Bridging Standards: Transforming Consolidated Clinical Document Architecture (C-CDA) Data via Health Information Networks to OMOP

Xiaohan Tanner Zhang MD, MS[1], Chris Roeder MS[2], Stephanie Hong MS, FAMIA[1], Thanaphop Na Nakhonphanom MS, MMedSc, MD[2], Adam Lee PhD[2], Richard Moffitt PhD[3], Josh Lemieux BA[4], James Cavallon BS[1], Monique Bangudi MPH[1], Lakshmi Anandan MPH[2], Rob Schuff MS[4], Bill Hogan MD, MS[5], Chris Chute MD, DrPH[1], Emily Pfaff MS, PhD[2], Melissa Haendel PhD[2]

1. Johns Hopkins University, 2.University of North Carolina at Chapel Hill, 3. Emory, 4. OCHIN, 5. Medical College of Wisconsin

Background

The *All of Us* (AoU) Research Program (8) is a historic effort to gather rich, multimodal health data from one million people across the United States, including Electronic Health Record (EHR) data in OMOP format to enable research. The Center for Linkage and Acquisition of Data (CLAD) has been tasked with passive data collection, including collecting EHR data from Health Information Networks and Exchanges (7) (HINs and HIEs, respectively). HINs and HIEs are part of the Trusted Exchange Framework and Common Agreement (TEFCA) (5) framework for health data sharing. The common Consolidated Clinical Document Architecture (C-CDA) (6) format is rooted in HL7 Reference Information Model (RIM) (3), but is difficult to use for analytics due to its complex structure and pervasive variability in source data. However, a significant amount of data is available in the C-CDA format in the US, making the transformation of clinical data from C-CDA to OMOP a potential mechanism to acquire missing EHR data for AoU participants. Here, we demonstrate the first scalable, national acquisition of participant-authorized EHRs via eHealth Exchange by a research entity; the development of a robust, repeatable ETL pipeline for transforming C-CDA documents into OMOP; and an evaluation of data quality and consistency compared to OMOP data received directly from provider organizations in AoU.

Methods

Data Retrieval. Using Master Patient Index (MPI) identifiers for AoU participants, participant-mediated SOAP calls using the research purpose of use were made to retrieve structured clinical data in the form of C-CDA XML documents. Documents were retrieved from the HIN eHealth Exchange and subsequently via the HIE, Manifest MedEx, with records retrieved in C-CDA. Some records were also retrieved in FHIR format from the individual Health Provider Organization (HPO) Cedars-Sinai.

Overview of the C-CDA to OMOP ETL Pipeline. This solution implements a rule-based Python-SQL-mixed ETL pipeline for normalizing and mapping C-CDA source codes to OMOP standard concepts. The multi-stage process includes vocabulary mapping and structural mapping.

Vocabulary Mapping. The architecture involves multiple facets, including the creation of value set mapping tables and an OID mapping table, handling of unstructured vocabulary names/codes, and the creation of mapping logics using standard OMOP vocabulary tables.

1. C-CDA Value Set Mapping Table Creation. This table addresses specific mapping challenges not resolved by standard OMOP vocabulary resources. Examples include HL7 Administrative Gender, HL7 Race & Ethnicity, and EPIC internal vocabularies.

2. HL7 OID to OMOP Vocabulary ID Map Creation. This table bridges Object Identifiers (OIDs), used in HL7 and C-CDA to identify terminology systems, to OMOP standardized vocabulary identifiers (e.g., OID 2.16.840.1.113883.6.96 for SNOMED CT is mapped to "SNOMED"). This maintains terminology mapping between the HL7 and the OMOP vocabulary IDs.

3. Handling Unstructured and Misclassified Data

- **Code System Normalization:** This stage standardizes vocabulary references, handling variations in naming conventions (e.g., "CPT" vs. "CPT4"), case sensitivity, OID references, and mapping non-standard vocabularies to OMOP equivalents. This ensures consistent matching to the correct OMOP vocabulary.
- **Advanced Code Pattern Recognition:** This innovative feature uses regular expressions to analyze code structural patterns and automatically correct misclassifications, such as ICD-10-PCS codes mislabeled as ICD-10-CM or mixed ICD-9/ICD-10 codes from ICD dual coding time that cannot be rendered by OMOP tables automatically. This proactive correction prevents inaccurate concept assignments.
- **Code Format Normalization:** Addresses variable NDC code (e.g., 5-4-2, 5-3-2, 4-4-2 segments, with or without hyphens) and ICD code formatting (with or without dots and coordinations) by detecting patterns and applying standardization using SQL string functions. This ensures correct matching regardless of original formatting.

