PharmaSUG 2022 - Paper RW-160

RWD (OMOP) to SDTM (CDISC): A primer for your ETL journey

Ashwini Yermal Shanbhogue

ABSTRACT

Real-World Data (RWD) is defined as data relating to patient health status and/or the delivery of health care routinely collected from a variety of sources like electronic health records (EHRs), medical claims and billing data, data from product and disease registries, patient-generated data etc. The increase in availability of this observational data and the evolution of tools to analyze it has piqued the interest of global regulatory agencies to use this data to support regulatory decision making. Since this data and its collection was not designed for this purpose however, it does not have uniform structure and vocabulary, making it difficult and time consuming to compare or exchange between computer systems. To circumvent this problem, RWD can be put into a common format or common data model (CDM) with common representation (terminologies, vocabularies). OMOP (Observational Medical Outcomes Partnership), which is a part of one of the four largest RWD networks, Observational Health Data Sciences and Informatics (OHDSI) is one such CDM. To be available for regulatory submissions, RWD in CDM must then be extracted, transformed and loaded (ETL) into an FDA supported data standard like Clinical Data Interchange Standards Consortium's (CDISC's) Study Data Tabulation Model (SDTM). In this presentation, I will explore the mapping of an open access RWD in OMOP CDM to appropriate variables and ontologies in CDISC SDTM and any associated challenges. This will enable sponsors to kick start their ETL journey with a blueprint to the process and equip them to deal with forthcoming challenges.

INTRODUCTION

It takes an average of 17 years for a drug to go from discovery to market (Morris et al., 2011). In order to accelerate this process and bring innovations to patients faster, the 21st Century Cures Act (Cures Act) was signed into law in December, 2016. One of the provisions of the Cures Act was adding section 505F to the Federal Food, Drug, and Cosmetic Act (FD&C Act) (21 U.S.C. 355g), according to which, US Food and Drug Administration (FDA) created the framework for the Real World Evidence (RWE) Program. This program intends to explore the possibility of using RWE to get regulatory approvals for new indications of a drug that has been approved already or to fulfill post approval study requirements.

WHAT ARE RWD AND RWE?

Real World Data (RWD) is defined as, "data relating to patient health status and/or the delivery of health care routinely collected from a variety of sources like electronic health records (EHRs), medical claims and billing data, data from product and disease registries, patient-generated data, including from in-home-use settings, and data gathered from other sources that can inform on health status, such as mobile devices", whereas Real-World Evidence (RWE) is defined as, "the clinical evidence about the usage and potential benefits or risks of a medical product derived from analysis of RWD." (FDA, 2018, CDISC, n.d).

It is important to note that data obtained from traditional clinical trials are not considered RWD and the evidence obtained from their analysis will not be considered RWE. However, it is permissible to combine data from traditional clinical trials with RWD to create hybrid clinical trials, which could then potentially generate RWE (FDA, 2018).

The increase in availability of RWD and the evolution of tools to analyze it has piqued the interest of global regulatory agencies to use this data to support regulatory decision making. RWD, however, is collected regionally and globally, from disparate sources, in disparate formats, using various methods and algorithms, de- identified using various methods and is finally aggregated into disparate kinds of data sets. Therefore, this data does not have uniform structure, format, or vocabulary, making it difficult and time consuming to compare or exchange it between computer systems (FDA, 2021). However, sponsors are expected to implement SDTM standard structure for human clinical trials tabulation data sets and this continues to be the expected format of input data that FDA, along with other regulatory agencies consider in their review of submissions.

ENTER THE COMMON DATA MODEL (CDM)

To circumvent the first step of the above broader problem, i.e., the heterogeneity of the data, RWD can be put into a common data model (CDM). CDMs are a defined group of data elements that have defined relationships among them. Thus, they have a standardized or common structure, format, and vocabulary.

The Observational Medical Outcomes Partnership (OMOP) is one such CDM. OMOP was the result of a public- private partnership between US FDA, NIH, pharmaceutical companies, academic researchers, and health data partners. The OMOP project has also given rise to the open Science community, Observational Health Data Sciences and Informatics (OHDSI), which is one of four large RWD networks (OHDSI, 2021).

