Paper DS02

Using CDISC USDM and BCs to select data from alternative data sources for regulatory purposes

Berber Snoeijer, ClinLine, Leiderdorp, The Netherlands

ABSTRACT

As the use of alternative data sources (ADS), including real-world data (RWD), grows in regulatory decision-making, standardized methods are needed to assess their suitability. The CDISC Unified Study Definitions Model (USDM) and Biomedical Concepts (BCs) provide a structured approach to identify and evaluate relevant data. This paper describes how USDM and BCs support two regulatory use cases: study feasibility assessments and RWD studies. USDM enables alignment between protocol requirements and available data, facilitating faster, reproducible feasibility evaluations. BCs ensure consistent interpretation of clinical concepts across sources, enhancing the validity of RWD analyses. Together, these tools help streamline data selection, improve transparency in data lineage, and support regulatory-grade evidence generation.

INTRODUCTION

Alternative data sources to clinical trial data include:

- real-world data,
- data collected from digital health technologies (DHT), and
- historical clinical trial data.

Note that the definition of Real-world data according to the FDA is “data relating to patient health status and/or the delivery of health care routinely collected from a variety of sources” (1).

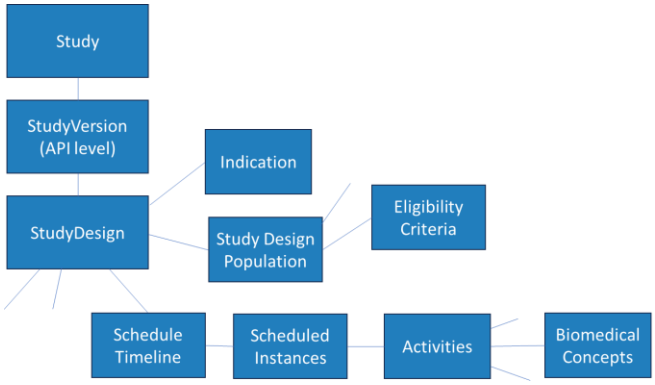
To improve the efficiency of clinical trials and correspondingly reduce costs and reduce the timelines for new drug submissions, these alternative data sources become more and more important in achieving that goal. For example, pragmatic trials, trials with an external control arm and decentralized trials are likely to include these alternative data sources. However, they may have different structures, use different coding dictionaries and may have different levels of completeness than we normally expect in a fully controlled clinical trial setting.To link the study design with these alternative data sources is therefore an imperative step in achieving the expected added value and efficiency.

The CDISC Unified Study Definition Model can hold any study design in a digitized format. This includes the digitized selection criteria, timing information, and all assessments and corresponding data needed for the evaluations in the form of Biomedical Concepts. Moreover, the model enables the linkage of the standardized concept coding we use for clinical trials with coding used by the alternative data sources.

USDM

The CDISC USDM data model can hold a complete study design in digitized format. In addition to that, multiple documented versions of the design can be stored as well in the USDM including cross-references to the digitized information. The figure below pictures the backbone of the data model starting with the overarching study concept, via a study version to one or more study designs. Each study design includes a number of features including eligibility criteria and the scheduled timeline representing the content of the study schedule of activities. The USDM model information, implementation guide and examples can be found on the CDISC ddf github: https://github.com/cdisc-org/DDF-RA.

Figure 1: USDM Backbone



PARAMETERIZED ELIGIBILITY CRITERIA

The USDM data model includes a parameterization feature in the form of syntax templates. This enables systems to tag pieces of information presented in the objectives, endpoints, footnotes and eligibility criteria. Via the corresponding tag mapping feature, they can be linked to the corresponding parameterized items stored elsewhere in the data model. The figure below shows what eligibility criteria can be tagged in order to automatically select the corresponding data from alternative data sources. Relevant time frames are tagged in blue, measurements and assessments are tagged in green, and expected results are tagged in pink. The corresponding information is stored respectively in the timeline and timing part of the USDM, the biomedical concept part of the USDM and in a corresponding expected response extension of the USDM. The Biomedical Concepts stored in USDM are linked via more general activity descriptions to the timeline and corresponding timing. This combination of information enables us to precisely select the information required from the source.

Figure 2: Tagging options for eligibility criteria.

2. Age of ≥ 18 years
3. Positive SARS-CoV-2 (2019-nCoV) test within 14 days prior to randomization
4. Hospitalized for SARS-CoV-2 (2019-nCoV)
5. Bilateral pneumonia on chest x-ray or CT
6. *[Original Criteria Deleted]*
7. At least one of the following within 7 days prior to randomization:
   - Ferritin > 500 ng/mL
   - CRP > 5 mg/dL
   - D-dimer > 1,000 ng/mL
   - LDH > 250 U/L
   - Fever - a measured temperature of at least 100.4 °F [38 °C]

https://cdn.clinicaltrials.gov/large-docs/69/NCT04447469/Prot_SAP_000.pdf

SOURCE DATA SELECTION

The information stored in USDM can drive the selection of eligible patients based on the source data model and format. The availability of information and timing of assessments are key to enabling the selection of the right patients. For each of the required selection criteria, it needs to be defined at each available timepoint whether the value is still available and valid and whether if available it fulfills the eligibility criteria. To enable that, the information needs to be aligned and ordered in time. Based on that ordering, for each of the criteria, the status at that timepoint can be assessed:

- Green: Information available and according to expected response.
- Orange: Information not available.
- Red: Information available but not according to expected response.

