PHUSE EU Connect 2025

Paper RE01

Close Encounters with Real World Data

Jeff Abolafia, Certara, Radnor, PA, USA

Sarah Ferko, IBM Consulting, USA

Ingeborg Holt, Orizaba Solutions, USA

ABSTRACT

There is a rapid increase in the use of RWD to support marketing applications. However, submitting RWD may pose numerous challenges for regulatory reviewers in their analysis of the data. In our six previous papers, we outline these challenges and present several high-level solutions. This current paper builds on our previous work by taking a deep dive into CDISC SDTM example data and highlighting updates needed to the SDTM to better accommodate RWD. Specifically, we examine the gaps in the Subject Visits (SV) domain for its use in regulatory review. Considering how other models such as HL7 FHIR, OMOP, and PCORnet handle certain elements, we propose updated domains and data elements to represent patient encounters that are a key component of RWD. These ideas are intended to help sponsors submit RWD in the currently required CDISC format and to continue the conversation with SDOs such as CDISC around the updates needed for more efficient regulatory review.

INTRODUCTION

Over the past several years, there has been an increase in the number of New Drug Applications (NDAs) and Biologics License Applications (BLAs) that include Real World Data (RWD) and data from other observational studies. Approximately twenty years ago when Clinical Data Interchange Standards Consortium (CDISC) standards were initially developed to support marketing applications, NDAs and BLAs were primarily a collection of randomized controlled trials (RCTs). The rise in the use of electronic healthcare records (EHRs) was spurred by regulation including the Health Information Technology for Economic and Clinical Health Act (HITECH) Act and the Affordable Care Act (ACA) which increased the use of EHRs and thus the availability of electronic healthcare data. With the increasing volume of EHR data, there has been an increasing push to make use of this data for other purposes, including bringing medical products to market. Consequently, there has been a rapid increase in the trend to include RWD and other non-interventional study designs in marketing applications. While study designs other than RCTs have been used to provide regulatory authorities with additional evidence of efficacy and/or safety, they currently present numerous challenges for both regulatory reviewers and sponsors. In our previous six papers (1, 2, 3, 4, 5, 6), we examined the current regulatory landscape for submitting RWD and the gaps and limitations of using the current data standards for submitting RWD. We also acknowledged the challenges with providing the traceability required for regulatory review and offered recommendations for submitting RWD using the current standards. In this paper, we build on our previous work and demonstrate how our recommendations can be applied when submitting RWD encounter data as part of a marketing application. These recommendations are based on U.S. Food and Drug Administration (FDA) Real World Evidence (RWE) guidance, our previous papers, and how other common data models (i.e., OMOP, HL7 FHIR, PCORnet, and Sentinel) model RWD data elements. We propose new and updated data elements and/or domains to more thoroughly represent RWD for a regulatory submission. While our focus is on subject visit or subject encounter data, many of our recommendations can be applied to other domains or topic areas.

BACKGROUND

REAL WORLD EVIDENCE (RWE)

RWE can come from a variety of sources including EHRs, payer administration claims data, 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 (7) 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 (8) 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 recognizes that 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, and these considerations may vary depending on the use of the RWE. 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 healthcare routinely collected from a variety of sources; and

- **Real-World Evidence (RWE)** is 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 healthcare that is not collected under a protocol and RWE as the data derived from the RWD to be submitted to the FDA. Note that using previously collected data for use in a 'RWE Study', a protocol should be developed for the study.

REGULATORY LANDSCAPE

The 21st Century Cures Act (9), enacted in 2016, mandated that FDA create a framework to evaluate RWE to support marketing applications for new drugs or new indications for drugs already approved. Subsequently, in 2018, FDA published its framework for RWE, the *Framework for FDA'S Real World Evidence Program* (10). The framework focused on three areas:

(1) Whether RWD is 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, which is similar to RCTs. FDA guidance notes there is no change in the standards expected to be used in the review of RWD versus RCT data. Starting in 2021, the FDA expanded its RWE program by issuing multiple guidance documents that provided additional considerations, as well as implementation strategies for using RWD to support marketing applications for drugs or biological products. The *Real-World Data: Assessing Electronic Health Records and Medical Claims Data to Support Regulatory Decision Making for Drug and Biological Products* guidance, which was finalized in 2024, explains how sponsors can use routinely collected data from EHRs and medical claims in clinical studies to support submissions for drug or biologic safety and effectiveness. It emphasizes the importance of selecting data sources that are fit-for-purpose, developing and validating definitions of exposures, outcomes and covariates, and

maintaining data provenance, quality, and traceability throughout accrual, curation, and analysis (11).

A second guidance document to note is the *Data Standards for Drug and Biological Product Submissions Containing Real-World Data Guidance for Industry*, which was finalized in 2023 (12). This guidance outlines how study data derived from RWD sources should be formatted and submitted to support regulatory decision making. It emphasizes that when submitting RWD to support a new drug or biologic application, sponsors should use the data standards specified in the FDA's Data Standards Catalog (13). The guidance acknowledges that currently available standards may not fully reflect all RWD sources and encourages documentation of any deviations, mappings, or transformations applied to align the data with FDA-supported formats. It also provides recommendations for early engagement with FDA and using available resources to improve interoperability, consistency, and regulatory readiness of RWD submissions. FDA guidance documents related to RWE can be found on FDA's website: Real-World Evidence | FDA

Based on FDA guidance, RWD must be submitted using CDISC standards; thus, sponsors need to convert real-world datasets into CDISC Study Data Tabulation Model (SDTM) and Analysis Data Model (ADaM) format when submitting this information in support of a marketing application. This guidance also provides advice on mapping RWD to CDISC format and considerations for the required documentation. While the agency acknowledges that its current catalog of standards does not necessarily reflect data standards developed for real-world sources, it has indicated that it is considering updates, including recommendations for mapping. With respect to access to the source RWD, the guidance advises that sponsors confirm their ability to submit source patient-level data for any RWD used in studies submitted to the FDA in support of a marketing application.

Finally, in response to the FDA acknowledging that the Data Standards Catalog does not reflect standards for RWD, the FDA also published a Federal Register Notice in April 2025 exploring the Health Level Seven (HL7) Fast Healthcare Interoperability Resources (FHIR) for submission of data collected from RWD sources titled *'Exploration of Health Level Seven Fast Healthcare Interoperability Resources for Use in Study Data Created From Real-World Data Sources for Submission to the Food and Drug Administration; Establishment of a Public Docket; Request for Comments'*. The notice served as a request for public comments on the opportunities and challenges of using HL7 FHIR to standardize and structure clinical study data collected from RWD sources (14). The public comment period closed on June 23, 2025.

CURRENT STATE

CDISC Domain Overview

CDISC is a development effort that began in 1999 to standardize the data commonly collected for RCTs. CDISC has worked closely with both the pharmaceutical industry and the FDA to develop tabulation and analysis standards to meet the needs of both stakeholders. Since December 2016, it has been recognized by FDA within the Data Standards Catalog as the required standard for submitting clinical data to support a marketing application, where it is specified along with other standards. CDISC publishes a framework for both a tabulation model to represent collected data (i.e., the SDTM) and an analysis model to represent analysis data (i.e., the ADaM) as well as implementation guides and associated metadata for these standards (Define.xml).

The SDTM is organized into topic-related domains whereas ADaM is organized into standardized data structures for analysis purposes. 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 (15). MedDRA and WHODrug are two additional dictionaries leveraged in CDISC standards, and LOINC and SNOMED dictionaries may also be used when applicable in the data and the submission. CDISC produces other standards related to the clinical research process, including non-clinical data (SEND), clinical study planning (Protocol), and clinical data collection (CDASH). Further, data exchange standards are supported by CDISC such as dataset metadata submitted using the Define-XML, the Operational Data Model (ODM),

and the recently released Dataset-JSON standard.

Read more about CDISC

Visit/Encounter Data within CDISC

Within the CDISC SDTM, encounter information is represented within different domains including Subject Visits (SV), Demographics (DM), and Healthcare Encounters (HO). Within this paper, the current state is assessed using SDTMv2.0 (16) and SDTMv3.4 (17).

In an RCT, much of the subject visit information is contained in the study protocol and related study documentation, such as the Schedule of Activities (SoA). As a result, the CDISC domains for representing visits are sparse and do not contain many variables that provide context about the encounter. However, given that RWD is not collected under a protocol, additional qualifiers are useful to interpret individual patient encounters, and suggestions for additional qualifiers are provided in the Recommendations section. Information is provided below that summarizes the current state for encounter details provided within the SV, DM, and HO domains.

