CLAD Sourced EHR Data in *All of Us* (CDR v9)

Overview

CDR v9 includes a new stream of electronic health record (EHR) data sourced from Consolidated Clinical Document Architecture (C-CDA) documents. These records were acquired through the *All of Us* Center for Linkage and Acquisition of Data (CLAD) pilot, which used the eHealth Exchange health information network to retrieve participant-authorized records from Manifest MedEx, a California statewide health information exchange. This source returned C-CDA payloads that were mapped by CLAD to the OMOP Common Data Model and added by the *All of Us* Data and Research Center to the curated data repository alongside existing *All of Us* EHR data.

The pilot EHR data is complementary as it fills gaps in medications, visits, procedures, and diagnoses for participants who previously had none. This article describes where the data comes from, how it was transformed at a high level, and what researchers will find in the OMOP tables.

Where the Data Comes From

In February and March of 2026, the pilot sent record requests for participants who had a valid authorization to share their EHRs with *All of Us* and a California ZIP code, seeking data dating back to January 1, 2000. Requests were routed through the eHealth Exchange hub pursuant to each participant's HIPAA authorization, which is permitted under the network's trust framework. To our knowledge, the pilot was unprecedented in its scale of use of eHealth Exchange, the nation’s largest HIN, to broker participant-authorized medical record exchange for research. The network brokered the exchange but did not store the data; payloads were returned into CLAD's secure environment for mapping to OMOP.

Table 1 below summarizes what the two responding organizations returned.

Table 1. Records returned by the responding organization.

Responding Organization	Match Rate for Californian Participants	Participants With Records Processed by CLAD	Of Those, New to *All of Us* CDRv8
Manifest MedEx (CA statewide HIE)	66%	48,787	17,344

To avoid duplication with existing records, only participants with no other EHR data were added to CDRv9. Methods for aggregating data from multiple sources are currently being developed and will be included in future data releases. Merged with existing CDRv9 data, data from 15,371 participants (with no other EHR data, who were still consented) were added to CDRv9 with the source label “Participant Mediated EHR: CLAD”.

How the Data Was Transformed

C-CDA documents organize clinical data into XML sections such as problems, encounters, results, medications, and procedures. Each section was mapped into the OMOP Common Data Model. Mapping has two parts. Structural mapping decides which source fields populate which OMOP columns. Vocabulary mapping decides which source codes correspond to which standard OMOP concepts.

Vocabulary mapping follows the standard OHDSI approach used elsewhere in the program: source codes are resolved through the OMOP (Athena) vocabularies to obtain a standard concept_id and a target domain. The code's vocabulary determines which domain table it lands in, so a single C-CDA section can feed more than one domain. A procedures-section entry, for example, may become a procedure, a measurement, an observation, or a device exposure depending on the code it carries.

After mapping, a deduplication step keeps the most complete copy of each clinical event. The same event often appears in more than one document with different levels of detail. The pipeline groups duplicates and retains the row with the most populated fields, so completeness is preserved when sources overlap.

One scope note matters for interpretation: the pipeline reads structured, coded entries only. Free-text narrative blocks inside each section are not extracted.

What Researchers Will Find in the OMOP Tables

The C-CDA data populate the same OMOP domain tables researchers already use in the Researcher Workbench. Field definitions are unchanged from the standard OMOP model; refer to the CDR data dictionaries for column-level detail.

Records That Carry Two Codes

Some observation and procedure entries carry two codes: a primary code (for example, a SNOMED procedure) and a separate result or finding code (for example, a LOINC measurement). The primary code drives the standard concept_id and the domain assignment. The secondary code populates value_as_concept_id on the resulting row. Researchers reading value_as_concept_id should expect it to carry the finding or result, not the event itself.

Considerations for Working With This Data

- **Recency and source bias:** data retrieved through a HIN/HIE for care purposes tends to favor recent records and problem lists, and may be lighter on older encounters.
- **Variability across systems:** EHR vendors place the same coded value in different parts of a C-CDA document and differ in structure. The pipeline absorbs this, but the upstream variability is the main source of structural complexity in the source data.

- **Completeness and correctness vary:** these records were generated for clinical care, not research. Individual records differ in completeness. Known issues in the pilot data include missing vital-sign units and some missing values.
- **Complementary, not identical:** values do not always match across the OMOP files originally sent to the DRC, the FHIR records, and the C-CDA records. Visit dates, immunizations, and medications can differ in timing and occurrence. Treat the C-CDA stream as a complementary source that increases completeness, not as a duplicate of existing data.
- **Deduplication within the source:** when the same event appeared in multiple documents, only the most complete row was retained.

For OMOP table and field definitions, see the CDR data dictionaries. For background on *All of Us* EHR collection and transformation generally, see *Introduction to All of Us Electronic Health Record (EHR) Collection and Data Transformation Methods*.