PHUSE US Connect 2024

Paper RE03

Future Clinical Data Submission Standards: CDISC, FHIR, OMOP, or Hybrid Model

Jeff Abolafia, P21, Blue Bell, PA, USA
Sarah Ferko, IBM Consulting, USA
Ingeborg Holt, IBM Consulting, USA

ABSTRACT

There is a rapid increase in the use of real world data (RWD) and observational studies to support marketing applications. However, current submission standards for these non-traditional data sources have not evolved. All study data must be submitted using CDISC standards. While CDISC standards are optimized for randomized clinical trials (RCTs), alternate standards may better be a better fit for other study designs.

This presentation examines the future of data submission standards. First, we present an environmental scan of possible standards for submitting RWD and RCT data. Second, we discuss the pluses and minuses of submitting data using CDISC as opposed to alternative standards. Third, we look at the pros and cons of submitting data under a single versus a hybrid approach, where each data type is submitted using the standard for which it is best suited. Finally, we provide short term and long recommendations for submitting all types of study data.

INTRODUCTION

In the past several years, there has been a sharp increase in the number of New Drug Applications (NDAs) and Biologics License Applications (BLAs) that include RWD and data from other observational studies. Approximately twenty years ago, when CDISC standards were initially developed to support marketing applications, NDAs and BLAs were primarily comprised of a collection of RCTs. Due to changes in the regulatory environment, there is a rapid increase in the trend to include RWD and other non-interventional study designs in marketing applications. Most NDAs and BLAs that were approved in the past two years include at least one real world or non-interventional study. While study designs other than RCTs can be used to provide regulatory authorities with additional evidence of efficacy or safety, they currently present numerous challenges for both regulatory reviewers and sponsors. In our previous three papers (1,2,3), we examined the current regulatory landscape for submitting RWD, the gaps and limitations for submitting RWD using the current data standards, and the challenges of providing the traceability required for regulatory review, as well as offered recommendations for submitting RWD using the current standards. In this paper, we consider the primary question: In future submissions, should RCT and non-interventional study designs be submitted using CDISC standards? The primary goal of this paper is to reframe the discussion around future submission standards by considering the following:

1) Current State: How do we fit RCT, RWD, and other non-interventional designs into CDISC standards?
2) Future State: Should RCT, RWD, and other non-interventional designs be submitted using CDISC standards, or should we choose a separate data standard to support regulatory submissions, such as HL7 FHIR or Observational Medical Outcomes Partnership (OMOP)? Alternatively, should a hybrid approach be considered?

In this paper, we expand on this discussion by first summarizing the issues and challenges when submitting RWD and other non-interventional study designs using the current submission standards. Secondly, we examine the regulatory landscape for submitting clinical data in support of a marketing application. Thirdly, we provide an environmental scan of the current data standards that may be viable in the future to submit clinical data. As part of our environmental scan, we also review possible exchange standards for representing and submitting clinical data, such as SAS® Transport and JSON. Fourthly, we do a deep dive into three separate standards: CDSIC, HL7 FHIR, and OMOP, outlining the advantages and disadvantages of each respective standard to serve as the standard for submitting clinical data. Finally, we provide short- and long-term recommendations for submitting clinical data in the future. While our discussion primarily focuses on standards for source or collected data, we also offer recommendations for submitting analysis

data. Our recommendations are intended to encourage collaboration and discussion between sponsors, standard organizations, and regulatory authorities.

BACKGROUND

REAL WORLD EVIDENCE (RWE)

RWE can come from a variety of sources including EHRs, payer administration claims, digital health technologies (DHTs), and patient registries. Sponsors are submitting an increasing amount of RWD to support marketing applications, and this trend is likely to accelerate. Aetion (4) estimates that approximately 78% of approved NDAs and BLAs in 2020 included a study containing RWD to provide evidence of safety and/or efficacy, compared to only 53% in 2019. Further, a study by Purpura in 2022 (5) estimated that 95% of approved NDAs and BLAs in 2021 included an RWE study.

Section 505F(b) of the FD&C Act defines RWE as "*data regarding the usage, or the potential benefits or risks, of a drug derived from sources other than traditional clinical trials*" (21 U.S.C. 355g(b)). In developing its RWE program, FDA believes it is helpful to distinguish between the sources of RWD and the evidence derived from that data. Evaluating RWE in the context of regulatory decision-making depends not only on the evaluation of the methodologies used to generate the evidence, but also on the reliability and relevance of the underlying RWD; these constructs may raise different types of considerations. For the purposes of this framework, FDA defines RWD and RWE as follows:

- **Real-World Data (RWD)** are data relating to patient health status and/or the delivery of health care routinely collected from a variety of sources; and

- **Real-World Evidence (RWE)** is the clinical evidence about the usage and potential benefits or risks of a medical product derived from analysis of RWD.

For the purposes of this paper, we will use the working definition of RWD as data relating to patient health status and/or the delivery of health care that is not collected under a protocol.

