Paper RE01

Harmonization and Interoperability between OMOP CDM and CDISC SDTM

Dhruba Sikdar, Johnson and Johnson, High Wycombe, UK

ABSTRACT

The growing interest in observational data, a.k.a Real World Data (RWD) to assess safety, effectiveness and economic value of healthcare solutions have created some practical challenges of using this in research, whereby disparate data sources make it difficult and time consuming to compare and exchange between various systems. Conversely, the ever-increasing use of RWD in support of regulatory submissions has created similar challenges for the Health Authorities across the globe. This problem prompted the need for a common data model (CDM) to help researchers and to aid regulatory decision making. While Clinical Data Interchange Standards Consortium's (CDISC's) Study Data Tabulation Model (SDTM) is the de facto standard for submission to US FDA and many other major global regulatory authorities, it has its limitation while dealing with observational data. A common data model such as the Observational and Medical Outcomes Partnerships (OMOP) common data model can help overcome many of these limitations by standardizing the structure, format, and terminologies of disparate data sources, enabling the execution of common analytical code across a federated data network.

This paper examines the harmonization and interoperability between two widely used data models, the CDISC SDTM and OMOP CDM and how they can complement each other.

INTRODUCTION

Real World Data (RWD) comprises of variety of sources like electronic health records (EHRs), medical claims and billing data, data from product and disease registries, patient-generated data etc. Analyses of these data can generate huge amount of insight and evidence to get regulatory approvals for new indications of a drug that has been approved already or to fulfil post approval study requirements. Also, the collection and analyses of real-world data to provide real-world evidence has been identified as a major priority in the twenty-first Century Cures Act of 2016 and in the concept paper for the proposed Cures Act 2.0 in the US, with the aim of accelerating drug development and innovation.

However, RWD as the name suggests, are collected from disparate sources, in disparate formats, languages, vocabularies all using various methods and algorithms, de- identified using various methods and is finally aggregated into disparate kinds of data bases. Therefore, this data does not have uniform structure, format, or terminology, thus making it difficult and time consuming to compare or exchange it between information systems.

A logical solution to address this problem would be to store data in a standardized format, such as a common data model (CDM). A CDM is an informational model that allows transformation of data contained in different databases to a common format, in which all coding and vocabulary are pre-specified and standardized and can be applied to all data irrespective of product or therapy area.

Several data networks have been established with a view to improving the standardization of observational healthcare data and to establish shared open-source tools that facilitate collaborative research. Examples include the Sentinel initiative of the US Food and Drug Administration (FDA) and the European Union Adverse Drug Reactions (EU-ADR), the National Patient-Centered Clinical Research Network (PCORnet). Similarly, the Observational Medical Outcomes Partnership (OMOP), a public-private partnership between the FDA, academic institutions, database owners, and pharmaceutical companies, and administered by Observational Health Data Sciences Informatics (OHDSI), is one such data network. It is an open-source electronic healthcare database system, aimed to utilize existing observational healthcare data for study treatment pathways, comparative effectiveness, safety, and patient-level prediction. The European Medicines Agency (EMA) had used the OMOP common data model to conduct multicenter cohort studies on the use of medicines in patients with coronavirus 2019 (COVID-19) [1].

Most of the above-mentioned networks have developed their own CDM to meet their aims. 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. These data networks, however, were developed for different purposes and contain data from different sources. Therefore, though they each have their own common data model, the data model is 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. US Food and Drug Administration (FDA) has mandated the use of Clinical Data Interchange Standards Consortium's (CDISC's) Study Data Tabulation Model (SDTM) for the purpose of regulatory submission for marketing approval of medical products. CDISC-SDTM is also the de facto standard for many other major health authorities across the world. Hence there is a need, at least for the purpose of regulatory submission that to extract, transform, and load (ETL) data from these networks' CDM to CDISC SDTM.
Studies comparing different CDMs showed that the OMOP CDM met most of the desired criteria (including content coverage, integrity, flexibility, ease of querying, standards compatibility, and ease/extent of implementations, privacy, and linkage) for data sharing and across use cases [2]. In fact, OMOP is fast becoming the choice for standardization of real-world health care data to enable observational research in a standardized way.
Therefore, it is imperative to establish harmony and interoperability between OMOP CDM and CDISC SDTM for RWD and RWE to be considered submission grade to various health authorities.

TALE OF TWO DATA STANDARDS: CDISC SDTM AND OMOP CDM

CDISC SDTM (Study Data Tabulation Model) and OMOP CDM (Observational Medical Outcomes Partnership Common Data Model) are both data models used in healthcare and clinical research settings, but they have different focuses and applications.

