

RWD-OMOP: A cost aware generative AI workflow automation tool

Pankaj Attri, Sherrine Eid
SAS Institute Inc., SAS Institute Inc.

Background

Harmonizing heterogeneous real-world data (RWD) into the OMOP Common Data Model (CDM) enables scalable, reproducible analyses across the OHDSI ecosystem. The conversion is however impeded by missingness, PII/PHI risk, bias, complex vocabulary mapping, evolving specifications, and performance and cost constraints. We embarked on a mission to create and validate a framework that addresses these challenges by profiling data quality and bias, producing and curating LLM-assisted concept mappings, auto-generating dependency-aware ETL for creating OMOP CDM tables, and instituting iterative quality and cost/performance gates.

Methods

Our framework uses generative AI to accelerate and strengthen OMOP. It begins with comprehensive data profiling to assess completeness, conformance, plausibility, PII/PHI risk, and potential biases. It then leverages large language models (LLMs) to draft mapping rules from source codes and free text to OMOP standard and custom concepts. A user interface enables expert review, curation, and approval of mappings, with full provenance. Once mappings are accepted, the system generates executable, dependency-aware transformation code to create CDM tables. Implementations support in-process SQL engines such as DuckDB and can target scalable platforms as needed. Quality gates are embedded throughout. The framework allows OHDSI tools including DQD and Achilles to automatically evaluate conformance, completeness, and plausibility. Results feed a closed-loop process that can refine mappings or regenerate code until configurable thresholds are met before release for analysis. Performance and cost are addressed through incremental processing, columnar open formats (e.g., Parquet, Vortex), and query pushdown to reduce storage footprint and ETL run times.

Results

The prototype demonstrated RWD-to-OMOP CDM table construction with dependency aware orchestration and successful execution of data quality checks under configurable thresholds. Agent-generated candidates provided high-coverage starting points for common source codes, and alternate concept suggestions accelerated expert curation while preserving provenance. High level table mapping (Figure 1), field level pseudo ETL code (Figure 2a, 2b), and SQL code (Figure 3) generated by the agents offered actionable starting points for expert verification and materially shortened CDM table build time. At every step, human experts could edit, revise, and audit the reasoning applied by agents, ensuring that decisions remained transparent and governed. Use of columnar formats and in-process execution decreased I/O and improved ETL latency relative to row-oriented baselines, while partitioning and pushdown reduced storage footprint and cost.