The three other, large RWD networks are- Informatics for Integrating Biology & the Bedside (i2b2/Accrual to Clinical Trials (ACT)), Patient-Centered Outcomes Research Network (PCORnet), and Sentinel. Each of these networks has its own CDM. Each of these data networks, however, was developed for different purposes and contains data from different sources. Therefore, though they each have their own common data models, the data models are common within the network and not between networks i.e., it is difficult to compare or exchange data between the data networks. In addition to this, RWD available within CDMs cannot be used for regulatory submissions as is. Hence, the subsequent need arises to extract, transform, and load (ETL) it into an FDA supported data standard like SDTM, which is a tedious and time-consuming process. Additionally, there is no clear transformation logic across industry and very few initiatives are underway to solve this problem. The Common Data Model Harmonization (CDMH) project is one such initiative, but we still don't have industry standard guidelines to achieve this transformation. To make a humble attempt at contributing to standardization, I started my journey of transformation of RWD in OMOP CDM to SDTM with the CDMH project as a jumping-off point and augmented it with manual curation, while identifying and documenting the challenges and limitations associated with the process.

CDMH AND BRIDG

The CDMH Project, begun in February 2017 and completed in August 2020 managed to harmonize the four disparate CDMs of the four large data networks, by mapping them to an intermediary data model, Biomedical Research Integrated Domain Group (BRIDG). The BRIDG international standard had already been mapped to the Clinical Data Interchange Standards Consortium Study Data Tabulation Model (CDISC SDTM) and HL7 Fast Healthcare Interoperability Resources (HL7 FHIR) standards (FDA, NIH, ONC, 2020) (Figure 1).

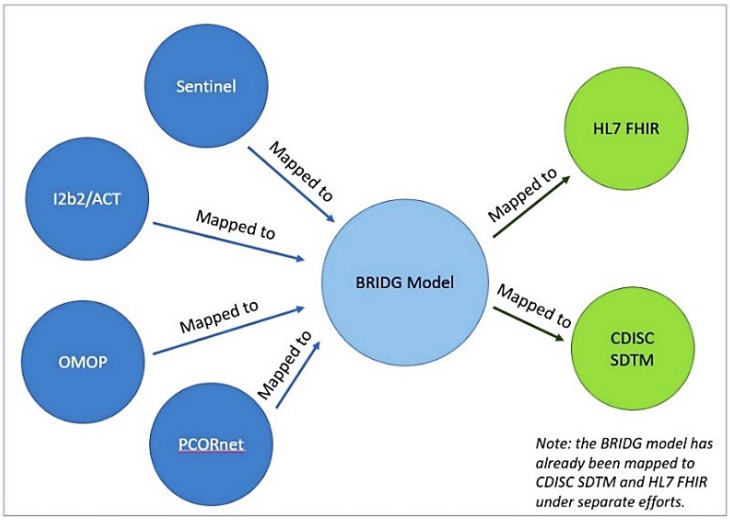


Figure 1 BRIDG- The Bridge between CDMs and SDTM (FDA, NIH, ONC, 2020)

The BRIDG domain information model is a product of collaborations between CDISC, US FDA, HL7 Biomedical Research and Regulation Work Group (HL7 BR&R WG), the International Organization for Standardization (ISO) and the US National Cancer Institute (NCI). BRIDG seeks to harmonize the semantics of basic, pre-clinical, clinical, and translational research domains and to bring semantic interoperability between the computer systems associated with these domains (ver Hoef et al., 2019).

Among the several ways BRIDG can be used (reference model, physical database, exchange format, Ontology (*BRIDG | Biomedical Research Integrated Domain Group*, n.d.)), CDMH has adopted BRIDG as a data integration/ mapping solution i.e since several standards have been mapped to the central standard, BRIDG, each of these standards can be mapped to one another more easily and efficiently. It is this use case that I have utilized, to facilitate my extraction, transformation, and loading (ETL) of Medical Information Mart for Intensive Care (MIMIC IV v0.9) demo data in OMOP CDM v5.3.1 to CDISC SDTM v3.2.

THE PROCESS

DATA SOURCE

The data set I have ETL'ed is a MIMIC IV v0.9 demo data set of 100 patients in the OMOP CDM v5.3.1 (Kallfelz et al., 2021, Goldberger et al., 2020). MIMIC is a large, permissibly accessible database containing de- identified electronic health records of patients admitted to the critical care units of Beth Israel Deaconess Medical Center, Boston, MA. The data recorded between 2008 and 2019 is stored in MIMIC IV v1.0, of which, MIMIC IV v0.9 is a subset. (Johnson et al., 2021).

From the open access webpage, https://physionet.org/content/mimic-iv-demo-omop/0.9/, I downloaded the folder, '1_omop_data_csv', which contains the csv files for the OMOP data (one file per table). The tables are of 6 types as per OMOP CDM- Clinical Data Tables, Health System Data Tables, Health Economics Data Tables, Standardized Derived Elements, Metadata Tables and Vocabulary Tables. The first row of each csv file is the column header.

