Leveraging the OMOP CDMv5 for CDISC SDTM RCT Data

Joshua Ransom, PhD[1], Eldar Allakhverdiiev[2]*, Gregory Klebanov[2]*, Jim Singer[2], Kirill Eitvid[2],
Rayhnuma Ahmed[1]**, Emelly Rusli, MPH[1]**, Alexandra Shilnikova[1]**
[1]SHYFT, a Medidata company, Waltham, MA, US; [2]Odysseus Data Services, Inc., Cambridge, MA, US
*/** These authors contributed equally
Contact: *jransom@mdsol.com, contact@odysseusinc.com*

Background

Randomized controlled clinical trials (RCTs) are the gold standard for demonstrating treatment efficacy. However, the treatment is investigated in a well-controlled setting not usually representative of the real world. This limits the assessment of treatment effectiveness in broader real world settings, including more complex treatment paradigms, different patient demographics, disease severities, and genotypes. These are of critical concern to payers, providers, regulators, life science companies and other stakeholders.

Emerging real world data (RWD) in the healthcare industry provides a means to evaluate epidemiology and burden of a disease in the context of clinical practice. Real-world evidence (RWE) derived from RWD can then be directly compared to RCT findings to better explore patient outcomes while accounting for factors that are not otherwise observed during the trial. This effort is not without a challenge. The lack of uniformity between the data frameworks and structures in clinical trials versus clinical practice makes comparative analyses between these data sources challenging, time consuming, and not scalable. This is where OMOP common data model (CDM) plays a pivotal role by providing a well-established, tailored framework for observational research data.

We describe our effort to convert RCT data from CDISC SDTM into OMOP CDM. We chose SDTM as it is a standard method of organizing and formatting clinical trial data and is one of the requirements for data submissions to the FDA[3]. By converting the CDISC SDTM into research-quality OMOP CDM, we were able to rapidly conduct mirrored patient profile and outcomes analyses on both clinical trial data and RWD.

Methods

Vocabulary Mappings

The raw CDISC dataset contained no standardized concept codes from medical vocabularies (e.g., MedDRA, LOINC, UCUM, etc.). First, we programmatically matched the descriptions from CDISC records against OMOP concepts based on description similarity. The match list was supplemented with human review for accuracy and additional concept matches. About 90% of CDISC records were successfully mapped to standard concepts in this initial step.

The remaining trial-specific concepts lacked corresponding codes in standardized vocabularies, such as investigational treatments or research-oriented PRO instruments (i.e., Alzheimer's Disease Assessment Survey). In order to address this issue, we extended the OMOP Vocabulary by adding custom concepts that follow OMOP Standardized Vocabularies rules.

SDTM Data Classification

In the CDISC SDTM, medical observations collected during a study are divided among three record classes - Events (EV), Findings (FD), and Interventions (IV). Each class can encompass various record types (i.e., Events can be Medical History or Adverse Reactions, etc.). We used the three parent classes to simplify the conversion process and reduce the risk of data duplication **(Table 1).**

Fact Relationship

We used the OMOP FACT_RELATIONSHIP table in order to maintain full fidelity to the SDTM data record relationships. The types of SDTM relationships that were converted are as follows:

- Two independent records' relationship, such as a concomitant medication taken to treat an adverse event.
- Dependent relationships between comments/notes and a parent record (or records), such as a comment recorded in association with an adverse event.
- Relationship between a subject and a pool of trial associated subjects.
- Relationship between a subject and trial associated person(s).
- Relationship between a subject and non-trial associated person(s).

Standardized Derived Elements

The OMOP Common Data Model structure is not intended to capture RCT design information, such as: type of trial, number of arms, inclusion/exclusion screening criteria for entering the trial, etc. Therefore, we stored trial design information inside of the OMOP Standardized Derived Elements tables **(Fig. 1, 2)**.

Mirrored Patient Profile Analyses

We loaded the transformed CDISC data into the SHYFT (a Medidata company) Quantum Solution V6.7.0 alongside an existing RWD source - Medicare Synthetic Public Use Files (SynPUF)[4]. We studied all patients with an Alzheimer's Disease diagnosis within the same observation period window. We generated descriptive statistics for patient demographics, comorbidity prevalence, concomitant medication prevalence, and PRO survey. In addition, we ran Kaplan-Meier survival analysis for the time from Index to first drug related adverse events (e.g., Application site disorder, Erythema, Drug-induced erythema, Application site rash, Application site irritation, Application site pain, and Application site edema)[5]. Given that both data sets were converted into OMOP CDM, we were able to complete the analyses within 2 working days.