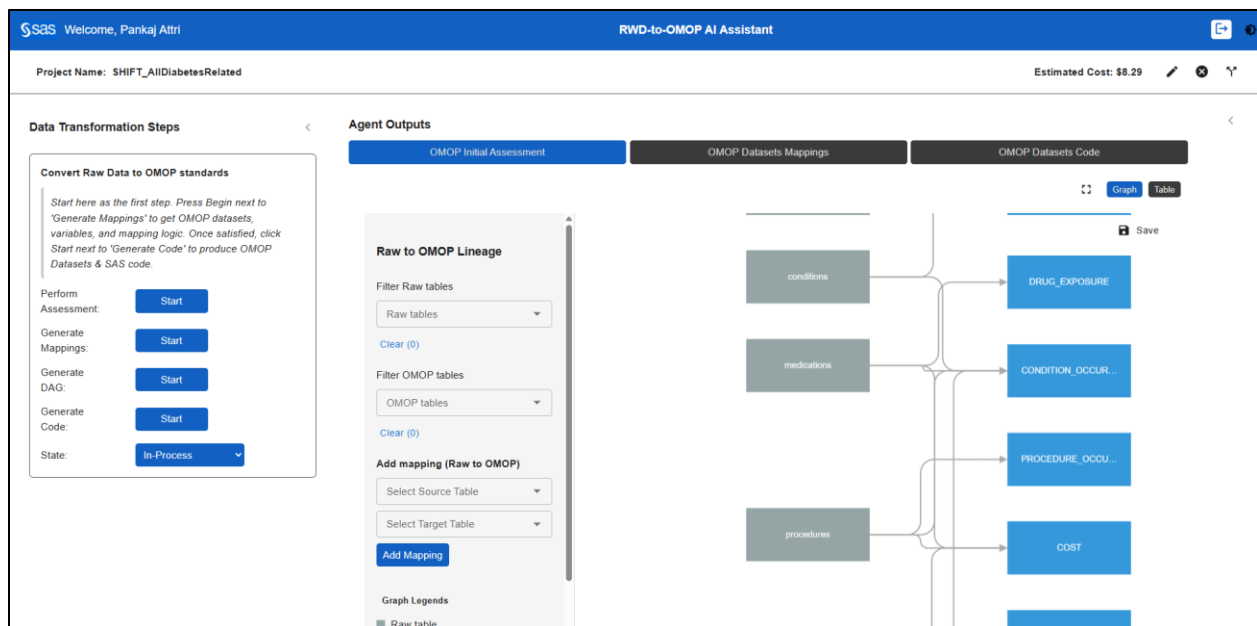


Figure 1. Table Level mapping generated by agent

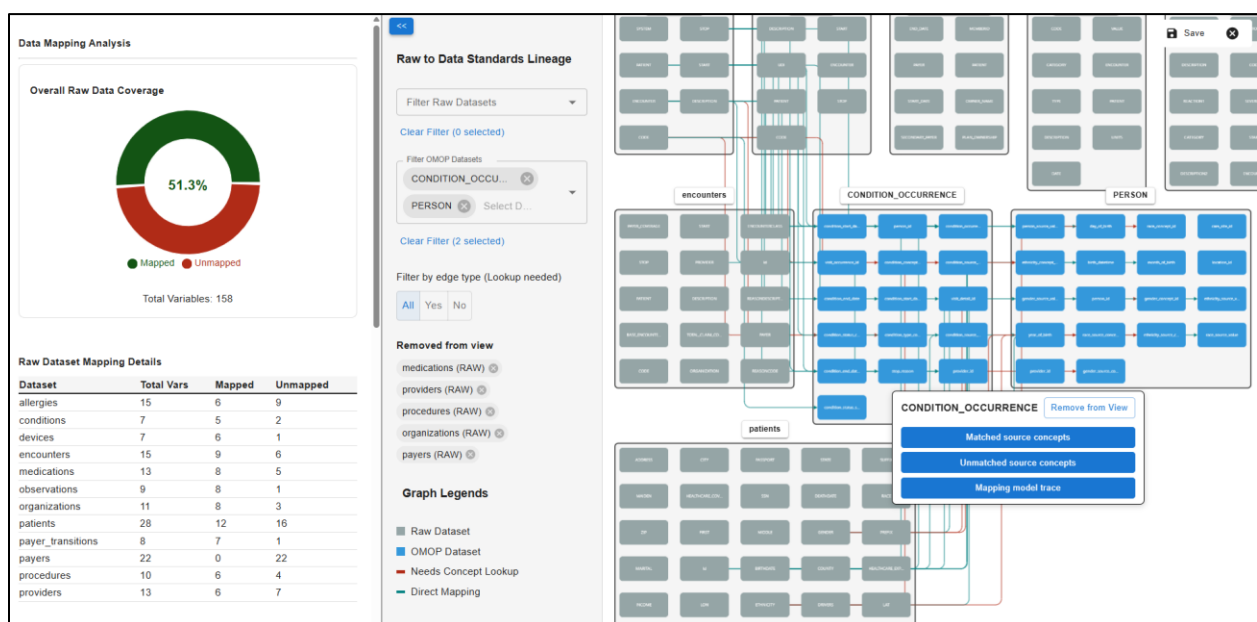


Figure 2a. Field level ETL rules and clinical concepts mappings generated by agent

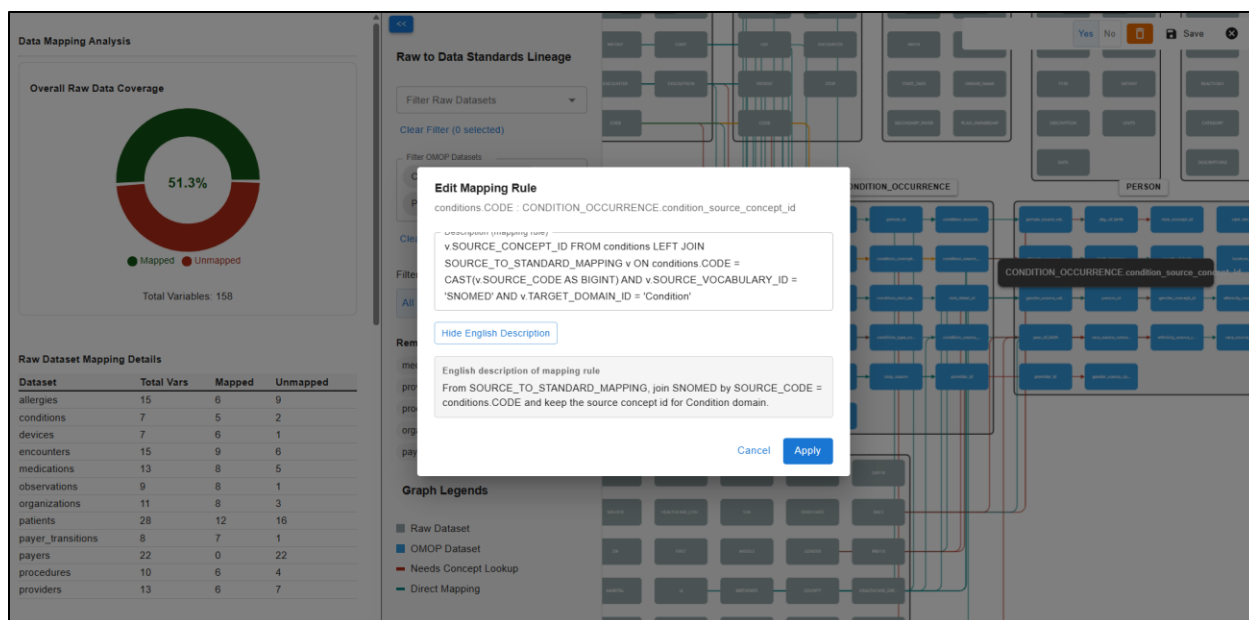


Figure 2b. Field level ETL rules and clinical concepts mappings generated by agent