Subject Visits

The SDTM SV domain, provided below in Figure 1, contains limited information about the visit beyond timing variables. Within SV, Visit Number (VISITNUM) is the topic variable that represents the clinical encounter number. Additional qualifiers are provided within SV, including Reason for Occur Value (SVREASOC), Contact Mode (SVCNTMOD), and Epi/Pandemic Related Change Indicator (SVEPCHGI). However, other qualifiers for an encounter that are often found in an EHR to describe the RWD are not currently accounted for within the SV domain (see Recommendations section below).

Figure 1: SV Domain Specification

SV – Specification

sv.xpt, Subject Visits — Special Purpose. One record per actual or planned visit per subject, Tabulation.

Variable Name	Variable Label	Type	Controlled Terms, Codelist or Format[1]	Role	CDISC Notes	Core
STUDYID	Study Identifier	Char		Identifier	Unique identifier for a study.	Req
DOMAIN	Domain Abbreviation	Char	SV	Identifier	Two-character abbreviation for the domain most relevant to the observation. The domain abbreviation is also used as a prefix for variables to ensure uniqueness when datasets are merged.	Req
USUBJID	Unique Subject Identifier	Char		Identifier	Identifier used to uniquely identify a subject across all studies for all applications or submissions involving the product.	Req
VISITNUM	Visit Number	Num		Topic	Clinical encounter number. Numeric version of VISIT, used for sorting.	Req
VISIT	Visit Name	Char		Synonym Qualifier	Protocol-defined description of a clinical encounter.	Perm
SVPRESP	Pre-specified	Char	(NY)	Variable Qualifier	Used to indicate whether the visit was planned (i.e., visits specified in the TV domain). Value is "Y" for planned visits, null for unplanned visits.	Exp
SVOCCUR	Occurrence	Char	(NY)	Record Qualifier	Used to record whether a planned visit occurred. The value is null for unplanned visits.	Exp
SVREASOC	Reason for Occur Value	Char		Record Qualifier	The reason for the value in SVOCCUR. If SVOCCUR="N", SVREASOC is the reason the visit did not occur.	Perm
SVCNTMOD	Contact Mode	Char	(CNTMODE)	Record Qualifier	The way in which the visit was conducted. Examples: "IN PERSON", "TELEPHONE CALL", "IVRS".	Perm
SVEPCHGI	Epi/Pandemic Related Change Indicator	Char	(NY)	Record Qualifier	Indicates whether the visit was changed due to an epidemic or pandemic.	Perm
VISITDY	Planned Study Day of Visit	Num		Timing	Planned study day of VISIT. Should be an integer.	Perm
SVSTDTC	Start Date/Time of Observation	Char	ISO 8601 datetime or interval	Timing	Start date/time of an observation represented in IS0 8601 character format.	Exp
SVENDTC	End Date/Time of Observation	Char	ISO 8601 datetime or interval	Timing	End date/time of the observation represented in IS0 8601 character format.	Exp
SVSTDY	Study Day of Start of Observation	Num		Timing	Actual study day of start of observation expressed in integer days relative to the sponsor-defined RFSTDTC in Demographics.	Perm
SVENDY	Study Day of End of Observation	Num		Timing	Actual study day of end of observation expressed in integer days relative to the sponsor-defined RFSTDTC in Demographics.	Perm
SVUPDES	Description of Unplanned Visit	Char		Record Qualifier	Description of what happened to the subject during an unplanned visit. Only populated for unplanned visits.	Perm

Demographics

There is limited information within the SDTM DM domain regarding encounter information. Specifically, the SDTM DM domain contains details for site and investigator, including:

- Study Site Identifier (SITEID), a unique identifier for the site within a study;

- Investigator Identifier (INVID), an identifier to describe the Investigator for the study; and
- Investigator Name (INVNAM), the name of the investigator for a site.

Given the one record per subject structure of the DM domain, this domain is designed for an RCT and assumes one site and one investigator per patient.

Healthcare Encounters

The HO domain is an events domain that contains data for inpatient and outpatient healthcare events by describing the location or place of the healthcare encounter (e.g., hospitalization, nursing home stay, rehabilitation facility stay, ambulatory surgery). Other details for the encounter, such as the name of the provider or the facility where the encounter took place, are not included in the HO domain, given HO is derived from the CDISC Events class. Instead, if collected, these details are currently presented as supplemental qualifiers (i.e., a SUPPHO dataset). Similar to the SV domain, certain qualifiers often found in an EHR to describe the RWD are also not included in the HO domain, as discussed in the Recommendations section below.

Visit/Encounter Data within BIMO

The Bioresearch Monitoring (BIMO) Program is a component of the FDA's oversight of clinical research supporting new drug and biologics applications (18). For submissions to the Center for Drug Evaluation and Research (CDER) and the Center for Biologics Evaluation and Research (CBER), BIMO data packages are included as part of an NDA or BLA submission.

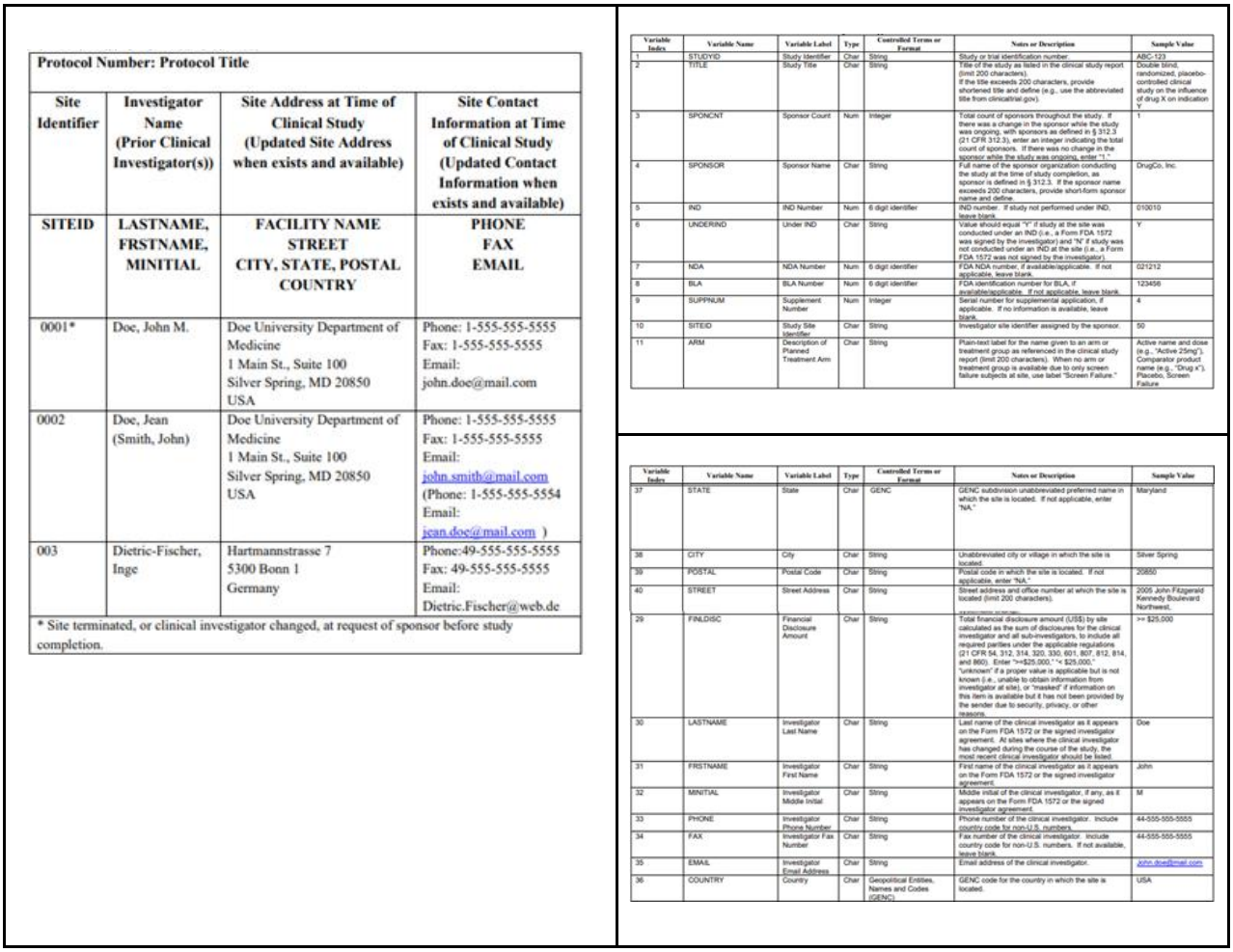
The BIMO program encompasses on-site inspections, data audits, and remote regulatory assessments designed to evaluate the conduct and reporting of FDA-regulated research. These activities ensure the integrity of data submitted to the agency, the protection of research subjects, and compliance with applicable regulations. Importantly, the scope of the BIMO program extends beyond traditional clinical trials to include RWE used in regulatory submissions.

