Pathways for advanced transformation of CDISC SDTM data sets into OMOP CDM

Alexander Davydov, MD[1], Alexandra Orlova[1], Eva-Maria Didden, PhD[2], Rose Ong, PhD[2], Patricia Biedermann, MSc[2], Graham Wetherill, PhD[2], Gregory Klebanov, MSc[1], Michael Kallfelz, MD[1]

[1]Odysseus Data Services, Inc., Cambridge, MA, USA;
[2]Actelion Pharmaceuticals Ltd, Allschwil, Switzerland

Abstract

The Observational Medical Outcome Partnership Common Data Model (OMOP CDM) is aimed at making observational real world evidence research possible making multiple sources of diverse format accessible. The data from these sources can then be analyzed using one unified data model, allowing to share queries and results. The Clinical Data Interchange Standards Consortium (CDISC) Study Data Tabulation Model (SDTM) is widely used in clinical trials and post-marketing surveillance registries, but its conversion to OMOP CDM is still uncommon. We describe challenges faced and possible solutions for conversion from CDISC SDTM to OMOP CDM in order to map study-specific variables in a way that is accurate and minimizes data loss by imputation and introduction of new customized concepts. The Medical Dictionary for Regulatory Activities (MedDRA) vocabulary wasn't recognized as a source vocabulary, so we propose its development and application model in OMOP CDM. Mapping SDTM data to OMOP CDM allows comparing and combining results with outcomes generated in other types of mapped healthcare data assets, such as commercial claims or electronic health records.

Research Category

Observational data management

Introduction

SDTM was developed by CDISC as a general conceptual model for harmonizing clinical study data in its collection, management, analysis and reporting such as submission to regulatory authorities (FDA, PMDA, etc.)[1]. SDTM is widely used in clinical trials and post-marketing surveillance registries. It is built around the concept of observations, which consist of discrete pieces of information collected during a study.

SDTM to OMOP CDM conversions are rarely done[2]. To avoid loss of information as well as misrepresentation of study-specific variables, approaches for conversions need to be established and standardized among the OHDSI community.

Materials/methods

Data from two post-marketing, US-based, multi-center studies of Pulmonary Hypertension (PH) patients newly treated with Opsumit (macitentan) were converted from STDM to OMOP CDM: OPUS – a prospective observational registry from 04/2014 to 03/2020 and OrPHeUS – a retrospective chart review from 10/2013 to 03/2017. Standardized vocabularies (version 4-SEP-2019) and OMOP CDM (version 5.3.1) were used as a target model[3].

Results

Imputation rules. Most source tables have incomplete dates (month and/or day missing) or entirely missing dates. The OMOP CDM requires complete dates3. In order to prevent substantial data loss, imputation rules were introduced. Implementation of these rules resulted in reduction of data loss to 2-3% overall instead of 30% in the case if no imputation rules were applied. Imputation rules were mainly based on the temporal order of reference events, such as date of birth/death, previous start/end dates (e.g. date of Opsumit initiation), or date of last available information.

SDTM-specific information rescuing. SDTM registry data may contain specific information that does not completely match current OMOP CDM design. Table 1 contains some examples how this information could be stored in the OMOP CDM without any adjustments and additional tables/fields.

Table 1. Examples of SDTM-specific data mapping to OMOP CDM.

Case	OMOP CDM Current Solution	Comment
Additional characteristic of a medical event	Separate record in the proper event table. ● Link characteristic and assessed medical event with FACT_RELATIONSHIP record.	Characteristic could be severity, seriousness, status (stable/unstable) or outcome of a medical event.
Study-specific variables	Record in the OBSERVATION table: ● event _concept_id = specific fact; value_as_string = name of the study.	Variables could be study enrolment date, date of informed consent, reason for study discontinuation, etc.

Custom Mappings. Except for the MedDRA vocabulary, the raw SDTM data did not contain concept codes from standardized OMOP vocabularies. Currently, all-level MedDRA terms are Classification concepts in OMOP vocabularies since MedDRA wasn't recognized as a source vocabulary before. That is why MedDRA terms have no 'Maps to' links to the Standard concepts provided by the OHDSI. In order to map 5,856 MedDRA codes to Standard concepts, mapping automation (UMLS MedDRA - SNOMED/ICD10/ICD10CM crosslinks, name matching) with further expert review and additional manual mapping was performed and resulted in 6,613 source_to_concept_map records. The approach we used may be applied for further MedDRA vocabulary development and application. Classification concepts can be used as source concepts populating the source_concept_id field of the event tables (there is no such limitation in OMOP CDM), hence not just Standard equivalent concepts (with SNOMED hierarchy), but source MedDRA concepts (with different MedDRA hierarchy) are still accessible in analytics.

Concomitant and study medication data were encoded in the source data by the WHODrug vocabulary, which is not a part of the OMOP CDM yet. To address this issue, the extracted free text from medication tables as well as non-MedDRA adverse events, laboratory tests, death reasons, medical history facts, clinical events associated signs, etc. were custom mapped.

PH is a complex disease with several rare subgroups and subtypes of subgroups. Standardized OMOP CDM vocabularies do not cover all disease subtypes as well as some study-specific concepts. To further differentiate the subtypes, a combination of standard SNOMED concepts was used, e.g. Pulmonary Arterial Hypertension (PAH) associated with connective tissue disease + Connective tissue disease overlap syndrome; PAH associated with congenital heart disease + History of surgically corrected congenital heart defect[4]. Whenever mapping was not possible, customized concepts were introduced, e.g. for drug- and toxin-induced PAH, PH with unclear multifactorial mechanism, PAH WHO Functional Class. In order to make cohort definition and associated vocabulary exercises easier to use, all custom concepts were incorporated into the concept's hierarchy (custom CONCEPT_RELATIONSHIP and CONCEPT_ANCESTOR tables).

Conclusion

In this study, we developed and applied novel approaches when converting SDTM to OMOP CDM resulting in a significant reduction of source data loss. To better support SDTM conversions, MedDRA vocabulary development and application model might be extended to OMOP Standardized vocabularies. To further improve medication mapping in research data conversion, the WHODrug vocabulary should be added to OMOP CDM.

References

1. CDISC Submission Data Standards Team. Study Data Tabulation Model Implementation Guide: Human Clinical Trials Version 3.2. Accessed from https://cdisc.org/standards/foundational/sdtm.
2. Joshua Ransom, Eldar Allakhverdiiev, Gregory Klebanov, Jim Singer, Kirill Eitvid, Rayhnuma Ahmed, et al. Leveraging the OMOP CDMv5 for CDISC SDTM RCT Data. Accessed from https://www.ohdsi.org/web/wiki/doku.php?id=projects:workgroups:clinicalstudy.
3. Christian Reich, Patrick Ryan, Rimma Belenkaya, Karthik Natarajan, Clair Blacketer. OMOP Common Data Model Specifications v5.3.1. Accessed from https://github.com/OHDSI/CommonDataModel/releases/tag/v5.3.1
4. Galiè N, Humbert M, Vachiery J-L, Gibbs S, Lang I, Torbicki A, et al. 2015 ESC/ERS Guidelines for the Diagnosis and Treatment of Pulmonary Hypertension. Rev Esp Cardiol (Engl Ed). 2016 Feb;69(2):177.