Bridging FHIR and OMOP: Data Lineage for Observational Data Conversion

Berk BB, Benzie M, Bolisetty B, Favre S, Fortune J, Goslin J, Kheterpal V, Marinan K, Marsan A, Ramos E[2], Venumuddula S, Wyderko J ([1]All authors: CareEvolution, Ann Arbor, MI, USA; [2]Also affiliated with: Digital Trials Center, Scripps Research Translational Institute, La Jolla, CA, USA)

Background

The interoperability between Fast Healthcare Interoperability Resources (FHIR) and the Observational Medical Outcomes Partnership Common Data Model (OMOP CDM) has become increasingly important for healthcare data research [1]. While FHIR excels as a clinical data exchange standard, OMOP provides a standardized format optimized for observational research. Converting data between these models presents significant challenges in maintaining data integrity and traceability.

The HL7 Vulcan FHIR→OMOP Working Group was established to develop a standard implementation guide for this transformation [2-4]. Previous work, including the FHIR-to-OMOP Cookbook, CAMP FHIR, and NACHC's fhir-to-omop, has focused primarily on mapping between resources and tables [5-10]. Here we implement comprehensive data lineage tracking throughout the transformation process.

This paper presents our experience implementing a FHIR→OMOP transformation API [10] that captures data lineage, enabling tracing of data from source to destination across four dimensions:

- **Data Source Provenance**: Tracking source system provenance of each element.
- **Entity Lineage**: Maintaining traceability from FHIR resources to resulting OMOP table rows.
- **Concept Standardization**: Recording source codes to OMOP standard concepts.
- **Processing Events**: Capturing comprehensive logs of transformation decisions, including warnings, informational messages, errors, and documentation of FHIR resources that do not result in OMOP rows with detailed explanations.

Methods

We developed a REST-based FHIR→OMOP conversion API [11] that implements a comprehensive data transformation pipeline: data ingestion → concept standardization → domain identification → field mapping → OMOP output. Concept standardization follows established OMOP conventions, including THEMIS collaborative recommendations for standardized analytics [12] and quality criteria from the OHDSI Network Data Quality Dashboard [13]. Source codes are mapped to OMOP standard concepts using the CareEvolution Orchestrate Terminology API which performs exact code and

display matching and natural language processing. When standard concepts cannot be identified, source codes are preserved with concept_id = 0 and detailed rationale captured in the PROCESSING_LOG table.

Throughout this pipeline, we maintain comprehensive lineage tracking, populating three extension tables alongside standard OMOP CDM tables:

DATA_SOURCE Table: Captures data source lineage and entity tracking by maintaining unique resource IDs for each FHIR resource, source system metadata, and complete mapping between FHIR resource instances and resulting OMOP table rows.

SOURCE_CODING Table: Documents concept standardization, preserving original source codes and links to OMOP rows.

PROCESSING_LOG Table: Records all processing events including transformation decisions, data quality issues, exclusion rationales, and detailed explanations for FHIR resources that are not transformed to OMOP records. Events are categorized based on level (Error, Information, Warning) and a structured set of codes (See Table 2 below).

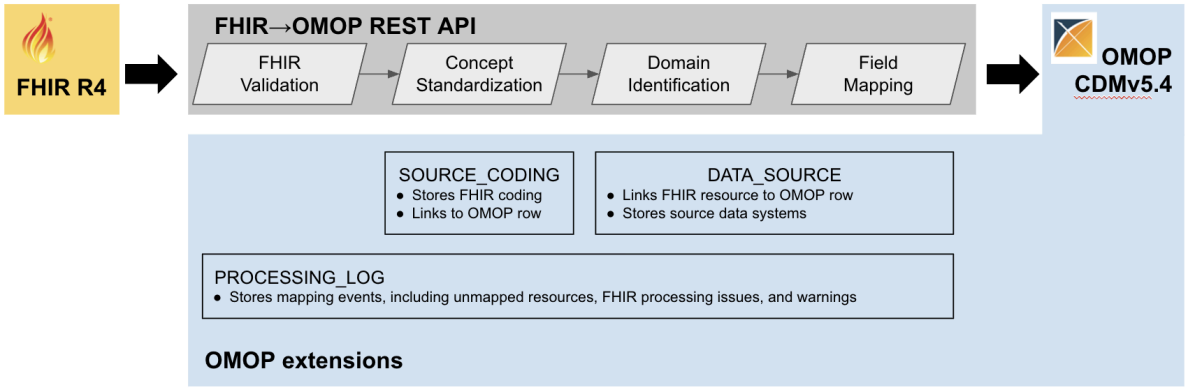

Figure 1: FHIR→OMOP transformation pipeline with comprehensive data lineage tracking.

