PharmaSUG 2019 - Paper AD-109

Tool Development Methods for Implementing and Converting to New Controlled Terminology in SDTM datasets

Martha O’Brien and Keith Shusterman, Reata Pharmaceutics, Inc.

ABSTRACT

Controlled terminology (CT) for SDTM datasets allows for easier review for committee members, other programmers, consultants, consulting companies, the FDA, and many others. This may ultimately reduce the time it takes to get the drug or device to market. Ensuring a proper method of choosing and implementing a newer version of CT is not only necessary but vital to submission acceptance.

Currently the FDA only requires CT versions of 2011-06-10 or later and with quarterly outputs there are many options to choose. This can make it difficult when starting a study or after a study starts a sponsor may decide that up versioning the CT is necessary. Sponsors will also need to harmonize the new CT with their own specific values that have been added to the extensible codelists prior to implementing new versions.

Having a partially automated process to convert to a newer CT eases not only the time constraint but also reduces the possibility of human error. This paper will provide a process for creating tools when up versioning CT during an ongoing study.

INTRODUCTION

It can seem like a daunting task to revisit all source-to-SDTM mappings when a new version of CDISC CT is released and needs to be applied. Manually reviewing every addition, deletion, and alteration to official codelist values, and then applying those updates to all SDTM variables that use CDISC CT, would be a very time consuming and inefficient task.

While some level of manual review will always be necessary, this paper offers an approach that reads in the entire CDISC SDTM codelist and outputs a lookup table (LUT) with columns for synonyms. By default, the synonym columns are populated from the official CDISC terminology. However, by outputting this LUT as an Excel spreadsheet, additional synonyms can be manually added to facilitate all necessary CDISC CT mapping. Additionally, rows can be manually added to the LUT to account for sponsor-specific extensions of extensible codelists.

This paper also presents a macro that reads in source datasets and the LUT. The output is a table that can be easily joined to the source dataset to give CDISC and sponsor-specific CT values. With this join, it is a simple matter to identify which raw values do not have an associated CDISC or sponsor-specific value in the LUT. From there, synonyms can be added to the CDISC codelist values, or new sponsor-specific values can be added as rows

For the case of already completed SDTM studies that need to be harmonized for an integrated analysis, we also present a modified macro that can read in SDTM (including associated SUPPQUAL) datasets instead of source datasets. In the event that source values are mapped to the SUPPQUAL domain, these source values can be used to quickly identify synonyms in the new CT version through the creation of the LUT.

With these tools in place, up-versioning CT becomes a (comparatively) simple matter of creating this LUT based on the new CT, running a macro to identify terms that don’t match the new CT, and updating the LUT accordingly.

CREATING AND USING LUT FROM CDISC CT WITH SOURCE DATA

MACRO TO CREATE LUT FROM CDISC CT

The following code creates a useful LUT based on the target CDISC CT list that contains the CT submission value and all “official” synonyms as individual columns. This table can be used to easily determine when the input data exactly matches a CT value or synonym.

```
/*The table “ct” here refers to the CDISC terminology Excel spreadsheet
converted to a SAS dataset using PROC IMPORT*/

data ctterm;
       set ct (where=(codelist_code ne ''));
run;

/*derive number of columns needed for lookup table based on number of synonyms
available*/
proc sql noprint;
       select max(count(CDISC_Synonym_s_,';')) + 1 into :cols
       from ctterm
;
quit;

/*split up synonyms into individual columns*/
data syns;
       set ctterm;
       %do i = 1 %to &cols.;
    syn&i = strip(scan(CDISC_Synonym_s_,&i.,';'));
       %end;
       keep codelist_code codelist_name cdisc_submission_value
nci_preferred_term cdisc_synonym_s_ syn:;
run;

/*output synonyms as an excel spreadsheet*/
proc export data = syns dbms = xlsx
       outfile = "<LUT folder path>"
       replace;
run;

%mend;
```

A small portion of the resulting LUT is shown below in Table 1.