The BIMO data files are submitted within the eCTD format in Module 5, which also contains clinical study reports and associated datasets. To guide sponsors in the preparation of these data packages, the FDA has issued the Bioresearch Monitoring Technical Conformance Guide (19), which provides non-binding recommendations on data standards and submission practices. Unlike study data, these files can be submitted in a variety of formats, including as PDF files. The technical conformance guide outlines three key components of the BIMO data package:

1. **Clinical Study-Level Information:** Provides details about each study included in the submission.
2. **Subject-Level Data Line Listings by Clinical Site:** Offers participant-level data organized by site to facilitate inspection planning and data verification.
3. **Summary-Level Clinical Site Dataset:** Summarizes key data by site to enable identification of patterns or outliers that may warrant inspection.

Figure 2 containing examples of these three listings is provided below.

Figure 2: Example BIMO Data Files



Protocol Number: Protocol Title

Site Identifier	Investigator Name (Prior Clinical Investigator(s))	Site Address at Time of Clinical Study (Updated Site Address when exists and available)	Site Contact Information at Time of Clinical Study (Updated Contact Information when exists and available)
SITEID	LASTNAME, FRSTNAME, MINITIAL	FACILITY NAME STREET CITY, STATE, POSTAL COUNTRY	PHONE FAX EMAIL
0001*	Doe, John M.	Doe University Department of Medicine 1 Main St., Suite 100 Silver Spring, MD 20850 USA	Phone: 1-555-555-5555 Fax: 1-555-555-5555 Email: john.doe@mail.com
0002	Doe, Jean (Smith, John)	Doe University Department of Medicine 1 Main St., Suite 100 Silver Spring, MD 20850 USA	Phone: 1-555-555-5555 Fax: 1-555-555-5555 Email: john.smith@mail.com (Phone: 1-555-555-5554 Email: jean.doe@mail.com)
003	Dietric-Fischer, Inge	Hartmannstrasse 7 5300 Bonn 1 Germany	Phone:49-555-555-5555 Fax: 49-555-555-5555 Email: Dietric.Fischer@web.de

* Site terminated, or clinical investigator changed, at request of sponsor before study completion.

Variable Index	Variable Name	Variable Label	Type	Controlled Terms or Format	Notes or Description	Sample Value
1	STUDYID	Study Identifier	Char	String	Study or trial identification number.	ABC-123
2	TITLE	Study Title	Char	String	Title of the study as listed in the clinical study report (limit 200 characters). If the title exceeds 200 characters, provide shortened title and define (e.g., use the abbreviated title from clinicaltrial.gov).	Double blind, randomized, placebo-controlled clinical study on the influence of drug X on indication Y
3	SPONCNT	Sponsor Count	Num	Integer	Total count of sponsors throughout the study. If there was a change in the sponsor while the study was ongoing, with sponsors as defined in § 312.3 (21 CFR 312.3), enter an integer indicating the total count of sponsors. If there was no change in the sponsor while the study was ongoing, enter "1."	1
4	SPONSOR	Sponsor Name	Char	String	Full name of the sponsor organization conducting the study at the time of study completion, as sponsor is defined in § 312.3. If the sponsor name exceeds 200 characters, provide short-form sponsor name and define.	DrugCo, Inc.
5	IND	IND Number	Num	6 digit identifier	IND number. If study not performed under IND, leave blank.	010010
6	UNDERIND	Under IND	Char	String	Value should equal "Y" if study at the site was conducted under an IND (i.e., a Form FDA 1572 was signed by the investigator) and "N" if study was not conducted under an IND at the site (i.e., a Form FDA 1572 was not signed by the investigator).	Y
7	NDA	NDA Number	Num	6 digit identifier	FDA NDA number, if available/applicable. If not applicable, leave blank.	021212
8	BLA	BLA Number	Num	6 digit identifier	FDA identification number for BLA, if available/applicable. If not applicable, leave blank.	123456
9	SUPPNUM	Supplement Number	Num	Integer	Serial number for supplemental application, if applicable. If no information is available, leave blank.	4
10	SITEID	Study Site Identifier	Char	String	Investigator site identifier assigned by the sponsor.	50
11	ARM	Description of Planned Treatment Arm	Char	String	Plain-text label for the name given to an arm or treatment group as referenced in the clinical study report (limit 200 characters). When no arm or treatment group is available due to only screen failure subjects at site, use label "Screen Failure."	Active name and dose (e.g., "Active 25mg"), Comparator product name (e.g., "Drug x"), Placebo, Screen Failure

Variable Index	Variable Name	Variable Label	Type	Controlled Terms or Format	Notes or Description	Sample Value
37	STATE	State	Char	GENC	GENC subdivision unabbreviated preferred name in which the site is located. If not applicable, enter "NA."	Maryland
38	CITY	City	Char	String	Unabbreviated city or village in which the site is located.	Silver Spring
39	POSTAL	Postal Code	Char	String	Postal code in which the site is located. If not applicable, enter "NA."	20850
40	STREET	Street Address	Char	String	Street address and office number at which the site is located (limit 200 characters).	2005 John Fitzgerald Kennedy Boulevard Northwest.
29	FINLDISC	Financial Disclosure Amount	Char	String	Total financial disclosure amount (US$) by site calculated as the sum of disclosures for the clinical investigator and all sub-investigators, to include all required parties under the applicable regulations (21 CFR 54, 312, 314, 320, 330, 601, 807, 812, 814, and 860). Enter ">=$25,000," "< $25,000," "unknown" if a proper value is applicable but is not known (i.e., unable to obtain information from investigator at site), or "masked" if information on this item is available but it has not been provided by the sender due to security, privacy, or other reasons.	>= $25,000
30	LASTNAME	Investigator Last Name	Char	String	Last name of the clinical investigator as it appears on the Form FDA 1572 or the signed investigator agreement. At sites where the clinical investigator has changed during the course of the study, the most recent clinical investigator should be listed.	Doe
31	FRSTNAME	Investigator First Name	Char	String	First name of the clinical investigator as it appears on the Form FDA 1572 or the signed investigator agreement.	John
32	MINITIAL	Investigator Middle Initial	Char	String	Middle initial of the clinical investigator, if any, as it appears on the Form FDA 1572 or the signed investigator agreement.	M
33	PHONE	Investigator Phone Number	Char	String	Phone number of the clinical investigator. Include country code for non-U.S. numbers.	44-555-555-5555
34	FAX	Investigator Fax Number	Char	String	Fax number of the clinical investigator. Include country code for non-U.S. numbers. If not available, leave blank.	44-555-555-5555
35	EMAIL	Investigator Email Address	Char	String	Email address of the clinical investigator.	john.doe@mail.com
36	COUNTRY	Country	Char	Geopolitical Entities, Names and Codes (GENC)	GENC code for the country in which the site is located.	USA

The BIMO data facilitates the FDA's ability to make data-driven, risk-based decisions regarding inspection prioritization and regulatory compliance. It includes site-level datasets that link to the CDISC DM domain and enable reviewers to evaluate site-specific efficacy and ensure a single site is not driving the results for the study. These datasets include:

1. A single summary-level clinical site dataset that contains data from all major (i.e., pivotal) studies supporting safety and efficacy in the application, including studies with different treatment indications.

2. For clinical investigator sites involved in multiple studies supporting an application, site data should be reported independently for each study within the dataset.

3. For each study and investigator site, the following variables to support efficacy should be submitted:

    a. Endpoint (ENDPOINT)

    b. Treatment Arm (ARM)

    c. Safety Population (SAFPOP)

    d. Primary Efficacy Population (EFFPOP)

e. Treatment Efficacy Results for each objective measured (TRTEFFR1, TRTEFFR2, …)

As BIMO contains investigators and site information that can be linked to CDISC's Demographic (DM) domain, these datasets could be expanded for use with RWD to contain multiple sites and providers. However, the data would need to be capable of storing multiple sites and investigators, along with investigator specialty, and link these to the Subject Visit domain. Any change or standardization of the data would need to ensure the files stay usable for the investigations for which they were designed, and therefore using alternative datasets may be a better option.

Subject Visits: RCTs vs RWD

RCTs and RWD studies are similar in that patients go to healthcare facilities to be treated by healthcare providers, but *how* the visit is conducted and documented and *who* they see in these two cases can be very different. In an RCT, a sponsor company develops a protocol that dictates patient care, which includes, at a minimum, specifying the treatment to administrate and the timing of the administration at each visit. It may also require subjects not to take specific concomitant medications and require assessments at specific visits to evaluate the patient. Investigators are trained on the study protocol and how the patient should be evaluated to ensure consistency with other investigators at other study sites. Each visit has a purpose and a list of activities that must take place as documented in the protocol's SoA; for example, labs may be drawn at baseline to determine if a subject meets the inclusion/exclusion criteria, and if the subject is enrolled, drawn at regular intervals after the baseline visit. The protocol's SoA ensures all patients participating in the clinical trial can be analyzed the same way at the same timepoints for the evaluation of efficacy and safety.

