# Creating a Standardized EHR Analytics Data Source for the National Cancer Institute's *Connect for Cancer Prevention Study*

**Edward A. Frankenberger[1], Jacob M. Peters[1], Nicole M. Gerlanc[1]**
[1]**National Cancer Institute (NCI), National Institutes of Health (NIH)**

## Background

The Connect for Cancer Prevention Study ("Connect") is a new, multi-site prospective cohort study enrolling up to 200,000 cancer-free adults receiving care within integrated U.S. healthcare systems. The cohort will be followed for 25+ years, during which time comprehensive data and biospecimens are continuously collected. These data will be used to enhance our understanding of cancer etiology and inform precision strategies for prevention and early detection[1]. The Connect Coordinating Center (CCC) manages electronic health record (EHR) data in the Observation Medical Outcomes Partnership (OMOP) Common Data Model (CDM) to enable harmonized, high-throughput epidemiologic analyses.

Each participating site independently develops and executes their own ETL processes, managing local OMOP databases and submitting data to the CCC in CSV or Parquet formats via Google Cloud Storage. To lower the barrier for sites to contribute, the CCC does not mandate specific OMOP CDM nor vocabulary versions and requires only a core set of tables (Clinical Data and Vocabulary Tables) most often populated in EHR-only OMOP instances.

Despite adopting a standardized data model, variations in CDM and vocabulary versions, site-specific code mappings, differing ETL strategies, and implementation errors have complicated the creation of a unified OMOP analytics database for Connect. To address these challenges, we developed an API-driven, file-centric, cloud-based pipeline designed to harmonize incoming OMOP data, ensuring consistent structural and semantic content across all contributing sites.

## Methods

The OMOP pipeline comprises discrete tasks orchestrated by Apache Airflow and executed on single-node instances, using DuckDB for data transformations. The pipeline source code (link) and Airflow directed acyclic graph (DAG) (link) are available on GitHub.

Each OMOP data file is processed individually, without cluster-based parallelism, through the following sequential tasks:

1. Conversion of CSV to Parquet, resolving formatting issues (non-escaped quotes, text encoding, etc.)
2. Schema validation against OMOP CDM standards.
3. Normalization of column types, handling missing/additional columns, and isolating non-conforming rows.
4. Structural upgrades from OMOP CDM version 5.3 to 5.4, as needed.
5. Vocabulary harmonization through concept remapping and resolving domain-table mismatches.
6. Derived data table generation (observation_period, condition_era, etc.)

During each step of the pipeline, detailed metadata artifacts are automatically generated to document processing outcomes. In the final steps, these metadata are aggregated into a comprehensive data delivery report that summarizes aspects of the data delivery, including site-specific table completeness, row counts, column-level discrepancies, vocabulary version usage, and schema compliance results. Harmonized data are ultimately loaded into Google BigQuery, forming a standardized analytics database accessible to analysts.

Vocabulary harmonization posed particular challenges due to:
- Domain discrepancies between site and target vocabulary versions.
- Source-to-target concept mapping updates.
- Deprecated standard codes in newer vocabulary versions.
- Missing source_concept_id values in provided OMOP tables.
- Domain-table mismatches (e.g., condition concepts in measurement table).

Building upon prior OMOP-to-OMOP ETL methods by Blacketer et al.[2], we expanded this process to update concept identifiers and appropriately relocate data points to the correct table. Figure 1 illustrates the logic underlying vocabulary harmonization.