WHITE RABBIT

I then prepared the downloaded tables for ETL using an open source, OHDSI, Java tool, White Rabbit (OHDSI, n.d.). White Rabbit scans input data present in csv files or databases and creates a scan report in the form of an excel file with a 'Field Overview' tab, a 'Table Overview' tab and a tab for each table in the data set. Each tab/ spreadsheet contains information about the structure of and values in each table and some metrics related to them. The 'Field Overview' table, in particular, makes it easier to perform the next step in the process, which is mapping, since all the table and column names from a data set are present in one, easily accessible spreadsheet. Figure 2 shows a snapshot of the White Rabbit scan report.

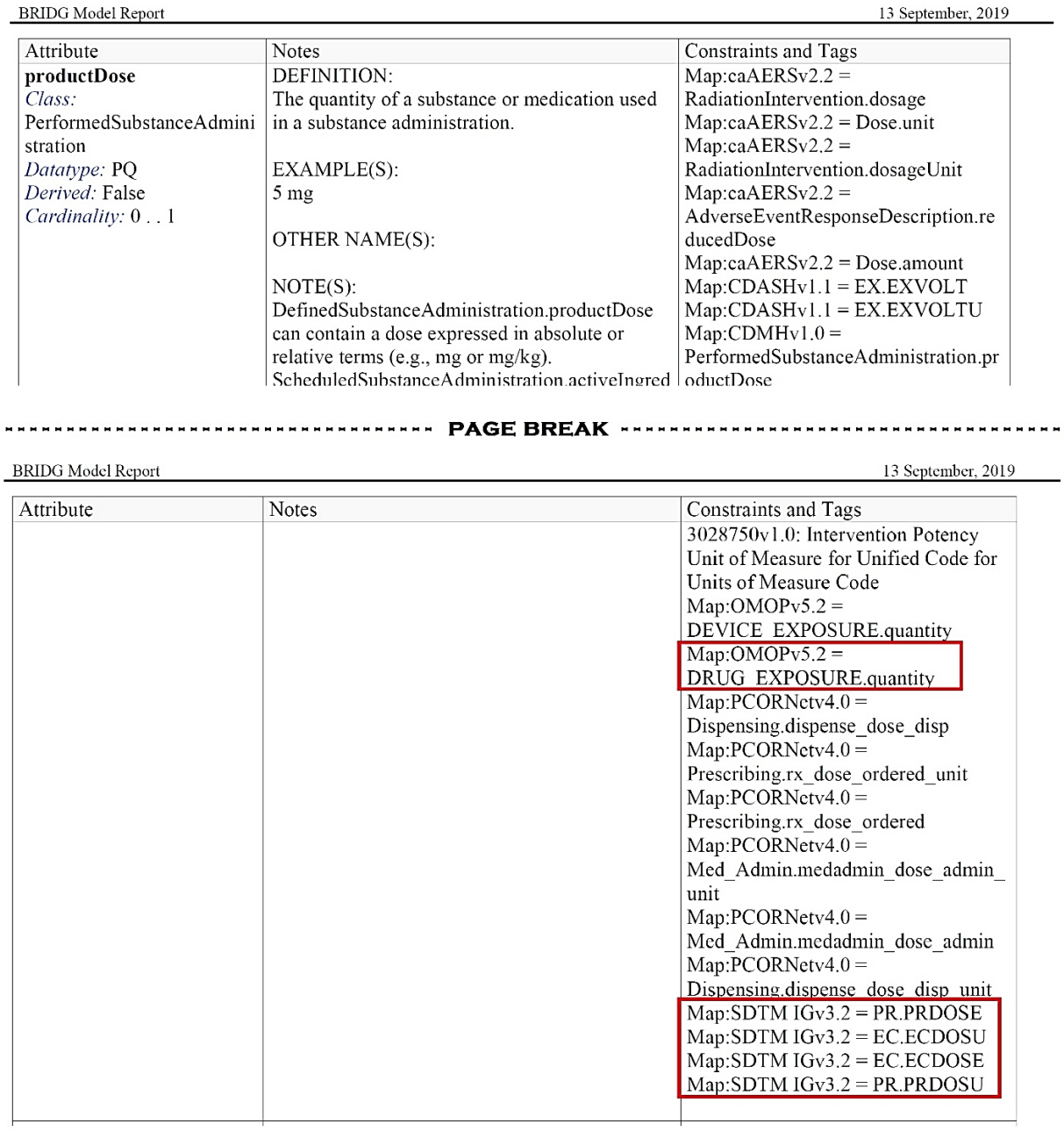
A	B	C	D	E	F	G	H	I	J
Table	Field	Description	Type	Max length	N rows	N rows checked	Fraction empty	N unique values	Fraction unique
fact_relationship.csv	domain_concept_id_1		INT	2	-1	1752	0.0%	2	0.1%
fact_relationship.csv	fact_id_1		INT	20	-1	1752	0.0%	1032	58.9%
fact_relationship.csv	domain_concept_id_2		INT	2	-1	1752	0.0%	2	0.1%
fact_relationship.csv	fact_id_2		INT	20	-1	1752	0.0%	1032	58.9%
fact_relationship.csv	relationship_concept_id		INT	6	-1	1752	0.0%	4	0.2%