4. Utilization of Athena Vocabulary Tables for Mapping creation. This table is created after the codes have been normalized. The codes are then mapped to standard OMOP concept IDs using ‘maps to’ relationships. Expired codes are retained to preserve historical data integrity. Routine vocabulary updates are expected to enhance accuracy.

Structural Mapping. The structural mapping component uses the Python lxml.etree package (1) and XPath (2) to parse C-CDA XML documents. A rule-based system, defined in Python dictionaries, is applied to map data from specific C-CDA sections (e.g., problems, medications) to OMOP tables (e.g., Conditions, Drug_exposure), and is inspired by the readability of Google’s Whistle data transformation language (4).

Figure 1 details how C-CDA observation elements from the document's Results section are mapped to the OMOP measurement table. The process begins from a defined 'root' path of .../component/observation. Each mapping rule then builds upon this path to populate a specific OMOP field. For example, the rule for value_as_number uses element: 'value' to navigate to the <value> element and attribute: 'value' to extract the number 10.2. Other OMOP fields are populated in two steps. See measurement_concept_id and measurement_id in the figure.

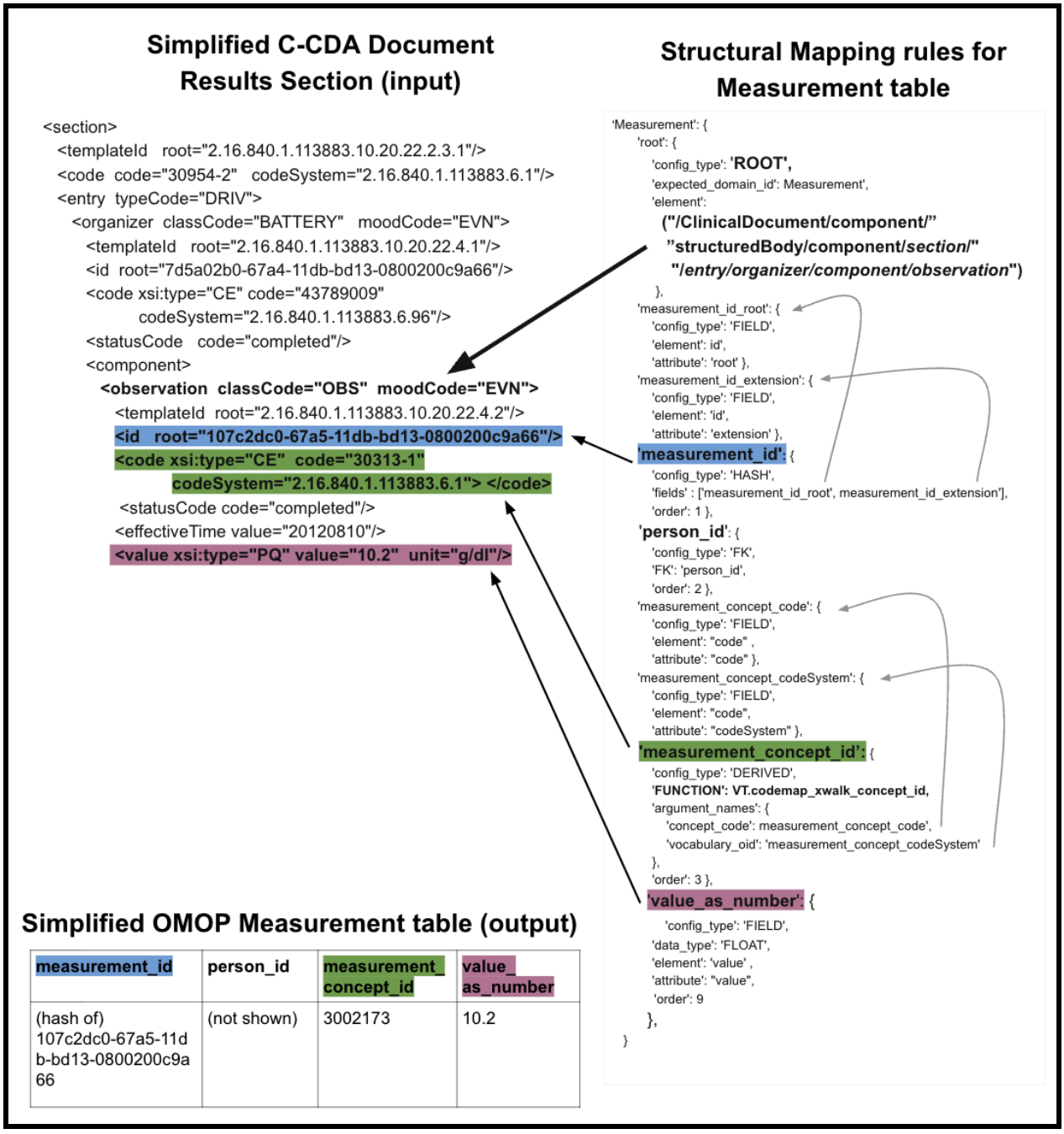

Figure 1. Simplified Mapping for OMOP measurement table. Shown are the links between XML elements in a C-CDA document (left), the mapping rules (right), to create (parts of) an OMOP measurement table row (bottom).

Results

The CLAD ETL pipeline processed over 56,000 production files from Cedars-Sinai (an HPO) and Manifest MedEx (a California HIE) via retrieval from eHealth Exchange (the HIN) with high performance, handling the files in approximately 30 hours. The mapping achieved high success rates across key OMOP domains,

including near-100% completion for Drug Exposure, Measurement, Condition, and Procedure records from both the HPO and HIE sources. Qualitatively, the transformed data accurately reflects clinical nuances, vocabulary cleaning, and yields reliable, research-ready datasets.

Figure 2 shows how the C-CDA and FHIR feeds add detail to the data collected by other methods. Table 1 shows vocabulary mapping rates.

