

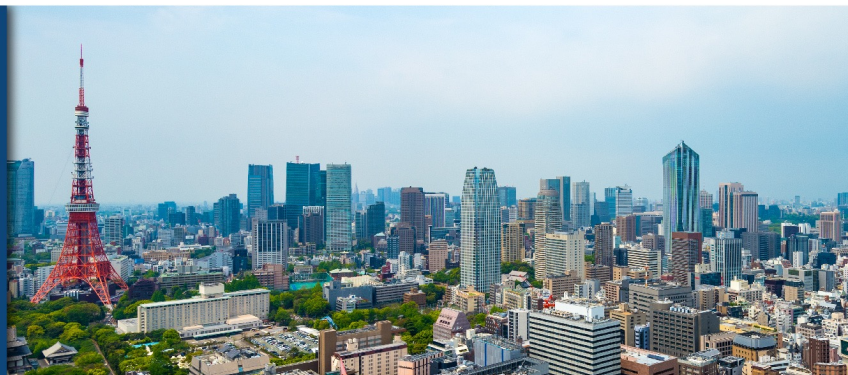


2023

JAPAN

INTERCHANGE

TOKYO | 10-11 JULY



Updates on CDISC Data Science Projects

Sam Hume, DSc
VP, Data Science
CDISC

Meet the Speaker

Sam Hume

Title: VP, Data Science

Organization: CDISC

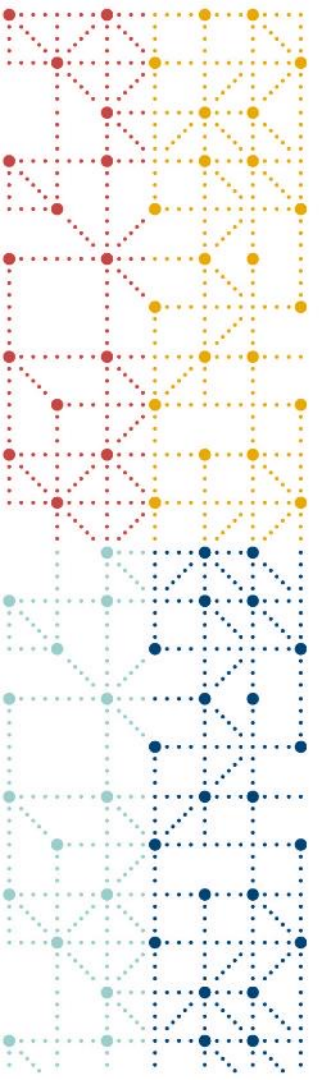
Sam Hume leads the CDISC Data Science team, which collaborates with CDISC staff and stakeholders to develop tools and standards that support clinical and translational data science. Sam directs delivery of the CDISC Library metadata repository that houses all CDISC standards, co-leads the CDISC Data Exchange Standards team, co-leads CORE, and leads the technical CDISC RWD efforts. He has 25 years' experience in clinical research informatics and has held a number of senior technology positions in the biopharmaceutical industry. He holds a doctorate in information systems.



A decorative graphic on the left side of the slide, consisting of a grid of small dots connected by thin lines. The dots are colored in a gradient from yellow at the top to blue at the bottom. The lines are also colored in a gradient from yellow to blue. The overall effect is a stylized, abstract pattern.

Agenda

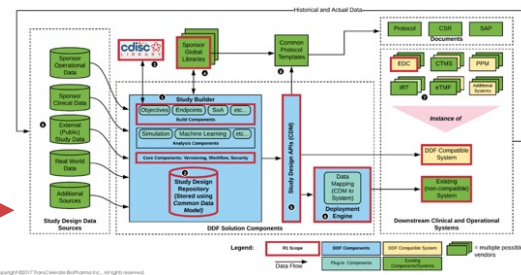
1. CDISC Library
2. ODM v2.0
3. Dataset-JSON Pilot
4. COSA
5. CORE
6. Biomedical Concepts
7. OAK – SDTM Automation
8. CDISC Data Exchange Framework