Parent Classes	Example Child Classes	CDISC Parent Identifier Column
EV	AE, MH, DS	xxTERM
IV	CM, TX	xxTRT
FD	LB, GN, QS	xxTEST

Table 1. SDTM Parent to Child class mapping

ItemGrou	STUDYID	DOMAIN	ARMCD	ARM	TAETORD	ETCD	ELEMENT	TABRANCH	TATRANS	EPOCH
1	CDISCPILOT01	TA	Pbo	Placebo	1	SCRN	Screen	Randomized to Placebo		Screening
2	CDISCPILOT01	TA	Pbo	Placebo	2	PBO	Placebo			Treatment
3	CDISCPILOT01	TA	Xan_Hi	Xanomeline High Dose	1	SCRN	Screen	Randomized to High Dos		Screening
4	CDISCPILOT01	TA	Xan_Hi	Xanomeline High Dose	2	HIS	High_Start			Treatment
5	CDISCPILOT01	TA	Xan_Hi	Xanomeline High Dose	3	HIM	High_Middle			Treatment
6	CDISCPILOT01	TA	Xan_Hi	Xanomeline High Dose	4	HIE	High_End			Treatment
7	CDISCPILOT01	TA	Xan_Lo	Xanomeline Low Dose	1	SCRN	Screen	Randomized to Low Dose		Screening
8	CDISCPILOT01	TA	Xan_Lo	Xanomeline Low Dose	2	LO	Low			Treatment

Figure 2. Sample Trial Arm description from SDTM format

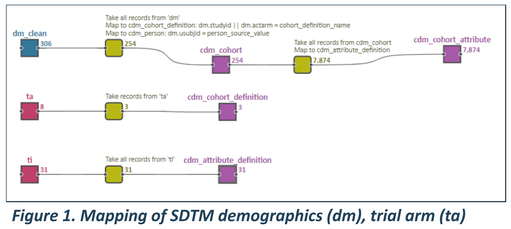

Figure 1. Mapping of SDTM demographics (dm), trial arm (ta) and trial inclusion (ti) tables into OMOP CDM

Results

Variable	Result	CDISC	SynPUF
Patient Cohort	*Total Inclusion*	254	372,727
Drug Reaction AEs	*FALSE, N (%)*	127 (50.0%)	347,182 (93.1%)
	TRUE, N (%)	127 (50.0%)	25,545 (6.9%)
Age At Index	*Mean (SD)*	75.08 (8.24)	73.00 (13.18)
	Median	77.00	74.00
Gender	*FEMALE, N (%)*	143 (56.3%)	217,859 (58.5%)
	MALE, N (%)	111 (43.7%)	154,868 (41.5%)
Race	*American Indian, N (%)*	1 (0.4%)	0 (0.0%)
	Black or African American, N (%)	23 (9.1%)	37,817 (10.1%)
	White, N (%)	230 (90.6%)	314,106 (84.3%)
Ethnicity	*Hispanic or Latino, N (%)*	12 (4.7%)	20,804 (5.6%)
	Not Hispanic or Latino, N (%)	242 (95.3%)	8,009 (2.1%)
Age Subgroup: 10 years bins	*50-60, N (%)*	17 (6.7%)	25,852 (6.6%)
	60-70, N (%)	49 (19.3%)	94,206 (23.9%)
	70-80, N (%)	111 (43.7%)	131,613 (33.4%)
	80-90, N (%)	77 (30.3%)	90,660 (23%)
Count of any HCP Visit	*Mean (SD)*	13.81 (4.92)	80.14 (40.97)
	Median	16.00	75.00
Lorazepam Exposure	*FALSE, N (%)*	232 (91.3%)	344,299 (92.4%)
	TRUE, N (%)	22 (8.7%)	28,428 (7.6%)
Donepezil Exposure	*FALSE, N (%)*	249 (98.0%)	359,341 (96.4%)
	TRUE, N (%)	5 (2.0%)	13,386 (3.6%)
Baseline ADAS Score	*Mean (SD)*	23.73 (12.40)	-
	Median	21.00	-
ADAS Delta Score	*Mean (SD)*	1.85 (5.20)	-
	Median	1.00	-

Table 3. Mirrored patient profile statistics across RCT and RWE data sets

CDM Target Table	Conversion Type	Record Counts	
		CDISC SDTM	OMOP CDM
cdm_person	*Total , N (%)*	306	**306 (100%)**
cdm_visit_occurrence	*Total, N (%)*	3,559	**3,559 (100%)**
cdm_death	*Total, N (%)*	3	**3 (100%)**
cdm_condition_occurrence	*Total, N (%)*	1,581	**1,581 (100%)**
cdm_drug_exposure	*Standard, N (%)*	-	1,781
	Custom, N (%)	-	365
	Unmapped, N (%)	-	13
	Total, N (%)	2,159	**2,146 (99.4%)**
cdm_procedure_occurrence	*Total, N (%)*	11	11 (100%)
cdm_observation	*Total, N (%)*	6,188	**6,188 (100%)**
cdm_measurement	*Standard, N (%)*	-	85,017
	Custom, N (%)	-	123,235
	Total, N (%)	208,252	**208,252(100%)**
Overall Conversion Total	*Standard, N (%)*	-	**98,446 (44.3%)**
	Custom, N (%)	-	**123,600 (55.7%)**
	Total Unmapped, N (%)	-	**13 (0.0%)**
	Total Mapped, N (%)	222,059	**222,046 (100%)**