We considered leveraging existing OMOP tables for lineage data but concluded that they were not appropriate or adequate. We found the existing OMOP CDM tables to be inadequate for comprehensive FHIR→OMOP lineage tracking because the METADATA table is designed only for high-level CDM instance information rather than individual resource transformation tracking, while the FACT_RELATIONSHIP table captures relationships between OMOP facts rather than lineage back to source FHIR resources. The SOURCE_TO_CONCEPT_MAP table, while useful for vocabulary mappings, requires pre-loading vocabularies into the OMOP vocabulary tables and has column length constraints that cannot reliably accommodate FHIR codesystem URLs as vocabulary IDs, making it impractical for FHIR data sources that contain a wide variety of site-specific code systems in a non-database-based ETL approach. Therefore, we implemented three purpose-built extension tables to capture the four-dimensional lineage framework and handle the diverse vocabulary landscape with full URL preservation that standard OMOP CDM tables were not designed to accommodate.

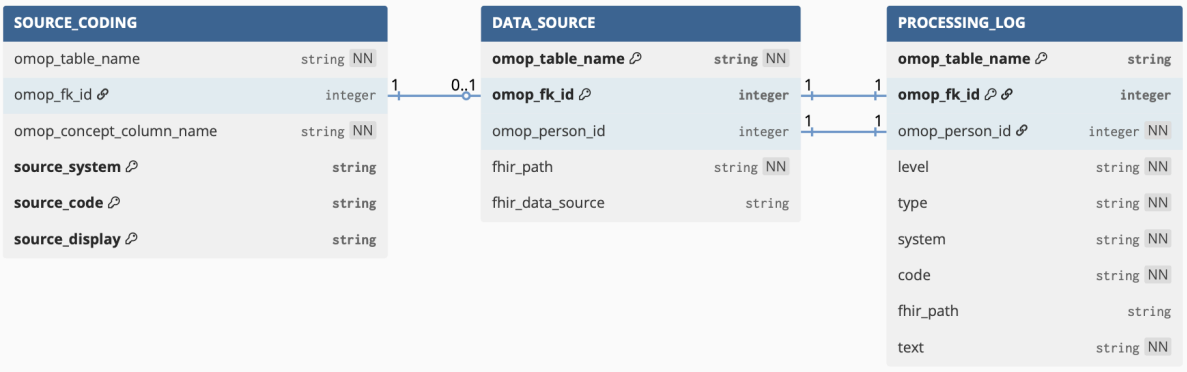

Figure 2: Extension table schemas.

The dataset was drawn from the PRediction Of Glycemic RESponse Study (PROGRESS) study conducted by Scripps Digital Clinical Trials Center using CareEvolution's MyDataHelps research platform [14, 15]. The data for 665 individuals was sourced via participant-mediated exchange from 2170 unique zip codes across 46 US states and a diverse set of EMRs and other FHIR providers. On average, each individual had data from 6.3 source systems.

To ensure adherence to the OMOP CDM standard, we ran output through the OHDSI Data Quality Dashboard and deployed a managed instance of OHDSI's Atlas tool connected to the OMOP CDM output.

data_source_type	person_id
Other FHIR	630
Epic	563
Payor (Other)	27
Cerner	24
Payor (CMS)	19
Athena	19
NextGen	7
eClinicalWorks	5
Allscripts	1

Table 1: Individuals by Data Source

To ensure adherence to the OMOP CDM standard, we ran output through the OHDSI Data Quality Dashboard and deployed a managed instance of OHDSI's Atlas tool connected to the OMOP CDM output.