	CDISC SDTM	OMOP CDM
Purpose and Scope	SDTM is primarily designed for organizing and standardizing clinical trial data. It provides a standardized structure for the representation and exchange of data collected during clinical trials, facilitating data integration and analysis.	CDM, on the other hand, is focused on observational healthcare data. It aims to standardize and harmonize real-world data from various sources, such as electronic health records and insurance claims databases, for observational studies and comparative effectiveness research.
Data Structure	Organizes data into specific domains, such as demographics, adverse events, medical history, and laboratory findings. It defines the structure and metadata for each domain, ensuring consistency and interoperability across clinical trial datasets.	Structures data into specific entities, such as patients, observations, conditions, procedures, and drug exposures. It establishes relationships and hierarchies among these entities, allowing for in-depth analysis of healthcare data and longitudinal studies.
Data Elements	Specifies a predefined set of standardized variables and controlled terminology to represent common data elements collected in clinical trials. It provides clear definitions for these variables and ensures that the data is organized and presented consistently across trials.	CDM does not mandate specific data elements, but rather provides a framework or schema for organizing diverse healthcare data. It allows for flexibility in data representation and enables researchers to map their specific data elements to the CDM structure.
Use Case	Is widely used in clinical research, pharmaceutical industry, and regulatory submissions. It facilitates the analysis and reporting of clinical trial data, ensuring consistency and compliance with regulatory requirements.	Is primarily used for real-world data analysis, population health research, and comparative effectiveness studies. It enables researchers to integrate and analyse large-scale observational healthcare data from various sources.

It's important to note that SDTM and CDM serve different purposes in the healthcare and research domains. The choice of which model to use depends on the specific requirements and objectives of the study or analysis being conducted.
.

WHY INTEROPERABILITY AND HARMONIZATION MATTER BETWEEN DATA STANDARDS

FDA has published their guidance document for data standards required for submitting RWD [3]. A high-level summary of this indicates that the expectations for submission of study data derived from RWD is not very different to that of traditional clinical trials. Submission of RWD still need to be transparent in data curation, data transformation, mapping and to be well documented in SDRG and /or Define XML. They still need to adhere to Study Data Guidance and the supported data standards listed in Data Standards Catalog just as in traditional clinical trials. This creates additional challenges for sponsors when it comes to submitting RWD using CDISC standard. Example of such challenges include management of semantic concepts (terms), inconsistent format, inconsistent coding or miscoding (e.g., drugs or diagnoses), changes in data collection or coding practices during study conduct (e.g., International Classification of Diseases-9 (ICD-9) and ICD-10 codes), handling missing information (either because information is not typically recorded in health care settings or due to inconsistent data entry) [3]. Most of these challenges stem from the fact that RWD, by definition, is not collected under the supervision of a protocol and by research study staff. In addition, the business cases for using RWD and RCTs are different. RWD is not designed or collected for research purposes. As a result, real world databases are purpose built for health care providers, not researchers. Furthermore, these databases utilize different data standards that contain different concepts and coding systems compared to those defined within CDISC and used in regulatory review [4]. Thus, fostering interoperability and harmonization between different standards can potentially provide seamless transition from one standard to another. Cases where multiple data sources (EHR, Registry, claims, DHT, clinical trials etc) are involved, establishing ways to harmonize various data standards on the above aspects will serve greater good for all parties – healthcare professionals, researchers, sponsors, regulators – alike.

SOME CONSIDERATIONS AND CHALLENGES TO ACHIEVE INTEROPERABILITY BETWEEN CDISC SDTM AND OMOP CDM

Interoperability between CDISC SDTM and OMOP CDM is possible through mapping and transforming the data from one model to the other. While SDTM and CDM have different structures and purposes, their interoperability can facilitate data integration and analysis across clinical trial and real-world data.
Following general steps can be taken to achieve interoperability between CDISC SDTM and OMOP CDM.

Data Mapping: Data elements from source SDTM domains can be mapped to the corresponding entities and attributes in the OMOP CDM schema and vice versa. This involves identifying equivalent or related concepts between the two models and establishing a mapping between SDTM variables and CDM entities. This process can be replicated on the other way

(OMOP to SDTM) too. There is wide divergence in terminologies used and their precise meaning between OMOP and SDTM. Examples range from the meaning and specific terms used for race/ethnicity, terminology systems for medications, and interpretation of the measurements of vital signs from health care records. Sponsors should select an appropriate mapping approach that best fits the characteristics of the data and the nature of the study [3].

For an OMOP to SDTM mapping, the first step is to extract, transform and load (ETL) data from their native standards into OMOP CDM by using some automated tool. W. Anderson et al (2023) used Edge Tool suite - an open-source software project from the OHDSI - on COVID-19 data from a pilot healthcare site for ETL purposes. Then they mapped this OMOP-CDM into SDTM using the OHDSI tools, including Athena and ACHILLES, along with the OHDSI ICD-10 code standardized vocabularies, derivation logic was generated to map a total of 21 concepts from the pilot dataset to a subset of variables and their SDTM counterparts [5].

