

# Mapping Raw Data to SDTM Define using SAS Clinical Standards Toolkit

Andrew Bullivant

PHASTAR

# Drivers for work

- Project with many studies requiring mapping to CDISC to generate submission ready documentation but no further reporting
- Evaluate SAS Clinical Standards Toolkit for define generation
- Define was key deliverable so desire to have this available early in the reporting process
- Raw metadata not complete or easily accessible
- Raw data not consistent across studies

# Raw to SDTM Mapping Information

- Spreadsheet for one experienced programmer to quickly run through a study raw data and record the following information:
  - SDTM variable(s) mapped to.
  - Where condition on SDTM variable e.g. VSTESTCD="SYSBP".
  - Supplemental QNAM and QLABEL.
  - NOT SUBMITTED and reason why.
  - Other useful information spotted while doing the mapping.
  - Record CRF page

## Raw to SDTM example

| A       | B        | C    | D      | E                           | F      | G          | H        | I           | J             | K                                                 | L                   | M               | N        | O         | P                  | Q                              | R                                                                                                                                                                                                                                                                                                                                                                    |
|---------|----------|------|--------|-----------------------------|--------|------------|----------|-------------|---------------|---------------------------------------------------|---------------------|-----------------|----------|-----------|--------------------|--------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| MEMNAME | NAME     | TYPE | LENGTH | LABEL                       | FORMAT | eDT or CRF | CRF Page | SDTM Domain | SDTM Variable | Other details                                     | Other SDTM Variable | Where Condition | Assigned | SUPP QNAM | SUPP QLABEL        | SDTM Variable Label            | CDISC Notes                                                                                                                                                                                                                                                                                                                                                          |
| ▼       | ▼        | ▼    | ▼      | ▼                           | ▼      | ▼          | ▼        | ▼           | ▼             | ▼                                                 | ▼                   | ▼               | ▼        | ▼         | ▼                  | ▼                              | ▼                                                                                                                                                                                                                                                                                                                                                                    |
| DMOG    | AGE      | 1    | 8      | Age (Years)                 |        | CRF        | 6        | DM          | AGE           |                                                   |                     |                 |          |           |                    | Age                            | Age expressed in AGEU. May be derived from RFSTDTC and BRTHDTC, but BRTHDTC may not be available in all cases (due to subject privacy).                                                                                                                                                                                                                              |
| DMOG    | ALC_AMT  | 1    | 8      | Alcohol Amount              |        | CRF        | 6        | SU          | SUDOSE        |                                                   |                     | SUTRT="ALCOHOL" |          |           |                    | Substance Use Consumption      | Amount of SUTRT consumed. Not populated if SUDOSTXT is populated.                                                                                                                                                                                                                                                                                                    |
| DMOG    | ALC_UNIT | 2    | 15     | Alcohol Amount Unit         | \$     | CRF        | 6        | SU          | SUDOSU        |                                                   |                     | SUTRT="ALCOHOL" |          |           |                    | Consumption Units              | Units for SUDOSE, SUDOSTOT, or SUDOSTXT. Examples: OUNCES, CIGARETTE EQUIVALENTS, GRAMS.                                                                                                                                                                                                                                                                             |
| DMOG    | ALC_USR  | 1    | 8      | Alcohol User                | YN     | CRF        | 6        | SU          | SUOCCUR       |                                                   | SU.SUTRT            | SUTRT="ALCOHOL" |          |           |                    | SU Occurrence                  | When the use of specific substances is solicited, SUOCCUR is used to indicate whether or not (Y/N) a particular pre-specified substance was used. Values are null for substances not specifically solicited.                                                                                                                                                         |
| DMOG    | A_NO     | 2    | 6      | Volunteer A Number          | \$     | CRF        | 5        | DM          |               |                                                   |                     |                 |          | ANUMBER   | Volunteer A Number |                                |                                                                                                                                                                                                                                                                                                                                                                      |
| DMOG    | COHORT   | 2    | 5      | Cohort                      | \$     | CRF        |          | DM          | ARM           |                                                   |                     |                 |          |           |                    |                                |                                                                                                                                                                                                                                                                                                                                                                      |
| DMOG    | DOB      | 1    | 8      | Date of Birth               | DATE   | CRF        | 5        | DM          | BRTHDTC       |                                                   |                     |                 |          |           |                    | Date/Time of Birth             | Date/time of birth of the subject.                                                                                                                                                                                                                                                                                                                                   |
| DMOG    | DTSTOP   | 1    | 8      | Date Stopped Smoking        | DATE   | CRF        | 6        | SU          | SUENDTC       |                                                   |                     | SUTRT="TOBACCO" |          |           |                    | End Date/Time of Substance Use |                                                                                                                                                                                                                                                                                                                                                                      |
| DMOG    | DTSTOPT  | 2    | 9      | Date Stopped Smoking (text) | \$     |            |          |             |               | NOT SUBMITTED - Same information as other field   |                     |                 |          |           |                    |                                |                                                                                                                                                                                                                                                                                                                                                                      |
| DMOG    | ENTRYCOM | 2    | 300    |                             | \$     |            |          |             |               | NOT SUBMITTED - Empty in all cases and not on CRF |                     |                 |          |           |                    |                                |                                                                                                                                                                                                                                                                                                                                                                      |
| DMOG    | RACE     | 1    | 8      | Race                        | RACE   | CRF        | 6        | DM          | RACE          |                                                   |                     |                 |          |           |                    | Race                           | Race of the subject. Sponsors should refer to "Collection of Race and Ethnicity Data in Clinical Trials" (FDA, September 2005) for guidance regarding the collection of race ( <a href="http://www.fda.gov/RegulatoryInformation/Guidances/ucm126340.htm">http://www.fda.gov/RegulatoryInformation/Guidances/ucm126340.htm</a> ) See Assumption below regarding RACE |