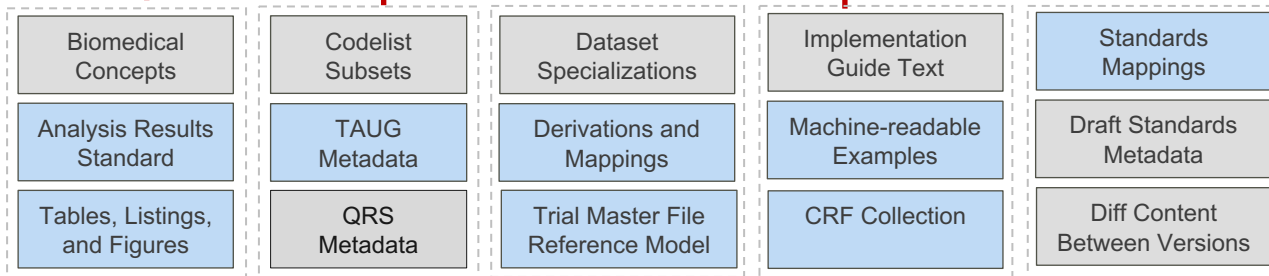
CDISC Library

CDISC Library: Standards as a Service

Software Applications Consume Standards Metadata via the API



Executable Conformance Rules



REST architecture principles at work

CDISC Library Data Standards Browser



Data Standards Browser

Search



Dashboard

Expand All

Filter Products

Data Collection

Data Tabulation

SDTM v2.0

SDTM v1.8

SDTM v1.7

SDTM v1.6

SDTM v1.5

SDTM v1.4

SDTM v1.3

SDTM v1.2

SDTMIG v3.4

SDTMIG-MD v1.1

SDTMIG v3.3

SDTMIG-AP v1.0

SDTMIG v3.2

SDTMIG-MD v1.0

SDTMIG v3.1.3

SDTMIG v3.1.2

SENDIG v3.1.1

SENDIG-AR v1.0

SENDIG-DART v1.1

SENDIG v3.1

SENDIG v3.0

Data Analysis

QRS Instruments

Terminology

Draft Content

SDTMIG v3.4

Status: Final
Effective Date: 2021-11-29
Implements: SDTM v2.0

Export

Classes

General Observations Interventions Events Findings Findings About Special-Purpose Trial Design Study Reference Relationship

Data Sets

BS CP CV DA DD EG FT GF IE IS LB MB MI MK MS NV
OE PC PE PP QS RE RP RS SC SS TR TU UR VS

Findings VS

Name: Structure
Vital Signs: One record per vital sign measurement per time point per visit per subject

Description: A findings domain that contains measurements including but not limited to blood pressure, temperature, respiration, body surface area, body mass index, height and weight.

Vital Signs

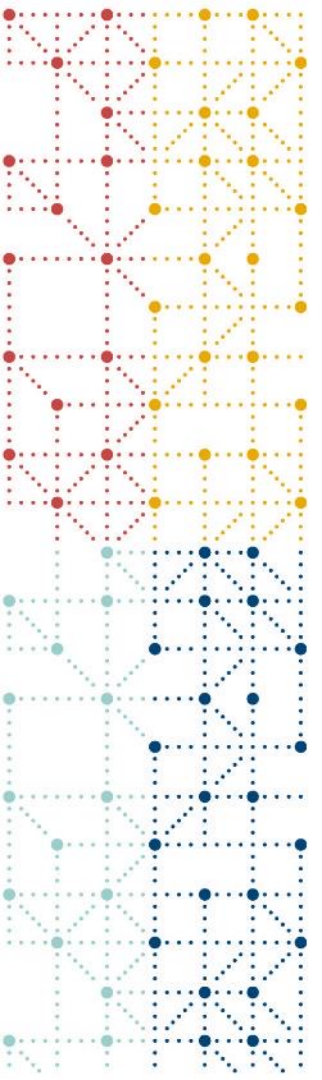
Filter results

Ordinal ↑	Name	Label	Description	Data Type	Role	Core	Code List	Described Value Domain	Implements	Value List
1	STUDYID	Study Identifier	Unique identifier for a study.	Char	Identifier	Req			STUDYID	
2	DOMAIN	Domain Abbreviation	Two-character abbreviation for the domain.	Char	Identifier	Req			DOMAIN	"VS"
3	USUBJID	Unique Subject Identifier	Identifier used to uniquely identify a subject across all studies for all applications or submissions involving the product.	Char	Identifier	Req			USUBJID	

