From OMOP to CDISC SDTM: Successes, Challenges, and Future Opportunities of using EHR Data for Drug Repurposing in COVID-19

Wesley Anderson[1], Ruth Kurtycz, Tahsin Farid[2], Shermarke Hassan[3], Kalynn Kennon[3], Pam Dasher[1], Danielle Boyce[4], Will Roddy[1], Smith F. Heavner[1,5]
[1]CURE Drug Repurposing Collaboratory, Critical Path Institute, [2]U.S. Food and Drug Administration, [3]Infectious Disease Data Observatory, [4]Johns Hopkins University, [5]Department of Public Health Sciences, Clemson University

Background

Real-World Evidence (RWE) generated from the analysis of Real-World Data (RWD) is a vital source of information regarding the effectiveness and efficacy of treatments and therapeutics. It can be pivotal in gaining regulatory approvals of new indications for a drug that has already been approved (e.g., drug repurposing). While regulatory agencies such as the FDA have published guidelines on utilizing RWD in observational settings, gathering data from multiple healthcare institutions to analyze this RWD is often difficult because of differences in proprietary models within and between these institutions.

The CURE Drug Repurposing Collaboratory (CDRC) is a public-private partnership with the U.S. Food and Drug Administration (FDA) and the National Center for Advancing Translational Sciences (NCATS). It has brought together key partners including the Society of Critical Care Medicine (SCCM) and the Infectious Disease Data Observatory (IDDO) to share their expertise in automated data extraction and standardization from real-world data sources such as Electronic Health Records (EHRs). This partnership has led to the development of the Edge Tool suite[1,2] - a group of public and free Docker container deployments of open-source software projects from the OHDSI stack. The Edge Tool enables the conversion of relevant data stored in proprietary models to the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM). The use of the docker container is especially important in that it is able to support the use of these tools at sites with lower technical resources. With COVID-19 as a use-case, the Edge Tool suite was previously able to reduce the number of hours necessary to extract, transform, and load (ETL) structured components of the medical record from an alpha site into the OMOP CDM from over 2,000 hours to fewer than 90 hours[2]. Although the OMOP CDM is sufficient in its ability to standardize and harmonize data from various institutions, it is not in an accepted CDM for regulatory submissions.

The Clinical Data Interchange Standards Consortium's (CDISC) Study Data Tabulation Model (SDTM) is an FDA supported data standard which data must adhere to before such submissions can take place. Therefore, it is necessary to map the data extracted with the Edge Tool suite into the CDISC SDTM format in settings where such a submission is desired. There is also a need for collaboration on observational data through converting CDISC standards to other existing standards[3]. Previously, there has been work on the conversion of data from SDTM to OMOP[4,5] but less has been done to convert data in OMOP to SDTM, although there have been various efforts in this space[6,7]. This study partners the findings in Roddy et. al.[4] to offer valuable perspectives on the data model conversion process. Moreover, it offers a longitudinal dataset available in both OMOP CDM format at SCCM and CDISC SDTM format at IDDO, both of which are available upon request to the respective groups. This unique configuration allows for the opportunity to do a comparative analysis on the relative utility between datasets in their respective formats along with a documented crosswalk between OMOP and SDTM; this may enable determination of the relative utility of the two formats for different purposes, while also increasing the interoperability between data models, and by extension, data integration opportunities between clinical trial and observational data. The goal of this project was to map critical data variables from the OMOP CDM into CDISC SDTM and construct a limited, cross-sectional analysis dataset.

Methods

The data mapped into CDISC SDTM from the OMOP CDM were from a pilot healthcare site which utilized the Edge Tool to ETL their COVID-19 data into the OMOP CDM. After the ETL, this healthcare site transferred over 12,000 patients' data. Through the use of the OHDSI tools, including Athena and ACHILLES, along with the OHDSI ICD-10 code standardized vocabularies, derivation logic was generated to map a total of 21 concepts from the pilot dataset to a subset of variables (Figure 1) and their SDTM counterparts. This logic included lists of OMOP CDM concept ids related to each variable, as well as detailed logic on the timing and calculation of the different variables (e.g., specifying the maximum serum creatinine level recorded for a patient in the first 48 hours of hospitalization as the measurement of interest).

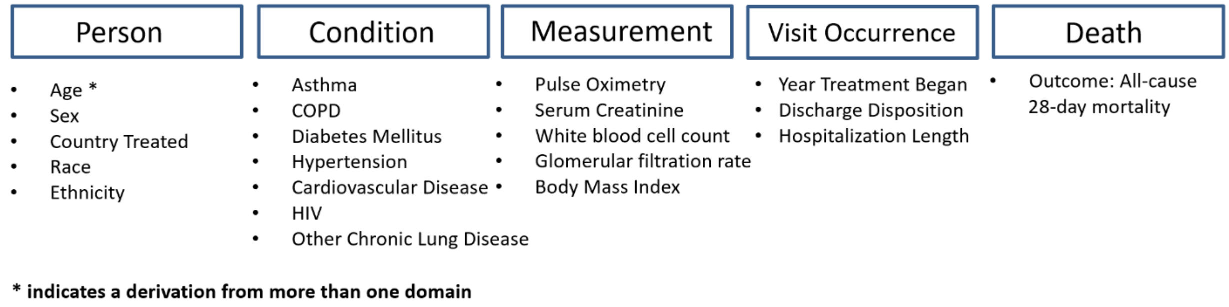


Figure 1. Open-source variable list that will be mapped from OMOP to SDTM.

A few examples of a mapping between OMOP and SDTM are shown in Table 1.