Patients are routinely assessed in person according to the protocol within an RCT, and any identified issues are recorded and managed according to the protocol. The drug of interest is often administered at the site, and it is tracked closely to ensure compliance. Unscheduled visits may occur, and in some cases, the protocol may not be followed. These events are carefully recorded as protocol deviations, and any deviations are captured and submitted to the FDA. All systems are designed to evaluate the patients and record the data as timely and as consistently as possible.

Encounters that take place in the real-world setting differ from those in RCTs. In the US, a patient experiences a symptom or problem and makes an appointment with either their Primary Care Physician (PCP) or a specialist, often depending on their insurance. Given interoperability challenges across healthcare systems, the patient's PCP may or may not be alerted to the fact that they are also seeing a specialist or had a visit to the ER. Although efforts may be made for patients to have at least one physical a year, there is no requirement for this, and patients may go years without seeing a doctor. For healthy patients, visits are likely very sparse, and assessments are not conducted in a systematic way across healthcare providers. Table 1 below highlights the main differences between RCTs and RWD for encounters.

Table 1: RCTs vs. RWD for Subject Visits

Topic	Randomized Controlled Trial	Real World Data
Randomization	**Randomized**	Can be randomized or not randomized; currently most RWD studies are not randomized.
Scheduled Patient Provider Interaction / Healthcare Delivery	**Protocol-defined** visit	Encounter (**patient-initiated**)
Unscheduled Patient Provider Interaction / Healthcare Delivery	Unscheduled visit (**not protocol-defined**)	Encounter (**patient-initiated**)

Treatment Facility and Provider	Subject assigned to **one site and one investigator within SDTM DM**	Site and care provider may vary widely across encounters
Schedule Cadence	Pre-specified and **protocol driven**	Driven by **patient need**, standard of care, reimbursement, etc.
Care Structure	**Driven by a protocol**, upon which all investigators are uniformly trained	Heterogenous and **can vary by site, system, payer, plan type within payer, etc.**
Recording Structure	Highly standardized **Case Report Form (CRF),** upon which study site staff have been trained	Source dependent, often a **non-standardized electronic system** that is customized by the individual provider site
Timing	Nominal VISIT/VISITNUM, actual dates, study days based on reference dates as defined in the protocol	Calendar dates when an encounter took place, **encounters may span multiple dates,** or **multiple encounters may occur on the same date**
Outcomes and Safety	**Prospectively defined endpoints; solicited adverse events** (AEs)	**Opportunistic outcomes; AEs not systemically solicited** and rely on patient and/or physician to initiate the report
Visit Content / Actions Taken	**Protocol-mandated** assessments (e.g., labs with specified panels, PROs, imaging)	Varies per patient **based on provider judgement** and may be influenced by payer decisions

Gaps in SV

In an RCT, many of the details for a visit are described directly in the protocol and are not presented in the datasets, as they are the same for each patient in the RCT. As described above, the information describing an RCT encounter using CDISC domains including SV, DM, and HO, along with the BIMO file, is not robust enough to represent a RWD encounter, where more details often exist within a patient's EHR. Specifically, the SV domain primarily contains timing variables without other visit details to provide context for the encounter. Only one provider (investigator) can be specified in the SDTM DM domain, which includes investigator identifier (DM.INVID) and investigator name (DM.INVNAM). Provider specialty is not specified in DM but is known by the sponsor. Similarly, only one location/facility can be specified in the DM domain within SITEID. Some additional information can be provided within the HO domain, but site or investigator is not standardized in the parent domain.

Within CDISC, DM.SITEID and DM.INVID are identifiers assigned by the sponsor and are not standardized except for Type = 'Char'. SITEID can be used as a unique key in BIMO and in DM, but this is currently not enforced by the standard. This enables a single site to have more than one investigator and a single investigator to support more than one site, but a different SITEID is required for each investigator. To identify the healthcare provider at the visit level, SITEID and INVID would need to be linked to the SV dataset. There is currently no stated rule that the BIMO SITEID and investigator information must match DM.SITEID, DM.INVID, and DM.INVNAM, although this can be assumed. Matching BIMO and DM SITEID could be electronically implemented, but INVNAM would be harder to implement electronically given the field formatting differences between BIMO and the CDISC DM dataset. As of 2024, FDA is working to better standardize the BIMO file, but the results of this work have yet to be announced.

RWD MODELS AND ENVIRONMENTAL SCAN

Existing RWD models can be leveraged to understand how encounter data is represented across other models and to address gaps identified within CDISC. A summary is provided below for four data models, including HL7 FHIR, OMOP, PCORnet, and Sentinel.

FHIR

Fast Healthcare Interoperability Resources (FHIR) is a standard for healthcare data exchange published by HL7 in 2012 to replace HL7 V2. It began being adopted within EHR systems after the 2020 mandate by the Assistant Secretary for Technology Policy/Office of the National Coordinator for Health Information Technology (hereafter referred to as 'ASTP'). FHIR provides a framework for the exchange, integration, sharing, and retrieval of healthcare information across healthcare systems. The primary use cases for FHIR are to exchange healthcare and medical claims data. In 2019, over 80% of hospitals implemented API technology enabled with FHIR (20). Within FHIR, data elements are grouped into resources in a manner that supports business objectives for medical health records and claims data, and the FHIR implementation guides specify rules for the variables that are required, in addition to specifying terminology to be used. For those familiar with CDISC, resources are similar to domains. One key difference between the two standards is that FHIR allows resources to be linked directly to other resources without the use of additional 'support' data (e.g., the CDISC Related Records (RELREC) dataset). However, this requires some decisions to be made if this data is later represented in tabular format or in a relational database.

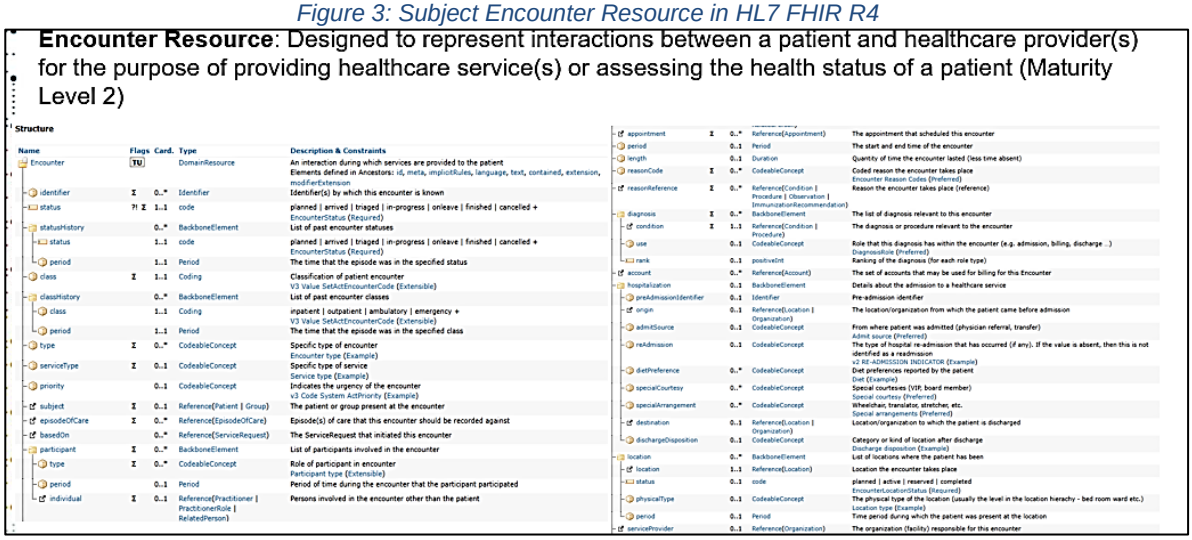
Read more about HL7 FHIR

Encounter Data in HL7 FHIR

Given that the primary use case for FHIR is exchanging EHR and claims data, FHIR contains detailed information about patients' interactions with healthcare systems, associated providers, and healthcare sites. FHIR contains three resources that are particularly applicable to patient encounters: ENCOUNTER, ORGANIZATION, and LOCATION. These resources are presented in Figures 3 and 4.

The ENCOUNTER resource is used to document the actual activities that occurred during a patient's encounter with a healthcare provider (HCP). This resource captures a range of information, including encounter status (e.g., planned, in-progress, on-hold, discharged, completed), class of encounter (e.g. inpatient, outpatient), priority, type of encounter (e.g. e-mail consultation, surgical day-care, skilled nursing, rehabilitation), services provided, reason for the encounter, what the encounter was based on or part of, the organization responsible for the encounter, the encounter period (start and end date/times), providers involved in the encounter, and admission and discharge information. The ENCOUNTER resource also contains a unique identifier for the encounter and data elements linking to ORGANIZATION and LOCATION resources.