Figure 2 Snapshots of Field Overview Tab from the White Rabbit Scan Report

USING BRIDG TO FIND SDTM VARIABLES

Following this, I mapped the MIMIC IV v0.9 data in OMOP CDM v5.3.1 to SDTM v3.2 with the help of BRIDG Release 5.3.1 (ver Hoef et al., 2019b). Please note that though the RWD is present in OMOP CDM v5.3.1, the latest version of OMOP available within BRIDG Release 5.3.1, is OMOP v5.2 and this is what I have used for the mapping. Each attribute of each class of each sub- domain within the BRIDG model is mapped to several data standards. If a particular attribute is mapped to standard A as well as standard B, it follows that standards A and B can be mapped to each other. I used this logic to find MIMIC IV- OMOP

variables and their corresponding SDTM variables within the BRIDG model. Figure 3 shows a snapshot of the BRIDG 5.3.1 model with this logic applied.



BRIDG Model Report 13 September, 2019

Attribute	Notes	Constraints and Tags
productDose *Class:* PerformedSubstanceAdministration *Datatype:* PQ *Derived:* False *Cardinality:* 0 . . 1	DEFINITION: The quantity of a substance or medication used in a substance administration. EXAMPLE(S): 5 mg OTHER NAME(S): NOTE(S): DefinedSubstanceAdministration.productDose can contain a dose expressed in absolute or relative terms (e.g., mg or mg/kg). ScheduledSubstanceAdministration.activeIngred	Map:caAERSv2.2 = RadiationIntervention.dosage Map:caAERSv2.2 = Dose.unit Map:caAERSv2.2 = RadiationIntervention.dosageUnit Map:caAERSv2.2 = AdverseEventResponseDescription.reducedDose Map:caAERSv2.2 = Dose.amount Map:CDASHv1.1 = EX.EXVOLT Map:CDASHv1.1 = EX.EXVOLTU Map:CDMHv1.0 = PerformedSubstanceAdministration.productDose

PAGE BREAK

BRIDG Model Report 13 September, 2019

Attribute	Notes	Constraints and Tags
		3028750v1.0: Intervention Potency Unit of Measure for Unified Code for Units of Measure Code Map:OMOPv5.2 = DEVICE_EXPOSURE.quantity Map:OMOPv5.2 = DRUG_EXPOSURE.quantity Map:PCORNetv4.0 = Dispensing.dispense_dose_disp Map:PCORNetv4.0 = Prescribing.rx_dose_ordered_unit Map:PCORNetv4.0 = Prescribing.rx_dose_ordered Map:PCORNetv4.0 = Med_Admin.medadmin_dose_admin_unit Map:PCORNetv4.0 = Med_Admin.medadmin_dose_admin Map:PCORNetv4.0 = Dispensing.dispense_dose_disp_unit Map:SDTM IGv3.2 = PR.PRDOSE Map:SDTM IGv3.2 = EC.ECDOSU Map:SDTM IGv3.2 = EC.ECDOSE Map:SDTM IGv3.2 = PR.PRDOSU

Figure 3 Snapshot of BRIDG Model Showing an Attribute Mapped to an OMOP Variable as well as Corresponding SDTM Variables

In this manner, I created an Excel file containing an initial mapping of MIMIC IV- OMOP variables to SDTM variables. Figure 4 shows a snapshot of this file.

