Process for Expediting ETL to the OMOP Common Data Model

Richard Starr[1], Myung Choi[2], Michael Riley[2], Jon Duke[2]
Georgia Institute of Technology[1] and Georgia Tech Research Institute[2] , Atlanta GA

ABSTRACT: *The process of implementing the OMOP CDM as a new user of the OHDSI platform or as an existing user needing to add a new data source in a different format can be a challenging and time consuming task. We have created an open source ETL software package[1] that simplifies the task of importing new datasets into the OMOP CDM. This software allows the complex OMOP transformation logic to be abstracted from the ETL creation. This reduces the possibility of errors in the transformation by standardizing the OMOP logic. The software is configurable to run on large batches of data or on single records.*

Background

The OMOP Common Data Model provides a standardized vocabulary and data model to allow collaborative research to occur across different health data sources. There is a significant learning curve when undertaking the transformation of an existing data source to the OMOP CDM. While there are publicly available code snippets and example ETLs, these are difficult to implement without a detailed understanding of the vocabulary and data model.

Methods

As we work with multiple data providers using different data formats, we implemented an ETL system that uses staging tables to normalize the disparate incoming data formats into a standardized generic format. The OMOP specific transformations occur on datasets from this standardized format. Any dataset to be ingested only has to be transformed to meet this simple generic format. This allows new dataset formats to be ingested into the OMOP CDM with a minimum of effort. This way the complex OMOP transformations are abstracted from the process of creating the ETL for each new dataset.

This approach has been applied in several use cases:

- Bulk import of Claims and EHR records for longitudinal datasets
- A real-time backend process to ingest data for the GT OMOP-on-FHIR project[2]
- Nightly batch processing to update a Clinical Data Warehouse from a EHRs

ETL Staging Tables

The ETL Staging tables are designed to be similar to potential source datasets. The most common data sources are from Claims records or EHR systems. The internal patient identifiers, visit identifiers, etc. are used as the OMOP source values to link the staging records for the transformations. It is preferable for the coding systems for the source values to be explicitly defined, but if this is not available, the OMOP transformation will assume common systems.

