**Training materials**

# Introduction

This training provides principal investigators and study team members an overview of existing standards for sharing individual participant data (IPD). If they are not familiar with those standards, it allows them to assess whether the standards are relevant for their study.

# CDISC standards

Clinical Data Interchange Standards Consortium (CDISC) is a Standard Development Organization (SDO) that develops standards for clinical research. CDISC standards are used by FDA in order to process large volume of clinical trials data without dealing with a new and study-specific format each time a new trial has to be reviewed in detail.

# CDISC Study Data Tabulation Model (SDTM) standard

### Description

SDTM divides study data into several domains. Each domain is essentially a spreadsheet with standardized set of columns. For example, demographic domain (DM) is a table with a single row for each study subject.

### Example

Clinical study NIDA-CTN-19 ([NCT00084188](https://clinicaltrials.gov/ct2/show/NCT00084188); HIV/STD Safer Sex Skills Groups for Women in Drug Treatment Programs - 1) provides data in SDTM format at [NIDA Data Share](https://datashare.nida.nih.gov) platform. Table below shows an example of Trial Arm (TA) domain. This domain describes the study. We see that the study has two arms titled ‘Safer sexual skills building’ and ‘Single session’. Within those arms, there are several elements defined, such as screening, treatment or 90 days follow up. Columns are typically labeled with 8-character column title (shown in parentheses; e.g., ETC). To increase readability, we expanded the 8-character abbreviations to expanded title (shown prior the abbreviation; e..g., Element Code).

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Study ID  (STUDYID)** | **Domain (DOMAIN)** | **Arm Code (ARMCD)** | **Arm (ARM)** | **Element Order (TAETORD)** | **Element Code  (ETCD)** | **Element (ELEMENT)** | **Epoch (EPOCH)** |
| NIDA-CTN-0019 | TA | SSB | SAFER SEXUAL SKILLS BUILDING | 1 | SCREENING | SCREENING | SCREENING |
| NIDA-CTN-0019 | TA | SSB | SAFER SEXUAL SKILLS BUILDING | 2 | BASELINE | BASELINE | BASELINE |
| NIDA-CTN-0019 | TA | SSB | SAFER SEXUAL SKILLS BUILDING | 3 | TREATMENT | FIVE SESSIONS | TREATMENT |
| NIDA-CTN-0019 | TA | SSB | SAFER SEXUAL SKILLS BUILDING | 4 | POST | POST INTERVENTION | POST INTERVENTION |
| NIDA-CTN-0019 | TA | SSB | SAFER SEXUAL SKILLS BUILDING | 5 | FU1 | FOLLOWUP 90 DAY | FOLLOW-UP 90 DAYS |
| NIDA-CTN-0019 | TA | SSB | SAFER SEXUAL SKILLS BUILDING | 6 | FU2 | FOLLOWUP 180 DAY | FOLLOW-UP 180 DAYS |
| NIDA-CTN-0019 | TA | CONTROL | SINGLE SESSION | 1 | SCREENING | SCREENING | SCREENING |
| NIDA-CTN-0019 | TA | CONTROL | SINGLE SESSION | 2 | BASELINE | BASELINE | BASELINE |
| NIDA-CTN-0019 | TA | CONTROL | SINGLE SESSION | 3 | TREATMENT | SINGLE SESSION | TREATMENT |
| NIDA-CTN-0019 | TA | CONTROL | SINGLE SESSION | 4 | POST | POST INTERVENTION | POST INTERVENTION |
| NIDA-CTN-0019 | TA | CONTROL | SINGLE SESSION | 5 | FU1 | FOLLOWUP 90 DAY | FOLLOW-UP 90 DAYS |
| NIDA-CTN-0019 | TA | CONTROL | SINGLE SESSION | 6 | FU2 | FOLLOWUP 180 DAY | FOLLOW-UP 180 DAYS |

Another example from the same study is communicating laboratory measurements. SDTM groups all laboratory results into a single spreadsheet (LB.csv [=laboratory SDTM domain]). It consist of 5,196 rows. Unlike Trial Arm domain, laboratory data are on patient level. Selected columns that each row contains are:

* subject identifier (USUBJID)
* category of lab test (LBCAT); For example <DRUG SCREEN>
* laboratory test code (LBTESTCD) and name (LBTEST); For example <blood drug screen for amphetamines>
* result of the test (LBORRES) and units (LBORRESU)
* when the test was done (relative to study start date for the participant) (LBDY); For example, -6 (6 days prior index date) and at what visit (VISITNUM)

Table below has only selected columns from the data.

|  |  |  |  |
| --- | --- | --- | --- |
| **LBTESTCD** | **LBTEST** | **LBCAT** | **LBORRES** |
| AMPHET | AMPHETAMINES | DRUG SCREEN | NEGATIVE |
| BARBIT | BARBITURATES | DRUG SCREEN | NEGATIVE |
| BENZOD | BENZODIAZEPINES | DRUG SCREEN | NEGATIVE |
| COCAINE | COCAINE | DRUG SCREEN | NEGATIVE |
| METH | METHAMPHETAMINES | DRUG SCREEN | NEGATIVE |
| METHADON | METHADONE | DRUG SCREEN | NEGATIVE |
| MORPHINE | OPIATES/MORPHINE | DRUG SCREEN | NEGATIVE |
| PCP | PCP | DRUG SCREEN | NEGATIVE |
| TCA | TRICYCLIC ANTIDEPRESSANTS | DRUG SCREEN | NEGATIVE |
| THC | CANNABINOIDS (THC) | DRUG SCREEN | NEGATIVE |
| AMPHET | AMPHETAMINES | DRUG SCREEN | NEGATIVE |
| BARBIT | BARBITURATES | DRUG SCREEN | NEGATIVE |
| BENZOD | BENZODIAZEPINES | DRUG SCREEN | UNCLEAR OR EQUIVOCAL |
| COCAINE | COCAINE | DRUG SCREEN | POSITIVE |
| METH | METHAMPHETAMINES | DRUG SCREEN | NEGATIVE |
| METHADON | METHADONE | DRUG SCREEN | NEGATIVE |
| MORPHINE | OPIATES/MORPHINE | DRUG SCREEN | NEGATIVE |

Additional examples (not from an actual past study but using demo data) for CDISC SDTM format are available at <https://github.com/lhncbc/r-snippets-bmi/tree/master/cdisc/inst/extdata/cdisc01/csv>

# CDISC Define-XML data dictionary standard

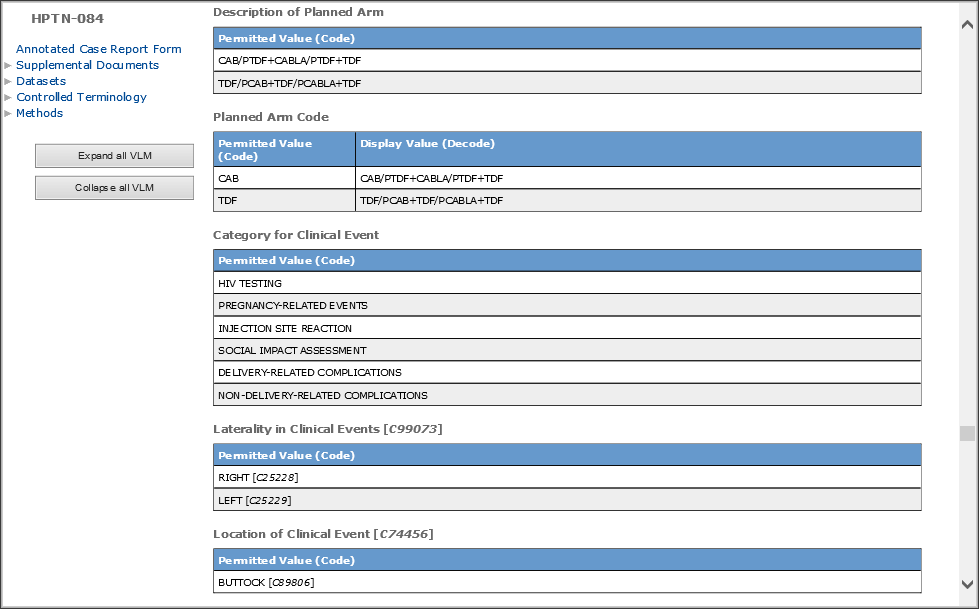
### Description

Define-XML is a data dictionary standard. It provides a way to describe what data elements are present in an SDTM dataset. Typically, a single Define-XML file is provided for a study. The define-XML data dictionary file is provided in conjunction with patient level data.