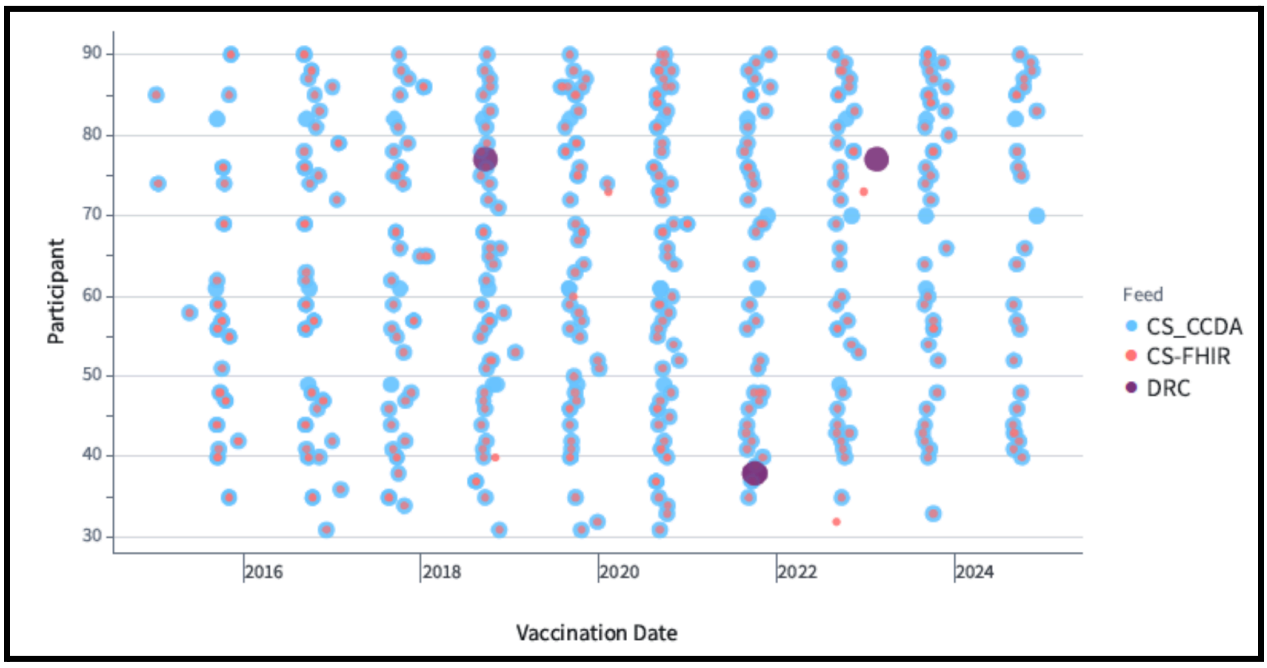

Figure. 2 Example complementary content coming via the HIN/HIE/HPO retrieval process. Shown are influenza vaccination counts by date for a sample of random participants for each of three data sources: the HPO Cedars Sinai FHIR, Cedars Sinai CCDA, and the AllOfUs direct submission.

Data Source	OMOP Table	Total Records	Mapping Rate	Notes
Site A	Drug Exposure	2,020,862	100%	Complete mapping achieved.
Site A	Measurement	842,392	100%	Complete mapping achieved.
Site A	Condition	49,610	100%	Complete mapping achieved.
Site A	Procedure	383,513	100%	Complete mapping achieved.
Site A	Observation	163,124	99.96%	Minor unmapped due to data quality.
Site A	Visit Occurrence	708,297	0.01%	Significant mapping limitation identified.

Site B	Drug Exposure	52,823	100%	Complete mapping achieved.
Site B	Measurement	5,329,582	100%	Complete mapping achieved.
Site B	Condition	235,216	100%	Complete mapping achieved.
Site B	Procedure	193,634	100%	Complete mapping achieved.
Site B	Visit Occurrence	903,854	96.20%	High mapping success.
Site B	Observation	529,567	6.80%	Unmapped due to non-standard encodings.

Table 1. Term mapping rates by site and OMOP table – The percentages displayed in the table represent the final counts of successfully ingested records into the OMOP Common Data Model, divided by the total counts of source records that were intended to be mapped for each respective OMOP table and site. While most tables have high mapping rates, poorer mapping rates are due to both complexities at the source and in the maturity of the CLAD mappings.

Conclusion

CLAD was able to successfully demonstrate national participant-mediated retrieval of EHR data for research via a HIN/HIE process. The CLAD C-CDA to OMOP code mapping solution significantly advances transforming C-CDA documents to OMOP through a multi-faceted, rule-based approach. The successful pilot offers a flexible and viable mechanism to convert C-CDA