Patients with 1 or more red flags at a timepoint cannot be included in the study at that timepoint according to the eligibility criteria. For patients without red flags, an increasing number of green flags are indicative of their potential to be eligible for the study. The table below shows how the eligible status can change through time. At the start, in March 2018, no indication is available (orange flag) and the patient is still too young to be eligible (red flag). In addition, one of the lab values is out of the expected range (red flag). Another lab value is within range (green flag). A few months later, in August 2018, the patient gets the study indication diagnosis (green flag) but has a score which is not within the expected inclusion range (red flag). In the next year (February 2019), the patient ages to become within the expected age range (green flag) and 1 month later all other lab scores are within range (green flag). Finally, in June 2019, the indication score is also within range which indicates that the patient is potentially eligible for the study although still some information is missing (number of orange flags). To complicate the process, lab values and other assessments might have an expiration data as indicated by criteria specifications like ‘within the last 6 months’. Therefore, if no new measurements are taken within that time it results in green and red flags turning orange and the certainty that a patient might be eligible for a study is decreasing.

Patient	Date	Event	Number of Red flags	Number of Green flags	Number of Orange flags
001	25-03-2018	Lab measurements	2	1	5
001	04-08-2018	Indication and score	3	2	3
001	02-02-2019	Ageing	2	3	3
001	25-03-2019	Lab measurements	1	4	3
001	24-06-2019	Indication score	0	5	3
001	25-09-2019	Lab measurement values expired	0	3	5

Using this method for each patient and each timepoint, it can be assessed whether and when a patient could be eligible for a study.

The number of potentially eligible patients and the corresponding certainty can be calculated and can directly inform the study design team about potentially severely limiting criteria. Doing this in the study design phase, may prevent costly and cumbersome study amendments later in the process. Another use case for this is the direct selection of potentially eligible data patients and their corresponding data.

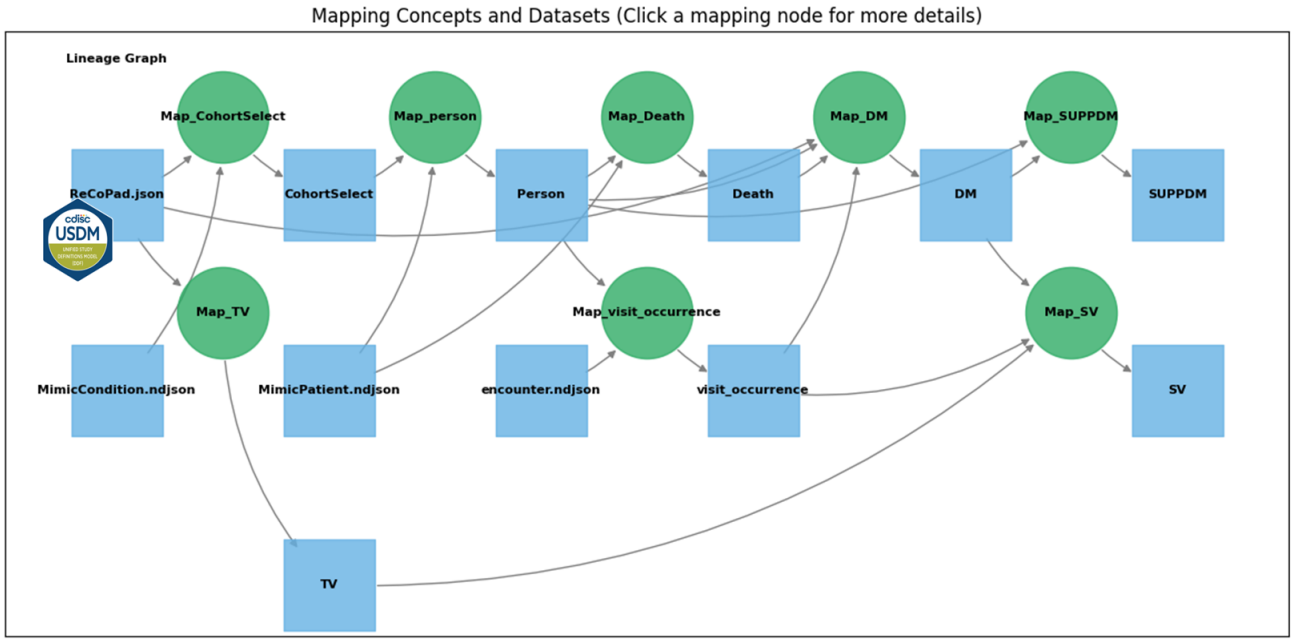
DATA LINEAGE

Data lineage refers to the process of tracking how data is generated, transformed, transmitted and used across a system over time. It documents data's origins, transformations and movements, providing detailed visibility into its life cycle. This process simplifies the identification of errors in data analytics workflows, by enabling users to trace issues back to their root causes. (Wikipedia). When using data from alternative data sources for regulatory requirements, regulatory agencies requirements include to provide sound and complete lineage information (2) (3).
This includes information of how data is converted from source to submission and all the steps in between like data selection, matching and imputations. The mapping information needs to include the

- location, names, variables and formats for data entities in each step of the process, and
- data relationships and transformations in each step of the process.

