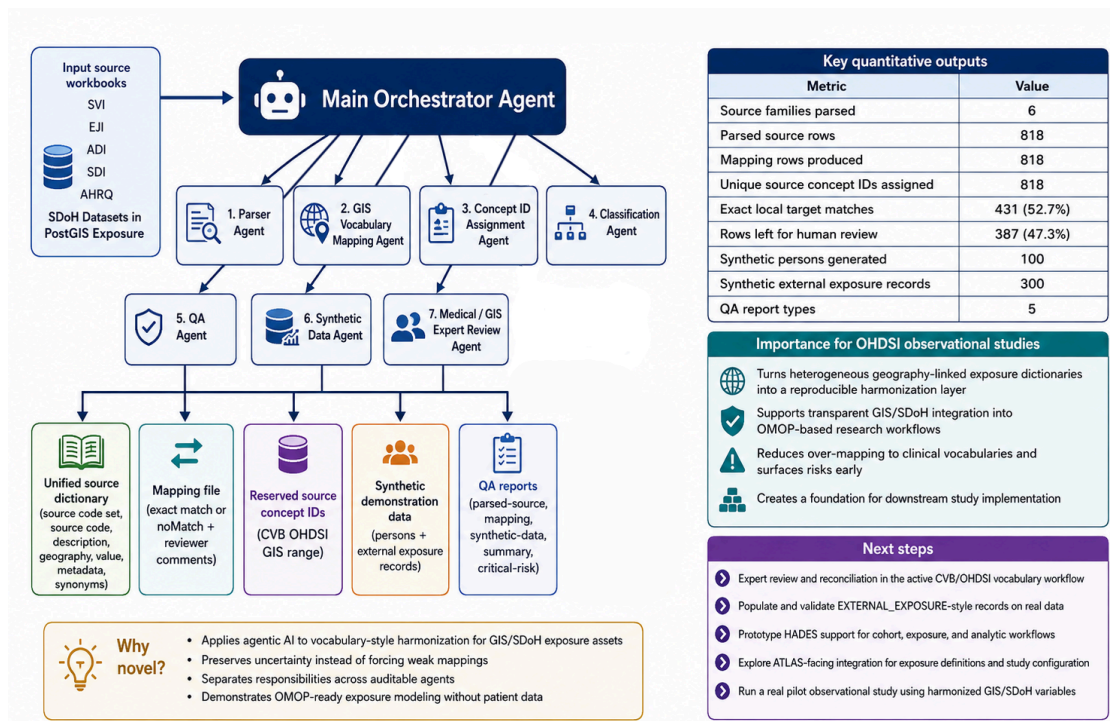


Figure 1. Agentic AI orchestration for exposure data harmonization.

Across the current run, 818 parsed source rows produced 818 mapping rows and 818 distinct assigned source concept IDs. Of these, 431 (52.7%) were supported by exact local-vocabulary matches, while 387 (47.3%) remained noMatch or were left for human review. QA flagged 0 structural issues across the predefined error categories.

The most important result was separation of machine-executable steps from expert-governed decisions. The SME-in-the-loop review layer identified risks central to OHDSI GIS implementation: confusing exposure variables with result values, treating geography as a clinical concept, converting SDoH metadata into patient-level facts, over-mapping local exposure variables to clinical vocabularies, and overstating synthetic-data demonstrations as analytic validation. These findings support the design principle that agentic AI is most valuable when it accelerates structured preparation, comparison, and error discovery, while preserving human authority over semantic acceptance. The workflow also revealed an orchestration limitation. A main orchestrator agent cannot remain a passive dispatcher if the task involves heterogeneous scientific documentation. It requires explicit self-prompting behavior: checking whether each intermediate output satisfies the governing schema, asking whether assumptions were introduced, testing whether uncertainty was preserved, and deciding when to escalate unresolved semantics to reviewers. Without this reflective control loop, multi-agent workflows can produce fluent but brittle artifacts that appear complete while silently encoding mapping errors.