CDISC Library API

SDTM Implementation Guide (SDTMIG) ^

GET	<code>/mdr/sdtmig/{version} /mdr/sdtmig/{version}</code>	▼ 🔒
GET	<code>/mdr/sdtmig/{version}/classes /mdr/sdtmig/{version}/classes</code>	▼ 🔒
GET	<code>/mdr/sdtmig/{version}/classes/{class} /mdr/sdtmig/{version}/classes/{class}</code>	▼ 🔒
GET	<code>/mdr/sdtmig/{version}/classes/{class}/datasets /mdr/sdtmig/{version}/classes/{class}/datasets</code>	▼ 🔒
GET	<code>/mdr/sdtmig/{version}/datasets /mdr/sdtmig/{version}/datasets</code>	▼ 🔒
GET	<code>/mdr/sdtmig/{version}/datasets/{dataset} /mdr/sdtmig/{version}/datasets/{dataset}</code>	▼ 🔒
GET	<code>/mdr/sdtmig/{version}/datasets/{dataset}/variables /mdr/sdtmig/{version}/datasets/{dataset}/variables</code>	▼ 🔒
GET	<code>/mdr/sdtmig/{version}/datasets/{dataset}/variables/{var} /mdr/sdtmig/{version}/datasets/{dataset}/variables/{var}</code>	▼ 🔒
GET	<code>/mdr/root/sdtmig/datasets/{dataset}/variables/{var} /mdr/root/sdtmig/datasets/{dataset}/variables/{var}</code>	▼ 🔒



ODM v2.0 and the Dataset-JSON Pilot

ODM v2.0 Data Exchange Standard

- Completed Public Review (final publication August 2023)
- Includes Dataset-JSON and a new version of Define-JSON is coming
- Major update to ODM v1.3.2 that breaks backwards compatibility

Study Setup	Integration	Data Exchange	End-to-end Standards
<ul style="list-style-type: none">• Study Design Model• Flexible metadata beyond CRFs• Matrix forms	<ul style="list-style-type: none">• Enhanced semantics• RWD / HL7 FHIR support• Data Queries	<ul style="list-style-type: none">• Dataset-JSON• JSON support• REST API*	<ul style="list-style-type: none">• Biomedical Concepts• Enhanced MethodDef• Traceability enhancements

What is Dataset-JSON and Advantages

What is JSON?

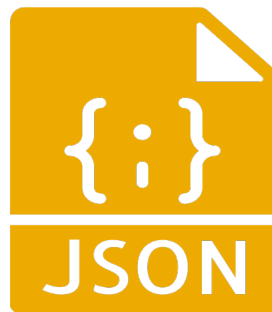
An open standard file format and data interchange format that uses human-readable text to store and transmit data objects consisting of attribute–value pairs and arrays

What is Dataset-JSON?

A dataset exchange standard for exchanging tabular data leveraging JSON designed to meet the regulatory submission needs and eliminating limitations of legacy formats

Dataset-JSON advantages...

- Based on the JSON standard used worldwide
- Open-source and truly human readable
- Same or smaller file sizes relative to current required format
- Remove variable naming, width, or format limitations
- Simple transformation to/from SAS data



Proposed Dataset-JSON Pilot

Milestone 1: Short Term

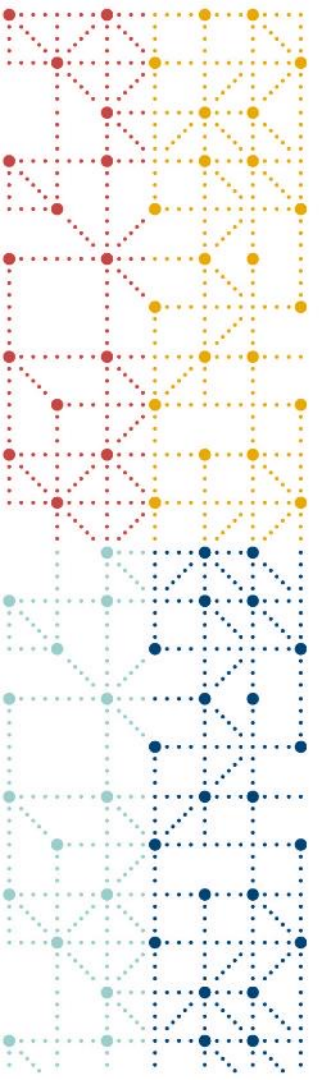
- Pilot submissions using JSON format with existing XPT ingress/egress to carry the same data
- Same content, different suitcase, no disruption to business process on either side
- In parallel, evaluate how FDA toolset can support JSON format and identify tool upgrade roadmap

➔ **Success Criteria: Accept Dataset-JSON as a transport format option (in addition to existing XPT format)**