REGULATORY LANDSCAPE

The 21st Century Cures Act (6) was enacted on December 13, 2016, and mandated that the FDA create a framework to evaluate RWE to support marketing applications for new drugs or new indications for drugs already approved. Subsequently, in December 2018, the FDA published its framework for RWE, the *Framework for FDA'S Real World Evidence Program* (7). The framework focused on three areas:

(1) Whether RWD are fit for use,

(2) Whether the study design can provide adequate evidence, and

(3) Whether the study conduct meets regulatory requirements.

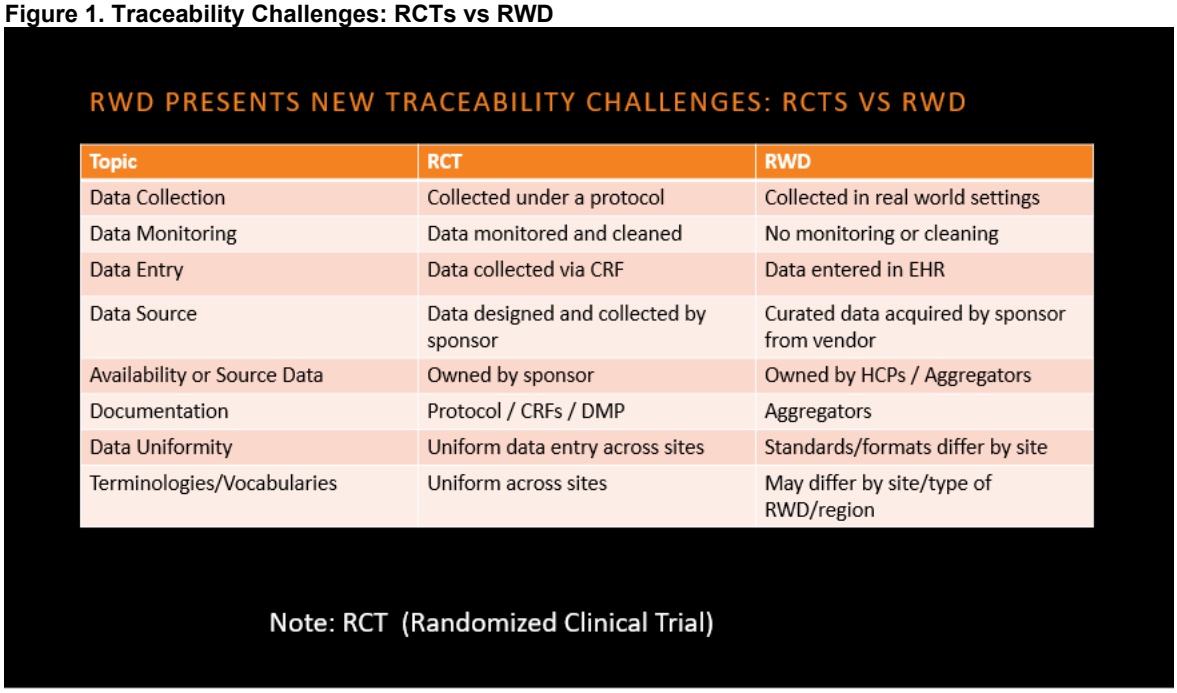
The FDA will focus on these three areas when evaluating RWD submitted in support of a marketing application. In 2021, the FDA expanded its program by issuing four more guidance documents that provided additional detail, as well as implementation strategies for using RWD to support marketing applications for drugs or biological products. The first guidance, *Real-World Data: Assessing Electronic Health Records and Medical Claims Data to Support Regulatory Decision Making for Drug and Biological Products* (8) focused on the validation of definitions for study design elements and data provenance and quality throughout the study lifecycle. This guidance was finalized in 2023 (9).

In October 2021, the FDA issued a second guidance, *Data Standards for Drug and Biological Product Submissions Containing Real-World Data Guidance for Industry* (10), which outlined the data standards required when submitting RWD in support of a marketing application. A final version of this guidance was published in 2023 (11). The primary takeaway described in this guidance is that RWD and data from other non-interventional study designs must be submitted using the standards documented in the FDA Data Standards Catalog (12). To date, this means that RWD *must* be submitted using CDISC standards; thus, sponsors need to convert real-world datasets into CDISC format when submitting this information in support of a marketing application. This guidance also provides advice on mapping RWD in CDISC format and considerations for the required documentation. While the agency acknowledges that its current catalog of standards does not necessarily reflect data derived from real-world sources, it has indicated that it is considering updates, including recommendations for mapping. With respect to RWD access, the guidance advises that sponsors should confirm that they are able to submit patient-level data for any RWD used in studies in support of a marketing application required under an NDA and/or BLA. Additional RWD/RWE guidance documents published by FDA are not summarized in this paper but are publicly available on the FDA RWE site: https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence.

