Paper RE01

Mapping OMOP Data to SDTM
User Case for RWD in Registry Study

Zhouzhou (Joe) Xi, Bristol Myers Squibb, New Jersey, USA
Kai (Kevin) Deng, Bristol Myers Squibb, New Jersey, USA
Lorraine Fang, Bristol Myers Squibb, New Jersey, USA

ABSTRACT

This paper shares our experience in mapping RWD (Real World Data) in OMOP (Observational Medical Outcomes Partnership) to CDISC SDTM-like model. Below perspectives are included:

- How our OMOP data (CAR-T specific) are organized
- How we convert and remap these OMOP data into SDTM-like domains
- Decoding, traceability and data masking issues

At the end, we summarize the advantage of using OMOP as potential RWD collection model for pharmaceutical industry.

INTRODUCTION

Over the past years, pharmaceutical industry is discussing on Real World Data (RWD) submission standards. Several models including CDISC, FHIR, OMOP, have been proposed as potential standards.

Based on recent guidance published by either FDA or CDISC committee on this topic, it is assumed that CDISC (particularly SDTM) is still considered as the preferred data standards for clinical analysis and HA(Health Authority) submission. But on the other hand, CDISC was designed for clinical trial data, and may cause inconvenience when it is used for RWD collection/mapping, this is the reason why OMOP or FHIR (designed for observational data collection) are proposed as potential future standards.

In our CAR-T registry studies, we use RWD as data source, which was collected externally and delivered to BMS server using OMOP model. To facilitate clinical analysis and support future submission, as well as to maintain consistency with other clinical data, we decide to convert these data into CDISC SDTM-like model, and to explore the feasibility to use OMOP-to-SDTM approach for other studies.

In this paper, we will cover topics of how source data are organized, the experience and issues in converting or mapping them into SDTM-like structure. We hope that our experience can benefit the entire pharmaceutical industry.

HOW OUR OMOP DATA ARE ORGANIZED

OMOP (Observational Medical Outcomes Partnership) model is an open-source data standard, maintained by OHDSI (Observational Health Data Science and Informatics). OHDSI targets to harmonize disparate healthcare data sources into one relational database with standardized tables and vocabularies, such as:

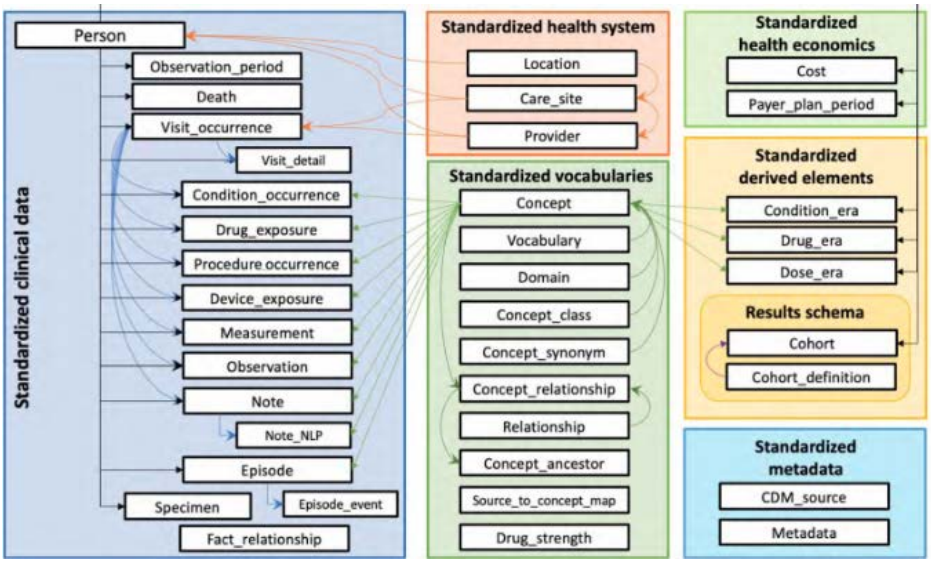

Figure 1: OMOP CDM Schema (from OHDSI website)

In our registry study, the data vendor follows OMOP model but adds one non-standard table (Form_option) to collect additional contents/details supplemental to each standard table.

In OMOP model, the actual data are categorized and stored into different tables/domains by their nature. For example, patient level information goes to Person table, AE (adverse events) goes to Observation table and lab values go to Measurement table, etc. The columns/variables xx_concept_id in each table contain standard coding of corresponding variables, and the coding value can be decoded either by linking Concept using concept_id or from OHDSI website https://athena.ohdsi.org/search-terms/start. Every record from each table can also be linked to form_option table through form_option_id for additional supplemental information. Below are examples of OMOP data (Person and Observation).