Codelist_Code	Codelist_Name	CDISC_Submission_Value	CDISC_Synonym_s_
C66767	Action Taken with Study Treatment	DOSE INCREASED	
C66767	Action Taken with Study Treatment	DOSE NOT CHANGED	
C66767	Action Taken with Study Treatment	DOSE RATE REDUCED	
C66767	Action Taken with Study Treatment	DOSE REDUCED	
C66767	Action Taken with Study Treatment	DRUG WITHDRAWN	
C66767	Action Taken with Study Treatment	NOT APPLICABLE	NA; Not Applicable
C66767	Action Taken with Study Treatment	UNKNOWN	U; UNK; Unknown

NCI_Preferred_Term	syn1	syn2	syn3
Dose Increased			
Dose Not Changed			
Dose Rate Reduced			
Dose Reduced			
Drug Withdrawn			
Not Applicable	NA	Not Applicable	
Unknown	U	UNK	Unknown

Table 1: LUT of CDISC terms and synonyms for the ACN codelist

This macro parses out the pre-specified synonyms given by CDISC in the column CDISC_Synonym_s_. Note that the number of synonym columns the macro creates is equal to the largest number of synonyms given in the CDISC_Synonym_s_ column across all codelists. The LUT is output as an Excel spreadsheet, allowing for easy sponsor updates where needed.

SOURCE DATA INPUT CODELIST

AEACN is an expected variable that uses the non-extensible codelist “ACN". When the associated CT list is updated, it is important to be aware of the changes and to appropriately incorporate them when up-versioning your datasets. Consider the source adverse event data shown in Table 2 below.

STUDY	SUBJECT	VERBATIM	ACTION
XYZ	001	Ate too much cheese	Dose not changed
XYZ	001	Something bad	Dose reduced
XYZ	002	Something really bad	Dose rate reduced
XYZ	032	Headache	Dose unchanged
XYZ	097	Tooth ache	Not applicable

Table 2: Example source AE data.

Here, we have a source variable called ACTION that will be mapped to CDISC CT and then migrated into the SDTM variable AEACN in the AE domain. Notice how one of the values is “Dose rate reduced”. The term “DOSE RATE REDUCED” was only recently added into the ACN codelist in the 2018-06-29 version of CDISC CT. Since ACN is a non-extensible codelist, this may have previously been mapped to the next closest CDISC value of “DOSE REDUCED”. When up-versioning, this raw value should now be mapped to the new exactly-matching CDISC term of “DOSE RATE REDUCED”.

MACRO TO APPLY LUT TO SOURCE DATA

The following macro, utilizing the LUT and source data as inputs, can be used to generate a list of source values and associated CT content. Macro variables are used to identify the target SDTM variable, codelist reference, source dataset name, and source variable name. In our internal process, the CDISC codelist associated with the SDTM variable is read in from the specifications. For the purposes of this demonstration, the CDISC codelist code is specified as an input macro variable.

```
%macro ct_update
            (var,   /*SDTM variable to which CT is being applied. e.g. AEACN*/
            code,   /*CDISC code for the applicable codelist. e.g. C66767 for
ACN codelist*/
            in,    /*Name of input raw dataset.*/
            rawvar /*Variable in the raw dataset that is being mapped to ct8*/
);
```

```
/*import CT with synonyms*/
PROC IMPORT DBMS=xlsx
            DATAFILE= "<LUT folder path>"
            OUT= WORK.syns
                    REPLACE;
RUN;
/*isolate codelist of interest*/
data syns2;
       set syns (where=(codelist_code = "&code."));
run;

/*derive number of columns needed for lookup table based on number of synonyms
available*/
proc sql noprint;
       select max(count(CDISC_Synonym_s_,';')) + 1 into :cols
       from syns2
;
quit;


/*get distinct values*/
proc sql;
create table dvals as
       select distinct &rawvar
              from &in
;
quit;

/*join to list of controlled terms and synonyms*/
proc sql;
create table &var.lut_full as
select *
       from dvals a
       left join syns2 b
       on

              %do i = 1 %to &cols.;
              strip(upcase(a.&rawvar.)) = strip(upcase(b.syn&i.)) or
              %end;

              strip(upcase(a.&rawvar.)) =
strip(upcase(b.cdisc_submission_value)) or
              strip(upcase(a.&rawvar.)) = strip(upcase(b.nci_preferred_term))
;
quit;

/*drop synonym variables to only contain the raw and CT values*/
data &var.lut;
       attrib &rawvar length=$200;
       attrib VARIABLE length=$200;
       attrib CDISC_SUBMISSION_VALUE LENGTH=$200;
       set &var.lut_full;
       variable = "&var.";
       keep &rawvar variable cdisc_submission_value;
run;

%mend;
```

Using this code, the output in Table 3 was produced.