### Example

Study HPTN084 (NCT03164564, Evaluating the Safety and Efficacy of Long-Acting Injectable Cabotegravir Compared to Daily Oral TDF/FTC for Pre-Exposure Prophylaxis in HIV-Uninfected Women) provides data (after request) using CDISC (including data dictionary in Define-XML format).

Define-XML is an XML format than can be rendered in human readable form using a stylesheet file. Figure below shows the Codelist section of the Define-XML file for this trial. Is provides values (using CDISC Controlled Terminology codes (prefixed with letter C; for example C25229 for left laterality)) for



# REDCap Data Dictionary standard

REDCap is an electronic data capture system for clinical research. Because it is widely used, the data dictionary format it uses for export and import of data elements is highly influential.

It is not currently being used by data sharing platforms. However, it is included in this training because it provides useful guidance on how data a dictionary may be communicated.

## Example

Hepatitis C target registry (NCT01474811) provides data dictionary in redcap format (available at <https://github.com/hcv-target/data_dictionary/blob/master/HCVTARGET2030_DataDictionary_2016-06-06.csv>).

Each data element is described in many metadata fields. For example: short name (Field Name), description (Field Label), or data type (columns Field and Text Validation Type). The column Choices, Calculations, OR Slider Labels provides list of permissible values (separated by pipe). For each permissible value, a code and display text is defined (separated by comma;e.g., NOT\_AVAILABLE, No subtype available for data element <HCV secondary subtype>).

