Translating medical vocabularies via the Unified Medical Language System

William T. Roddy[1], Daniel Olson,[1], Diane Corey[1], Ian Braun[1], Terrence R. McHugh[1], Emily Hartley[1], Smith Heavner[1], Ramona L. Walls[1]

[1] *Critical Path Institute, Tucson, AZ*

Keywords
data management, OMOP, SDTM, CDISC, UMLS, terminology, vocabulary

1. Background

The Rare Disease Cures Accelerator – Data and Analytics Platform (RDCA-DAP) of the Critical Path Institute (C-Path) is an FDA-funded effort to facilitate drug development for rare diseases (https://portal.rdca.c-path.org/). RDCA-DAP helps researchers leverage existing knowledge and analyze data to inform and optimize clinical trial design with new sources of evidence. The platform supports the use of data to improve the quantitative characterization of rare disease progression, define novel biomarkers and endpoints, and provides analytical tools to inform the design of innovative trial protocols.

One of the key deliverables is the creation of a knowledge graph from the natural history, registry and clinical trial data received. The data, however, must be cleaned and standardized prior to knowledge graph ingestion which has presented us with opportunities to implement novel (to our organization) automation procedures of certain data management activities (e.g. vocabulary mappings).

Recent advances in clinical research data standards have resulted in the development of several Common Data Models (CDMs) which are leveraged to support the sharing and reuse of data[1]. Critical Path Institute has chosen to map legacy data to the Observational Medical Outcomes Partnership (OMOP) model prior to data integration. The relevant concepts within the integrated data will be further mapped to OBO ontologies (https://obofoundry.org/) prior to knowledge graph ingestion (which is outside the scope of this submission). The OMOP standardized vocabulary includes many biomedical terminologies that enable standardization of source data; however, these terminologies do not currently include the Study Data Tabulation Model (SDTM) controlled terminology which is frequently used for submissions to regulatory authorities. Generating mappings between the SDTM terminology and the OMOP standardized vocabulary will further expand the capabilities of data sharing and reuse between real-world data sources and clinical trial data sources. We demonstrate an implementation of translating the terminology used in these two Common Data Models.

1.1. Methods

We have developed mappings between SDTM terminology and the OMOP standardized vocabulary by using the Unified Medical Language System (UMLS). The SDTM terminology is published by the National Cancer Institute Enterprise Vocabulary Services (NCI

ICBO 2022, September 25-28, 2022, Ann Arbor, MI, USA
EMAIL: wroddy@c-path.org (A. 1); dolson@c-path.org (A. 2); dcorey@c-path.org (A. 3) ibraun@c-path.org (A. 4) terrence.r.mchugh@gmail.com (A.5) ehartley@c-path.org (A.6) sheavner@c-path.org (A. 7) rwalls@c-path.org (A. 8)

ORCID: 0000-0002-8453-520X (A. 1); 0000-0002-8134-1207 (A. 2); 0000-0003-3840-2315 (A. 3) 0000-0002-2389-9288 (A. 4) 0000-0002-3805-0359 (A.5) X (A.6) 0000-0003-0912-0407 (A. 7) 0000-0001-8815-0078 (A. 8)

EVS), where concepts are identified by concept codes (C-Codes) which are included within the UMLS. We identified all UMLS Concept Unique Identifiers (CUIs) by searching for atoms (the smallest unit of naming in a source) with a source abbreviation of NCI and a source code containing the C-Code in the SDTM terminology. The UMLS CUIs associated with the SDTM terminology were used to retrieve source codes originating from terminologies that are included within the OMOP vocabulary. We primarily focused on the SNOMED-CT, LOINC, RxNorm, and UCUM vocabularies because they are standard within the OMOP vocabulary, but we also expanded the search to MedDRA and MeSH due to their appreciable representation in SDTM and OMOP. The source codes from these vocabularies were used to identify the equivalent OMOP vocabulary standard concept.

1.1.1. Status of Mapping Results

The 2021-12-17 release of the SDTM terminology included 22,132 unique C-Codes. Of these, 84.2% were available within the UMLS 2021AB release and 57.4% were only indexed in an NCI EVS terminology or the Metathesaurus vocabulary. Within the UMLS searched vocabularies 25.5% of the C-Codes were present and 1.3% were present in other vocabularies. We used all possible vocabulary codes to query the OMOP vocabularies (release v5.0 28-JAN-22) and found that 19.2% of the C-Codes mapped to a standard OMOP concept. There were no corresponding OMOP concepts for 4.6% of C-Codes; however, nearly 90% of these are UCUM concepts and this is expected based on the OMOP documentation[2].

To evaluate the applicability of this approach, we used SDTM data from the C-Path Online Data Repository[3] to identify submission values from controlled terminology codelists. We found that a majority of observations mapped to at least one standard concept. The appropriateness of initial mappings was assessed by comparing the SDTM codelist domain to the target concept domain. In many cases the source-to-target domain were appropriate; for example, codes in the laboratory data (LB) domain had target concepts in the Measurement domain. This preliminary mapping shows potential in the ability to extract translations between SDTM and OMOP vocabularies through the UMLS. These mappings require further refinement based on subject-matter expert review and additional transformation logic, to ensure that context appropriateness of mappings. For example, we are exploring further refinement by including the source SDTM domain in the mapping logic. Additionally, it may be possible to bolster the mappings with additional resources such as CDISC's LOINC to LB Mapping Files.

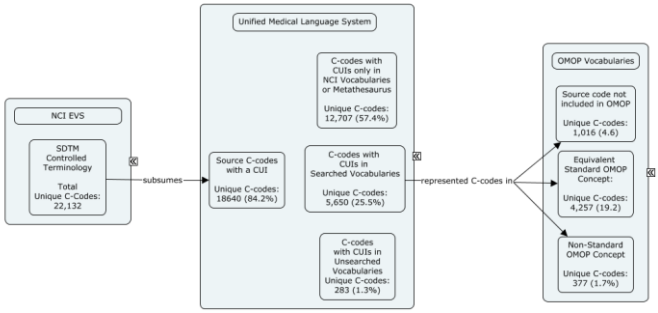

Figure 1: data flow with the number of unique C-Codes present in each

2. Conclusion

We show that it is feasible to aid the transformation process between CDMs by utilizing the UMLS to generate mappings between the SDTM terminology and the OMOP vocabularies. Future work will expand and ensure accuracy of the mappings, outline improvements of data standards for interoperability, and publish source code.

3. References

[1] Garza M, Del Fiol G, Tenenbaum J, Walden A, Zozus MN. Evaluating common data models for use with a longitudinal community registry. J Biomed Inform. 2016 Dec;64:333-341.

[2] Available from: https://www.ohdsi.org/web/wiki/doku.php?id=documentation%3Avocabulary%3Aucum

[3] Critical Path Institute Online Data Repository (CODR). Available from: https://codr.c-path.org/