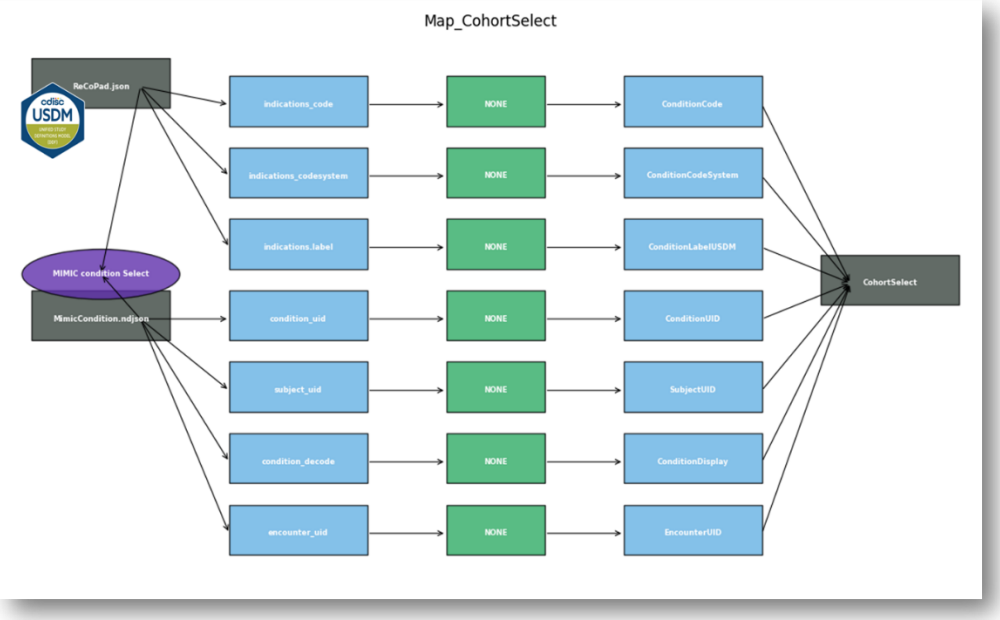
By documenting this information in a consistent manner, automated visualizations can be build upon it to show the lineage process in different levels of detail. Figure 3 shows the output from the ClinLine Open Source data lineage tool which can be accessed via https://github.com/ClinLine/RWD_Lineage_Tool. It shows how information from a USDM API file informs the selection of patients (CohortSelect) in real-world data source files in JSON format (MimicCondition, MimicPatient and Encounter). The follow-up steps in creating intermediate (OMOP-like) datasets to finally CDISC SDTM is shown in the lineage diagram as well.

Figure 3: Visualized data lineage of demographic and visit information



Each selection and/or mapping step is an action in the process (visualized by the green circles in Figure 3). The details of these mapping steps are further visualized as presented in Figure 4 for the selection of the cohort. It indicates that a data limiting join between the USDM input file and the condition file is used by the purple ellipse. Furthermore, the source variables used for the output as well as the resulting target variables are displayed. For each transformation the type of action is displayed.

Figure 4: Visualized data mapping from source to next step



Further information on the dataset linkage and transformations is visualized when clicking on the corresponding objects in the view.

CONCLUSION

The USDM is considered to be the starting point of the digital data flow (DDF). This enables us to automate the data selection from alternative data sources to submission. The corresponding data lineage needs to be documented and available for regulatory purposes. Visualizations like with ClinLine's open source data lineage tool help in understanding the different steps taken and in improving and optimizing this data flow.

REFERENCES

1. FDA. Considerations for the Use of Real-World Data and Real-World Evidence to Support Regulatory Decision-Making for Drug and Biological Products, Guidance for Industry. [Online] 8 2023. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/considerations-use-real-world-data-and-real-world-evidence-support-regulatory-decision-making-drug.

2. FDA CDER CBER OCE. Real-World Data: Assessing Electronic Health Records and Medical Claims Data to Support Regulatory Decision-Making for Drug and Biological Products - guidance for Industry. [Online] July 2024.

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

ACKNOWLEDGMENTS

With this paper I like to acknowledge CDISC for providing me the chance to take part in the USDM development process and our trainee Vedika Grover who created the first version of the RWD lineage tool.

RECOMMENDED READING

CDISC USDM implementation guide: https://github.com/cdisc-org/DDF-RA/blob/main/Deliverables/IG/USDM-IG.pdf
TransCelerate white paper: Practical Approach to implementing Digital Data Flow: https://transcelerate.github.io/ddf-home/documents/white_paper/DDF_Practical_Approach_to_Implementation.pdf

CONTACT INFORMATION

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

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