Challenges Submitting RWD Using CDISC Standards

The landscape for submitting RWD is still evolving. FDA guidance documents cited above indicate that the data standards and documentation required for submitting RWD are the same as for submitting RCT data, which presents numerous challenges for sponsors. In our previous three papers (1,2,3), we examined the gaps and limitations for submitting RWD using the current data standards and the challenges of providing the traceability required for regulatory review. These challenges are summarized in Figure 1. Note that 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 differ. RWD is typically not designed or collected for research purposes. As a result, real world databases are organized and configured in a way that makes sense to health care providers, not researchers. Further, these databases utilize different data standards that contain different concepts and coding systems compared to those defined within CDISC and used in regulatory review.

Figure 1. Traceability Challenges: RCTs vs RWD



Environmental Scan

In this section, we review common standards that are used to represent clinical data from both RCTs, RWD, and other study designs. A summary of selected attributes of these standards are presented in Figure 2. below. In the following section, we evaluate a subset of these standards as a possible submission standard to support future marketing applications.

Figure 2. Comparison of Clinical Data Standards

	CDISC	HL7	OMOP
Attribute			
Tabulation Standard	Yes (SDTM)	Yes (FHIR)	No, uses a relational database model
How organized	Topic Related Domains in SDTM	Resources	Relational database tables for clinical data, vocabulary, and other healthcare concepts
Typical exchange format	SAS V5 Transport / Metadata ODM XML	JSON	Each organization or user builds their own CDM, tools are provided for assistance with ETL to transform data
Use case(s)	Submission of RCT data to regulatory authorities including FDA and PMDA	Electronic health records, claims data	Enable efficient analyses of observational data

Supported FDA clinical data standard	Yes	No	No
Supports Non-rectangular structures	No	Yes	Not directly, but data could be extracted in non-rectangular format defined by user
Record linkage	Implicit (RELREC domain in SDTM)	Explicit (Data elements are linked via the model)	Yes, identifiers are used to link the data as needed.
Non-standard data elements	Supplemental variables	Extensions	No
Analysis standard	Yes (ADaM)	No	No
Typical dictionaries	MedDRA, WHODrug, CDISC Controlled Terminology	SNOMED, LOINC, ICD, RxNorm	Over 100 supported, see https://github.com/OHDSI/Vocabulary-v5.0/wiki/Standardized-Vocabularies
FDA review tools	Yes	No	Not applicable, but tools to support data analyses using the CDM do exist

CDISC

Clinical Data Interchange Standards Consortium (CDISC) is a data standard for clinical research and RCT data that is recognized by FDA as the standard for submitting clinical data in support of a marketing application, as outlined within the FDA Data Standards Catalog. CDISC publishes a framework for both a tabulation model to represent collected data (i.e., the Study Data Tabulation Model (SDTM)) and an analysis model to represent analysis data (i.e., the Analysis Data Model (ADaM)). The SDTM is organized into topic-related domains whereas the ADaM is organized into standard data structures. CDISC datasets utilize CDISC Controlled Terminology where applicable, which is developed by CDISC in collaboration with the National Cancer Institute's Enterprise Vocabulary Services (EVS). It represents the set of codelists and valid values used within CDISC-defined datasets, many of which are required to submit CDISC-compliant datasets to regulatory agencies.[1] MedDRA and WHODrug are both additional dictionaries leveraged in CDISC datasets, and LOINC and SNOMED dictionaries may be used when applicable to the data and to the submission. CDISC includes other standards related to the clinical research process, including for non-clinical data, clinical study planning, and clinical data collection. Further, data exchange standards are supported by CDISC such as dataset metadata submitted using the Define-XML standard.

Read more about CDISC

FHIR

Fast Healthcare Interoperability Resources (FHIR) is a standard for health care data exchange published by HL7. FHIR provides a framework for the exchange, integration, sharing, and retrieval of healthcare information across systems. The primary uses cases for FHIR are to exchange healthcare and medical claims data. In 2019, over 80% of hospitals implemented API technology enabled with FHIR. Data elements in FHIR are organized or grouped into resources. For those familiar with CDISC, resources are similar domains where data elements are grouped by topic area. In FHIR, data elements are grouped in a manner that supports business objectives for medical health records and claims data. Example resource types include Observation for vital signs and Encounter for an interaction with a health care provider. A FHIR resource defines a set of data elements, the relationship between the data elements, and constraints on the data. All resources have metadata and a human-readable component, and each resource of the same type is formatted in the same way. Resources or data elements each have a tag that acts as a unique identifier, similar to the URL of a webpage. This means that data elements are intrinsically linked to each other by the FHIR data model/structure.

At this time, the FHIR standard provides a data model for collected data. Unlike CDISC, FHIR has not developed a model for analysis data. While data modeled using CDISC standards is typically represented in SAS Transport format, data exchanged using the FHIR standards are usually represented in JSON format. This provides additional flexibility, allowing for non-rectangular data structures. FHIR also can leverage additional modern web technologies such as XML, HTTP, Atom, OAuth, and others. Furthermore, the FHIR standard includes RESTful protocols, where REST is the same architecture that underlies many enterprise tools, the internet, and modern APIs.