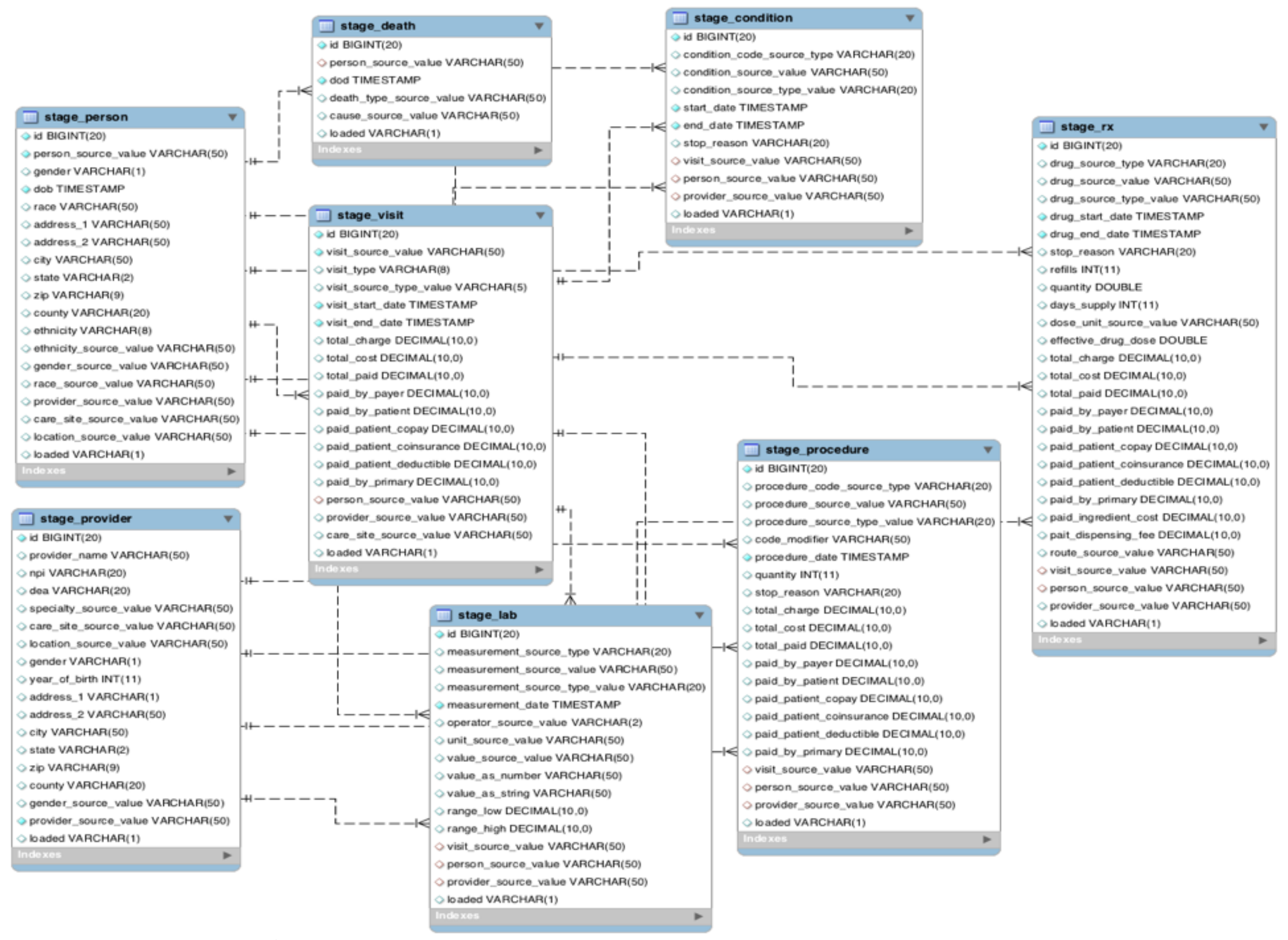

Figure 1. ETL Staging Tables

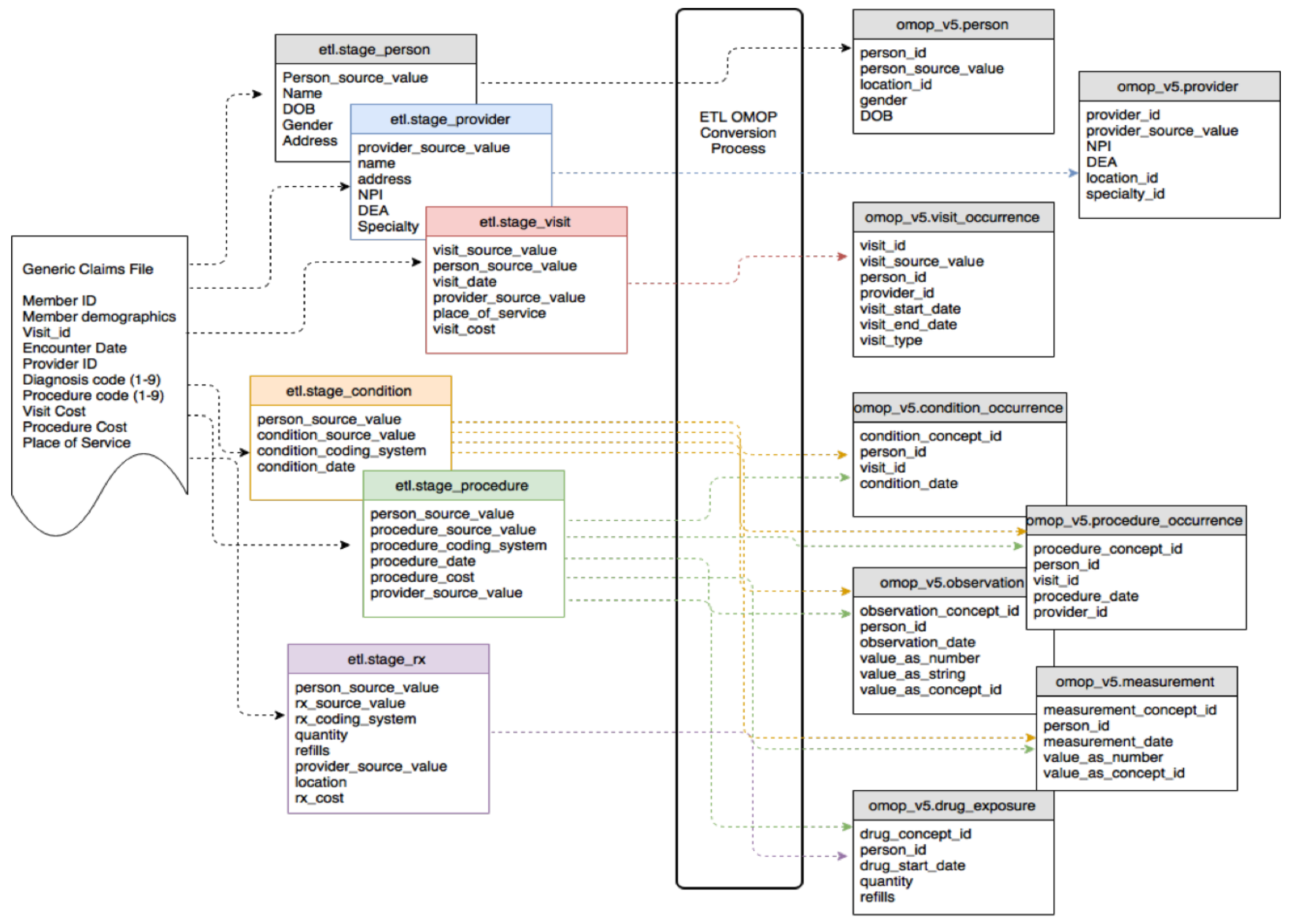

Figure 2. Example of ETL from Claims to OMOP CDM

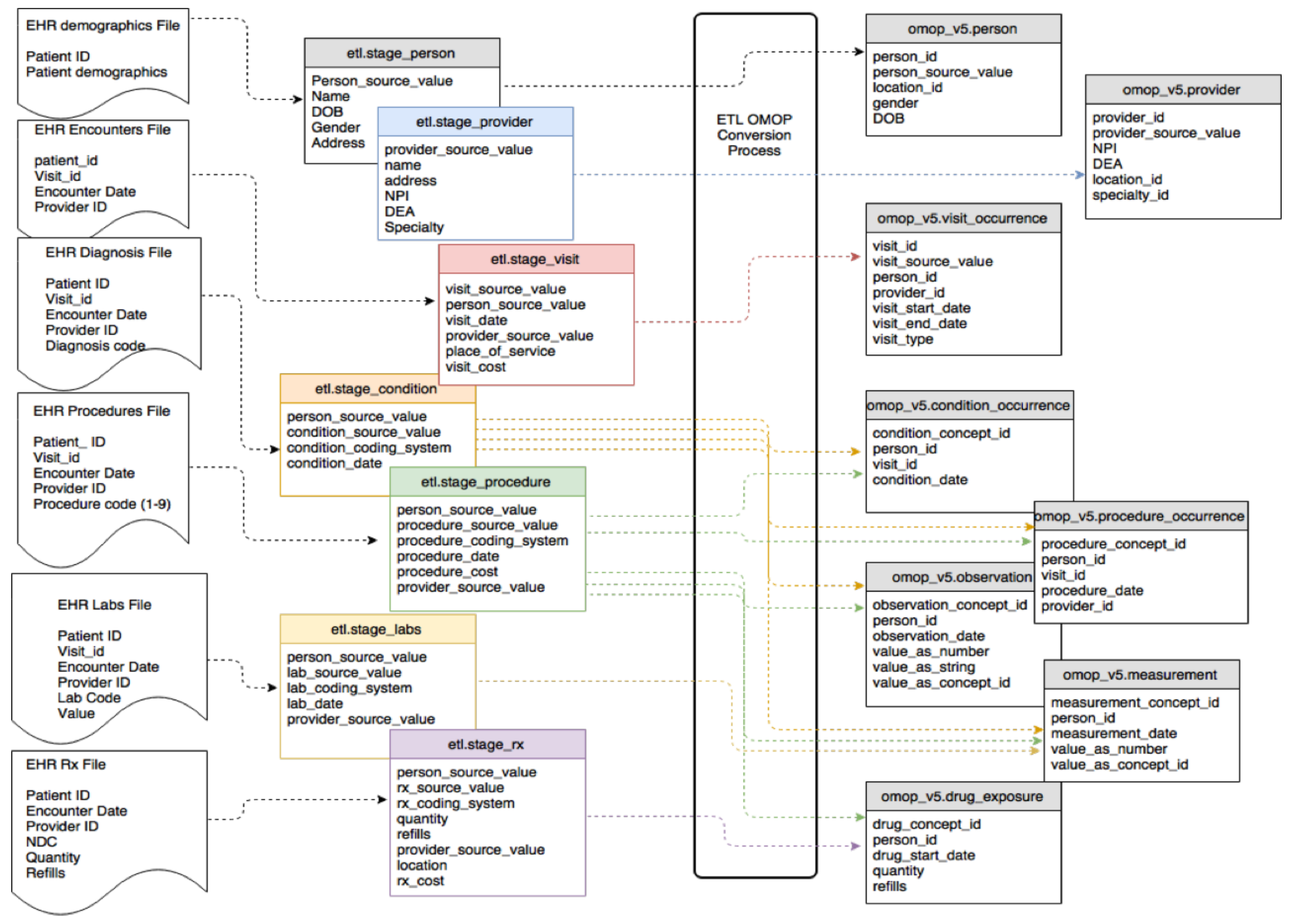

Figure 3. Example of ETL from EHR to OMOP CDM

Transformation to OMOP CDM

The logic to transform data from staging tables to the OMOP CDM is encapsulated in database stored procedures. These procedures can operate as a bulk loading process or as a single record (patient) process. The conversion process contains the logic to determine the proper destination table and links the concept ids from the CDM definitions.

One of the difficulties of a transformation into the CDM is that a single source data type can create records in different CDM tables. For example, a CPT4 