ACTION	VARIABLE	CDISC_Submission_Value
Dose not changed	AEACN	DOSE NOT CHANGED
Dose rate reduced	AEACN	DOSE RATE REDUCED
Dose reduced	AEACN	DOSE REDUCED
Not applicable	AEACN	NOT APPLICABLE
Dose unchanged	AEACN	

Table 3: Up-version support LUT targeting AEACN.

Notice for record 3 the CDISC submission value correctly pulls in 'DOSE RATE REDUCED', which is exactly what we wanted based on the updated CT. Currently one record does not match directly with a CDISC submission value or any of the synonyms which is why CDISC_Submission_Value is null. However, the LUT can be manually updated to contain any desired synonyms. In this example, an SME can manually add "Dose unchanged" as a synonym to "DOSE NOT CHANGED" as shown below in Table 4.

Codelist_Code	Codelist_Name	CDISC_Submission_Value	CDISC_Synonym_s_
C66767	Action Taken with Study Treatment	DOSE INCREASED	
C66767	Action Taken with Study Treatment	DOSE NOT CHANGED	
C66767	Action Taken with Study Treatment	DOSE RATE REDUCED	
C66767	Action Taken with Study Treatment	DOSE REDUCED	
C66767	Action Taken with Study Treatment	DRUG INTERRUPTED	
C66767	Action Taken with Study Treatment	DRUG WITHDRAWN	
C66767	Action Taken with Study Treatment	NOT APPLICABLE	NA; Not Applicable
C66767	Action Taken with Study Treatment	UNKNOWN	U; UNK; Unknown

NCI_Preferred_Term	syn1	syn2	syn3
Dose Increased			
Dose Not Changed	Dose unchanged		
Dose Rate Reduced			
Dose Reduced			
Drug Interrupted			
Drug Withdrawn			
Not Applicable	NA	Not Applicable	
Unknown	U	UNK	Unknown

Table 4: LUT of CDISC terms and synonyms for the ACN codelist with a custom synonym added (highlighted)

Once the Excel file is updated with the new synonym, running the macro again will pull in the updated LUT and produce the output shown below in Table 5.

ACTION	VARIABLE	CDISC_Submission_Value
Dose not changed	AEACN	DOSE NOT CHANGED
Dose unchanged	AEACN	DOSE NOT CHANGED
Dose rate reduced	AEACN	DOSE RATE REDUCED
Dose reduced	AEACN	DOSE REDUCED
Not applicable	AEACN	NOT APPLICABLE

Table 5: Up-version support LUT targeting AEACN with custom synonym applied (highlighted).

APPLYING APPROACH TO SDTM DATA

SDTM DATA INPUT CODELIST

A similar technique can be used for the purpose of harmonizing terminology in already completed SDTM domains that are to be pooled together for an integrated analysis. This is also a use case for storing source CRF values in the SUPPQUAL domains. If source values are available in the SDTM data, then terminology can be harmonized to a more recent CT version without any possible need for the source datasets. Suppose the source AE data shown in Table 1 was mapped to an SDTM AE domain shown below in Table 6, with raw values captured in SUPPAE shown below in Table 7.

STUDYID	DOMAIN	USUBJID	AESEQ	AETERM	AEACN
STUDY_XYZ	AE	STUDY_XYZ-001	1	ATE TOO MUCH CHEESE	DOSE NOT CHANGED
STUDY_XYZ	AE	STUDY_XYZ-001	2	SOMETHING BAD	DOSE REDUCED
STUDY_XYZ	AE	STUDY_XYZ-002	1	SOMETHING REALLY BAD	DOSE REDUCED
STUDY_XYZ	AE	STUDY_XYZ-032	27	HEADACHE	DOSE UNCHANGED
STUDY_XYZ	AE	STUDY_XYZ-097	4	TOOTH ACHE	NOT APPLICABLE

Table 6: AE domain mapped from the source AE data in Table 1.