While FHIR is not a currently supported standard by FDA for submitting clinical data, it is currently being used for

[1] More information is available at CDISC's Controlled Terminology webpage: https://www.cdisc.org/standards/terminology/controlled-terminology.

several separate activities at FDA (Galvez, 2023) (13). For example, SPL is based on HL7's Version 3 standard and FHIR is the emerging HL7 standard that CDER is exploring to support SPL use cases. In eCTD Module 3, FHIR is also being used as the exchange standard for submitting Pharmaceutical Quality (PQ) and Chemistry & Manufacturing Controls (CMC) data (Galvez).

Read more about HL7 FHIR

OMOP

The Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) is an open community data standard, designed to standardize the structure and content of observational healthcare data. It was originally developed under a grant from NIH to enable the merging of EHR and claims data and facilitate efficient analyses that can produce reliable evidence. It was designed to address the challenges associated with the heterogeneity of healthcare data by providing a standardized structure for organizing and representing diverse observational databases. It is currently maintained by Observational Health Data Sciences and Informatics (OHDSI) consortium.

At its core, OMOP CDM is designed to harmonize disparate healthcare data sources, ranging from electronic health records (EHRs) and claims data to registries and other real-world evidence. The primary objective is to create a consistent framework that facilitates the integration and analysis of data from different healthcare systems and settings. By establishing a common language, researchers and analysts can navigate through a wealth of information with greater efficiency, promoting cross-study collaboration and data sharing.

Designed to focus on observational data and reflecting real-world clinical practices and patient outcomes, it stands in contrast to the CDISC data standard which was developed for representing data from clinical trials with controlled environments and selected patient populations. By leveraging observational data, the OMOP CDM is designed to enable researchers to gain insights into the effectiveness and safety of medical treatments, patient characteristics, and disease patterns in diverse and representative populations.

The model comprises a set of standardized tables and concepts, providing a consistent structure for storing healthcare data elements such as patients, diagnoses, medications, and procedures. This standardization facilitates the translation of complex clinical information into a format that is easily accessible and analyzable. Moreover, OMOP CDM incorporates standardized terminologies, such as SNOMED CT and LOINC, and seeks to ensure semantic consistency to enhance interoperability across the various systems housing healthcare data.

A central component of the OMOP CDM is the OHDSI standardized vocabularies. The OHDSI vocabularies facilitate organization and standardization of medical terms used across the clinical domains of the OMOP CDM and enable standardized analytics across the database when constructing exposure and outcome phenotypes and other features within characterization, population-level effect estimation, and patient-level prediction studies.

OMOP CDM supports large-scale collaborative research efforts and enables the development of advanced analytical methods and tools within the OHDSI community. OHDSI offers a wide range of open-source tools designed for use with one or more of the databases of the Common Data Model (CDM) to support an array of data-analytics use cases on observational patient-level data. These tools facilitate performing analyses by filling in standard templates improving the reproducibility and transparency of analyses by using standard algorithms to compute commonly used metrics, such as incidence rate.

Read more about the OMOP Common Data Model
Read more about OHDSI's standardized vocabularies

Exchange Standards

While data standards specify common standardized data elements and associated terminologies, data exchange standards denote the file formats for submitting and/or exchanging data. Common exchange formats include SAS® Version 5 Transport, JSON (JavaScript Object Notation), and XML. The SAS transport format was originally developed in the late 1980s and is an openly documented specification maintained by SAS. It was developed to support data transfers between systems, especially those running different operating systems. Since 1999, the FDA has required sponsors to submit study datasets using the SAS V5 transport file format. SAS V5 transport format is an older specification which imposes several restrictions on submission datasets, including a rectangular flat data structure and limited data types. In addition, variable names must be alphanumeric, variable names are limited to a maximum length of eight characters, variable labels are restricted to a maximum length of 40 characters, character variable data widths are limited to 200 characters, there is no Unicode support, and the format has an inefficient use of storage space. While CDISC datasets are currently exchanged in SAS V5 transport format, CDISC is not inherently tied to this format. CDISC datasets can also be exchanged using XML, JSON, or other file formats. Thus, if the FDA and other regulatory authorities agree, CDISC datasets may be exchanged using another format.

CDISC has produced the Dataset-XML specification which can be used to exchange study data as an alternative to SAS V5 Transport format. In addition, there is an ongoing project in collaboration with FDA, PHUSE, CDISC to pilot the use of JSON for submitting study data. Dataset-JSON was adapted from the Dataset-XML specification, but instead uses JSON format for regulatory submission needs. JSON has evolved into one of the most popular formats for exchanging data, and in particular, has become the primary standard for exchanging data using the FHIR model. JSON is a modern-day format that can meet the regulatory requirements for submitting study data. It consists of human-readable text to store and transmit data objects comprised of attribute-value pairs. JSON is a vendor and technology neutral file format and has the become the de facto standard for data exchange using APIs. It overcomes most of the limitations of SAS V5 transport format noted above, and additionally provides for smaller file sizes and additional metadata.