Milestone 2: Long Term

- Enhance the CDISC SDTM and ADaM standards beyond XPT limitations (e.g., Variable names > 8, labels > 40, data > 200)
- New Define-XML / Define-JSON based on ODM v2.0
- Enhanced conformance rules
- Collaborate with FDA to develop plan to retool their environment to natively consume JSON

➔ **Success Criteria: accept advanced Dataset-JSON as the only transport format option and deprecate XPT**

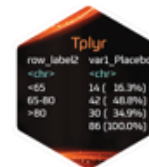


CDISC Open-Source Alliance (COSA)

CDISC Open-Source Alliance (COSA)

Community Driven Development

Supports and promotes open-source software projects that create tools for implementing or developing CDISC standards to drive innovation in the CDISC community



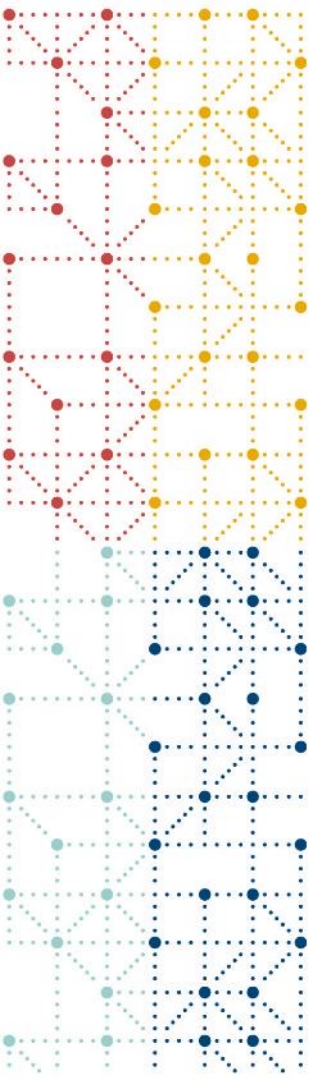
<https://cosa.cdisc.org>

COSA – OAK initiative



- OAK – SDTM Automation

- Targets the automated generation of SDTM from CDASH
 - Roche has achieved 80% automated SDTM transformations
- Creates language neutral algorithms that function as transformation rules
- Algorithms can be combined to perform all SDTM transformations
- Algorithms can be authored to transform non-CDASH conformant raw data to SDTM
- Plan to load the algorithms into the CDISC Library
 - Accessible via the Library API
 - Related to CDASH and SDTM content
- Authoring an R package to automate the transformation
- Currently > 200 volunteers have signed up to contribute
- Planning a demonstration at the US Interchange



CORE

CORE Software: Engine and Rule Editor

- Each project
 - Has a public GitHub repository on the cdisc-org account and is listed on the COSA Directory
 - Has been released under the MIT open-source license
 - Development is led by CDISC
 - Still under development, but are being actively used
 - Can be extended (supports the development of software extensions)
- CORE Engine
 - Written in Python
 - Makes use of the Venmo Business Rule Engine
- CORE Rule Editor
 - Written in TypeScript
 - Makes use of the VSCode editor



Running the CORE Engine

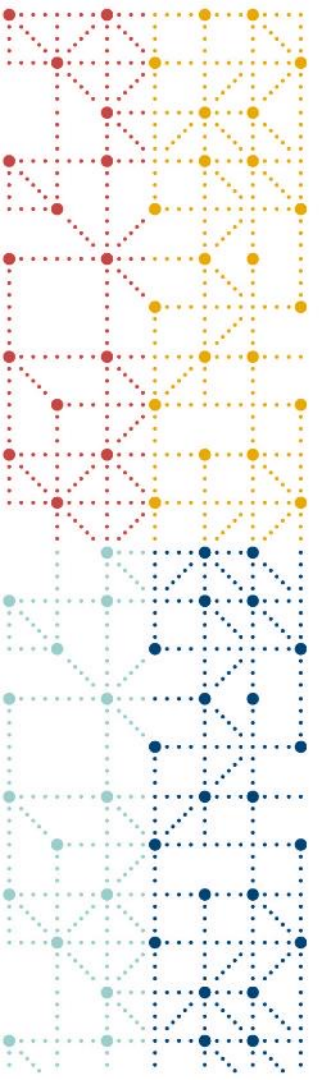
- Source Code
 - Available on GitHub using the MIT open-source license
- CLI executable available in GitHub
 - Cached rules
 - Windows, Mac, and Linux install packages
 - Unzip and run
 - Will need datasets to validate
- Engine available on PyPI
 - Engine is a component that can be used in your own code
- Desktop versions
 - Vendor released versions of CORE
 - Includes a user-friendly UI
 - Easier for non-technical users to evaluate
- View a short CORE demonstration
 - <https://www.cdisc.org/core>
 - See **CORE on GitHub** tab