Variable	OMOP Concept ID	OMOP Domain/Column	SDTM CT Submission Value	SDTM C-Code	SDTM Domain/Column
Age*	N/A	person.birth_datetime	N/A	N/A	dm.age
Sex	8532	person.gender_concept_id	M	C20197	dm.sex
Sex	8507	person.gender_concept_id	F	C16576	dm.sex

* indicates a derived element

Fig 1: An example of a mapping from OMOP CDM to SDTM. Source: https://www.ohdsi.org/wp-content/uploads/2023/10/3-AndersonBriefreport-Wes-Anderson.pdf

Challenges to this process include missing data at source, inconsistencies between OMOP and SDTM concepts, inconsistent measurement units from the pilot site and complex derivations of critical variables.

Transformation and Conversion: Once the mapping is established, data from source datasets (OMOP/SDTM) can be transformed and converted into the destination (SDTM/OMOP) format. This involves restructuring the data, conforming to the destination data structure, and ensuring that the required fields and relationships are maintained. Some examples of challenges that can occur at this stage as described by P. Biedermann et al (2021) [2] while transforming Pulmonary Hypertension (PH) registry data formatted in SDTM into OMOP CDM, is that the MedDRA codes that are used in the source SDTM have no direct translation to OMOP standardized vocabulary for conditions and laboratory data (SNOMED, LOINC) and, thus, a degree of manual mapping was required. Also, custom concepts were needed to capture information such as adverse event severity and causality, for which the OMOP CDM lacks functionality to capture concisely. Mapping of partial dates, mapping free text (e.g., from medication tables) are some other challenges faced by the authors.

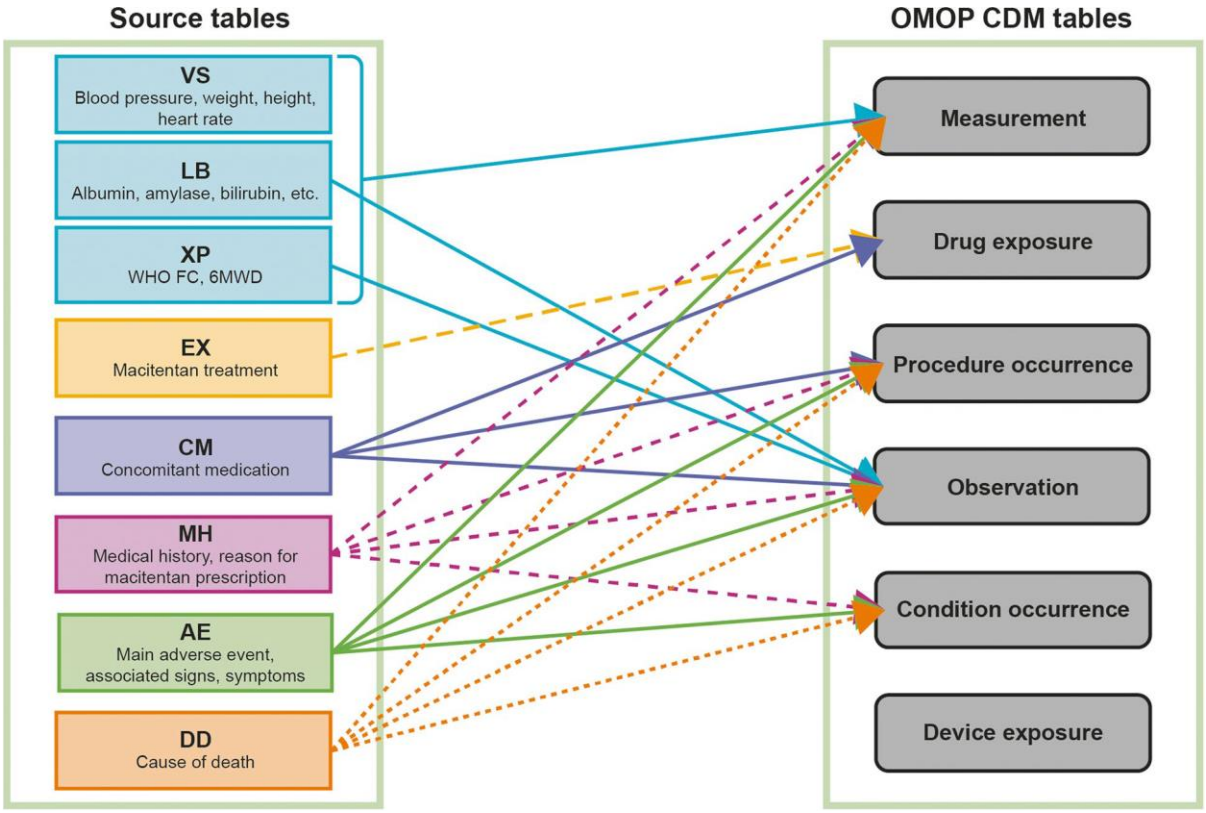

Fig 2: An example of mapping from SDTM to OMOP CDM. Source: https://pmc.ncbi.nlm.nih.gov/articles/PMC8565035/