TABLE_NUM	OMOP_TABLE	VARIABLE_NUM	OMOP_VARIABLE	SDTM_DATASET	SDTM_VARIABLE	BRIDG_SEARCH_RESULT	DATA_AVAILABLE	IN_OMOP_CDM	OMOP_TABLE_TYPE
27	PERSON	1	person_id	ALL	USUBJID		Y	Y	Clinical Data Table
27	PERSON	2	gender_concept_id	DM	SEX		Y	Y	Clinical Data Table
27	PERSON	3	year_of_birth	DM	BRTHDTC		Y	Y	Clinical Data Table
27	PERSON	4	month_of_birth	DM	BRTHDTC		N	Y	Clinical Data Table
27	PERSON	5	day_of_birth	DM	BRTHDTC		N	Y	Clinical Data Table
27	PERSON	6	birth_datetime	DM	BRTHDTC		N	Y	Clinical Data Table
27	PERSON	7	race_concept_id	DM	RACE		Y	Y	Clinical Data Table
27	PERSON	8	ethnicity_concept_id	DM	ETHNIC		Y	Y	Clinical Data Table
27	PERSON	9	location_id			No SDTM match found	N	Y	Clinical Data Table
27	PERSON	10	provider_id			No SDTM match found	N	Y	Clinical Data Table
13	DEATH	1	person_id	ALL	SUBJID		Y	Y	Clinical Data Table
13	DEATH	1	person_id	ALL	USUBJID		Y	Y	Clinical Data Table
13	DEATH	2	death_date	AE	AEDUR; AEENDTC; AESTDTC		Y	Y	Clinical Data Table
13	DEATH	2	death_date				Y	Y	Clinical Data Table
13	DEATH	3	death_datetime	AE	AEDUR; AEENDTC; AESTDTC		Y	Y	Clinical Data Table
13	DEATH	3	death_datetime				Y	Y	Clinical Data Table

Figure 4 Snapshot of Results of Initial Mapping Process

Key to column names and some values

Columns sourced from White Rabbit

TABLE_NUM: Unique number associated with each OMOP_TABLE.

OMOP_TABLE: Name of the OMOP table sourced from White Rabbit scan report.

VARIABLE_NUM: Unique number assigned to each OMOP_VARIABLE within an OMOP_TABLE.

OMOP_VARIABLE: Name of the OMOP variable sourced from White Rabbit scan report.

Columns sourced from BRIDG

SDTM_DATASET: The SDTM data set/s a particular OMOP variable should be mapped to as per BRIDG.

SDTM_VARIABLE: The SDTM variable/s a particular OMOP variable should be mapped to as per BRIDG.

BRIDG_SEARCH_RESULT: This column is blank if a corresponding SDTM match was found for an OMOP variable. If there was no SDTM match for an OMOP variable, it is indicated with, ‘No SDTM match found’. If an OMOP variable was missing from BRIDG, it is indicated with, ‘No OMOP match found’.

Columns added to support manual curation

DATA_AVAILABLE: Indicates if each OMOP variable in an OMOP table contains any data.

IN_OMOP_CDM: Indicates if the OMOP table is available in OMOP CDM v5.3

OMOP_TABLE_TYPE: The type of OMOP table.

MANUAL CURATION

Upon reviewing the mapping, I found that this initial mapping alone was not complete and accurate enough to allow the transformation of the RWD to SDTM. I then reviewed the data once again to verify the accuracy of the mappings suggested by BRIDG, by reviewing available values within variables and created more columns in the original Excel file to document the final mapping. Figure 5 shows a snapshot of the final mapping table.

TABLE_NUM	OMOP_TABLE	VARIABLE_NUM	OMOP_VARIABLE	SDTM_DATASET	SDTM_VARIABLE	BRIDG_SEARCH_RESULT	SDTM_DATASET_M	SDTM_VARIABLE_M	NA_REASON	OTHER_DETAILS
27	PERSON	1	person_id	ALL	USUBJID		ALL	USUBJID		
27	PERSON	2	gender_concept_id	DM	SEX		DM	SEX		
27	PERSON	3	year_of_birth	DM	BRTHDTC		DM	BRTHDTC		
27	PERSON	4	month_of_birth	DM	BRTHDTC		NA	NA	No Data in Source	
27	PERSON	5	day_of_birth	DM	BRTHDTC		NA	NA	No Data in Source	
27	PERSON	6	birth_datetime	DM	BRTHDTC		NA	NA	No Data in Source	
27	PERSON	7	race_concept_id	DM	RACE		DM	RACE		
27	PERSON	8	ethnicity_concept_id	DM	ETHNIC		DM	ETHNIC		
27	PERSON	9	location_id			No SDTM match found	NA	NA	No Data in Source	
27	PERSON	10	provider_id			No SDTM match found	NA	NA	No Data in Source	
13	DEATH	1	person_id	ALL	SUBJID		ALL	SUBJID		
13	DEATH	1	person_id	ALL	USUBJID		ALL	USUBJID		
13	DEATH	2	death_date	AE	AEDUR; AEENDTC; AESTDTC		AE	AEDUR; AEENDTC; AESTDTC		
13	DEATH	2	death_date				DM	DTHDTC; DTHFL		
13	DEATH	3	death_datetime	AE	AEDUR; AEENDTC; AESTDTC		AE	AEDUR; AEENDTC; AESTDTC		
13	DEATH	3	death_datetime				DM	DTHDTC; DTHFL		