Figure 3: Subject Encounter Resource in HL7 FHIR R4

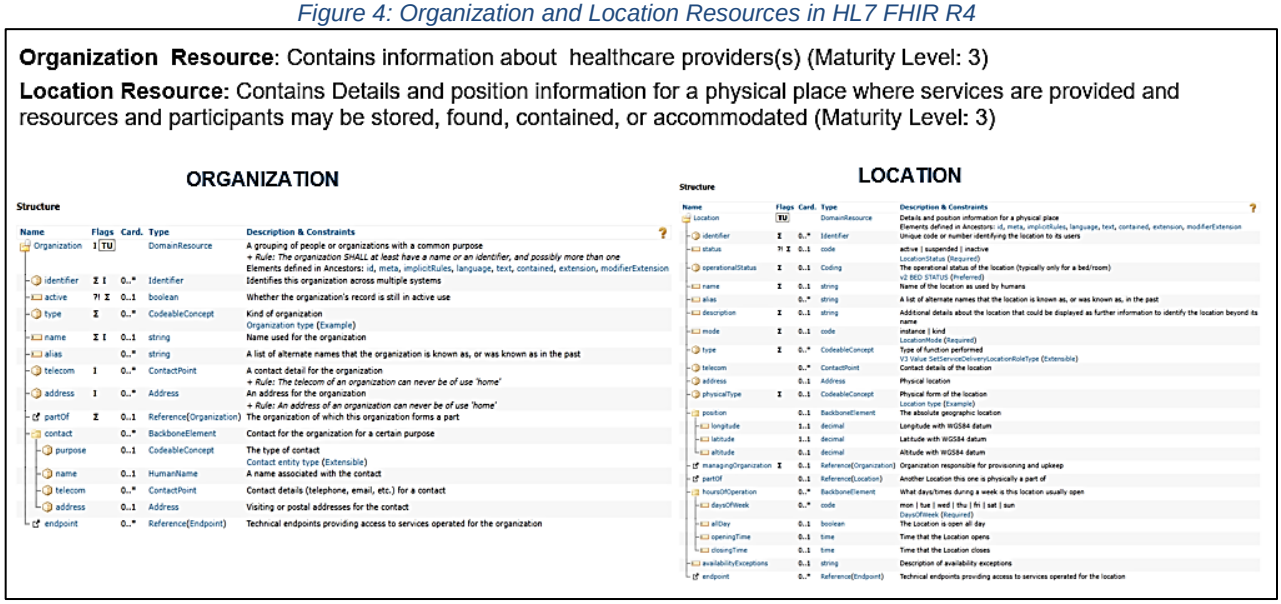


The ORGANIZATION resource is used to specify details about organizations involved in healthcare such as HCPs and payers/insurers. It provides data on the type of organization, name and description of the

organization, contact details, and qualifications/accreditations of the organization. The ORGANIZATION resource also has data elements to link to both the ENCOUNTER and LOCATION resources.

The LOCATION resource provides details of where the services were provided. It contains information on the physical location (address) as well as the geographic location (longitude, latitude, and altitude) of an HCP. The LOCATION resource also has data elements to link to both the ENCOUNTER and ORGANIZATION resources.

Figure 4: Organization and Location Resources in HL7 FHIR R4



Like CDISC, FHIR establishes resources that can be customized for a specific purpose; thus, not all the elements need to be included in an implementation. In addition, there are certain variables that must be supported in FHIR and others that are considered optional. For example, the LOCATION resource could be implemented without the geographic location variables such that longitude, latitude, and altitude are excluded. This provides some flexibility but also requires a group to create a specific implementation guide for the intended use.

OMOP

The Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM), currently maintained by Observational Health Data Sciences and Informatics (OHDSI) consortium, is an open community data standard, designed to standardize the structure and content of observational healthcare data. Originally developed under an NIH grant, the OMOP CDM provides a standardized framework for integrating and analyzing heterogeneous healthcare data, including EHRs, claims, and registries. Its goal is to harmonize diverse data sources through a common structure and language, enabling efficient, reliable analyses, cross-study collaboration, and data sharing.

Designed to focus on observational data and reflect real-world clinical practices and patient outcomes, it stands in contrast to the CDISC data standard which was developed to represent RCT data. 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 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. OMOP also provides tools to evaluate the quality of data formatted in the OMOP CDM. A framework established by Kahn et al. defines data quality (DQ) as consisting of conformance, completeness, and plausibility. More on OMOP's data quality can be found in The Book Of OHDSI, Chapter 15 (21).

Read more about the OMOP Common Data Model

Read more about OHDSI's standardized vocabularies

Encounter Data in OMOP

OMOP contains four tables especially relevant to capturing a subject's engagement with the healthcare system: VISIT_OCCURENCE, LOCATION, PROVIDER, and CARE_SITE. These tables are described in Figure 5.

Figure 5: Subject Encounter Related Data in OMOP v5.4

visit_occurrence

Table Description

This table contains Events where Persons engage with the healthcare system for a duration of time. They are often also called "Encounters". Visits are defined by a configuration of circumstances under which they occur, such as (i) whether the patient comes to a healthcare institution, the other way around, or the interaction is remote, (ii) whether and what kind of trained medical staff is delivering the service during the Visit, and (iii) whether the Visit is transient or for a longer period involving a stay in bed.

location

Table Description

The LOCATION table represents a generic way to capture physical location or address information of Persons and Care Sites.

provider

Table Description

The PROVIDER table contains a list of uniquely identified healthcare providers; duplication is not allowed. These are individuals providing hands-on healthcare to patients, such as physicians, nurses, midwives, physical therapists etc.

care_site

Table Description

The CARE_SITE table contains a list of uniquely identified institutional (physical or organizational) units where healthcare delivery is practiced (offices, wards, hospitals, clinics, etc.).

Source : https://ohdsi.github.io/CommonDataModel/cdm54.html#visit_occurrence

The VISIT_OCCURENCE table is the primary table documenting the visit or encounter. The table captures each encounter that a person has with the healthcare system and represents a span of time during which a person receives medical services at a care site in a specific setting. Key concepts include start date/time, end date/time, type of visit (inpatient, outpatient, emergency, etc.), where the patient was admitted from and discharged to, a unique identifier for the encounter, and the provenance of the record. This table also includes especially helpful linking information: a provider identifier that links to the PROVIDER table, a care site (or facility) identifier that links to the CARE_SITE table, and an identifier that links to the visit that occurred prior to the given visit.

The PROVIDER table specifically contains information pertaining to the healthcare professionals providing hands-on healthcare to patients (e.g., physicians, nurses, physical therapists). In addition to a provider identifier, this table includes data elements for provider name, National Provider Identifier (NPI), provider specialty, provider demographic information, and a care site identifier that links to the CARE_SITE table.

The primary purpose of the CARE_SITE table is to provide information on healthcare facilities. These can include hospitals, clinics, departments, or other organizational units. This table stores the unique list of hospitals, clinics, physician offices, pharmacies, nursing homes, or any other location where healthcare is rendered. The CARE_SITE table includes a unique care site identifier, the full name of the care site, the type of care site (e.g., Hospital, Clinic, Pharmacy, Emergency Room), and a location identifier, linking the CARE_SITE table to the LOCATION table.

The LOCATION TABLE captures information about the physical location or address of a care site. It supports the identification of physical places such as care sites, patient residences, or provider offices. This table has a unique identifier for a given location as well as data elements for the address, city, state, zip code, and county for a care site.

PCORNET Common Data Model (CDM)

The Patient-Centered Outcomes Research Institute (PCORI), authorized by Congress in 2010, is an independent, nonprofit research organization that funds patient-centered comparative clinical effectiveness research (CER). The PCORnet CDM is a data standards specification that defines the organization and representation of data for the PCORnet Distributed Research Network (DRN). The PCORnet DRN allows multi-site patient-centered research across the Clinical Research Networks (CRNs) and other interested contributors. Like Sentinel and OHDSI, the PCORnet distributed network facilitates the conduct of observational research while allowing each participating organization to maintain physical and operational control over its own data. The PCORnet DRN operations center and infrastructure, including the CDM, is led by the PCORnet Coordinating Center and overseen by the governance established by PCORnet's stakeholders.