Table 1: Example mapping between OMOP and CDISC concepts.

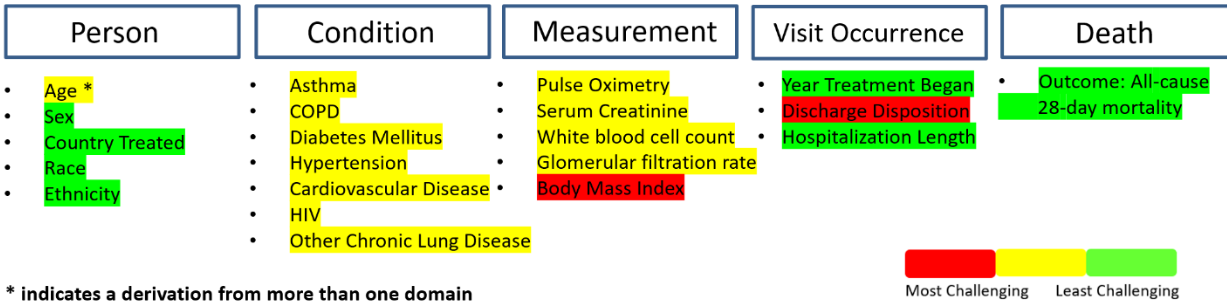
Variable	OMOP Concept ID	OMOP Domain/Column	SDTM CT Submission Value	SDTM C-Code	SDTM Domain/Column
Age*	N/A	person.birth_datetime	N/A	N/A	dm.age
Sex	8532	person.gender_concept_id	M	C20197	dm.sex
Sex	8507	person.gender_concept_id	F	C16576	dm.sex

* indicates a derived element

Results

Concepts related to the 21 variables of interest were successfully mapped to CDISC SDTM. Figure 2 stratifies the concepts by categorical level of difficulty, with green being the most straightforward concepts to map, yellow being moderately difficult, and red being the most difficult.

Several factors influenced the difficulty level, including missing data from the pilot site, inconsistency between OMOP and SDTM concepts, and complex variable derivation. BMI presented challenges due to high missingness in weight and height, inconsistent measurement units from the pilot site, and a complex derivation process. Discharge disposition posed challenges due to coding schema misalignment, resulting in limited exact matching between the two data models. An absence of discharge disposition recording in clinical trials likely hinders CDISC's RWE-forward mapping, which underscores the necessity for harmonization to utilize CDISC SDTM for RWE beyond the scope of clinical trials.



*** indicates a derivation from more than one domain**

Figure 2. Categorical measures of difficulty in mapping the different variables between OMOP and CDISC.

Note that due to missing variables during data transfer, some originally intended mappings (e.g., WHO Ordinal Scale Score) could not be derived. Also of note is that updates will be necessary for comorbidity concepts to accommodate future healthcare sites' additional conditions or concepts. Through this work, all patient records were successfully made available in CDISC SDTM format through IDDO.

Conclusion

This work shows the feasibility and success of mapping data in an OMOP CDM to the CDISC SDTM model. Further improvement of the process will be evaluated with the Usagi tool in the future. Our long-term goal is to increase the size of the publicly available COVID-19 dataset through the process outlined here and in previous works by Heavner et. al. [1,2]. We would also like to expand this work outside of COVID-19 and into other critical care and rare disease spaces.

References

1. Heavner SF, Anderson W, Kashyap R, Dasher P, Mathé EA, Merson L, Guerin PJ, Weaver J, Robinson M, Schito M, Kumar VK, Nagy P. A Path to Real-World Evidence in Critical Care Using Open-Source Data Harmonization Tools. Crit Care Explor. 2023 Apr 3;5(4):e0893. 10.1097/CCE.0000000000000893
2. Heavner SF, Llano T, Schito M, Stone H, Dasher P, Russel T, Kumar V, Saeks B, Cooke M, Kashyap R, Robinson M, Nagy P. Lowering the OMOP ETL Barrier for Clinical Registries. OHDSI Symposium, 2022.
3. Facile R, Muhlbradt EE, Gong M, Li Q, Popat V, Pétavy F, Cornet R, Ruan Y, Koide D, Saito TI, Hume S, Rockhold F, Bao W, Dubman S, Jauregui Wurst B. Use of Clinical Data Interchange Standards Consortium (CDISC) Standards for Real-world Data: Expert Perspectives From a Qualitative Delphi Survey. JMIR Med Inform. 2022 Jan 27;10(1):e30363. 10.2196/30363.
4. William T. Roddy, Daniel Olson, Diane Corey, Ian Braun, Terrence McHugh, Emily Hartley, Smith Heavner, Ramona Walls. Translating SDTM Terminology and OMOP Vocabularies using the UMLS. Poster presented at: *AMIA 2022 Annual Symposium*; November 5-9, 2022; Washington, DC.
5. J. Ransom, E. Allakhverdiiev, G. Klebanov, J. Singer, K. Eitvid, R. Ahmed, E. Rusli, A. Shilnikova. Leveraging the OMOP CDMv5 for CDISC SDTM RCT Data. OHDSI Symposium, 2018.

6. COVID-19 Real World Data (RWD) Elements Harmonization Project. Food and Drug Administration Web site. https://www.fda.gov/drugs/coronavirus-covid-19-drugs/covid-19-real-world-data-rwd-data-elements-harmonization-project. Accessed June 14, 2007.
7. Shanbhogue, AY. RWD (OMOP) to SDTM (CDISC): A primer for your ETL journey. PharmaSUG, RWD-106, 2022.