Paper DS04

From Source to Submission: Getting the Best of Multiple Standards

Berber Snoeijer, ClinLine, Leiderdorp, The Netherlands
Jules van der Zalm, OCS Life Sciences, Den Bosch, The Netherlands

ABSTRACT

Enabling an end-to-end data flow of real-world data from source to regulator involves a number of different data standards. As FHIR's [1] main use is the exchange of electronic health record information through JSON formatted files, it falls short in data selection and fit-for-purpose evaluation. On the other hand, direct transfer to CDISC SDTM [2] would entail extensive mapping, which might not be necessary for the majority of patients. Moreover, SDTM format is not optimal for fit-for-purpose assessments and imputation.

We propose aligning and standardizing the essential intermediary stage of data transformation. The OMOP Common Data Model [3] serves this purpose very well as it includes a number of source traceability variables and standardized mapping of concepts. In this presentation, we will demonstrate how we add custom variables to the OMOP standard to allow for full end-to-end traceability and seamless mapping to SDTM.

INTRODUCTION

Real-World Data (RWD), according to the FDA definition, is data that is related to patient health status or the delivery of health care routinely collected from a variety of sources [4].

This data is collected in various systems and formats such as Electronic Health Records (EHR) or claims. The data may come from different entities like hospitals and general practitioner practices or more specialized sources like nursing homes, thrombosis services or physiotherapy practices. Each of them uses their own fit for health care purpose information systems and corresponding data model in which their data is stored. This implies that there is a lot of variation between and within domains regarding the data collection practices and data models used. Combining the data from all these different sources to one common data model (CDM) will aid the ability to select patients, assess trial feasibility and assess fit for purpose in a consistent and efficient manner.

After patients and data are selected according to trial specific requirements, they can be optionally matched to interventional control arm patients and then included in the final analysis. For this purpose, it is helpful if both the data from the real-world external control arm and the data from the interventional arm are in the same format (e.g. standard). Therefore, the selected data of the real-world data patients needs to be transformed to CDISC SDTM.

As stated in the FDA guidance, sponsors should address the procedures used to ensure completeness and accuracy of study data, as well as processes for data accrual, curation, and transformation over the data life cycle [4]. This indicates that all data mappings and all transformations need to be documented. The EMA data quality framework states that traceability refers to data presenting the knowledge of how data came to be, what source it originated from, and what processing it went through before appearing in its current form [5]. With that, traceability falls within reliability in that it connects what is measured at the source with the actual data presented and used for analysis.

Our paper shows how we efficiently document and perform data mappings ensuring transparency and traceability from source to submission datasets.

STANDARDISATION

For submission and integrated analysis purposes, CDISC SDTM and ADaM are required and beneficial. However, they are less suited for patient selection, data selection or fit for purpose assessments. Contrary to that, the OMOP data model [3] is specifically created for observational research and therefore well-fit for RWD standardization. This OMOP data model includes source as well as standardized variables aiding the traceability to the source, efficiently select or even automate the selection of patients, and address any biases that may occur due to different data collection practices.

Although well suited for this purpose, we do not need all the datasets of the OMOP Common Data Model (CDM) in the data selection phase. For example, specific observational research datasets like observation_period, cohorts or episodes are not needed in the data selection phase. On the other hand, although OMOP is flexible and covers source as well as standardized variables, not all information from source that is beneficial for study design is included in OMOP. Therefore, our proposed standardization mimics OMOP by reducing the datasets and variables that are not needed and adding some additional source-specific traceability variables. This OMOP like data can accurately be transformed to full OMOP and/or to CDISC SDTM. To indicate the difference with OMOP, we called it OMOP+. This OMOP+ version very well resembles standards used by real-world data vendors. So, for the purpose of this paper, if applicable, OMOP+ can be replaced by any data vendor CDM.

Apart from the variation in data models and storage, a variation in dictionaries can be used for real-world data as well which is different from clinical study standards. While in clinical research, we use MedDRA for conditions, WHODrug for medications, and CDISC for assessments and units, in the real-world setting, standards like SNOMED, LOINC, RxNorm and others are used. There is overlap in terminology between these dictionaries. However, there are also a lot of alternate variations and different levels of concept coverage between these dictionaries. For example, the relatedness or severity could optionally be included in a SNOMED term while it is a separate variable in SDTM. Therefore, terminology or code mapping is a separate activity in addition to variable mapping which is very important for the interoperability and reliability of the data. The documentation of this mapping and showing the lineage from source to submission values in this mapping is key to adequately be able to utilize the data.