Table below shows example data elements in REDCap format.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Variable / Field Name** | **Form Name** | **Section Header** | **Field Type** | **Field Label** | **Choices, Calculations, OR Slider Labels** | **Field Note** | **Text Validation Type OR Show Slider Number** | **Text Validation Min** | **Text Validation Max** | **Identifier?** | **Branching Logic (Show field only if...)** | **Required Field?** |
| il28bgt\_lbtestcd | il28b\_hcv\_genotypes |  | descriptive | IL28B |  |  |  |  |  |  | 0 = 1 |  |
| il28bgt\_lbtest | il28b\_hcv\_genotypes |  | descriptive | IL28B genotype |  |  |  |  |  |  | 0 = 1 |  |
| il28bgt\_lboccur | il28b\_hcv\_genotypes |  | radio | IL28B genotype available for this subject? | N, NO | Y, YES |  |  |  |  |  |  | y |
| il28bgt\_lborres | il28b\_hcv\_genotypes |  | radio | IL28B genotype | CC, CC | CT, CT | TT, TT |  |  |  |  |  | [il28bgt\_lboccur] = 'Y' | y |
| hcvgt\_lbtest | il28b\_hcv\_genotypes |  | descriptive | HCV Genotype |  |  |  |  |  |  | 0 = 1 |  |
| hcvgt\_lbtestcd | il28b\_hcv\_genotypes |  | descriptive | HCVGT |  |  |  |  |  |  | 0 = 1 |  |
| hcvgt\_lboccur | il28b\_hcv\_genotypes |  | radio | HCV genotype available for this subject? | N, NO | Y, YES |  |  |  |  |  |  | y |
| hcvgt\_lbdtc | il28b\_hcv\_genotypes |  | text | Year of HCV genotype |  | YYYY | year\_only |  |  |  | [hcvgt\_lboccur]= 'Y' AND [hcvgt\_supplb\_nolbdtc] != 'Y' | y |
| hcvgt\_supplb\_nolbdtc | il28b\_hcv\_genotypes |  | radio | Is year of HCV genotype unknown? | N, NO | Y, YES |  |  |  |  |  | [hcvgt\_lboccur] = 'Y' AND [hcvgt\_lbdtc] = '' | y |
| hcvgt\_lborres | il28b\_hcv\_genotypes |  | dropdown | HCV genotype | 1, 1 | 2, 2 | 3, 3 | 4, 4 | 5, 5 | 6, 6 |  |  |  |  |  | [hcvgt\_lboccur]= 'Y' | y |
| hcvgt\_s\_lborres | il28b\_hcv\_genotypes |  | dropdown | HCV subtype | a, a | b, b | c, c | d, d | e, e | NOT\_AVAILABLE, No subtype available |  |  |  |  |  | [hcvgt\_lboccur]= 'Y' | y |
| hcvgt2\_lbtestcd | il28b\_hcv\_genotypes |  | descriptive | HCVGT2 |  |  |  |  |  |  | 0 = 1 |  |
| hcvgt2\_lbtest | il28b\_hcv\_genotypes |  | descriptive | HCV Secondary Genotype |  |  |  |  |  |  | 0 = 1 |  |
| hcvgt2\_lboccur | il28b\_hcv\_genotypes |  | radio | Does the subject have a mixed HCV genotype? | N, NO | Y, YES |  |  |  |  |  | [hcvgt\_lboccur]= 'Y' | y |
| hcvgt2\_lborres | il28b\_hcv\_genotypes |  | dropdown | HCV secondary genotype | 1, 1 | 2, 2 | 3, 3 | 4, 4 | 5, 5 | 6, 6 |  |  |  |  |  | [hcvgt2\_lboccur] = 'Y' | y |
| hcvgt2\_s\_lborres | il28b\_hcv\_genotypes |  | dropdown | HCV secondary subtype | a, a | b, b | c, c | d, d | e, e | NOT\_AVAILABLE, No subtype available |  |  |  |  |  | [hcvgt2\_lboccur] = 'Y' | y |
| q80k\_lbtestcd | il28b\_hcv\_genotypes |  | descriptive | Q80K |  |  |  |  |  |  | 0 = 1 |  |
| q80k\_lbtest | il28b\_hcv\_genotypes |  | descriptive | Q80K mutation present? |  |  |  |  |  |  | 0 = 1 |  |
| q80k\_lboccur | il28b\_hcv\_genotypes |  | radio | Q80K mutation available for this subject? | N, NO | Y, YES |  |  |  |  |  |  | y |
| q80k\_lborres | il28b\_hcv\_genotypes |  | radio | Q80K mutation present? | N, NO | Y, YES |  |  |  |  |  | [q80k\_lboccur] = 'Y' | y |
| ns5a\_lbtestcd | il28b\_hcv\_genotypes |  | descriptive | NS5ABLPM |  |  |  |  |  |  | 0 = 1 |  |
| ns5a\_lbtest | il28b\_hcv\_genotypes |  | descriptive | NS5A baseline polymorphism(s) present? |  |  |  |  |  |  | 0 = 1 |  |
| ns5a\_lboccur | il28b\_hcv\_genotypes |  | radio | NS5A baseline polymorphism(s) available for this subject? | N, NO | Y, YES |  |  |  |  |  | [hcvgt\_lborres] = '1' AND ([hcvgt\_s\_lborres] = 'a' OR [hcvgt\_s\_lborres] = 'b') | y |
| ns5a\_lborres | il28b\_hcv\_genotypes |  | radio | NS5A baseline polymorphism(s) present? | N, NO | Y, YES | UNKNOWN, Unknown |  |  |  |  |  | [ns5a\_lboccur] = 'Y' | y |

# Acknowledgement

This work was supported by the Intramural Research Program of the National Institutes of Health (NIH)/ National Library of Medicine (NLM)/ Lister Hill National Center for Biomedical Communications (LHNCBC) and by NIH Office of AIDS research.

The findings and conclusions in this article are those of the authors and do not necessarily represent the official position of NLM, NIH, or the Department of Health and Human Services.

# Authors

Vojtech Huser  
Craig Mayer