Figure 5 Snapshot of Final Mapping Table

Columns added following manual curation

SDTM_DATASET_M: Final SDTM data set/s decided upon after additional manual review. If no suitable data set was found, it is indicated with ‘NA’.

SDTM_VARIABLE_M: Final SDTM variable/s decided upon after additional manual review. If no suitable variable was found, it is indicated with ‘NA’.

NA_REASON: Reasons why no suitable SDTM data set or variable were found.

OTHER_DETAILS: Any additional details that help the transformation process.

Thus, I completed the process of mapping MIMIC IV data in OMOP CDM to SDTM. However, this exercise brought to light a lot of challenges that could come up during this process.

CHALLENGES

Some of the challenges I faced were due to the nature of the data and some due to gaps in the two models used (OMOP and BRIDG). These challenges are illustrated below.

MANUAL CURATION IS ALWAYS REQUIRED

In some instances, when I searched for an OMOP variable within the BRIDG model, I found a very long list of SDTM data sets and variables that it could map to. However, the mapping had to be reviewed manually to figure out which of these several possible data sets, the values of the source variable belong to. For example, in the 'DEATH' table, looking for the 'cause_concept_id' variable in BRIDG, brought up a long list of SDTM data sets that it could be mapped to- AE, CE, CM, DD, DV, EG, FA, IE, IS, LB, MB, MH, MI, MO, MS, PC, PE, PP, QS, RP, RS, SC, SR, SS, SU, TR, TU and VS. However, AE is the only data set this variable can be mapped to. Figure 6 illustrates this example.

TABLE_NUM	OMOP_TABLE	VARIABLE_NUM	OMOP_VARIABLE	SDTM_DATASET	SDTM_VARIABLE	BRIDG_SEARCH_RESULT	DATA_AVAILABLE	IN_OMOP_CDM	OMOP_TABLE_TYPE	SDTM_DATASET_M	SDTM_VARIABLE_M	NA_REASON	OTHER_DETAILS
13	DEATH	5	cause_concept_id	AE	AEDECOD; AEOUT; AEPTCD; AETERM		Y	Y	Clinical Data Table	AE	AEDECOD; AEOUT; AEPTCD; AETERM		All records have a value of '0', which means 'Unknown or Unmapped value' as defined in section 4.2.5 (The Book of OHDSI). For the purposes of this exercise we assume the Death Cause is 'unknown'
13	DEATH	5	cause_concept_id	CE	CEDECOD; CEOCCUR; CESTAT; CETERM		Y	Y	Clinical Data Table	NA	NA	No Data in Source present with relevance to Proposed variables by BRIDG	
13	DEATH	5	cause_concept_id	CM	CMOCCUR; CMSTAT		Y	Y	Clinical Data Table	NA	NA	No Data in Source present with relevance to	
13	DEATH	5	cause_concept_id	DD	DDORRES; DDRESCAT; DDSTRESC		Y	Y	Clinical Data Table	NA	NA	No Data in Source present with relevance to	
13	DEATH	5	cause_concept_id	DV	DVDECOD; DVTERM		Y	Y	Clinical Data Table	NA	NA	No Data in Source present with relevance to Proposed variables	
13	DEATH	5	cause_concept_id	EG	EGORRES; EGORRESU; EGSTAT; EGSTRESC; EGSTRESN; EGSTRESU		Y	Y	Clinical Data Table	NA	NA	No Data in Source present with relevance to Proposed variables	
13	DEATH	5	cause_concept_id	FA	FAORRES; FAORRESU; FASTRESC; FASTRESN; FASTRESU		Y	Y	Clinical Data Table	NA	NA	No Data in Source present with relevance to	
13	DEATH	5	cause_concept_id	IE	IEORRES; IESTRESC		Y	Y	Clinical Data Table	NA	NA	present with relevance to	
13	DEATH	5	cause_concept_id	IS	ISORRES; ISORRESU; ISSTAT; ISSTRESC; ISSTRESN; ISSTRESU		Y	Y	Clinical Data Table	NA	NA	present with relevance to Proposed variables	
13	DEATH	5	cause_concept_id	LB	LBORRES; LBSTRESC; LBSTRESN; LBSTRESU		Y	Y	Clinical Data Table	NA	NA	present with relevance to	
13	DEATH	5	cause_concept_id	MB	MBORRES; MBORRESU; MBRESCAT; MBSTAT; MBSTRESC; MBSTRESN; MBSTRESU		Y	Y	Clinical Data Table	NA	NA	No Data in Source present with relevance to Proposed variables	