STUDYID	RDOMAIN	USUBJID	IDVARVAL	QNAM
STUDY_XYZ	AE	STUDY_XYZ-001	1	CRFACN
STUDY_XYZ	AE	STUDY_XYZ-001	2	CRFACN
STUDY_XYZ	AE	STUDY_XYZ-002	1	CRFACN
STUDY_XYZ	AE	STUDY_XYZ-032	27	CRFACN
STUDY_XYZ	AE	STUDY_XYZ-097	4	CRFACN

QLABEL	QVAL
CRF Collected Action Taken	DOSE NOT CHANGED
CRF Collected Action Taken	DOSE REDUCED
CRF Collected Action Taken	DOSE RATE REDUCED
CRF Collected Action Taken	DOSE UNCHANGED
CRF Collected Action Taken	NA

Table 7: SUPPAE domain mapped from the source AE data in Table 1

Here, we see that the source value of “Dose rate reduced” was standardized to the CDISC value of “DOSE REDUCED” before the value “DOSE RATE REDUCED” was added to the CDISC ACN codelist. Since the source CRF value is preserved in SUPPAE, we can then use a similar programming technique to find CDISC values in the newer CT. The only difference is that now SUPPAE.QVAL where QNAM = ‘CRFACN’ is used as the input variable to be standardized instead of the variable from the original source dataset.

MACRO TO APPLY LUT TO SDTM DATA

The modified macro to perform this check is below.

```
%macro ct_update
              (var,   /*SDTM variable to which CT is being applied. e.g. AEACN*/
              code,   /*CDISC code for the applicable codelist. e.g. C66767 for
ACN codelist*/
              in,    /*name of input SUPPQUAL dataset. e.g. SUPPAE*/
);


/*import CT with synonyms*/
PROC IMPORT DBMS=xlsx
            DATAFILE= "<LUT folder path>"
            OUT= WORK.syns
                    REPLACE;
RUN;


/*isolate codelist of interest*/
data syns2;
       set syns (where=(codelist_code = "&code."));
run;

/*derive number of columns needed for lookup table based on number of synonyms
available*/
proc sql noprint;
       select max(count(CDISC_Synonym_s_,';')) + 1 into :cols
       from syns2
;
quit;

/*get distinct values*/
proc sql;
create table dvals as
       select distinct qval
              from &in
              where upcase(qnam) = upcase("&qnam.")
;
quit;
```

```
/*join to list of controlled terms and synonyms*/
proc sql;
create table &var.lut_full as
select *
       from dvals a
       left join syns2 b
       on

              %do i = 1 %to &cols.;
              strip(upcase(a.qval)) = strip(upcase(b.syn&i.)) or
              %end;

              strip(upcase(a.qval)) = strip(upcase(b.cdisc_submission_value)) or
              strip(upcase(a.qval)) = strip(upcase(b.nci_preferred_term))
;
quit;

/*drop synonym variables to only contain the raw and CT values*/
data &var.lut_partial;
       set &var.lut_full;
       keep qval cdisc_submission_value;
run;


/*join to SUPP domain*/
proc sql;
create table &in._syn as
select a.*, b.cdisc_submission_value
       from
       (
              select *
              from suppae
              where upcase(strip(qnam)) = upcase(strip("&qnam."))
       ) a
              left join &var.lut_partial b
                     on upcase(strip(a.qval)) = upcase(strip(b.qval))
;
quit;

/*define parent domain name*/
%let in2 = %substr(&in.,5,2);

/*join to parent domain*/
proc sql;
create table &var.lut as
       select a.&var., b.qval, b.cdisc_submission_value
              from &in2 a
                     left join &in._syn b
                            on a.usubjid = b.usubjid
                                   and a.&in2.seq = b.idvarval
;
quit;

%mend;
```

The output of this macro is a similar LUT that displays the value currently stored in AEACN, the CRF value retained in QVAL, and the associated CDISC synonym under the new CT. This is shown below in Table 8.