Limitations

The current evaluation is intentionally a prototype validation rather than an estimate of mapping accuracy or efficiency. No independent gold-standard mapping set, precision/recall evaluation, quantified post-SME error rate, or controlled curator-time comparison against manual harmonization was performed. Likewise, the synthetic `EXTERNAL_EXPOSURE` output demonstrates structural executability but does not yet establish successful downstream GAIA/HADES analysis, phenotype validity, or real-world exposure attribution.

Conclusions

This work extends prior OHDSI GIS infrastructure by placing an agentic harmonization layer between heterogeneous external exposure documentation and OMOP-compatible research workflows. Its novelty is an auditable orchestration pattern in which parsing, vocabulary mapping, concept ID assignment, QA, synthetic demonstration, and expert review are distributed across specialized agents that preserve uncertainty rather than masking it through forced mappings. The current output is a curated draft, not a production release; source concepts, mappings, and `EXTERNAL_EXPOSURE` assumptions require reconciliation with the active CVB/OHDSI vocabulary workflow and expert GIS review. This includes final identifier reconciliation, provenance and licensing review, semantic approval, and release-readiness governance before generated vocabulary content can be treated as reusable OHDSI vocabulary assets. The expected impact is a shift from artisanal, document-by-document harmonization toward reproducible, governed, and scalable preparation of GIS/SDoH assets for observational research.

The current prototype was evaluated only in OpenAI Codex. Portability to local or open-source LLMs has not yet been tested, and deployment safeguards for workflows involving restricted geospatial data, PHI, or institution-specific data-sovereignty requirements require separate evaluation. The modular, schema-bound architecture provides a natural target for such model substitutions, but model equivalence should not be assumed without testing.

Beyond harmonization, next steps should mature the agentic layer into a sequence of reviewable pipelines. An SME-review pipeline should triage uncertain mappings, surface semantic conflicts, prepare reviewer packets, capture decisions, and feed accepted rules back into agent configuration. A vocabulary-publication pipeline should align approved concepts and mappings with CVB/OHDSI staging, synonym management, provenance checks, relationship staging, and release-readiness QA. An executable QA pipeline should continuously test structural validity, duplicate handling, geography/value separation, unsupported mappings, provenance completeness, and `EXTERNAL_EXPOSURE` modeling assumptions. A GAIA validation pipeline should apply the approved layer to real CHORUS/PostGIS exposure data and verify spatial joins, temporal attribution, and exposure-record generation. An analytic-enablement pipeline should translate harmonized GIS/SDoH variables into HADES-compatible exposure cohorts, ATLAS-facing exposure-definition templates, and reusable study components. A study-execution pipeline should then support a pilot observational study through data preparation, cohort construction, covariate generation, diagnostics, provenance dashboards, and reproducible reporting. Across all stages, the main orchestrator should use self-prompting checkpoints to detect introduced assumptions, preserve uncertainty, trigger SME escalation, and confirm downstream fitness.

for OHDSI network research. The long-term objective is a trustworthy socio-technical workflow in which agentic AI becomes routine OHDSI infrastructure for environmental and social-context research.

References

1. OHDSI GIS Working Group. GIS Vocabulary Package [Internet]. OHDSI; [cited 2026 Feb 6]. Available from: <https://ohdsi.github.io/GIS/vocabulary.html>
2. OHDSI GIS Working Group. GIS Toolchain [Internet]. OHDSI; [cited 2026 Feb 6]. Available from: https://ohdsi.github.io/GIS/#GIS_Toolchain
3. OHDSI GIS Working Group. EXTERNAL_EXPOSURE Table - GIS Schema Extensions [Internet]. OHDSI; [cited 2026 May 15]. Available from: https://ohdsi.github.io/GIS/schema-extensions.html#external_exposure
4. OHDSI GIS Working Group. LOCATION_HISTORY Table - GIS Schema Extensions [Internet]. OHDSI; [cited 2026 May 15]. Available from: https://ohdsi.github.io/GIS/schema-extensions.html#location_history
5. Bridge2AI Consortium. CHoRUS Dataset - Data Generation Project [Internet]. NIH Bridge2AI; [cited 2026 May 16]. Available from: <https://bridge2ai.org/data-chorus/>.