Figure 6 Manual Curation of Proposed SDTM Mappings by BRIDG

UNCAPTURED MAPPING

In other instances, when I searched for an OMOP variable within the BRIDG model, the SDTM data set/s that the data belongs to did not appear among the list of SDTM variables that that variable could map to. For example, searching for the variable, 'death_date' from the 'DEATH' table will only bring up the SDTM data set AE (Adverse events) and the variables, 'AEENDTC', 'AESTDTC', and AEDUR but not the 'DM' data set and the variables, 'DTHDTC' and 'DTHFL', where the date of death of a patient also belongs. Figure 7 illustrates this uncaptured mapping.

TABLE_NUM	OMOP_TABLE	VARIABLE_NUM	OMOP_VARIABLE	SDTM_DATASET	SDTM_VARIABLE	BRIDG_SEARCH_RESULT	DATA_AVAILABLE	IN_OMOP_CDM	OMOP_TABLE_TYPE	SDTM_DATASET_M	SDTM_VARIABLE_M
13	DEATH	2	death_date	AE	AEDUR; AEENDTC; AESTDTC		Y	Y	Clinical Data Table	AE	AEDUR; AEENDTC; AESTDTC
13	DEATH	2	death_date				Y	Y	Clinical Data Table	DM	DTHDTC; DTHFL

Figure 7 Mapping Uncaptured by BRIDG

CONTROLLED TERMINOLOGIES

OMOP CDM stores data in the form of standardized values called concepts. They are coded into Concept IDs (variables end in 'concept_id') using standard vocabularies, which are available in the OHDSI vocabulary repository called ATHENA and can be decoded using ATHENA again. For example, the value 8516 in 'race_concept_id' column in the 'PERSON' table, can be decoded as 'Black or African American'. Although values from the original source, called source values (variables end in 'source_concept_id' or 'source_value') are available within the CDM, their use for analytic purposes is discouraged (Observational Health Data Sciences and Informatics, 2021). In some records, however, it appears that the source values could not be coded using OMOP standard vocabulary and coded source data (source concept ID) was used verbatim (the 'race_concept_id' column contains the same values as race_source_concept_id- 2000001401, 2000001402, 2000001405). These values could not be decoded using ATHENA and I had to rely on the 'race_source_value' column to decode them to 'Unknown', 'Other' and 'Unable to Obtain'. Figure 8 illustrates this challenge using a modified PERSON table (with selected columns and records) to highlight the challenge.

ATHENA Interface

ID	CODE	NAME	CLASS	CONCEPT	VALIDITY	DOMAIN	VOCAB
8516	3	Black or African American	Race	Standard	Valid	Race	Race
8515	2	Asian	Race	Standard	Valid	Race	Race

person_id	race_concept_id	race_source_value	race_source_concept_id
-7.7552E+17	8516	BLACK/AFRICAN AMERICAN	2000001406
-8.9708E+18	8527	WHITE	2000001404
7.70819E+18	2000001401	UNKNOWN	2000001401
7.48984E+18	2000001401	UNKNOWN	2000001401
3.49862E+17	2000001402	OTHER	2000001402
-1.095E+18	2000001402	OTHER	2000001402
5.89442E+18	2000001405	UNABLE TO OBTAIN	2000001405
-3.7805E+18	2000001405	UNABLE TO OBTAIN	2000001405

PERSON table

Figure 8 Some Records Contain Same Value as 'race_source_concept_id' in 'race_concept_id'

NON-STANDARD VARIABLES

As is natural in the SDTM mapping process, a lot of associated variables that cannot be mapped to the parent SDTM domain are moved to the supplementary (SUPP) domains. BRIDG does not recommend or predict which information must be moved to the supplementary domains. As seen above, OHDSI recommends mapping only concept IDs to the parent domain. In order to maintain full transparency of the transformation however, I have saved source variables in the SUPP domains. There is a need to come up with a consistent naming strategy for the non-standard variables (NSVs) that reside in these SUPP Domains.