MAPPING

Mapping data from one standard to another is something best done through programming. There's a variety of programming languages to choose from – most commonly used are SAS, R, and Python. Programming offers the benefit of traceability by nature; the source data remains intact, while the mapping steps are usually clear from the code. Programmed conversion is also a repeatable process so that when your data changes at the source (e.g. because more patients are included) all you do is run the same program and the conversion is done.

The mapping process from source to OMOP is an ETL (extract-transform-load) process like most other – traditional – data conversions. Most important differences when compared to data mapping in clinical trials is that real-world data is more heterogeneous due to the many different sources. To tackle this challenge, a recommended first step is to perform a pre-analysis of the source data to find differences and commonalities between various sources of data. This is best performed as a programmatic step, again in order to make this reproducible. The result of such a pre-analysis could look like this:

Dataset	Variable	Label	Data Type	Value	Data source 1	Data source 2	Remarks
DEMO	RACE	Race	Char	Asian	Y	Y	
DEMO	RACE	Race	Char	African American		Y	
DEMO	RACE	Race	Char	Black or African American	Y		
DEMO	RACE	Race	Char	Other	Y	Y	
DEMO	BIRTH_YY	Birth year	Char			Y	Values are too
DEMO	BIRTH_YY	Birth year	Num		Y		Values are too

While the code does provide insights in how data is mapped from source to target, it's not a convenient document to read. It is recommended – and often required – to further document all mapping in a mapping specification document, such as a spreadsheet. This spreadsheet will provide a variable-by-variable description of the conversion, detailing the derivation of each target variable.

Hardcoding of value conversions in the programs must be avoided. Any values in the source that need to be changed going into the output dataset (e.g. recordings, translations, abbreviations) should be documented in an external document, that is then programmatically referred to and used in the program. This provides full transparency of changes made to values and allows for programmatic checks to be embedded which warn the user of any mismatches or missing recode information.

DEFENSIVE PROGRAMMING

Defensive programming is a key practice in programming, especially in the field of clinical research and even more so when dealing with real-world data, due to its heterogenous nature. Defensive programming emphasizes writing code that anticipates and deals with potential errors or changes in the source data structure and its values. By assuming that unexpected events – such as missing data, out-of-range values, or system errors – cannot be avoided, defensive programming ensures that the code can handle these scenarios and warn the programmer or user proactively when such events occur.

One fundamental practice of defensive programming is data validation. Before processing the data, it should be checked for completeness and consistency, and any anomalies should be flagged early to prevent issues further down the line. Additionally, coding with flexibility in mind – such as having re-codings of values in an external file – helps future-proof programs against unforeseen changes in data structure or values. Furthermore, defensive programming for real-world data also includes the use of text recognition to find uncoded entries or unsuspected codes for data that potentially needs to be included.

Another important aspect is error handling. Rather than allowing the programs to crash or produce invalid outputs without warning, defensive programming helps catch data issues by providing clear error and warning messages and by performing a clean exit from the program run, so that no potentially invalid output is generated.

CHALLENGES IN REAL-WORLD DATA MAPPING

Working with real-world data imposes many challenges that programmers don't face working with most clinical data.

Real-world data often lacks the structured and consistent format of clinical trial data. Clinical programmers may be used to receiving their data in clearly defined separate datasets per case report form while Real-world data may come in a single large file or in a single file per patient. While OMOP and FHIR structures are comprehensive, they also have greater variability, reflecting the variety in data sources such as electronic health records, claims data, and registries. For example, in the FHIR structures, data may be duplicated, "nested" at different levels, and/or include unexpected categorization dedicated to therapeutic purposes. Data from different patients may come from different systems and thus have different data structures. This requires additional data preparation and processing to ensure consistency and compatibility across sources.

The controlled terminology and coding of events, medications and assessments is different from those that we use in clinical trials. A further complication is that a wide variety of codes representing different levels of detail can be used in real-world data. Identifying the information matching the intended assessments can therefore be challenging with the risk to incorrectly map source information to the target standardized variables or to mis information that should have been mapped to these variables.