CDISC, FHIR, OR OMOP?

CDISC

CDISC has many attributes that make it an exceptional data standard for clinical research. FDA first required the use of CDISC data standards for clinical and nonclinical studies that started after December 17, 2016, though the use of CDISC traces back to the early 2000s. Given that CDISC is still the current standard supported by the FDA as outlined in the FDA Data Standards Catalog, it continues to be used to submit data in support of marketing applications for medical products. Thus, FDA reviewers and analysts, SDOs, sponsors, and other industry groups are familiar with these standards. Apart from familiarity, these groups have also developed tools for the representation and analysis of CDISC data. Further, a sponsor company may have internal tools and processes established for the collection, transformation, and reporting of CDISC data. Thus, changing standards from CDISC would require a large overhaul on tools that use CDISC data for data collection, reporting, and analysis across these groups.

CDISC is a very comprehensive end-to-end standard, with data standards supporting SEND, SDTM, and ADaM data, as well as other standards such as those used for study planning (i.e., PRM) and data collection (i.e., CDASH). Unlike FHIR, CDISC also has a dedicated analysis standard to support the creation of analysis datasets and associated metadata (i.e., ADaM), and is harmonized with the SDTM. Further, there are a variety of CDISC exchange standards. Apart from the Dataset-XML and JSON-XML standards presented in the Exchange Standards section above, the Define-XML standards is used to exchange study metadata (e.g., SDTM and ADaM Define.xml files to support the submission of the study datasets). Other exchange standards include SDM-XML, ODM-XML, and CTR-XML.

Apart from data and exchange standards, additional CDISC standards include CDISC controlled terminology, originally presented in the CDISC section above and a variety of therapeutic area standards. Further, Therapeutic Area User Guides (TAUGs) are therapeutic area-specific guidance to aid in the submission of NDA/BLA data and documentation, and as of January 2024, 50 TAUGs are published on cdisc.org. In addition to TAUGs, there are numerous questionnaires, ratings, and scales (QRS) supplements published on the CDISC website which aim at the standardized representation of SDTM (and in some cases, ADaM) data for a multitude of QRS measurements. There are countless industry experts that contribute to CDISC and have provided their expertise when developing these documents. From the regulatory agency perspective, agency-authored documentation including technical specifications and the study data technical conformance guide include CDISC data. Thus, changing a data standard will result in numerous updates to the related standards and documentation for both SDOs and regulatory agencies.

CDISC data may be considered to be ‘siloed’ as submitting data specifically produced for clinical research, since it was developed for the submission of RCT data rather than for RWD. Though it has started to be used for the submission of RWD, it has not yet been optimized for the submission of RWD when RWD is submitted to support a premarket application. Further, it can be a great deal of work to convert RWD into CDISC format. Within CDISC, there are other cons to consider when considering the optimal standard for RWD. For example, the CDISC SDTM is very rigid, and it is cumbersome to add new elements not defined within the SDTM. This is difficult when considering the numerous data concepts common to EHRs and other sources of RWD that do not currently exist within the SDTMIG and makes converting RWD to submit in CDISC format difficult. Supplemental qualifier datasets are used to include additional captured data not previously defined within the CDISC SDTM. Secondly, linking is not easy within CDISC, where Related Records (RELREC) domain is needed to comprehensively link data across domains, which are both cumbersome to make as well as to analyze downstream from a regulatory perspective. Thirdly, CDISC maintains a very rectangular structure, unlike FHIR which supports non-rectangular structures.

FHIR

FHIR has many attributes which make it a viable future standard for submitting clinical data. The advantages of using FHIR for submitting all clinical data include the following:

- FHIR is optimized for RWD (and, for EHR and claims data). Over time, there is a trend for more RWD to be submitted as part of a marketing application.
- There is a trend toward collecting data for RCTs and other observational study designs in an HER.
- Most data in EHRs are already exchanged using FHIR, and in the future, it is likely that almost all EHR data will be represented in FHIR. As a result of the 21st Century Cures Act, FHIR has already become the leading API for exchanging healthcare data. The Office of the National Coordinator for Health IT (ONC) and the Centers for Medicare & Medicaid Services (CMS) have designated FHIR as the standard required to fulfill the requirements of the 21st Century Cures Act (6). The federal rule states that certified EHRs must support the standard FHIR APIs by the end of 2022.
- FHIR is a more "modern" standard, that leverages web standards such as JSON, HTTP, Atom, OAuth and others.
- Data exchanged using FHIR is usually represented in JSON and is not limited by the constraints on SAS Version 5 transport format.
- Linking to related data within a marketing application and to external data sources is much easier than in CDISC.
- As noted above, FHIR is already being used at FDA for several submission-related activities.