Person:

Column	Description
person_id	A unique ID for every person in the database
gender_concept_id	A concept_id code for the gender or sex of the person at birth
year_of_birth	Year in which person was born
month_of_birth	Month in which person was born
day_of_birth	Day on which person was born
birth_datetime	Not populated
race_concept_id	A concept_id to code the race of the person
ethnicity_concept_id	A concept_id to code for the ethnicity of the person
location_id	ID for the location that links to the location_id in the location table
provider_id	Not populated
care_site_id	ID for the care site that links to the care_site_id in the care_site table
person_source_value	An ID that links back to a unique identifier in the source database from which OMOP is populated
gender_source_value	A value from the source database representing gender or sex at birth
race_source_value	Not populated
ethnicity_source_value	Not populated
ethnicity_source_concept_	Not populated
donor	Not standard OMOP-CDM, implemented for the donor outcome registry. TRUE / FALSE indicating whether the person was a donor.

Observation:

Column	Description
observation_id	This column contains the id by which this observation is known in database
observation_concept_id	This column contains the a concept_id coding for the observation
observation_date	This column contains the date of the observation, most likely the event date of the event the observation was recorded on
observation_datetime	Not populated
observation_type_concept_id	This column indicates where the data was extracted from.
value_as_number	The numerical value of the observation that was done, e.g. the chronological number of an HCT
value_as_string	This column can contain non-numerical data pertaining to the observation, providing flexibility to give extra information when the value_as_concept_id or value_as_number might benefit from additional context.
value_as_concept_id	The concept id of the observation if it was a categorical observation, e.g. the donor's sex at birth in an allogeneic HCT
qualifier_concept_id	Not populated
unit_concept_id	This column contains the unit of the observation coded as a concept_id, if applicable
observation_source_value	This column contains a reference to the form, field and value that created the particular row in the table
qualifier_source_value	Not populated
observation_event_id	Not populated
person_id	This column can be used to link the observation to the patient using the person table
visit_detail_id	Not populated
visit_occurrence_id	Not populated
provider_id	Not populated
form_option_id	Non-OMOP-CDM column. This column contains the form option ID, that can be used to identify the form option using the form_option table
patient_event_id	Non-OMOP-CDM column. This column contains the identifier of the patient event in the application
group_index	Non-OMOP-CDM column. Where the data row comes from a repeater group in the form, this will increment as repeater groups are added.

Figure 2 Sample of OMOP Spec (Person and Observation)

Accordingly, the actual data for either Person or Observation table are shown in below sample. By linking different tables (e.g. person vs observation) using person_id, the patient's basic profile or information can be retrieved.

Person:

person_id	gender_concept_id	location_id	provider_id	care_site_id	gender_source_value	gender_source_concept_id	donor
10001	8532	1		601	1		FALSE
10002	8507	2		602	2		FALSE
10003	8532	3		603	1		FALSE
10004	8507	4		604	2		FALSE

Observation:

observation_id	observation_concept_id	observation_date	observation_datetime	observation_type_concept_id	value_as_number	value_as_string	value_as_concept_id	qualifier_concept_id	unit_concept_id
50604889	2000500007	1/1/2024		32879	1				
24626896	2000200067	1/1/2024		32879			45878245		
14987021	36676867	1/1/2024		32879			739862		
52388325	0	1/1/2024		32879	73				9529
15046581	2000500020	1/1/2024		32879			2005897225		
15114733	2000500002	1/1/2024		32879			45877994		
47064183	4211787	3/1/2022		32879			2000200063		
15006055	4211787	1/1/2024		32879			4165515		
24891858	2000500008	1/1/2024		32879			4230556		
15072691	2000500001	1/1/2024		32879			45877994		