CORE Engine extensibility

- Operations
 - Define an operation on a dataset, e.g., `variable_permissibility`, `mean`
- Dataset Builder
 - Used to define a dataset to match a rule type
- Dataset Reader
 - Used to define dataset formats for reading, e.g., SAS v5 XPORT, Dataset-JSON, CSV
- Data Service
 - Define the service from which the dataset will be read, e.g., local, Azure, AWS
- Checks
 - Used in rule tests, e.g., `equal_to`, `non_empty`, `matches_regex`
- Cache
 - Used to interface with a cache for rules and metadata, e.g., in memory, Redis
- Reporting
 - Defines a type of reporting, e.g., Excel, JSON
- Logging
 - Specifies what and to what level of detail logs are generated



Biomedical Concepts

CDISC Biomedical Concepts and SDTM Dataset Specializations

Pragmatic Implementation of Biomedical Concepts

3 Key pieces

- Conceptual Layer – abstract BC's
 - Provides semantics - aligned with NCI terminology
 - Supports **study design**, Schedule of Activities (SOA)
- Extend foundational standards
 - Add explicit relationships between variables
 - Additional operational metadata, e.g., data type, etc.
- Implementation Layer - Dataset Specializations with VLM definitions
 - Supports programmers
 - Pre-configured building blocks for **Define-XML**
 - Link to BCs with unambiguous semantics & definitions
 - Dataset specializations as an extended dataset structure

Common Semantics in the Data Pipeline

Representation of a BC in a specific standard with implementation details such as value level metadata, formats, terminology



Simplified Model Separates BCs and Dataset Specializations

Base VS Dataset Definition

vs.xpt, Vital Signs — Findings, Version 3.2. One record per vital sign measurement per time point per visit per subject, Tabulation

Variable Name	Variable Label	Type	Controlled Terms, Codelist or Format	Role	CDISC Notes	Core
STUDYID	Study Identifier	Char		Identifier	Unique identifier for a study.	Req
DOMAIN	Domain Abbreviation	Char	VS	Identifier	Two-character abbreviation for the domain.	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
VSSSEQ	Sequence Number	Num		Identifier	Sequence Number given to ensure uniqueness of subject records within a domain. May be any valid number.	Req
VSGRPID	Group ID	Char		Identifier	Used to tie together a block of related records in a single domain for a subject.	Perm
VSSPID	Sponsor-Defined Identifier	Char		Identifier	Sponsor-defined reference number. Perhaps pre-printed on the CRF as an explicit line identifier or defined in the sponsor's operational database.	Perm
VSTESTCD	Vital Signs Test Short Name	Char	(VSTESTCD)	Topic	Short name of the measurement, test, or examination described in VSTEST. It can be used as a column name when converting a dataset from a vertical to a horizontal format. The value in VSTESTCD cannot be longer than 8 characters, nor can it start with a number (e.g. "1TEST"). VSTESTCD cannot contain characters other than letters, numbers, or underscores. Examples: SYSBP, DIABP, BMI.	Req

Add explicit relationships between variables

Add operational metadata such as data type, length, significant digits, value

Add relationships to concept-based dataset definition specializations

VS.HEIGHT specialization

vs.xpt, Vital Signs — Findings, Version 3.2. One record per vital sign measurement per time point per visit per subject, Tabulation

Variable Name	Variable Label	Type	Controlled Terms, Codelist or Format	Role	CDISC Notes	Core
STUDYID	Study Identifier	Char		Identifier	Unique identifier for a study.	Req
DOMAIN	Domain Abbreviation	Char	VS	Identifier	Two-character abbreviation for the domain.	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
VSSSEQ	Sequence Number	Num		Identifier	Sequence Number given to ensure uniqueness of subject records within a domain. May be any valid number.	Req
VSGRPID	Group ID	Char		Identifier	Used to tie together a block of related records in a single domain for a subject.	Perm
VSSPID	Sponsor-Defined Identifier	Char		Identifier	Sponsor-defined reference number. Perhaps pre-printed on the CRF as an explicit line identifier or defined in the sponsor's operational database.	Perm
VSTESTCD	Vital Signs Test Short Name	Char	(VSTESTCD)	Topic	Short name of the measurement, test, or examination described in VSTEST. It can be used as a column name when converting a dataset from a vertical to a horizontal format. The value in VSTESTCD cannot be longer than 8 characters, nor can it start with a number (e.g. "1TEST"). VSTESTCD cannot contain characters other than letters, numbers, or underscores. Examples: SYSBP, DIABP, BMI.	Req