Results

For this study, we processed over 2.8 million FHIR resources, generating comprehensive lineage data for all transformations through the three extension tables. The DATA_SOURCE table maintained complete entity tracking with over unique FHIR resource IDs linked to 1.1 million corresponding OMOP table rows. The SOURCE_CODING table documented over 3.8 million source codings, preserving both successful standardizations and unmappable source codes. The PROCESSING_LOG table captured detailed transformation events.

level	type	code	fhir_resource
Error	Processing	DuplicateFhirElement	985
		FhirProcessingIssue	1
Information	Processing	NoActionTaken	1878
		ProviderMissing	1748
		ProviderNotReferenced	2178
		ResourceRemovedBecauseOfIgnoredCode	32
		ResourceRemovedBecauseOfStatusCode	4322
		ResourceRemovedBecauseOfVerificationCode	1644
		RowCreated	2160867
		RowDuplicatesMerged	522315
		ValueTruncated	405761
		VisitInFuture	2
		VisitMissingStartDate	342
Warning	Processing	FhirProcessingIssue	71665
		PatientMissingDateOfBirth	14
		PatientMissingDateOfDeath	1

Table 2: PROCESSING_LOG events

Lineage tracking revealed patterns in data transformation decisions. For example, nearly 522k resources were duplicates. ~19% of over 75,358 Condition resources were transformed to OBSERVATION rather than CONDITION_OCCURRENCE if no standard OMOP concept can be applied or missing onset date. 2% of Condition resources did not populate an OMOP row due to verification codes (OMOP requires only *confirmed* conditions). More than 99% of all non-duplicate resources generated rows in OMOP tables.

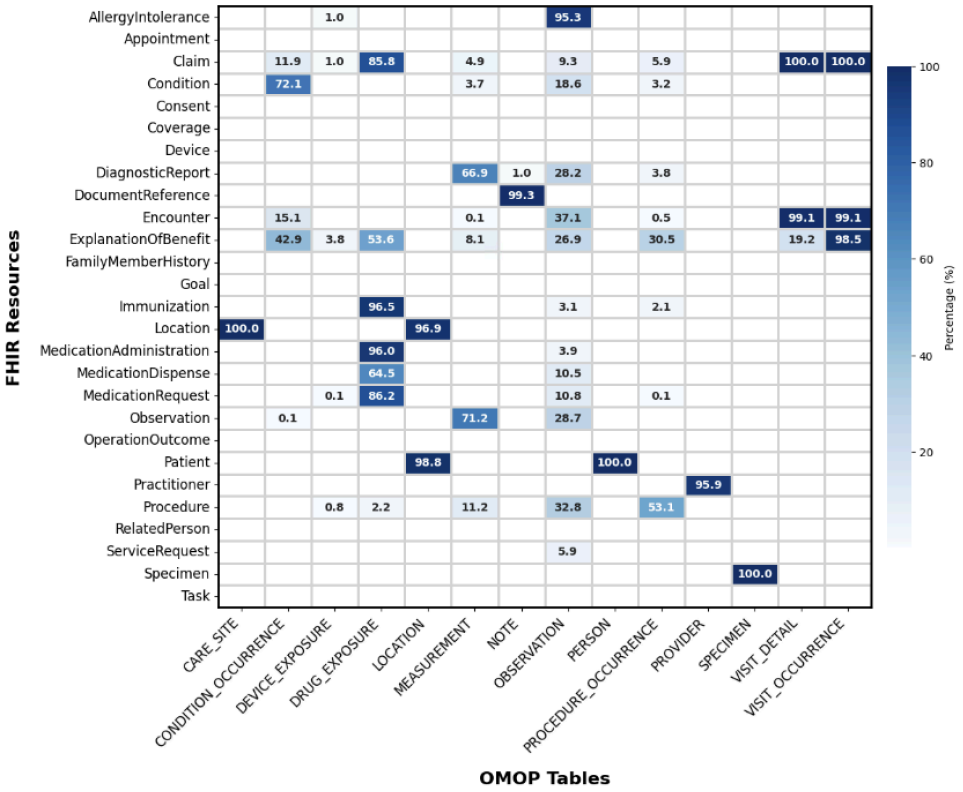

Figure 3: Heat map of % of FHIR resources by resourceType mapped to each OMOP table (FHIR resources can map to multiple OMOP rows)

To validate our lineage approach, we conducted trace-back tests on randomly selected OMOP records across all domains. Using our extension tables, we were able to reconstruct the complete data transformation chain back to the original FHIR resources.