to OMOP. Future work will expand support for more C-CDA templates, integrate diverse data sources, and automate mapping/validation processes. Finally, as the TEFCA matures, CLAD will be well positioned to interoperate with other HINs to acquire EHR and other data for research leveraging existing clinical data exchange mechanisms used for care.

References

1. Behnel S. The lxml.etree Tutorial [Internet]. Lxml.de. 2025 [cited 2025 Jul 1]. Available from: https://lxml.de/tutorial.html
2. XML Path Language (XPath) [Internet]. W3.org. 2016 [cited 2025 Jun 30]. Available from: https://www.w3.org/TR/xpath-10/
3. Boone KW. The CDA TM book. Springer; 2011.
4. GoogleCloudPlatform. GitHub - GoogleCloudPlatform/healthcare-data-harmonization: This is an engine that converts data of one structure to another, based on a configuration file which describes how. There is an accompanying syntax to make writing mappings easier and more robust. [Internet]. GitHub. 2020 [cited 2025 Jun 30]. Available from: https://github.com/GoogleCloudPlatform/healthcare-data-harmonization
5. ASTP. Trusted Exchange Framework and Common Agreement (TEFCA) | HealthIT.gov [Internet]. www.healthit.gov. 2024. Available from:

https://www.healthit.gov/topic/interoperability/policy/trusted-exchange-framework-and-common-agreement-tefca
6. Consolidated CDA Overview | HealthIT.gov [Internet]. www.healthit.gov. Available from: https://www.healthit.gov/topic/standards-technology/consolidated-cda-overview
7. HealthIt.gov. Health information exchange [Internet]. Healthit.gov. 2019. Available from: https://www.healthit.gov/topic/health-it-and-health-information-exchange-basics/health-information-exchange
8. National Institutes of Health. All of us research program [Internet]. Nih.gov. 2019. Available from: https://allofus.nih.gov/

Acknowledgement

Supported by NIH All of Us Research Program Award No. OT2 OD036113-01. Please contact info@cladteam.io for future collaboration work.