VS.SYSBP specialization

vs.xpt, Vital Signs — Findings, Version 3.2. One record per vital sign measurement per time point per visit per subject, Tabulation

Variable Name	Variable Label	Type	Controlled Terms, Codelist or Format	Role	CDISC Notes	Core
STUDYID	Study Identifier	Char		Identifier	Unique identifier for a study.	Req
DOMAIN	Domain Abbreviation	Char	VS	Identifier	Two-character abbreviation for the domain.	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
VSSSEQ	Sequence Number	Num		Identifier	Sequence Number given to ensure uniqueness of subject records within a domain. May be any valid number.	Req
VSGRPID	Group ID	Char		Identifier	Used to tie together a block of related records in a single domain for a subject.	Perm
VSSPID	Sponsor-Defined Identifier	Char		Identifier	Sponsor-defined reference number. Perhaps pre-printed on the CRF as an explicit line identifier or defined in the sponsor's operational database.	Perm
VSTESTCD	Vital Signs Test Short Name	Char	(VSTESTCD)	Topic	Short name of the measurement, test, or examination described in VSTEST. It can be used as a column name when converting a dataset from a vertical to a horizontal format. The value in VSTESTCD cannot be longer than 8 characters, nor can it start with a number (e.g. "1TEST"). VSTESTCD cannot contain characters other than letters, numbers, or underscores. Examples: SYSBP, DIABP, BMI.	Req

VS.HR specialization

vs.xpt, Vital Signs — Findings, Version 3.2. One record per vital sign measurement per time point per visit per subject, Tabulation

Variable Name	Variable Label	Type	Controlled Terms, Codelist or Format	Role	CDISC Notes	Core
STUDYID	Study Identifier	Char		Identifier	Unique identifier for a study.	Req
DOMAIN	Domain Abbreviation	Char	VS	Identifier	Two-character abbreviation for the domain.	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
VSSSEQ	Sequence Number	Num		Identifier	Sequence Number given to ensure uniqueness of subject records within a domain. May be any valid number.	Req
VSGRPID	Group ID	Char		Identifier	Used to tie together a block of related records in a single domain for a subject.	Perm
VSSPID	Sponsor-Defined Identifier	Char		Identifier	Sponsor-defined reference number. Perhaps pre-printed on the CRF as an explicit line identifier or defined in the sponsor's operational database.	Perm
VSTESTCD	Vital Signs Test Short Name	Char	(VSTESTCD)	Topic	Short name of the measurement, test, or examination described in VSTEST. It can be used as a column name when converting a dataset from a vertical to a horizontal format. The value in VSTESTCD cannot be longer than 8 characters, nor can it start with a number (e.g. "1TEST"). VSTESTCD cannot contain characters other than letters, numbers, or underscores. Examples: SYSBP, DIABP, BMI.	Req

For each dataset specialization update the variable definitions to match what is needed to represent the concept. A concept code and name is added to each dataset definition. A Where Clause for the specialization may be added.

Concept codes/name added to dataset metadata and used to provide the semantics for each specialization

Concept-specific codelist subsets created for use in the specializations. Maintained as part of the CT dictionary. A column value or default will be specified.



API Endpoints in CDISC Library

Biomedical Concepts (BC)

GET	/mdr/bc/packages	∨	🔒
GET	/mdr/bc/packages/{package}/biomedicalconcepts	∨	🔒
GET	/mdr/bc/packages/{package}/biomedicalconcepts /{biomedicalconcept}	∨	🔒

Study Data Tabulation Model Dataset Specializations (SDTM)

GET	/mdr/specializations/sdtm/packages	∨	🔒
GET	/mdr/specializations/sdtm/packages/{package}/datasetspecializations	∨	🔒
GET	/mdr/specializations/sdtm/packages/{package}/datasetspecializations /{datasetspecialization}	∨	🔒