Table 4. Conversion mapping statistics to CDM target tables, overall 99.99% records successfully converted

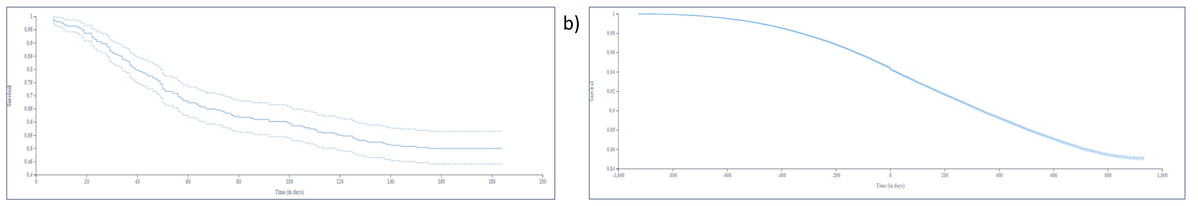

Figure 3. Drug Reaction AEs Free-Survival a) CDISC (N=254, Median survival = 184 days), b) SynPUF (N = 372,727, Median survival = N/A)

Conclusions

RCT data in the CDISC SDTM format generates high fidelity data but is unable to fully adhere to the OMOP CDM v5 vocabulary specifications. There were important lessons learned during the CDISC conversion process, including:

1. Aggregating similar tables to EV, FD, or IV parent tables (e.g., AE, MH, & DS all rolled up to EV)
2. Creating custom concepts for investigational events (e.g., new drugs without RxNorm codes)
3. Creating custom concepts for Survey/PRO data (may warrant new OMOP CDM tables)
4. Combining expert human review, medical thesaurus, and Natural Language Processing (NLP) of the trial record descriptions for optimal vocabulary mapping for trials that record descriptions instead of clinical codes
5. Parsing actual visits from patient and family medical history surveys in order to create trusted observation periods

Rapid comparisons of Alzheimer's Disease cohorts in RCT data versus RWD displayed a high degree of similarity for most comorbidities and concomitant medications analyzed. As expected, RCT-focused outcomes were more prevalent vs. the RWD cohort (e.g., PRO, drug-AEs, etc.). More trials need to be converted to improve generalizability of our methodology. In addition, OMOP to CDISC conversions could be explored depending on regulatory framework evolution. Future applications of this work will leverage the evidence generated from comparative analyses between RCT data and RWD to better inform healthcare guidelines, health technology assessments, and the use of synthetic control arms[6].

Finally, we acknowledge and thank members of the OHDSI community, the THEMIS working group, and Lev Zarakovich for their help and advice with the CDISC to OMOP conversion. We also thank and acknowledge the product engineers at SHYFT (a Medidata company) for creating the Quantum solution.

References

1. The Observational Health Data Sciences and Informatics (OHDSI). Retrieved 10/5/2018 from http://www.ohdsi.org
2. Christian Reich, Patrick Ryan, Rimma Belenkaya, Karthik Natarajan, Clair Blacketer and the OHDSI CDM and Vocabulary Development Working Group. OMOP Common Data Model V5.0. Retrieved 10/1/2017 https://github.com/OHDSI/CommonDataModel/releases/tag/v5.0.0
3. CDISC.org. CDISC SDTM v1.5. Dataset-XML datasets and a Define-XML file created by xml4pharma from the LZZT test datasets. Retrieved 10/1/2017 https://wiki.cdisc.org/display/PUB/CDISC+Dataset-XML+Resources
4. Christophe Lambert, Amritansh, Praveen Kumar. Transforming the 2.33M-patient Medicare synthetic public use files to the OMOP CDMv5: ETL-CMS software and processed data available and feature-complete DE_SynPuf output files on OHDSI website. Retrieved on 9/1/2018 from ftp://ftp.ohdsi.org/synpuf/
5. Concept IDs and concept codes lists available upon request
6. Ransom, J., PhD, Ahmed, R., Rusli, E., MPH, Stern, A. 2018. *Using Real-World Evidence to Optimize Clinical Trials* *https://www.shyftanalytics.com/whitepapers/7-ways-to-use-real-world-data-to-transform-therapeutic-rd/*