observation_source_value	unit_source_value	person_id	visit_detail_id	visit_occurrence_id	provider_	form_option_id	patient_event_id	is_date_defaulte	group_inde
Therapy and cell infusion(s) | Chronological number of cellular therapy treatment for this patient | 1		10001				251612	9c67abe3-0997-4e1b-b	FALSE	
Lymphomas | > 4 nodal sites involved | 1		10001				238555	9de650ad-4209-41a7-a	FALSE	
Planned cellular infusion products | Name of product | 2		10001				240349	9c67abe3-0997-4e1b-b	FALSE	
Patient status | Patient weight at initiation of HCT/CT/IST | 73	kg	10001				252136	9de650ad-4209-41a7-a	FALSE	
Basic information on the planned cellular therapy | Clinical setting | 1		10001				203540	9c67abe3-0997-4e1b-b	FALSE	
Main treatment description | Was the cellular therapy product infused during this treatment/procedure? | 2		10001				104207	9c67abe3-0997-4e1b-b	FALSE	
B-Cell Non-Hodgkin Lymphomas (NHL) - Classification | High-grade transformation of indolent B-cell lymphoma | 1		10001				74290	173ffe5b-7997-4433-9f	FALSE	
Preparative regimen | Total body irradiation (TBI) | 1		10001				50889	9c67abe3-0997-4e1b-b	FALSE	
Patient status | Survival status at HCT/CT/IST | 1		10001				238344	9de650ad-4209-41a7-a	FALSE	
Planned cellular infusion products | Is the planned cell infusion product a commercial product? | 2		10001				70858	9c67abe3-0997-4e1b-b	FALSE	

Figure 3: Sample of OMOP data from Person and Observation Table

HOW TO COVERT OR MAP OMOP DATA INTO SDTM DOMAINS

OMOP model stores the data in vertical structure, which is similar to CDISC model, and thus it is relatively easy to convert it into CDISC SDTM model comparing with other data standards (e.g. FHIR). In our case the key part for this conversion is to split the general tables from OMOP model into corresponding SDTM domains.

OMOP model generally categorizes data into some basic tables such as:

- **Patient level**: Person/Death, corresponding to DM, DS in SDTM
- **Observation/Condition**: corresponding to event-based domains such as AE, CM or efficacy (RS)
- **Measurement**: corresponding to BDS based domains such as LB, VS
- **Other**: including information such as site, visit/episode, exposure, standard vocabulary and supplemental data

Below diagram illustrates the general mapping between OMOP and SDTM.

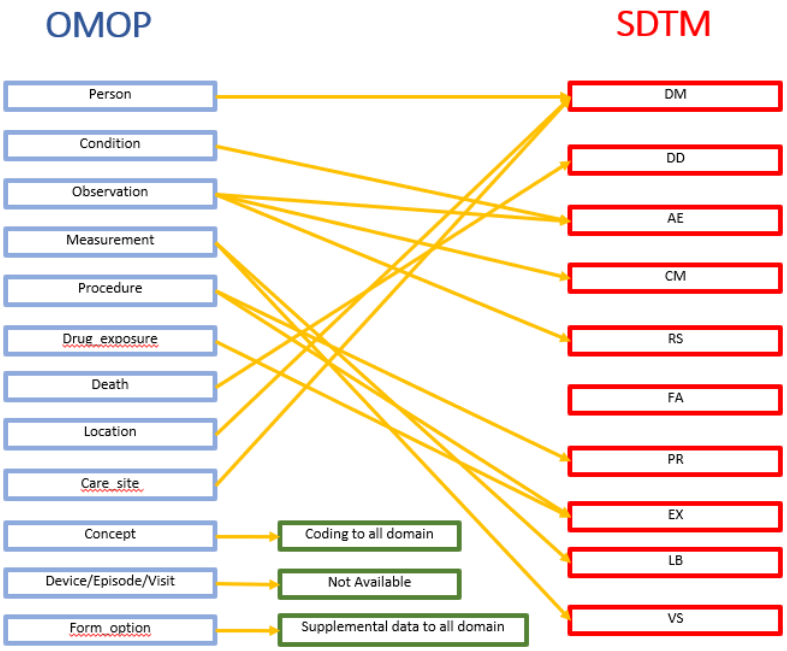

Figure 4: Mapping Spec from OMOP to SDTM

One disadvantage of using OMOP data is that some information important to clinical analysis may be missing in current model design. For example, in the collection of Adverse Events, key information such as AE outcome, AE start/end date, CTC grade are not collected in current Observation table. Our data vendor chooses to use form_option to document any additional information. but this adds extra complexity to parse the useful information and map the data to standard value.