Initial Use Cases

Assessments	Screening	Weeks from starting treatment pathway ³										
		-2 ¹	0 ¹	2 ¹	3 ¹	6 ¹	6 ^{1d}	9 ¹	16 ^{1*}	17 ¹		
Informed consent	X											
Blood Tests ^{2b}	X											X
ECG	X											
Medical History	X											
Physical and neurological assessment	X											
modified Toronto Clinical Neuropathy Score (mTCNS)	X											
Douleur Neuropathique 4 (DN4)	X											
Suicidal risk questionnaire	X											
Concomitant Medications	X	X	X	X	X	X	X	X	X	X	X	X
Vital Signs ¹	X											X
Pregnancy Test (for women of child bearing potential)		X ¹		X	X			X	X			
Randomisation (treatment allocation)		X ¹						X	X			
Dispense Study Medication		X	X	X	X	X	X	X	X	X		
Pain Diaries ¹		X	X	X	X	X	X	X	X	X		
Tolerability scale		X ¹				X			X			
Brief Pain Inventory-Modified Short Form (BPI-MSF)		X ¹				X			X			
Insomnia Severity Index (ISI)		X ¹				X			X			
Neuropathy Pain Symptom Inventory (NPSI)		X ¹				X			X			
Hospital Anxiety and Depression Scale (HADS)		X ¹				X			X			
RAND Short Form 36 (RAND SF-36)		X ¹				X			X			
EQ-5D-5L		X ¹				X			X			
Client Service Receipt Inventory (CSRI)		X ¹				X			X			
Pain Catastrophising Scale (PCS)		X ¹				X			X			
Adverse Events Assessment		X	X	X	X	X	X	X	X	X	X	X
Compliance Assessment		X	X	X	X	X	X	X	X	X	X	X
Patient Global Impression of Change (PGIC)												X

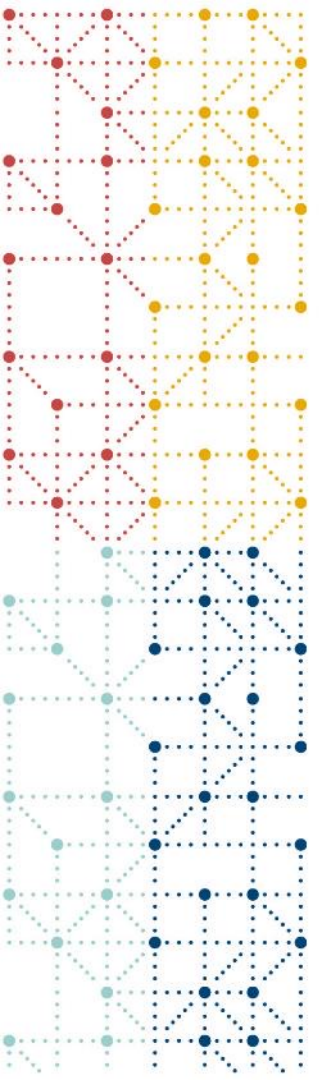
Retrieve a list of assessments for a study

VS (Vital Signs) - [SDTMIG 3.1.2]

Related Supplemental Qualifiers Dataset: [SUPPVS](#) (Supplemental Qualifiers for VS)

Variable	Where Condition	Label / Description	Type	Length or Display Format	Controlled Terms or ISO Format
VSORRES VLM		Result or Finding in Original Units	text	30	
	VSTESTCD = "DIABP" (Diastolic Blood Pressure)	Diastolic Blood Pressure in Orig U	integer	2	
	VSTESTCD = "FRMSIZE" (Body Frame Size)	Body Frame Size - Orig	text	6	Size • "SMALL" • "MEDIUM" • "LARGE"
	VSTESTCD = "HEIGHT" (Height)	Height in Orig U	float	5.1	

Publish BC content as Define-XML document including value level metadata



CDISC Data Exchange Framework

CDISC's Data Exchange Framework



Logical Data Model

The UML class diagram (normative) as well as SQL Data Dictionary, Entity Relationship Diagram and example JSON output (informative)



Application Programming Interface (API) Specification

The API definition (normative) in JSON and HTML forms



CDISC Controlled Terminology

The controlled terminology (normative) developed for the project and published quarterly in the CDISC Library.



JSON

The API returns an JSON payload by default. Examples provided as JSON files. The API may also support XML and other media types.