The PCORnet CDM, first released in 2014, originally based on the Mini-Sentinel Common Data Model v4.0 (MSCDM v4.0). It has evolved over time and is informed, in part, by the development of other distributed initiatives such as the HMO Research Network, the Vaccine Safety Datalink, various AHRQ Distributed Research Network projects, and the ONC Standards & Interoperability Framework Query Health Initiative. The PCORnet CDM uses standard healthcare terminologies including ICD, SNOMED, CPT, HCPCS, and LOINC® to enable interoperability. The most recent version of CDM, V6.1, was released in 2023. Version 4.0 of the PCORnet CDM introduced the concept of expanded value sets for fields with dozens or hundreds of allowable options, some related to exposure including DOSE_FORM, ROUTE, and UNIT as well as others such as LAB_RESULT_CM and RESULT_UNIT.

All data available through PCORnet is subject to a two-stage quality assessment process. The first stage checks data for conformance to ensure it adheres to the CDM and checks for completeness, including that diagnosis codes are aligned, plausible, and persistent, to ensure that records are maintained between refreshes. A list of all checks is provided on the PCORnet website:

https://pcornet.org/wp-content/uploads/2024/08/PCORnet-Data-Checks-v16.pdf Carefully selected data quality metrics accompanying RWE, and the underlying RWD when applicable, could enable the FDA to more efficiently assess and interpret the data under review. For example, it could aid in understanding whether missing values appear to be missing at random or if there is a pattern to missingness. It could also highlight if the selected data differs in quality from the source data. Data quality itself is a large topic that is not covered in detail in this paper but may possibly be covered in a subsequent presentation.

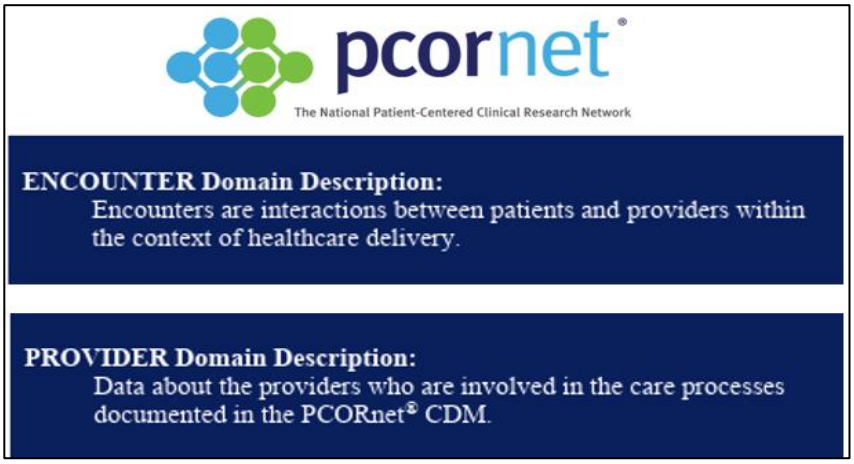
Read more about the PCORnet

Read more about PCORnet data

Encounter Data in PCORnet

PCORnet contains two tables related to a subject's engagement with the healthcare system: ENCOUNTER, and PROVIDER. These tables are described in Figure 6.

Figure 6: Subject Encounter Related Data in PCORnet v7.0



The ENCOUNTER table captures interactions between patients and healthcare providers in the context of healthcare delivery. Each record represents a distinct healthcare event, such as a hospital stay, clinic visit, or emergency department encounter. It is similar to the OMOP VISIT_OCCURENCE table and serves as the primary table capturing an encounter between a patient and an HCP. The ENCOUNTER table contains data elements representing unique encounter ID, start date/time, end date/time, discharge date/time and other discharge information, type of encounter (inpatient, outpatient, emergency, etc.), where the patient was admitted, and a unique provider ID that links to the PROVIDER table. In contrast to OMOP, the ENCOUNTER table also contains information about the facility or care site. This includes a facility identifier, the type of facility, and the geographic location (5-digit zip code) of the facility.

The PROVIDER table contains information about healthcare professionals involved in patient care. It supports linking encounters and clinical events to the individuals responsible for delivering services. Like OMOP, the PROVIDER table contains a provider ID that links to the ENCOUNTER table and concepts for provider name, NPI, provider demographic information, and provider specialty.

Sentinel

The FDA Sentinel System is a national electronic surveillance system designed to monitor the safety of medical products after they reach the market. It was established in response to a 2007 congressional mandate under the Food and Drug Administration Amendments Act (FDAAA), which directed the agency to enhance its capacity for active post-market surveillance of drugs, biologics, and medical devices.

Sentinel serves as a cornerstone of FDA's RWE and pharmacovigilance initiatives, demonstrating how a CDM can transform post-market safety monitoring into an agile, scalable, and scientifically rigorous system for protecting public health.

Launched in 2008, Sentinel represented a major evolution from passive, spontaneous adverse event reporting toward a proactive, data-driven approach. The system enables the FDA to assess the safety of regulated medical products using routinely collected healthcare data from millions of patients across the United States. The data initially came from claims and later, EHRs were added. The system enables FDA and other stakeholders to answer questions about the safety of marketed products using large-scale observational data containing real-world use and outcomes.

Read more about FDA's Sentinel Initiative.

Encounter Data in Sentinel

At the core of Sentinel's infrastructure is the Sentinel Common Data Model (SCDM) that is used to format disparate data sources into a consistent format. This allows distributed partners such as healthcare systems, insurers, and academic partners to maintain control of their data locally while participating in federated analyses using standardized queries. The SCDM currently includes 19 tables that represent information for the data elements needed for Sentinel activities. Several identifiers are used to link records across tables: these include, but are not limited, to a unique person identifier, PatID, a unique provider identifier, ProviderID, a unique facility identifier FacilityID, and a unique encounter identifier, EncounterID. An overview of the SCDM is shown in Figure 7, and the facility, provider, and encounter tables are shown in Tables 2-4.

Figure 7: Sentinel CDM Overview

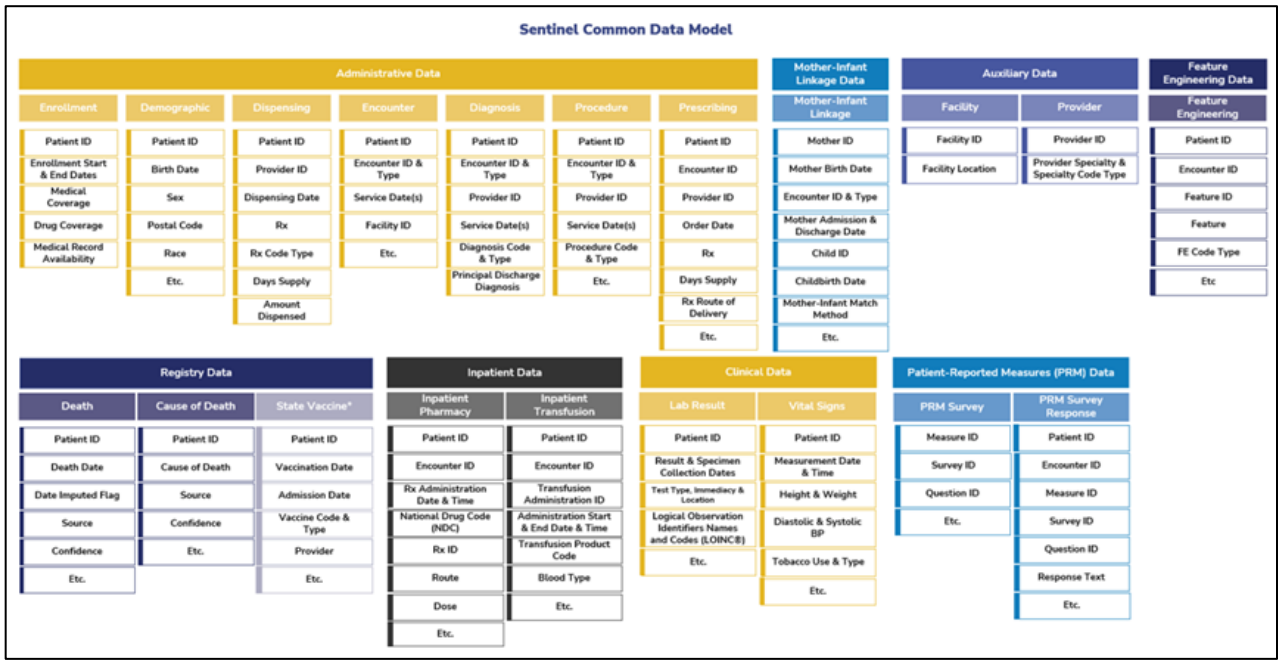


Table 2: SCDM Facility Table

VARIABLE NAME	VALUES	STATUS CORE
FACILITYID	SERVICING FACILITY IDENTIFIER	ONLY SAS SPECIAL MISSING ALLOWED
FACILITY_LOCATION	GEOGRAPHIC LOCATION	NOT REQUIRED