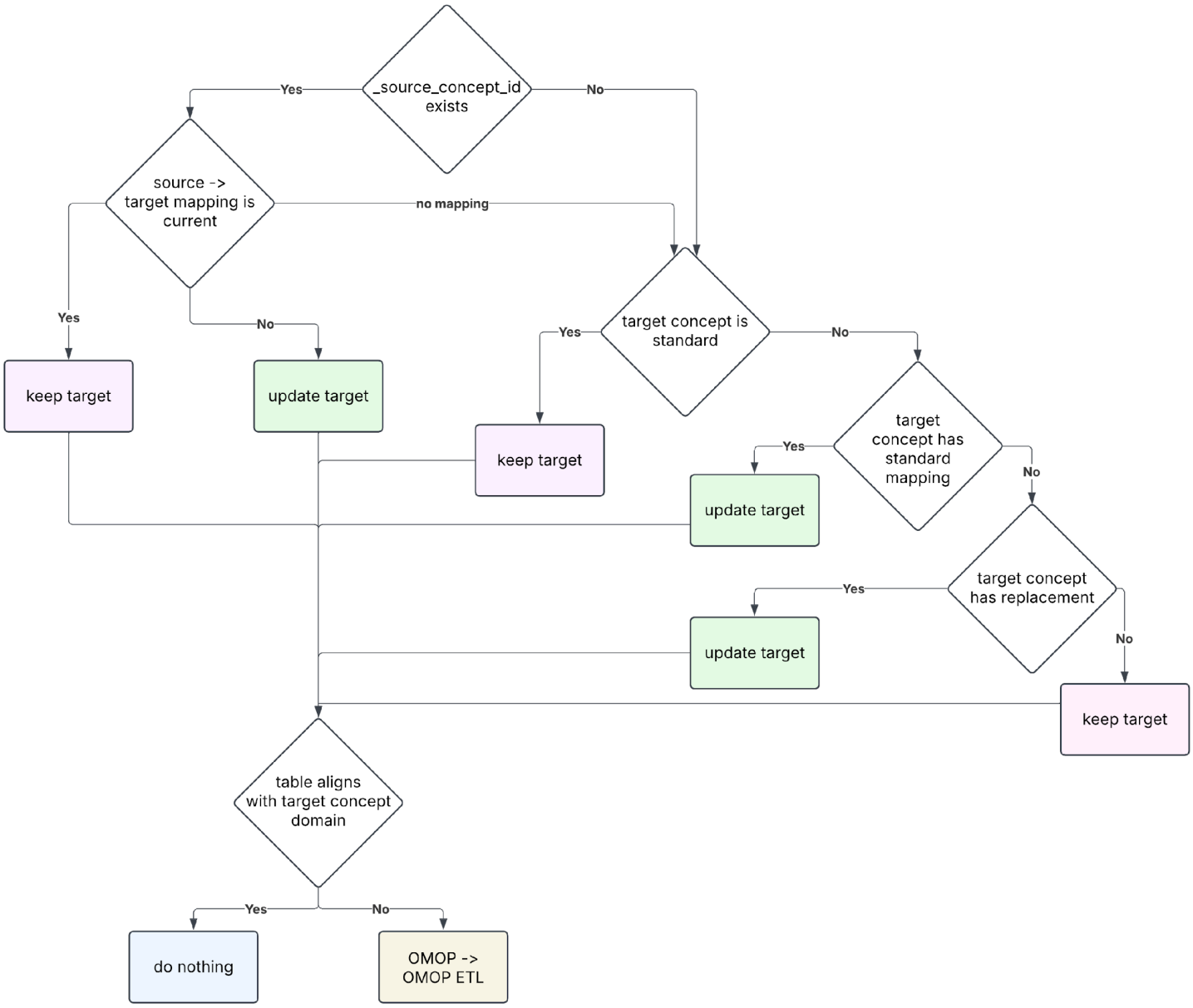


**Figure 1. Vocabulary harmonization decision tree**

**Results**

Our OMOP harmonization pipeline has been successfully deployed within a Google Cloud environment. With moderate computing resources, the pipeline processes a 100 GB site delivery in approximately 30 minutes.

To date, the CCC has received initial OMOP data deliveries from five participating sites, encompassing nearly 30,000 study participants and containing over 240 million rows from a cumulative total of 93 OMOP tables across all submissions. On average, each site delivered 52% (range: 33–72%) of the tables expected within a complete CDM instance.

Four sites provided data in CDM version 5.3, while one site used version 5.4. The pipeline automatically upgraded all incoming CDM v5.3 files to v5.4, ensuring structural consistency. Sites utilized a diverse range of vocabulary versions, from v20210402 to v20250227. The pipeline's vocabulary harmonization process effectively reconciled these differences without manual intervention. Notably, no rows violated data type constraints, allowing user access to all received data.

Furthermore, each participating site exhibited some level of table-domain mismatch, irrespective of vocabulary version. Addressing this challenge, the pipeline's vocabulary harmonization logic explicitly manages these discrepancies, ensuring semantic consistency across all data deliveries.

**Conclusion**

Our work demonstrates that a unified, standardized OMOP analytics database can be successfully created even without enforcing uniform CDM or vocabulary standards across contributing sites. This flexibility significantly enhances the scalability and applicability of large-scale, multi-site studies, providing a robust framework for harmonized data analysis in diverse research settings.

**References**

1. *About the Study*. National Cancer Institute Connect for Cancer Prevention Study. 1 July 2025. https://www.cancer.gov/connect-prevention-study/about

2. Blacketer C, Ivanov A, Burrows E, Dymshyts D, DeFalco F. *ETIng from your OMOP CDM to your OMOP CDM? An efficient solution to vocabulary migration* [abstract]. 2024 Global OHDSI Symposium. 2024 Oct.