Another way of conversion from one data model to another could be achieved through using an intermediary standard as depicted by A.Y Shanbhogue (2022) [6]. The author used Biomedical Research Integrated Domain Group (BRIDG) as an intermediary and mapped source data (OMOP CDM) to this data model first. The BRIDG international standard had already been mapped to the SDTM and thus OMOP to SDTM conversion was achieved, although a fair amount of manual curation was needed.

Terminology Alignment: SDTM uses CDISC-controlled terminology (CT) for standardizing data elements, while CDM uses various ID systems and coding vocabularies like SNOMED-CT, RxNorm, and LOINC. To achieve interoperability, the terminologies used in SDTM datasets must be aligned or mapped to the appropriate vocabularies used in the CDM. The concepts of semantic interoperability, incorporating standardized terminologies, promotes consistency and accuracy in data interpretation. However, the assumption behind this semantic interoperability needs to be clearly defined and validated. There is always a risk of losing information during mapping from one dictionary to other (e.g., SNOMED-CT to MedDRA) but this can be mitigated by employing standard, well reviewed mapping methods.

ETL Processes: Extract, Transform, Load (ETL) processes can be implemented to automate the conversion and loading of source format into the destination format. ETL pipelines can handle data extraction from SDTM datasets, perform the necessary transformations, and load the data into the OMOP CDM database or vice versa. For these processes, there are various open-source tools available within the community.

CONCLUSION

Although interoperability and harmonization between different standards is too big a subject to discuss here, but by considering the steps above and mitigating the associated challenges, it becomes feasible to integrate SDTM data from clinical trials with real-world data in the OMOP CDM format. This integration enables researchers to combine and analyze data from different sources, creating a comprehensive view of patient outcomes, treatment effectiveness, and safety profiles across clinical trial and observational data. It also helps regulators to have comprehensive evidence to make decisions.

It is worth noting that the specific mapping and transformation approaches may vary depending on the datasets, standards, and tools used. Additionally, expert input and careful consideration are necessary to ensure accurate mapping, maintain data integrity, and achieve meaningful interoperability between CDISC SDTM and OMOP CDM.

In future trials, where study design gets increasingly complex due to the availability of diverse data and clinical innovations, one single standard may not adequately represent a multitude of data sources. A hybrid approach where traditional clinical trial data in CDISC SDTM and observational RWD in OMOP CDM may have the potential to facilitate both broader and targeted research. In that scenario, establishing interoperability and harmonization between these two standards will be essential.

REFERENCES

[1] European Medicines Agency. COVID-19: EMA sets up infrastructure for real-world monitoring of treatments and vaccines. 2020. https://www.ema.europ a.eu/en/news/covid -19-ema-sets-infrastructure-real-world -monitoring -treatments -vaccines

[2] Biedermann P, Ong R, Davydov A, Orlova A, Solovyev P, Sun H, Wetherill G, Brand M, Didden E.M. (2021). “Standardizing registry data to the OMOP Common Data Model: experience from three pulmonary hypertension databases”. BMC Med Res Methodol Nov 2021. https://pmc.ncbi.nlm.nih.gov/articles/PMC8565035/

[3] FDA Guidance. (2023). "Data Standards for Drug and Biological Product Submissions Containing Real-World Data Guidance for Industry". December 2023. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/data-standards-drug-and-biological-product-submissions-containing-real-world-data

[4] 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.

[5] Anderson W, Kurtycz R, Farid T, Hassan S, Kennon K, Dasher P, Boyce D, Roddy W, Heavner S. (2023). “From OMOP to CDISC SDTM: Successes, Challenges, and Future Opportunities of using EHR Data for Drug Repurposing in COVID-19”. https://www.ohdsi.org/wp-content/uploads/2023/10/3-AndersonBriefreport-Wes-Anderson.pdf

[6] Shanbhogue A.Y. (2022). “RWD (OMOP) to SDTM (CDISC): A primer for your ETL journey”. Paper presented at the PharmaSUG Annual Conference 2022, Austin, TX.

CONTACT INFORMATION

Contact the author at:

Dhruba Sikdar
Johnson and Johnson
50-100 Holmers Farm Way
High Wycombe
HP12 4EG
UK
Email: dsikdar@its.jnj.com