While FHIR has many attractive features, there are many drawbacks to using the FHIR as a single standard to submit clinical data:

- While FHIR was developed and optimized for healthcare data, it lacks many of the concepts necessary to represent clinical research data. This makes converting RCT data to the FHIR standard difficult, and in some cases, not possible. When non-RCT data is submitted in CDISC format, many non-standard variables and domains must be created, and many validation/business rules do not apply. However, FHIR is developing resources/data elements to represent research study concepts. FHIR also has several current projects that are designing FHIR profiles that can facilitate representing data for clinical research and ultimately for submission.
- At this point in time, FDA and sponsors lack familiarity with FHIR.
- We do not know what a FHIR submission package would look like (e.g., how would the data be bundled for submission? What would be the counterpart CDISC's Define.xml file, a vital component needed for regulatory review?).
- FDA and sponsors would have to develop new tools or update existing tools to facilitate a FHIR submission. Furthermore, training, new processes, and substantial change management is needed to accommodate a FHIR submission. Many FDA documents and technical specifications were developed for a CDISC submission, and these would all have to be updated to accommodate a new submission standard.
- FHIR is primarily a collection/exchange standard. Currently, FHIR does not have a model for analysis data. Until new technologies and tools are developed, analysis datasets will continue to be a key component of regulatory review.
- CDSIC and FHIR lack harmonization; thus, FHIR and ADaM, the CDISC analysis standard, are not harmonized. As a result, it will be much more challenging to create ADaM datasets when FHIR is the source format.
- Many sponsors have designed end-to-end processes around CDISC standards. Switching to FHIR would place a short-term burden on sponsors.

OMOP

The use of the OMOP CDM for submitting data to the FDA would represent a big change to the current regulatory process which has been tuned over the last 20 years to use CDISC SDTM and more recently CDISC ADaM for analysis. OMOP CDM, like CDISC, has a standard structure and a variety of tools have been developed to use this structure to analyze the data. OMOP's relational database model means data can be easily extracted into multiple different output formats, and mappings from OMOP to CDISC SDTM have been performed (14). OMOP CDM's emphasis on semantic interoperability, incorporating standardized terminologies, promotes consistency and accuracy in data interpretation. However, the underlying assumptions that are made to enable this semantic interoperability would need to be surfaced and validated and may need to be done differently depending on the therapeutic area. One area of particular concern is the collection of Adverse Events. OHDSI's data structure is such that any terminology can be included and harmonized with existing terminologies; however, Diagnoses, Conditions, Problems, Procedures use SNOMED as the base terminology to harmonize other mappings. Given the granularity of SNOMED CT, which currently contains more than 300,000 medical concepts, compared to MedDRA's 80,000 lower-level terms, careful attention would need to be paid to any mapping to ensure condition and problems found in EHRs and claims were mapped in a way that assured FDA reviewers no signals could be lost in the mapping.

The OMOP CDM, like other standards, has both advantages and disadvantages for use a RWE standard. OMOP CDM provides a standardized and interoperable framework for organizing healthcare data and this provides the advantage

that use of any standard does in that it enhances consistency and interoperability across different data sources converted to the standard, making it easier to aggregate and analyze data from diverse sources.

The OMOP CDM was designed to facilitate studies conduct across multiple databased and thus allows researchers can verify results found in their own data by using the OMOP network to increase the data sources the analyses include with minimal effort. This provides the ability to conduct additional observational studies to verify findings if desired. Related is the benefit of reproducibility. Standardized codes and definitions for clinical concepts promote transparency which allows a reviewer or investigator the ability to understand the decisions made that could create bias within the study, for example ICD10 codes mapped to the inclusion exclusion criteria affecting the patient cohort studies. This transparency and reproducibility is important in all studies, but is very important in when using observational data from EHRs which can contain a lot of variability as the data is likely entered by multiple people at multiple organizations coming from diverse geographical areas.

A disadvantage of the OMOP CDM is that it can require significant effort to transform the data. This is particularly an issue if the original data sources use different formats and/or coding systems. This process can be resource-intensive and time-consuming, however, the OHDSI community provides tools and free support to help with the conversion and in addition, may service providers offer conversion services as well. In addition, the OMOP CDM is being used by some major health systems including The University of California Health System, with 18 health professional schools, six medical centers, and 10 hospitals and has built a secure central data warehouse using the OMOP CDM. It's uses include operational improvement, promotion of quality patient care, and to enable the next generation of clinical research (15). If the use of the OMOP CDM rises, the transformation should become more standardized and tools for use of the data, including tools for clinical review, will likely grow.

Loss of granularity of the original data during the mapping to the OMOP CDM can occur and this could impact the precision of certain analyses, especially if specific details from the original data are crucial to the study. As noted above, mapping from SNOMED CT to MedDRA presents a challenge, but this challenge will be present regardless of the CDM chosen. These risks can be mitigated by several methods, including using a standard, reviewed mapping from one dictionary to another.