A crucial task in working with real-world data involves identifying and managing data that may not be directly pertinent to the research objectives but could still be sensitive or excessive in nature. Unlike clinical trials, where data collection is confined to the specifications outlined in the protocol and SAP, RWD often encompasses a wider range of information obtained from healthcare records, claims, or registries. This can inadvertently lead to the inclusion of personally identifiable information (PII) or sensitive variables that do not necessarily contribute to the analysis. For instance, RWD sources may contain fields such as social security numbers, specific addresses, or unrelated medical histories, posing risks to patient privacy if not handled appropriately. As programmers, it is important to assess and filter data to retain only those variables that are directly relevant to the fit for purpose assessments and/or study aims. Adopting effective data minimization strategies, flagging sensitive fields, and closely collaborating with data privacy teams can help ensure compliance with privacy regulations while safeguarding data integrity.

Especially when registries are concerned, the corresponding documentation may be needed to fully understand the data that is stored. Some of the relevant traceability information is only available in the documentation or in standard dictionaries. For example, a response to a questionnaire might be filled as answer 1, 2, 3 in the database without the actual decoded answer is stored. The decoded answer then needs to be retrieved from the documentation.

After fit for purpose assessments are performed, the selected patients and corresponding data needs to be mapped to SDTM to allow for the potential comparison with an interventional arm and/or regulatory submission. The definition of the Index date is in that regard most crucial, which often is set to the date when the eligibility criteria are (first) met. Based on this index date SDTM variables like RFSTDTC (reference start date/time), VISITNUM (visitnumber) and decisions on target datasets can be made (e.g. MH: Medical History or AE: Adverse events).
For real-world data visit windows are to be used to identify the periods for which data is relevant for the study objective. However, for each intended visit it may be that there are 0, 1 or more visits within that window. In case of more visits in a window, it needs to be decided how visit numbers are handled. Especially since some measurements needed for evaluation might be available in one of the visits within the window while not in the other visit.

Finally, the mapping of information to SDTM controlled terminology might be challenging as well depending on the kind of data. The transformation from SNOMED to MedDra is handled quite well using the OMOP concept relationship dataset while the transformation to RxNorm is not included. The transformation of LOINC to CDISC controlled terminology is handled by the CDISC dictionary while the unit transformation from UCUM to CDISC CT is not handled.

CONCLUSION

Traceability of the data flow from source to submission is key in utilizing real-world data from source to submission ready datasets. Using an intermediate observation common data model like OMOP is beneficial for data selection and fit for purpose assessments. The data and patients needed for the specific study then can be used for SDTM mapping and regulatory submission purposes.
This process and corresponding traceability will be ensured by source data pre-analysis, defensive programming and standardized mapping documentation.

REFERENCES

[1] HL7, "HL7 FHIR," [Online]. Available: https://hl7.org/fhir/.

[2] CDISC, "CDISC SDTM," [Online]. Available: https://www.cdisc.org/standards/foundational/sdtm.

[3] OHDSI, "OMOP CDM v5.4," [Online]. Available: https://ohdsi.github.io/CommonDataModel/cdm54.html.

[4] C. a. O. U.S> Department of Health and Human Services Food and Drug Administration CDER, "Real-World Data: Assessing Electronic Health Records and Medical Claims Data to Support regulatory Decision-Making for Drug and Biological Products - Guidance for Industry," July 2024. [Online]. Available: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/real-world-data-assessing-electronic-health-records-and-medical-claims-data-support-regulatory.

[5] EMA, "Data quality Framework for EU medicines regulation," October 2023. [Online]. Available: https://www.ema.europa.eu/system/files/documents/regulatory-procedural-guideline/data-quality-framework-eu-medicines-regulation_en_1.pdf.

CONTACT INFORMATION

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

Author Name: Berber Snoeijer
Company: ClinLine
Address: Van Der Valk Boumanweg 15, 2352 JA Leiderdorp
Work Phone: +31 6 21197621
Email: b.snoeijer@clinline.eu
Website: www.clinline.org

Author Name: Jules van der Zalm
Company: OCS Life Sciences
Address: Ruwekampweg 2g
Work Phone: +31 73 523 6000
Email: jules.vanderzalm@ocs-consulting.com
Website: www.ocs-lifesciences.com