SOME SPECIAL DATA HANDLING

CODING

As shown in the sample data, much information (e.g. Person.gender) is provided in the form of standard code (in xx_concept_id), and the user can decode this value by linking the concept_id with Concept table or looking up at OHDSI website (Athena). For example, in order to check gender code :

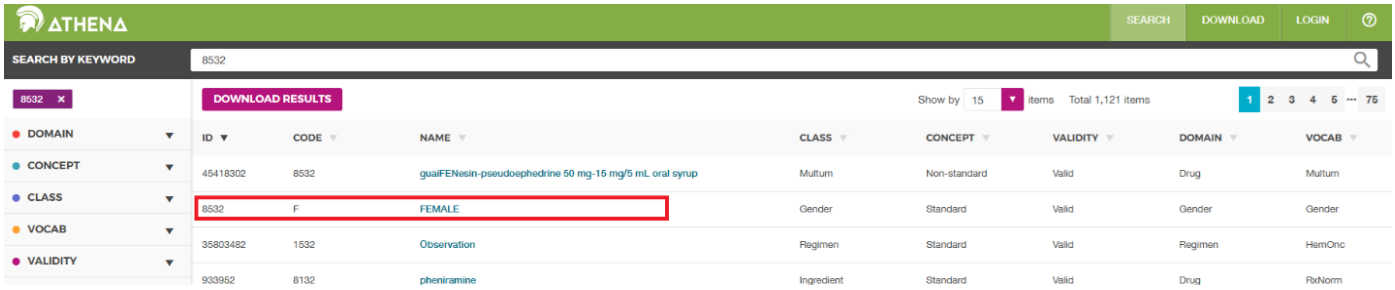

Figure 5: Sample of Decoding from Athena website

Whether the standard coding in OMOP is aligned with CDISC standard terms can be an issue, which is still being explored by our team.

AE or CM decoding using MedDRA or WHO Drug dictionary is another issue. In OMOP model, AE or CM verbatim terms are stored in either Condition or Observation table as source value, but their standard codes (Concept_id to link with Concept) may not be provided if the terms are not codable per investigator. To facilitate clinical analysis, we provide these terms to our internal coding team for a MedDRA or WHO Drug recoding regardless of their original coding was provided or not.

TRACEABILITY

CRF is usually not the standard way to collect RWD. It is difficult to maintain the traceability from the source data. In our study, data vendor added record ID when processing RWD into OMOP model, and we use this record ID (e.g. Observation_ID form Observation table) as the source ID (e.g. AE.AESPID). But given that OMOP data are still

different from original health records, this cutoff in traceability adds to difficulty in data managing or quality control, and the issue remains to be further discussed and aligned throughout industry.

DATA MASKING

Many RWD vendors have data privacy requirement based on HIPPA law, so sharing data with third party may trigger legal issues, and thus data masking is required. In our study, team decided to mask actual person ID with one-to-one matched dummy ID. At the same time, any actual date is replaced with relative day to patient's initial drug exposure. So far, the data masking strategy remains as study specific, and industry-wide guideline may be needed to align or guide this activity.

CONCLUSION

Currently many RWD data models have been proposed to support RWD standardization and submission to health authority. Industry is still working on finding the most proper approach. In this proceeding, we shared our experience from mapping data from OMOP model to SDTM-like format (in order to facilitate further analysis and submission). We also summarize some key issues we encountered during this process and hope to share lesson learned to our peers.

Based on our experience on OMOP data, we can conclude that OMOP model has following advantages:

1. It is in vertical structure which is close to clinical data standard such as CDISC;
2. It uses basic categorization which facilitates further mapping or converting to SDTM model;
3. It has flexible data structure to allow un-standardized data to be collected;
4. It is relatively easy for hospital or other investigator to quickly fit health care data into one common data.

In conclusion, OMOP data model can simplify real world data collection, and at the same time facilitate data standardization and analysis for clinical or pharmaceutical usage. Based on our experience in using OMOP data for our registry study, we think it can be a good option for the use of RWD in pharmaceutical industry.

REFERENCES

(1) Abolafia, J, Ferko, S, & Holt, I. (2024). "*Future Clinical Data Submission Standards: CDISC, FHIR, OMOP, or Hybrid Model*". Paper presented at the PHUSE US Connect 2024.

(2) Shiralkar, P & Rutkowski, M. (2024). "*Usage of OMOP-CDM along with CDISC in Clinical Trials: Challenges and Opportunities*". Paper presented at the PHUSE US Connect 2024.

(3) CDISC Guidance (2023) "*Considerations and Examples of Using SDTM for Observational Studies".*

(4) Shanbhogue, A.Y., (2022). "*RWD (OMOP) to SDTM (CDISC): A primer for your ETL journey*". Paper presented at the PharmaSUG Annual Conference 2022.

(5) FDA Guidance (2021) "*Data Standards for Drug and Biological Product Submissions Containing Real-World Data Guidance for Industry".*

(6) OHDSI https://www.ohdsi.org/data-standardization/ (accessed 12/01/2024).

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Author Name: Kevin Deng
Company: Bristol Myers Squibb
Email: kevin.deng@bms.com

Author Name: Joe Xi
Company: Bristol Myers Squibb
Email: joe.xi@bms.com