The OMOP CDM has limited flexibility outside of its standardized structure. Adaptations and/or extensions to the model are not able to be made thus studies that require data elements not currently in the model would introduce complexity. Note that although CDISC currently does allow extensions in the form of custom domains and variables added in the supplemental qualifier domains, these are not often able to be analyzed easily and so this problem exists to some extent with the current standard.

As with any new standard, the adoption of OMOP CDM would require substantial resources, both in terms of time and expertise and changes to the existing systems which are likely tuned to using the current standards, CDISC. **Organizations would need to invest in training and technology to effectively implement and use the standard effectively,** though as noted above, there is a large amount of support for the OMOP standard both with implementing it and with analyses tool development.

Single or Hybrid?

CDISC SDTM, OMOP and HL7 FHIR standards should be considered as possible future standards to represent RCT data, RWD, and data from other non-interventional study designs. However, can one of these standards adequately represent all designs, considering the increasing diversity of study designs as innovations in clinical research continue to be made? If so, which one? If not a single standard, would a hybrid approach be better?

The single standard option encompasses choosing either the CDISC SDTM, OMOP or HL7 FHIR standard to represent all study designs for a marking application. In this scenario, the collected data (or tabulation data) is submitted in the tabulation standard chosen by FDA, with input from industry. However, the pros and cons outlined in the preceding section must be considered for each standard, as these will be relevant for NDA/BLA data submitted under the selected standard.

A second option is the hybrid approach. In this scenario, RCT data are submitted using the SDTM standard, EHR and claims data are submitted using the FHIR standard, and data from registry studies and other observational study designs are submitted using OMOP standards. A pro of using a hybrid approach is allowing data to be submitted in the standard for which it is optimized (e.g., CDISC for RCT data, FHIR for RWD). However, it is important to note the lack of harmonization between CDISC SDTM, OMOP, and HL7 FHIR. Having multiple standards requires two sets of tools and processes and can make the analysis of RCT data and RWD difficult if data are intended to be analyzed together. At this point in time, we are only considering the CDISC ADaM standard for submitting analysis data given that neither FHIR or OMOP have a current standard or model to represent analysis data. Thus, creating ADaM datasets from FHIR as the source format will be difficult.

RECOMMENDATIONS

Our recommendations for the future are both short term and long term. First, we focus on short term recommendations. Currently FDA guidance (12) dictates that clinical data must be submitted using the standards in the documented in the FDA Data Standards Catalog (11). For now, submitting data using CDISC standards is required. In the short term, this requirement results in cumbersome data transformations, non-standard variables and domains, and business and validation rules which primary apply to data collected in RCTs. In our previous three papers (1,2,3), we outline the challenges of submitting RWD using CDISC standards as well as for providing traceability. We also provide short term solutions to overcome a subset of these challenges and provide recommendations to facilitate the use of RWD in regulatory review.

While OMOP and FHIR are viable standards, they are not submission ready. Both OMOP and FHIR do not have analysis data models, nor are they harmonized with the CDISC ADaM analysis standard. Secondly, we have not defined what a submission package would look like for either of these two standards. Thirdly, review tools and other processes do not exist and would have to be developed for a FHIR or OMOP submission. It will take significant time to define the specifications for an OMOP or FHIR submission, to define the validation and business rules to check a non-CDISC submission, and to update FDA guidance and technical specifications related to regulatory submissions, which currently focus on the submission of CDISC data.

Regarding data exchange standards, we recommend transitioning to the JSON file format. JSON also addresses many of the shortcomings of the SAS V5 transport format, which is considered by many to be old and outdated. Furthermore, JSON allows for non-rectangular structures which will allow sponsors to capture RWD more accurately. To date, the results of the JSON pilot look promising. Also in the short term, the industry can move to the JSON exchange standard without making modifications to existing CDISC standards or the associated business and validation rules. Transitioning to the JSON standard should have minimal impact on both FDA and sponsor tools and processes.

What else can we do, besides transforming clinical data for all study designs to comply with CDISC standards? The industry may first consider changing the conversation from ‘How we fit data from non-interventional study designs into CDISC?’ to ‘What should the paradigm be for submitting data, when considering the increasing diversity of study designs?’ Some topics for consideration include:

- Can one standard accommodate data from all study designs?
    - Alternatively, should a hybrid approach be employed?
- CDISC has developed core data elements for representing data from RCTs needed for regulatory submission. Can industry continue to develop core data elements and terminologies for non-interventional studies needed for regulatory submission and review?
- Can CDISC standards be modified to accommodate data from all study designs?
    - If so, what changes are needed?
    - What changes has CDISC made to date to accommodate non-RCT study designs?
- Is OMOP or FHIR a more viable future submission standard?
    - If so, what changes are needed?
    - How have these standards evolved in the past several years?
- How will changes in technology affect future submission standards?
    - Will we be able to eliminate the artificial packaging of clinical data and enable real time data steaming?
    - Can we develop technologies and tools to eliminate the need for analysis datasets?
    - Can we harmonize CDISC, FHIR, and OMOP, so any of these standards can easily be mapped to the other two standards?
- Over time, will more and more data be collected in EHRs?
    - If so, does it make sense to transform this data from FHIR to another standard?
- Can industry, regulatory authorities, and standards organizations work together more effectively on pilot projects related to the future of submission standards?
- How can mapping, especially dictionary to dictionary, be resolved to ensure issues related to loss of granularity as well as the introduction of granularity that may result in signals (both of safety and efficacy) being diluted and this not recognized in studies with limited populations.
- Can our industry and regulatory authorities move beyond their comfort zone?
    - Right now, CDISC is familiar to both sponsors and FDA. Sponsors and FDA are much less familiar with other standards such as FHIR and OMOP. Though CDISC standards are currently required, this should not inhibit our industry from considering the best future state and evaluating other alternatives. Collectively, we should continue to monitor how each standard is evolving and how this affects its ability to meet future submission needs.

CONCLUSIONS

There is a vast and growing amount of healthcare data exchanged in a variety of formats. The increasing volume of data has the potential to better inform us on the potential benefits and risks of a medical product. However, the increase in the amount of data available from non-interventional study designs results in data collected using both different data standards and different file formats. Due to changes in the regulatory environment, there is an expectation from industry that these data should be included in marketing applications for medical products. It is incumbent on sponsors, regulatory authorities, standards development organizations, and other industry-related groups to collaborate to design and agree upon data standards that will allow for more efficient and effective regulatory review of both RCT and real-world data. Whether our future is a single standard or a hybrid approach, regulatory reviewers require submission standards which cover multiple study designs and contain biomedical concepts, data elements, and terminologies that are harmonized across study designs that will allow for valid scientific conclusions.

REFERENCES

(1) Abolafia, J, Ferko, S, & Holt, I. (2022). “Submission Standards for RWD: The Good, the Bad and the Ugly”. Paper presented at the PHUSE Annual Conference 2022, Belfast, United Kingdom. https://phuse.s3.eu-central-1.amazonaws.com/Archive/2022/Connect/EU/Belfast/PRE_RE09.pdf

(2) Ferko, S., Holt, I., & Abolafia, J., (2023). “Challenges and Considerations for Submitting Real World Data”. Paper presented at the PHUSE US Annual Conference 2023, Orlando, FL. https://phuse.s3.eu-central-1.amazonaws.com/Archive/2023/Connect/US/Florida/PRE_RE05.pdf

(3) Abolafia, J, Ferko, S, & Holt, I. (2023). “Submission Standards for Real World Data: Gaps, Limitations and Recommendations”. Paper presented at the PHUSE Annual Conference 2023, Birmingham, United Kingdom. https://phuse.s3.eu-central-1.amazonaws.com/Archive/2023/Connect/EU/Birmingham/PAP_RE03.pdf

(4) Aetion. “Aetion_eBook_2021_The role of RWE in FDA Approvals”. https://aetion.com/evidence-hub/welcome-to-the-evidence-hub/

(5) Purpura CA, Garry EM, Honig N, Case A, Rassen JA. The Role of Real-World Evidence in FDA-Approved New Drug and Biologics License Applications. Clin Pharmacol Ther. 2022 Jan;111(1):135-144.

(6) 21st Century Cures Act. (2016). https://www.congress.gov/114/plaws/publ255/PLAW-114publ255.pdf

(7) FDA Guidance (2018). “Framework for FDA’s Real-World Evidence Program”. December 2018. https://www.fda.gov/media/120060/download

(8) FDA Guidance (2021). "Real-World Data: Assessing Electronic Health Records and Medical Claims Data to Support Regulatory Decision Making for Drug and Biological Products". September 2021. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/real-world-data-assessing-electronic-health-records-and-medical-claims-data-support-regulatory

(9) FDA Guidance. (2023). “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”. August 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

(10) FDA Guidance. (2021). "Data Standards for Drug and Biological Product Submissions Containing Real-World Data Guidance for Industry". October 2021. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/data-standards-drug-and-biological-product-submissions-containing-real-world-data

(11) FDA Guidance. (2023). “Data Standards for Drug and Biological Product Submissions Containing Real-World Data Guidance for Industry". December 2023. https://www.fda.gov/media/153341/download

(12) FDA Guidance. (2023). “Data Standards Catalog”. April 2023. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/data-standards-catalog

(13) Galvez, J., (2023). “Reimaging Regulatory Data Submissions – Modernizing, partnering and exploring how regulatory data could be shared with HAs””. Paper presented at the PHUSE US Annual Conference 2023, Orlando, FL.

(14) 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.

(15) UC Health Data Warehouse (UCHDW), https://ctsi.ucla.edu/uc-health-data-warehouse-uchdw (accessed 1/15/2024).

CONTACT INFORMATION

Jeffrey Abolafia
Certara/Pinnacle 21
mailto:jabolafia@pinnacle21.com