Conclusions

Our FHIR⟶OMOP transformation pipeline with data lineage tracking addresses a gap in healthcare data interoperability for research. By systematically tracking data source lineage, entity relationships, concept standardization decisions, and processing events through dedicated extension tables, we enable researchers to understand precisely how data has been transformed.

The OHDSI community should recognize comprehensive data lineage as an essential component of FHIR⟶OMOP transformation pipelines. Our four-component lineage framework (data source, entity tracking, concept standardization, and processing events) provides a potential approach for implementing data traceability. The three extension tables offer a practical implementation model that other organizations can adopt to enhance transparency and trust in their transformed data.

This approach has particular significance for regulatory compliance and research validation, where complete data provenance is essential. Building on methodological foundations established in the FHIR to OMOP Cookbook and extending prior work presented at OHDSI 2024, our implementation

demonstrates that comprehensive lineage tracking is achievable at scale and significantly enhances the trustworthiness of transformed data for research applications.

Future work will focus on enhancing granularity of the processing log based on researcher feedback and developing standardized reporting mechanisms that leverage lineage data to provide researchers with automated data quality assessments and transformation summaries.

References

1. Tsafnat G, Dunscombe R, Gabriel D, Grieve G, Reich C. Converge or Collide? Making Sense of a Plethora of Open Data Standards in Health Care. J Med Internet Res. 2024 Apr 9;26:e55779. doi: 10.2196/55779.
2. Vulcan FHIR-to-OMOP Project Team. Vulcan FHIR-to-OMOP: HL7 FHIR to OMOP CDM Data Transfer Best Practices. HL7 Confluence; 2023. https://confluence.hl7.org/display/FHIR/2023+-+01+Vulcan+FHIR-to-OMOP
3. Duteau J, Gabriel D. Vulcan FHIR to OMOP IG Project Value Proposition Summary. HL7 Vulcan Accelerator; 2024.
4. HL7 International. FHIR to OMOP Implementation Guide; 2024. http://build.fhir.org/ig/HL7/fhir-omop-ig/.
5. Hong N, Wen A, Shen F, et al. FHIR-Ontop-OMOP: Building clinical knowledge graphs in FHIR RDF with the OMOP Common data Model. J Biomed Inform. 2022;134:104187.
6. Livne G, Terry M, Yang Q. FHIR to OMOP Cookbook: A starter guide for implementers seeking to convert HL7 FHIR resources to the OHDSI OMOP Common Data Model. 2024.
7. Ahalt SC, Chute CG, Fecho K, et al. Fast Healthcare Interoperability Resources (FHIR) as a Meta Model to Integrate Common Data Models: Development of a Tool and Quantitative Validation Study. JMIR Med Inform. 2019;7(4):e15199.
8. OHDSI. FhirToCdm: Conversion from FHIR HL7 to OMOP CDM. GitHub Repository; 2023. https://github.com/OHDSI/FhirToCdm.
9. National Association of Community Health Centers (NACHC). fhir-to-omop: FHIR to OMOP Conversion Tools. GitHub Repository; 2023. https://github.com/NACHC-CAD/fhir-to-omop.
10. Georgia Tech. OMOP on FHIR: Implementation of mapping between OMOP and FHIR. GitHub Organization; 2023. https://github.com/omoponfhir.
11. CareEvolution. Rosetta API: FHIR (R4) to OMOP Conversion. CareEvolution Documentation; 2023. https://rosetta-api.docs.careevolution.com/convert/fhir_to_omop.html
12. https://ohdsi.github.io/Themis/
13. https://ohdsi.github.io/DataQualityDashboard/
14. Carletti M, Pandit J, Gadaleta M, Chiang D, Delgado F, Quartuccio K, Fernandez B, Garay J, Torkamani A, Miotto R, Rossman H, Berk B, Baca-Motes K, Kheterpal V, Segal E, Topol EJ, Ramos E, Quer G. Multimodal A.I. correlates of glucose spikes in people with normal glucose regulation, pre-diabetes, and type 2 diabetes. 2025. Nature medicine.
15. Pawelek J, Baca-Motes K, Pandit J, Berk B, Ramos E. The Power of Patient Engagement With Electronic Health Records as Research Participants. JMIR Med Inform 2022;10(7):e39145. https://medinform.jmir.org/2022/7/e39145