NON-AVAILABILITY OF DATA IN SOME OMOP TABLES

MIMIC IV v0.9 demo data in the OMOP CDM v5.3.1 has certain limitations that posed challenges during the mapping process. Some of the tables in the data set do not have any rows (data) in them other than column headers because MIMIC data was not transformed into those tables by Kallfelz et al (2021). These tables cannot be accurately mapped to SDTM because there is no data to provide context to the mapping.

DE-IDENTIFICATION OF DATA

The MIMIC IV OMOP data set contains de- identified data to protect patient privacy and data has been scrambled. This scrambling has introduced some artifacts. Therefore, the transformed data available in the MIMIC demo data is not always a true representation of real values. For example, SUBJID has some negative values and year has values very much into the future like 2030.

CONCLUSION

In this paper, I have elucidated the process of mapping MIMIC IV v0.9 demo data in OMOP CDM v5.3.1 to CDISC SDTM v3.2 using the CDMH project as a jumping-off point, documented several associated challenges and provided suggestions to address some of these challenges. This document, that shows the end-to-end mapping of RWD to SDTM in a stepwise manner, could serve as a primer to the process of ETL.

All files and associated details are available at the author's GitHub (see Contact Information).

REFERENCES

1. Morris, Z. S., Wooding, S., & Grant, J. (2011). The answer is 17 years, what is the question: understanding time lags in translational research. *Journal of the Royal Society of Medicine*, *104*(12), 510–520. https://doi.org/10.1258/jrsm.2011.110180
2. *Framework for FDA's Real World Evidence Program.* (2021, December). FDA. https://www.fda.gov/media/120060/download
3. CDISC. (n.d.). *Real World Data | CDISC*. Www.Cdisc.org. Retrieved April 2, 2022, from https://www.cdisc.org/standards/real-world-data
4. *Data Standards for Drug and Biological Product Submissions Containing Real-World Data: Guidance for Industry.* (2021, October). FDA. https://www.fda.gov/media/153341/download
5. Observational Health Data Sciences and Informatics. (2021, January 11). The Book of OHDSI. Www.Ohdsi.Org. Retrieved April 4, 2022, from https://ohdsi.github.io/TheBookOfOhdsi/
6. *Common Data Model Harmonization (CDMH) and Open Standards for Evidence Generation.* (2020, August). FDA, NIH, Office of the National Coordinator for Health Information Technology (ONC). https://aspe.hhs.gov/sites/default/files/private/pdf/259016/CDMH-Final-Report-14August2020.pdf
7. ver Hoef, W., Hastak, S., & Evans, J. (2019). *BRIDG Release 5.3.1 User's Guide*. CDISC, FDA, HL7, ISO, NCI.
8. *BRIDG | Biomedical Research Integrated Domain Group*. (n.d.). Https://Bridgmodel.Nci.Nih.Gov/. Retrieved April 7, 2022, from https://bridgmodel.nci.nih.gov/

9. Kallfelz, M., Tsvetkova, A., Pollard, T., Kwong, M., Lipori, G., Huser, V., Osborn, J., Hao, S., & Williams, A. (2021). MIMIC-IV demo data in the OMOP Common Data Model (version 0.9). *PhysioNet*. https://doi.org/10.13026/p1f5-7x35.
10. Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P. C., Mark, R., ... & Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation [Online]. 101 (23), pp. e215–e220.
11. Johnson, A., Bulgarelli, L., Pollard, T., Horng, S., Celi, L. A., & Mark, R. (2021). MIMIC-IV (version 1.0). PhysioNet. https://doi.org/10.13026/s6n6-xd98
12. OHDSI. (n.d.). GitHub - OHDSI/WhiteRabbit. GitHub. Retrieved April 9, 2022, from https://github.com/OHDSI/WhiteRabbit
13. ver Hoef, W., Hastak, S., & Evans, J. (2019b, September). BRIDG Model (Release 5.3.1). CDISC, FDA, HL7, ISO, NCI. https://bridgmodel.nci.nih.gov/

ACKNOWLEDGMENTS

I thank Bhargav Koduru for his support and guidance in reviewing this paper and providing valuable input in understanding the process of SDTM mapping and transformation.

CONTACT INFORMATION

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

Ashwini Yermal Shanbhogue
ash23shan@yahoo.com
https://github.com/ash23shan