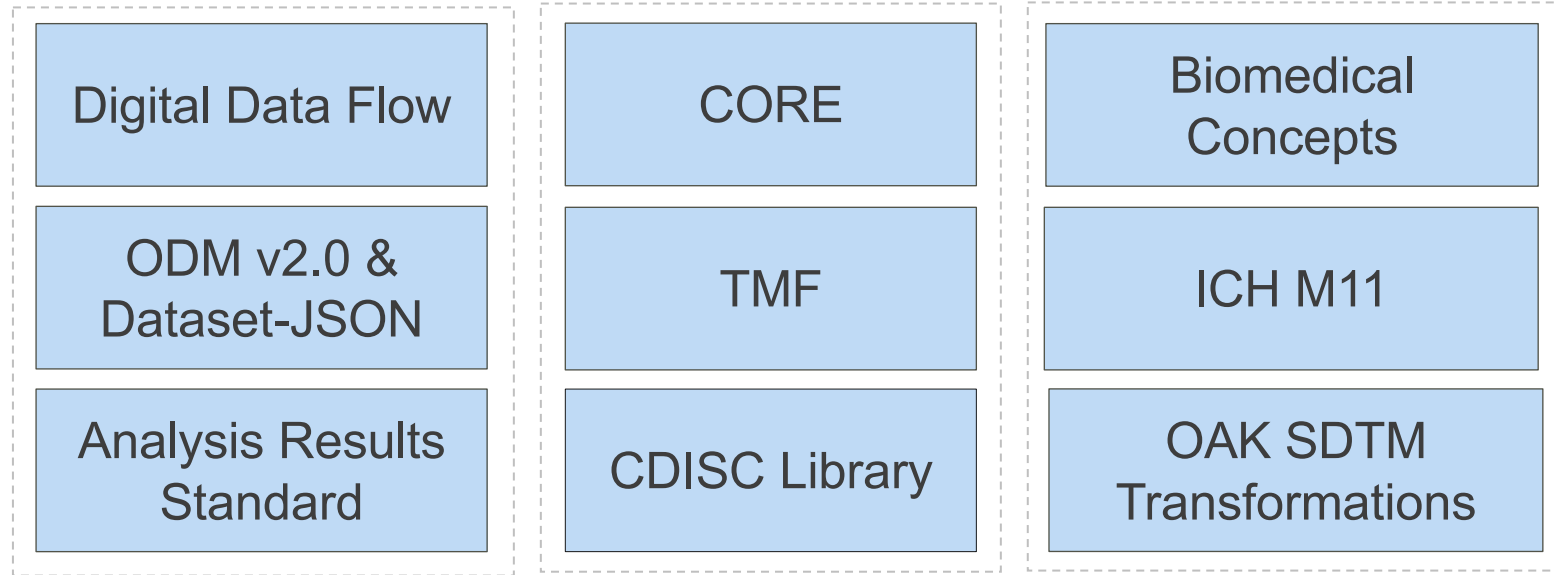


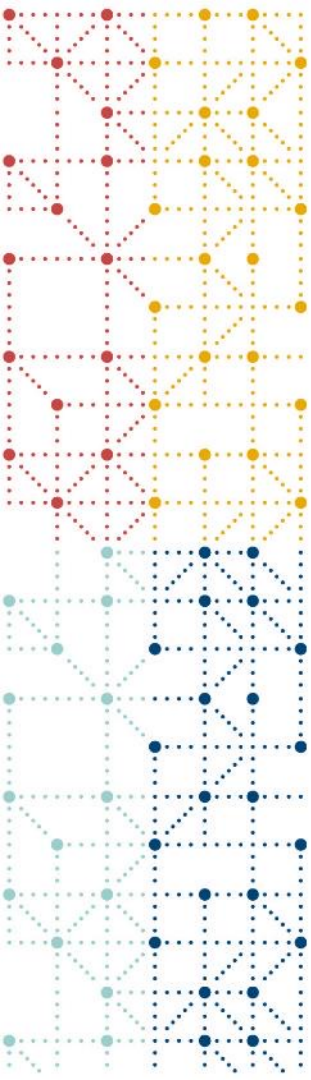
Biomedical Concepts

Semantics that work across standards, including RWD, coupled with dataset specializations that provide pre-configured standards.

Framework: Model + API + CT + JSON + BC

CDISC's Data Exchange Framework Today





Thank You!

Sam Hume, DSc

shume@cdisc.org

<https://www.linkedin.com/in/sam-hume-dsc>

