The Extract-Transform-Load Lessons for Loading Neonatal Healthcare Data to the OMOP-CDM

Bei Li, MS [1,2], Sifei Han, PhD[1], Lingyun Shi, MS[1], Lezhou Wu, PhD[1], Fuchiang (Rich) Tsui, PhD[1]

[1]Tsui Laboratory, Department of Biomedical and Health Informatics, Children's Hospital of Philadelphia, Philadelphia, PA, USA

[2]Department of Biomedical Informatics, Life Science School, Central South University, Changsha, Hunan Province, China

Abstract

The multi-center research collaboration faces a challenge: how to efficiently transform massive data from different sources into one general data model. The Observational Medical Outcomes Partnership Common Data Model (OMOP-CDM) can be a valuable solution. We performed an extract-transform-load (ETL) process that converted data from a regional women's hospital and a county department of human services into the OMOP-CDM. A key technical issue in the ETL process is the conversion from the event-centric clinical data to the person-centric CDM. With the extensive assessment of the validity, this ETL process will enable a standardized, agile, and accurate data transfer and integration across collaborative organizations. The effort is critical to facilitate the data share in the neonatal health research as the infant mortality continues to be a significant public health problem in the U.S.

Research Category (please highlight or circle which category best describes your research)

Observational data management, clinical characterization, population-level estimation, patient-level prediction, other (if other, please indicate)

Introduction

Infant mortality continues to be a significant public health problem in the U.S. [1] Based on CDC infant mortality rate statistics, Between 2003 and 2015, 1,223 infants died within their first year of life in Allegheny County, Pennsylvania, with the average mortality rate of 7.12 per 1,000 live births, which is higher than the United States' infant mortality average, 6.37 deaths per 1000 birth.[2] To solve this problem, it demands neonatal research collaboration and timely data share across multiple hospitals and organizations. However, the multi-center collaboration poses a challenge: how to efficiently transform massive data from different sources into one general data model. This study aims to build an Observational Medical Outcomes Partnership Common Data Model (OMOP-CDM) that standardizes complex, heterogeneous, and large multi-source data via an extract-transform-load (ETL) process.

Method

Data sources

We collected data from two sources: (1) clinical data from the UPMC Magee Obstetric Medical and Infant (MOMI) database, and (2) the death and birth certificates, and social demographics from the Allegheny County Department of Human Services (DHS). The study period is from January 1, 2003 to December 31, 2014. The MOMI data are curated maternal and neonatal electronic health records including 176 variables for all deliveries at the Magee Women's Hospital in Pittsburgh, PA. The honest broker performed data linkage between the MOMI and DHS data.

ETL process

To promote the data share between institutes, we built an OMOP-CDM for the linked database between MOMI and DHS using an ETL process. The study datasets include birth certificate, death certificate, delivery, personal alias, diagnosis, drug, encounter, laboratory, pharmacy, procedure, and etc. The software WhiteRabbit was used for the ETL preparation and the ETL process was assessed for its validity and efficiency. The designed ETL steps were described as follows.

1. The WhiteRabbit scans the local data tables, fields, and content, and then creates a detailed report containing necessary information on the tables, fields, and values in a field.
2. The function, Rabbit-in-a-Hat, read the scanned documents and generates new documentation for the ETL process.
3. To design and write the rule that maps the clinical data to the OMOP-CDM. The Structured Query Language (SQL) was used to connect different tables extracting the mapping from non-standard source codes to standard source codes. The patient's source records are modified and stored in a temporary data structure, and then are transformed into the related data fields in the CDM. Python scripts and SQL queries are developed for the ETL process.

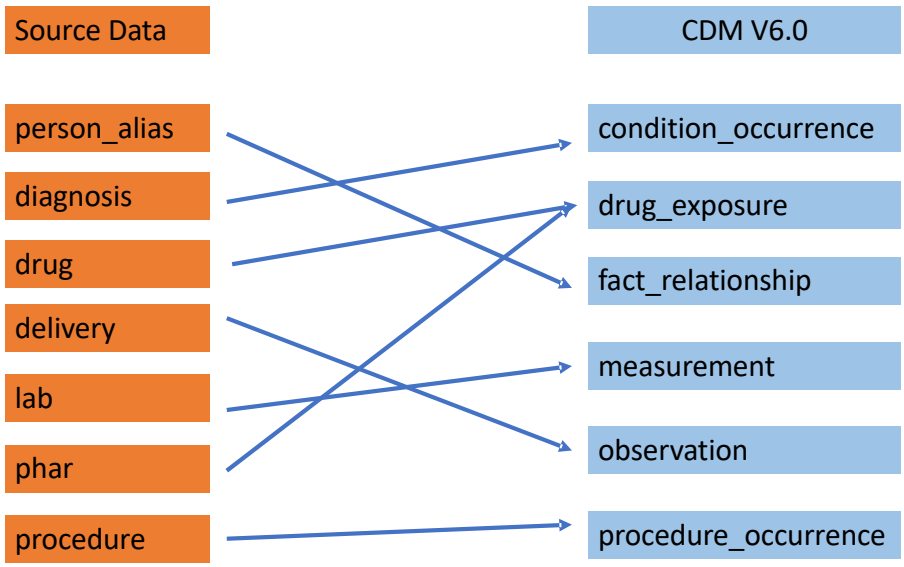

Figure 1. Data mapping between the source tables, from the UPMC Magee Obstetric Medical and Infant and the Allegheny County Department of Human Services, to the Observational Medical Outcomes Partnership Common Data Model (OMOP-CDM) version 6.0.

Results

Six standardized data tables in the OMOP-CDM can be mapped from the source data as shown in the Figure 1. The clinical data from UPMC MOMI database are event-centric (i.e., hospital encounter event or delivery event) while CDM is a person-centric model. Thus, one key technical issue in the ETL process is the conversion from the event-centric to the person-centric model. Figure 2 illustrates a simplified mapping logic indicating such conversion.

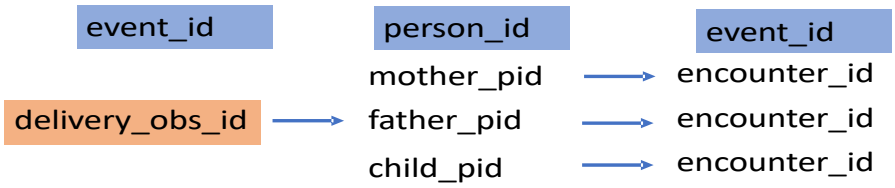

Figure 2. Mapping logic between event-centric model and person-centric model

Conclusion

A well-written ETL process can convert multi-source healthcare data into the OMOP-CDM. The conversion between event-centric clinical data to the person-centric CDM is critical. The effort of building an OMOP-CDM using ETL will enable a standardized, agile, and accurate data transfer and integration across multiple collaborative organizations.

Reference

1. Jacob JA. US infant mortality rate declines but still exceeds other developed countries. *JAMA - J Am Med Assoc*. 2016;315(5):451-452. doi:10.1001/jama.2015.18886
2. CDC. CDC Wonder. https://wonder.cdc.gov/lbd-current.html. Accessed November 15, 2018.

Email

Bei Li, libeiwu@126.com, LIB2@EMAIL.CHOP.EDU

Sifei Han, HANS2@EMAIL.CHOP.EDU

Lingyun Shi, SHIL2@EMAIL.CHOP.EDU

Lezhou Wu, WUL5@EMAIL.CHOP.EDU

Fuchiang(Rich) Tsui, TSUIF@EMAIL.CHOP.EDU