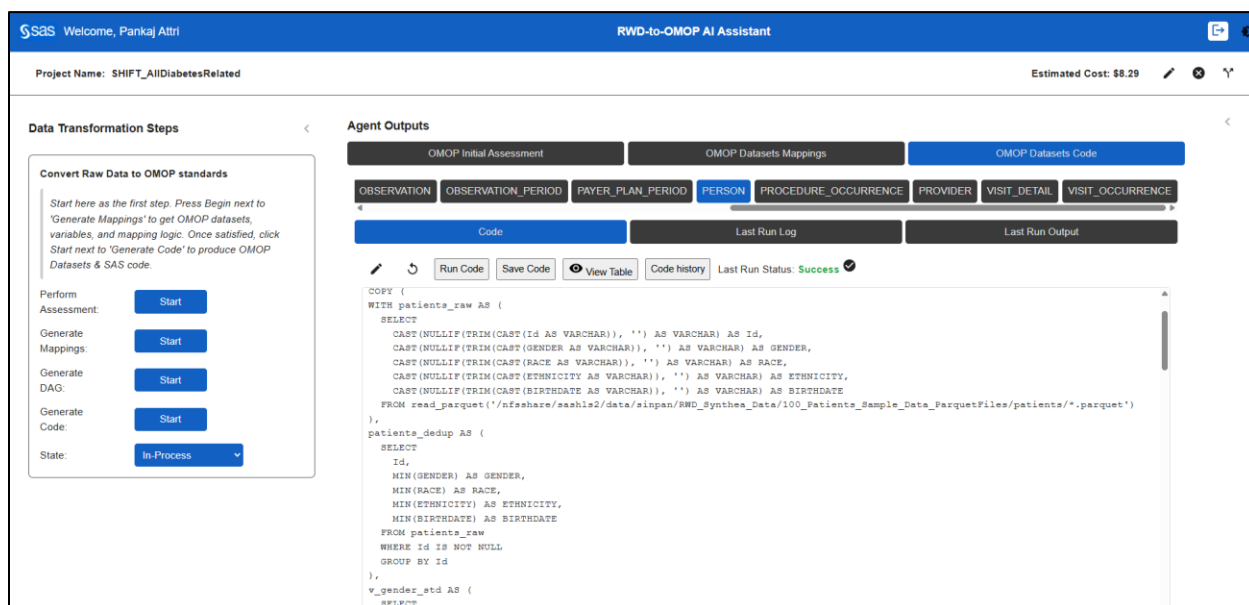


Figure 3. Agent generated executable SQL code (with error-feedback loop) for selected OMOP CDM tables

The results generated by running 100 patients' data from Synthea showed promising results. See below a snapshot of the DQD (Data Quality Dashboard) summarizing results.

DATA QUALITY ASSESSMENT

PROVIDED RWE SOURCE: RWE_MEMBER/ENROLL/OBSN/RX/SERVICES/SERVICE_CODES

DataQualityDashboard Version: 2.8.6

Results generated at 2026-04-27 01:48:03 in 8 mins

	Verification				Validation				Total			
	Pass	Fail	Total	% Pass	Pass	Fail	Total	% Pass	Pass	Fail	Total	% Pass
Plausibility	510	5	515	99%	291	0	291	100%	801	5	806	99%
Conformance	737	101	812	91%	119	0	119	100%	856	101	931	92%
Completeness	431	54	449	96%	16	3	19	84%	447	57	468	96%
Total	1678	160	1776	94%	426	3	429	99%	2104	163	2205	95%

819 out of 2104 passed checks are [Not Applicable](#), due to empty tables or fields.

62 out of 163 failed checks are SQL errors.

Corrected pass percentage for NA and Errors: 97% (1285/1324).

Figure 4. DQD Results

Conclusion

A transparent, AI-assisted, quality-gated framework can shorten time-to-OMOP, improve data fitness-for-use, and control operational costs without compromising governance. By uniting agent-assisted mapping, dependency-aware orchestration, and continuous OHDSI quality assessment, the approach makes high-quality, auditable OMOP conversions more attainable at scale and establishes a practical path to repeatable, cost-effective RWD standardization.