Table 3: SCDM Provider Table

Variable Name	Values	Status Core
ProviderID	Unique provider identifier	Only SAS special missing allowed
Specialty	Provider specialty	Not required
Specialty_CodeType	CMS or NPI	Not required

Table 4: SCDM Encounter Table

Variable Name	Values	Status / Core
EncounterID	Unique encounter identifier	Required
PatID	Unique patient identifier	Required
ADate	Encounter or admission date	Required
DDate	Discharge date	Conditional on 'EncType' value
EncType	Encounter Type	Required
FacilityID	Servicing provider identifier	Required
Discharge_ Disposition	Discharge Disposition Code	Conditional on 'EncType' value
Discharge_Status	Discharge Location or Type Information	Conditional on 'EncType' value
DRG	Diagnosis Related Group	Conditional on 'EncType' value
DRG_Type	DRG Version	Conditional on 'EncType' value
Admitting_Source	Admitting Facility or Healthcare Professional Type	Conditional on 'EncType' value

In the SCDM, the Facility and Provider tables each have unique identifiers, which are referred to as primary keys. Facility ID is used as a foreign key in the Encounter table, identifying treatment facility of the encounter. Provider ID is not a foreign key in the Encounter table but is instead a foreign key in the Diagnosis, Procedure, and Prescribing tables, allowing for identification of the healthcare provider who made the diagnosis, performed the procedure, and/or prescribed the drug in the encounter. This allows a level of granularity not available in the other RWD models. The Diagnosis, Procedure, and Prescribing tables also contain Encounter ID which identifies the encounter when the patient was treated.

RECOMMENDATIONS

Our recommendations are grounded in regulatory reviewer needs as documented in FDA guidance and as implemented based on our review of the data standards optimized for RWD presented in the previous section. Reviewer needs can be grouped into the following categories:

1. The need to understand the patient's journey during routine care;
2. The need to facilitate the analysis of potential bias from the provider and/or site;
3. The need for more comprehensive information surrounding a patient's encounter with an HCP;
4. The need for more information about the providers a patient interacts with;
5. The need for more information about the facility where the encounter occurred; and
6. The need for the ability to represent a patient encounter with more than one provider and/or site.

Given that CDISC standards are currently required by FDA to submit data supporting new marketing applications, we recommend submitting the collected data from all available sources in SDTM format. A short-term solution for RWD data is to use custom domains and/or supplemental qualifiers to address gaps. This may mean creating a custom domain for additional subject visit domain and/or adding supplemental variables to the SDTM SV domain within a SUPPSV dataset. It is recommended that specific options be discussed with the review team to see if there are recommendations from regulators based on what conventions may have been commonly used for representing nonstandard domains and variables. A mid-term solution would be for FDA or CDISC to predefine the names of the custom domains and/or supplemental qualifiers that can be used to represent the six recommendations above, as was done for the supplemental qualifier treatment emergent for AEs. For the long-term solution, we draw upon the FHIR HL7, OMOP, PCORnet, and Sentinel data models to improve the representation, traceability, and provenance needed to support the regulatory review of RWD.

Tables 5, 6, and 7 summarize our review of RWD models presented in the previous sections. For each key concept, we denote whether the concept is present or absent from a given model. These tables aim to facilitate the process of proposing new variables to enhance the CDSIC SDTM SV domain and to propose new domains that will store details about providers and healthcare facilities.

Table 5: Concepts to Address Gaps in the CDISC SDTM SV Domain

Concept	Definition	CDISC SDTMIG v3.4	HL7 FHIR R4	OMOP v5.4	PCORnet v7.0	Sentinel v8.2.1
Encounter ID	Source encounter identifier to identify unique interactions between an individual patient and the healthcare system	N/A (uses VISITNUM)	Encounter. Identifier	Visit_occurence. visit_ occurrence_id	Encounter. encounterid	EncounterID
Provider ID	Unique identifier for provider	N/A	Encounter. Service Provider	Visit_occurence. provider_id	Encounter. providerid	ProviderID
Facility ID	Unique identifier for facility (care site)	N/A	Organization. Identifier	Visit_occurence. care_site_id	Encounter. Facilityid	FacilityID
Appointment	Link to appointment that scheduled this encounter	N/A	Encounter. appointment	N/A	N/A	N/A
Part of	Link to another care	N/A	Encounter.	N/A	N/A	N/A

Concept	Definition	CDISC SDTMIG v3.4	HL7 FHIR R4	OMOP v5.4	PCORnet v7.0	Sentinel v8.2.1
care plan	plan encounter this encounter is part of		partOf			
Urgency	Indicates the urgency of the encounter	N/A	Encounter. priority	N/A	N/A	Derivable from Admitting_ Source
Part of	Link to another encounter this encounter is part of	N/A	Encounter. partOf	N/A	N/A	Derivable from Admitting_ Source
Based on	Request that initiated this encounter	N/A	Encounter. basedOn	N/A	N/A	Admitting_ Source
Status	Current state of the encounter (e.g., completed, in-progress)	N/A	Encounter. status	N/A	N/A	Discharge_ Status
Service	Broad categorization of the service that is to be provided (e.g., cardiology)	N/A	Encounter. serviceType	N/A	N/A	DRG (Diagnosis Related Group)
Reason for visit	List of medical reasons that are expected to be addressed during the episode of care	N/A	Encounter. Reason	N/A	N/A	No
Reason for missed visit	Reason for missed encounter	SVREASOC	Appointment. cancellation Reason	N/A	N/A	No
Class	Classification of patient encounter (e.g., ambulatory (outpatient), inpatient, emergency, home health)	N/A	Encounter. Class	N/A	N/A	EncType
Discharge status	Category or location after discharge (e.g., home, long-term care)	N/A	Encounter. admission. Discharge Disposition	Visit_ Occurence. discharged_to_ source_value	Encounter. discharge_ status	Discharge_ Status

Table 6: Concepts to Model Provider Information

Concept	Definition	CDISC SDTMIG v3.4	HL7 FHIR R4	OMOP v5.4	PCORnet v7.0	Sentinel v8.2.1
Provider ID	Unique Provider Identifier for the practitioner responsible for a given encounter, could be NPI	DM.INVID (limited to one provider per subject in DM)	Practitioner. identifier	Visit_occurrence. provider_id; Provider.NPI	Encounter. providerid; Provider.NPI	Unique provider identifier (UPI)

Concept	Definition	CDISC SDTMIG v3.4	HL7 FHIR R4	OMOP v5.4	PCORnet v7.0	Sentinel v8.2.1
Specialty	The specific type of healthcare provider or field of expertise	N/A	PractitionerRole. specialty	Provider. specialty_ source_vaue	Provider. provider_ specialty_ primary	Provider. specialty

Table 7: Concepts to Model Healthcare Facility Information

Concept	Definition	CDISC SDTMIG v3.4	HL7 FHIR R4	OMOP v5.4	PCORnet v7.0	Sentinel v8.2.1
Facility ID	Identifier for the care site	DM.SITEID (limited to one site per subject)	Organization. identifier	Visit_occurence. care_site_id	Encounter. facilityid	FacilityID
Facility Name	Name of the facility responsible for the encounter	N/A	Encounter. serviceProvider	Care_site.care_ site_name	N/A	Linkable from FacilityID
Facility Type	The type of facility (e.g., hospital, ER, urgent care, ambulatory visit)	N/A	Organization. Type	N/A	Encounter. facility_type	Can be derived from Encounter Type
Location	Geographic location or address of site where healthcare was received	N/A	Location. address	Location. Address_1	Encounter. facility_ location	Facility_ Location

The RWD encounter tables or resources reviewed above contain a unique identifier for the given encounter, and a facility identifier linked from a dataset containing facility information, such as location. Sentinel differs from the three other RWD models in that it does not have a provider identifier in the encounter table. Instead, it stores the provider identifier in the Diagnosis, Procedure, Prescribing, and Dispensing tables along with the associated encounter identifier. This representation allows different providers to perform different roles. For example, in the case where a psychologist works with a psychiatrist, the former may provide the diagnosis, and the latter may prescribe the drug. It is unclear how common it is to have multiple providers at a single visit, and ultimately there is only one physician responsible for an EHR entry, but this scenario should be considered. At a minimum, we recommend adding a provider table and a facility table and linking these to the CDISC SV domain. If the source of the data is EHRs, we also recommend adding as many of the variables that are available listed in Table 5 to provide useful contextual information about the visit. Submitting this additional information can help increase a reviewer's confidence and understanding of the patient's overall journey. For example, in a case where the patient goes directly from the PCP to the ER, the data quality may be lower than when a physician sees a patient for a physical and the patient goes home.

