Sharing is Caring – A Checklist to Facilitate ETL Dissemination and Replication

Clair Blacketer, MPH[1,2,3], Erica A. Voss, MPH[1,2,3]

[1]Janssen Research and Development, Titusville, NJ; [2]OHDSI collaborators, Observational Health Data Sciences and Informatics (OHDSI), New York, NY, [3]Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, Netherlands

mblacke@its.jnj.com; evoss3@its.jnj.com

Observational data management

Introduction

Based on recent estimates about 2 billion patient records in over 118 databases from around the world have been converted to the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) (1). Many studies using these datasets post results as well as the code used to generate them in an effort to be transparent and to allow the research community to fully replicate their methods (2-5). While this has long been the practice of the Observational Health Data Science and Informatics (OHDSI) initiative it is less common for collaborators to share the logic employed to convert the datasets that support these studies to the CDM. Recent retractions have brought to light the importance of data provenance, of which data transformation is a major component (6, 7). We aim to provide a checklist of items either created or utilized during the extract, transform, and load (ETL) process and that are necessary to fully describe the conversion of a dataset to the OMOP CDM. It is our hope that this list can provide guidance to those seeking a best practice approach for how to transparently share ETL logic.

ETL Checklist

ETL items that are produced or utilized during transformation and any OHDSI tools that support them are listed in Table 1 below with justification for each item.

ETL Component	OHDSI Tool	Discussion and Examples
Any documents that describe the native data, like database description or data dictionary		This information is critical to understand the version of the native data used at the time of ETL development. **Example**: The US National Health and Nutrition Examination Survey (NHANES) lists very detailed information about the tables and fields because the column names and values are coded in such a way that they are unusable without the documentation.
ETL specification document	RabbitInAHat	The plain language description of the logic used to convert the native data to the CDM allows reviewers to understand the choices and assumptions made and helps a replicator write code using the same methods as the original. **Example:** The way the OBSERVATION_PERIOD table is constructed is critical to research and can differ dataset to dataset. • Clinical Practice Research Datalink (CPRD) (spans from the first to last time a person is seen in the in a practice that is up-to-standard) • Optum® De-Identified Clinformatics® Data Mart Database (uses an insurance eligibility file to determine duration within the data)
Locally defined vocabulary	Usagi	To replicate the ETL build without information on decisions made around source code mappings to standard vocabularies is very difficult. Vocabulary is extremely important when defining analyses and so any replication must use the same vocabulary techniques. **Example**: Physician specialty codes are not usually coded with standard codes in source data and Usagi is used to map them to standardized vocabulary. Any analysis using specialty would need to refer to this mapping to be sure that specialties are appropriately mapped.

ETL Component	OHDSI Tool	Discussion and Examples
Test cases (R package)	RabbitInAHat	These unit test cases test the logic detailed in the document and show not only that the build is performing correctly but helps a replicator both understand and test the logic used in the original. **Example**: Institution A and Institution B have both developed a code set for transforming the Pharmetrics Plus database to the CDM. They cannot share the code due to differing infrastructures, yet the test cases are transferrable in that they use standardized, open-source technology.
Data quality process	Data Quality Dashboard	A file listing all data quality checks that were run and their pass/fail rate. This file is created based on a priori knowledge of the source and runs relevant data quality checks with pass/fail thresholds that take the content of the source data into account. A reviewer may want to see the choices and reasoning made in this file. **Example**: A dataset does not have race information so the failure threshold is set to 100% with a note indicating the rationale for this decision.
Source data description	WhiteRabbit	A file that gives an overview of the tables and columns in the source data at the time of ETL. **Example**: Data available for purchase is often versioned, including changes to column names. The scan report shows exactly the tables and columns in the dataset at the time of transformation.
ETL code base		This is the full code used to build the CDM from the native data. **Example**: The Synthea synthetic dataset and CDM transformation was built using entirely using open-source tools and anyone can download both the data and code to convert it.
CDM characterization	Achilles	These results characterize demographics and concepts in a CDM build. Due to restrictions on low cell counts these may not be able to be shared. **Example**: If analyst 2 runs a study developed by analyst 1 using the same data and they achieve radically different results the Achilles characterization can elucidate differences between the builds and potentially pinpoint the cause.

Table 1: ETL items essential for review and replication.

Conclusion

It is crucial to provide a standard set of documentation concerning the extract, transform, and load process used to convert a dataset from its native format to the OMOP Common Data Model. This transparency enables replication and increases reviewer trust in a dataset which is especially important in an era where retractions due to data provenance are threatening to undermine the credible (and incredible) work produced by the observational science community. The checklist provided here describes all items both utilized by and created during ETL and can be regarded as a best practice approach for what information should be shared about the process.

References

1. OHDSI Data Network: OHDSI; 2019 [Available from: https://www.ohdsi.org/web/wiki/doku.php?id=resources:2019_data_network.

2. Burn E, You SC, Sena A, Kostka K, Abedtash H, Abrahao MTF, et al. Deep phenotyping of 34,128 patients hospitalised with COVID-19 and a comparison with 81,596 influenza patients in America, Europe and Asia: an international network study. medRxiv. 2020:2020.04.22.20074336.

3. Morales DR, Conover MM, You SC, Pratt N, Kostka K, Duarte Salles T, et al. Renin-angiotensin system blockers and susceptibility to COVID-19: a multinational open science cohort study. medRxiv. 2020:2020.06.11.20125849.

4. Reps JM, Kim C, Williams RD, Markus AF, Yang C, Salles TD, et al. Can we trust the prediction model? Demonstrating the importance of external validation by investigating the COVID-19 Vulnerability (C-19) Index across an international network of observational healthcare datasets. medRxiv. 2020:2020.06.15.20130328.

5. Williams RD, Markus AF, Yang C, Duarte Salles T, Falconer T, Jonnagaddala J, et al. Seek COVER: Development and validation of a personalized risk calculator for COVID-19 outcomes in an international network. medRxiv. 2020:2020.05.26.20112649.

6. Mehra MR, Desai SS, Kuy S, Henry TD, Patel AN. Cardiovascular Disease, Drug Therapy, and Mortality in Covid-19. New England Journal of Medicine. 2020;382(25):e102.

7. Mehra MR, Desai SS, Ruschitzka F, Patel AN. RETRACTED: Hydroxychloroquine or chloroquine with or without a macrolide for treatment of COVID-19: a multinational registry analysis. The Lancet.