AEACN	QVAL	CDISC_Submission_Value
DOSE NOT CHANGED	DOSE NOT CHANGED	DOSE NOT CHANGED
DOSE REDUCED	DOSE REDUCED	DOSE REDUCED
DOSE REDUCED	DOSE RATE REDUCED	DOSE RATE REDUCED
DOSE UNCHANGED	DOSE UNCHANGED	
NOT APPLICABLE	NA	NOT APPLICABLE

Table 8: Up-version support LUT run using CRF values stored in SUPPAE as input

Whether up-versioning source or SDTM data, the same Excel synonym spreadsheet is used. As with the source input example, the value "DOSE UNCHANGED" can be manually added to the spreadsheet as a synonym of "DOSE NOT CHANGED". Once the synonyms spreadsheet is updated, running the macro gives a CDISC submission value for QVAL = "DOSE UNCHANGED" as well, as shown below in Table 9.

AEACN	QVAL	CDISC_Submission_Value
DOSE NOT CHANGED	DOSE NOT CHANGED	DOSE NOT CHANGED
DOSE REDUCED	DOSE REDUCED	DOSE REDUCED
DOSE REDUCED	DOSE RATE REDUCED	DOSE RATE REDUCED
DOSE UNCHANGED	DOSE UNCHANGED	DOSE NOT CHANGED
NOT APPLICABLE	NA	NOT APPLICABLE

Table 9: Up-version support LUT created using CRF values stored in SUPPAE as input with synonym added (Highlighted)

EXTENDING CODELISTS FOR SPONSOR-SPECIFIC TERMINOLOGY

The examples so far have been referencing a non-extensible codelist. In the case of an extensible codelist, it is possible to add sponsor-specific terms by adding rows to the LUT. This is very common need when mapping laboratory data. While the CDISC terminology for lab tests is extensive, it is not uncommon for studies to use a more obscure lab test that's not covered in the CDISC codelist.

In Table 6, we see an adverse event with AETERM = "ATE TOO MUCH CHEESE". Suppose the study has an endpoint of cheese toxicity, and the sponsor collects a corresponding lab test for serum cheese levels. This test does not exist in the CDISC codelist for LBTESTCD/LBTEST, so these codelists need to be extended to include the sponsor term of LBTESTCD = "CHEESELE" / LBTEST = "Cheese/Leukocytes". This can be manually added to the LUT so that the above process outputs the associated sponsor LBTESTCD/LBTEST values. A portion of the LUT showing this is below in Table 10.

Codelist_Code	Codelist_Name	CDISC_Submission_Value	CDISC_Synonym_s_
C65047	Laboratory Test Code	CHDW	Corpuscular HGB Conc Distribution Width; Corpuscular Hemoglobin Concentration Distribution Width
C65047	Laboratory Test Code	CHDWR	Ret Corpuscular HGB Conc Distr Width; Reticulocyte Corpuscular Hemoglobin Distribution Width
C65047	Laboratory Test Code	CHEESELE	Ratio of Cheese to Leukocytes
C65047	Laboratory Test Code	CHKAB	Chikungunya Virus Antibody

NCI_Preferred_Term	syn1	syn2	syn3
Corpuscular Hemoglobin Concentration Distribution Width	Corpuscular HGB Conc Distribution Width	Corpuscular Hemoglobin Concentration Distribution Width	Corpuscular Hemoglobin Concentration Distribution Width
Reticulocyte Corpuscular Hemoglobin Distribution Width	Ret Corpuscular HGB Conc Distr Width	Reticulocyte Corpuscular Hemoglobin Distribution Width	Reticulocyte Corpuscular Hemoglobin Distribution Width
	Cheese-to-Leukocytes	Brielirubin	Fetatin
Chikungunya Virus Antibody Measurement	Chikungunya Virus Antibody		Chikungunya Virus Antibody Measurement

Table 10: LBTEST portion of the LUT with extended value added (Highlighted)

CONCLUSION

It can seem like a daunting task to up-version study data to a new CDISC CT list. Manually reviewing every addition, deletion, and alteration to official and sponsor-specific codelist values, and then applying those updates to all SDTM variables that use CDISC CT, would be a very time consuming and inefficient task. While some level of manual review will always be necessary, building tools to provide support for this process is time well spent.

ACKNOWLEDGMENTS

The authors would like to thank Reata Pharmaceutics, Inc. and Angie Goldsberry for facilitating and supporting this paper. We would specifically like to acknowledge Steve Kirby for his input and assistance.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Martha O'Brien
Reata Pharmaceutics, Inc.
martha.obrien@reatapharma.com

Keith Shusterman
Reata Pharmaceutics, Inc.
keith.shusterman@reatapharma.com