procedure code may get transformed into the Procedure_Occurrence, Measurement, Observation or Drug_era tables. Having the intermediate staging tables unaware of this complication allows faster and more accurate implementation of new datasets into the CDM.

The transformation processes use additional auditing columns in the staging and OMOP tables to track the progress of the transformation. These provide metrics and error analysis for the dataset processing. Additionally, the auditing columns and process flow through system ensure that the original source data is preserved to allow Quality Control process to retrospectively inspect the data transformations for accuracy.

Conclusions

Separating the work of mapping an incoming dataset from the transformation into the OMOP CDM allows for faster integration of new dataset formats. Having a standardized OMOP transformation process from the staging tables into the CDM enables the logic to be independent of the individual datasets. This will help to create consistent and reproducible implementations of the OMOP CDM. And for new implementers of the OHDSI stack, it can be a kickstarter for the transformation of their existing dataset to the OMOP CDM.

References

1. Starr R., Generic OMOP ETL system used for raw datafile and FHIR ingestion, GitHub repository, https://github.com/gt-health/omop_etl_public
2. Choi M., Starr R., Braunstein M, Duke J. OHDSI on FHIR Platform Development with OMOP CDM mapping to FHIR Resources. Poster Session, OHDSI Symposium. 2017

Contact: richard.starr@gatech.edu