In RCTs, the protocol specifies *who* performs the assessment and *where* the assessment is completed. As a result, this information is usually not captured during data collection or presented in an SDTM dataset, nor has it been required by reviewers. Since RWD is not collected under a protocol, submitting this additional information surrounding the context of an encounter can better facilitate regulatory review. The rationale for providing these contextual variables is that all sites are not equal (e.g., a large medical center versus a corner drug store care clinic) and that clinical staff have different skillsets, experience levels, and specializations. In addition, coding is often not consistent from location to location, and healthcare reimbursement can also differ from state to state which can affect how coding is done. Data models such as HL7 FHIR, OMOP, PCORnet, and Sentinel take a similar approach to providing this information; encounter/visit resources or tables contain a primary (or uniquely identifying) key for Encounter ID, and that record contains the unique Facility or Location ID where the visit took place. The corresponding Provider

table and the Facility/Location table (or resource) contain more information about the provider and facility where the encounter took place. Note that the Sentinel model has the provider ID linked to the Diagnosis, Prescription, Procedure, and Dispensing tables, whereas the other three RWD models directly link provider to encounter as noted above. To facilitate regulatory review in the long-term, we recommend implementing the most common strategies used by these other models:

1. Create a PROVIDER domain as a custom domain to store details about providers, including provider specialty. This table should have a primary key for Provider ID that can be used to link the SV dataset to the PROVIDER domain.
2. Create a FACILITY or CARE SITE domain as a custom domain to store details about facilities. This table should have a primary key for Facility ID that can be used to link the SV dataset to the FACILITY or CARE SITE domain.
3. Add Provider ID and Facility ID to the SDTM SV dataset, as well as other applicable domains, as supplemental qualifiers to the parent domains.

At minimum, the PROVIDER table should contain a unique provider identifier for the practitioner responsible for a given encounter which could be the NPI, along with the specialty for the given provider (i.e., the specific type of healthcare provider or their field of expertise). The FACILITY table should contain at least the unique identifier for the facility, the name of the facility responsible for the encounter, the facility type (e.g., hospital, ER, urgent care, ambulatory visit), and the geographic location or address where healthcare was received. We recommend reviewing the utility of having the Provider ID linked to encounter or if a more granular view is needed to link to Diagnosis, Prescription, Procedure, and Dispensing tables. This may not be particularly useful given the granularity of RWD and adds complexity to the standard.

Our next set of recommendations suggest more updates to the CDISC SDTM DM domain. We recommend that CDISC redefine SITEID and INVID as primary site identifier and primary care physician or primary treating physician (e.g., in the case of cancer treatment). These could remain the site and investigator that also link to BIMO. Alternatively, DM could contain a Provider ID and Facility ID that link to the custom Provider and Facility tables, respectively.

A more long-term option would be to introduce cardinality to the DM, SV and other SDTM domains as done in HL7 FHIR. This would allow for multiple providers (or investigators) and facilities for a given subject or encounter without the need for non-standard variables. Currently, this is not an option as CDISC datasets are submitted in SAS Transport format, which is limited to rectangular data structures. However, The FDA is exploring JSON as a new exchange standard, with the goal of replacing SAS version 5 XPORT Transport Format (XPT) for submission of electronic study data to CDER and CBER. While the initial JSON specification is limited to rectangular representation of data, JSON in general does not have this limitation. Future specifications for submitting clinical data in JSON will most likely be extended to accommodate non-rectangular data structures.

CONCLUSION

Due to the mandates of the 21st Century Cures Act, other regulations including the HITECH Act and ACA, and the desire to learn from patients' experiences that are documented across various sources of RWD, there is an expectation from industry that RWD should be included in marketing applications for medical products. Over the past several years, marketing applications have evolved from a collection of RCTs to RCTs, RWD studies, and other non-RCT study designs. However, the standards for submitting study data have not evolved as CDISC currently remains the only available standard. While RWD has the potential to provide additional information about the benefits or risks of a medical product, RWD studies pose numerous challenges to regulatory reviewers as presented throughout this paper.

The data available from RWD sources is collected using both different data standards and different file formats. Further, the lack of monitoring during the data collection and the lack of documentation can lead to quality issues which may impact the evaluation of safety and efficacy, particularly if the study population is small and the subjects’ selection is not randomized. As a result, additional information is often required to

provide regulatory reviewers with confidence and understanding of the data. At this time, RWD must be submitted in CDISC format, which was developed primarily for RCTs and data collected under a protocol. CDISC may not be a great fit for RWD but can be supplemented by additional custom domains and non-standard variables. We can learn a great deal from RWD and non-RCT data standards such as HL7 FHIR, OMOP, PCORnet, and Sentinel. These standards can inform us of how RWD is represented in standards optimized for RWD. However, it is important to keep in mind the use case(s) and assumptions under which each standard was developed as well as what options are available for customization to ensure the best representation strategy is chosen. In this paper, we have illustrated how RWD standards model subject visit/encounter data and other associated data, and how these strategies could be implemented to supplement a CDISC SDTM submission to facilitate regulatory review.

While our recommendations provide a short term solution, it is incumbent on sponsors, regulatory authorities, standards development organizations, and other industry-related groups to collaborate and agree upon the data elements needed to ensure an accurate efficacy and safety profile of a drug based on RWD, as well as whether to design a new data standard, choose an existing one, or modify an existing one (either directly or via an implementation guide) that will allow for more efficient and effective regulatory review of real-world data. Whether the future is a single standard or a hybrid approach, regulatory reviewers are tasked with ensuring the safety of the public and require submission standards that allow them to efficiently understand the quality and certainty of RWD as well as other information that contributes toward regulatory decision making.

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. Ferko, S., Holt, I., & Abolafia, J., (2024). “Future Clinical Data Submission Standards: CDISC, FHIR, OMOP, or Hybrid Model”. Paper presented at the PHUSE US Annual Conference 2024, Bethesda, MD. https://phuse.s3.eu-central-1.amazonaws.com/Archive/2024/Connect/US/Bethesda/PAP_RE03.pdf

5. Abolafia, J, Ferko, S, & Holt, I. (2024). “Considerations for the Submission of RWD using CDISC with Insights from HL7 FHIR and OMOP”. Paper presented at the PHUSE Annual Conference 2024, Strasbourg, France. https://phuse.s3.eu-central-1.amazonaws.com/Archive/2024/Connect/EU/Strasbourg/PAP_RE03.pdf

6. Abolafia, J, Ferko, S, & Holt, I. (2025). “AE, CE, & MH Considerations for the Submission of RWD using CDISC, with Insights from HL7 FHIR and OMOP”. Paper presented at the PHUSE Annual Conference 2025, Orlando, FL. https://phuse.s3.eu-central-1.amazonaws.com/Archive/2025/Connect/US/Orlando/PAP_RE02.pdf

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

8. 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.

9. 21st Century Cures Act. (2016). https://www.congress.gov/114/plaws/publ255/PLAW-114publ255.pdf

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

11. FDA Guidance (2024). "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

12. FDA Guidance. (2023). "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

13. U.S. Food & Drug Administration. (2025, March). *Data Standards Catalog*. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/data-standards-catalog

14. “Exploration of Health Level Seven Fast Healthcare Interoperability Resources for Use in Study Data Created From Real-World Data Sources for Submission to the Food and Drug Administration; Establishment of a Public Docket; Request for Comments”. Accessed 2025 Oct 23. https://www.federalregister.gov/documents/2025/04/23/2025-06967/exploration-of-health-level-seven-fast-healthcare-interoperability-resources-for-use-in-study-data

15. CDISC’s Controlled Terminology webpage. 26 September 2025. https://www.cdisc.org/standards/terminology/controlled-terminology.

16. Clinical Data Interchange Standards Consortium. (2025). SDTM | CDISC. https://www.cdisc.org/standards/foundational/sdtm

17. Clinical Data Interchange Standards Consortium. (2025). SDTMIG | CDISC. https://www.cdisc.org/standards/foundational/sdtmig

18. U.S. Food & Drug Administration. (2022, August 4). *Bioresearch Monitoring Program Information*. https://www.fda.gov/inspections-compliance-enforcement-and-criminal-investigations/fda-bioresearch-monitoring-information/bioresearch-monitoring-program-information

19. U.S. Food & Drug Administration. (n.d.). *Bioresearch Monitoring Technical Conformance Guide: Standardized Format for Electronic Submission of NDA/BLA Content for BIMO Inspections* (Document no. 85061). https://www.fda.gov/media/85061/download

20. Posnack, Steven and Barker, W. (2021). “The Heat is On: US Caught FHIR in 2019”. The Heat is On: US Caught FHIR in 2019 - Health IT Buzz

21. The Book of OHDSI (2021). https://ohdsi.github.io/TheBookOfOhdsi/

CONTACT INFORMATION

Jeffrey Abolafia
Certara/Pinnacle 21
mailto:jeffrey.abolafia@certara.com