# Sheet characteristics

- First five columns populated from proc contents of raw data
- Conditional selection of SDTM domain and variable based on SDTMIG information
- Display of explanatory SDTMIG notes for selected variable

## Choose SDTM

|  | H        | I        | J         | K                 |
|--|----------|----------|-----------|-------------------|
|  | CRF Page | SDTM Dom | SDTM Vari | Other details     |
|  | 6        | DM       | AGE       |                   |
|  | 6        | SU       | DOSE      |                   |
|  | 6        | SU       | DOSU      |                   |
|  | 6        | SV       | DOCCUR    |                   |
|  | 6        | TA       |           |                   |
|  | 5        | TD       |           |                   |
|  | 5        | TE       |           |                   |
|  | 5        | TR       |           |                   |
|  | 5        | TS       |           |                   |
|  | 6        | TU       |           |                   |
|  |          |          |           | NOT SUBMITTED - S |
|  |          |          |           | NOT SUBMITTED - E |
|  | 6        | DM       | RACE      |                   |
|  | 6        | DM       |           |                   |

## Variable list conditional on domain

|   | I        | J         |           |
|---|----------|-----------|-----------|
|   | SDTM Dom | SDTM Vari | Other det |
| 6 | DM       | AGE       |           |
| 6 | SU       | SUDOSE    |           |
| 6 | SU       | SUDOSTXT  |           |
| 6 | SU       | SUDOSU    |           |
| 5 | DM       | SUDOSFRM  |           |
|   | DM       | SUDOSFRQ  |           |
| 5 | DM       | SUDOSTOT  |           |
| 6 | SU       | SUROUTE   |           |
|   |          | SUSTDTC   |           |
|   |          |           | NOT SUBM  |
|   |          |           | NOT SUBM  |

SDTMIG Label and Notes are shown

| Q                         | R                                                                                                                                                                                                            |
|---------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| SDTM Variable Label       | CDISC Notes                                                                                                                                                                                                  |
| Age                       | Age expressed in AGEU. May be derived from RFSTDTC and BRTHDTC, but BRTHDTC may not be available in all cases (due to subject privacy).                                                                      |
| Substance Use Consumption | Amount of SUTRT consumed. Not populated if SUDOSTXT is populated.                                                                                                                                            |
| Consumption Units         | Units for SUDOSE, SUDOSTOT, or SUDOSTXT. Examples: OUNCES, CIGARETTE EQUIVALENTS, GRAMS.                                                                                                                     |
| SU Occurrence             | When the use of specific substances is solicited, SUOCCUR is used to indicate whether or not (Y/N) a particular pre-specified substance was used. Values are null for substances not specifically solicited. |

# Sheet feedback

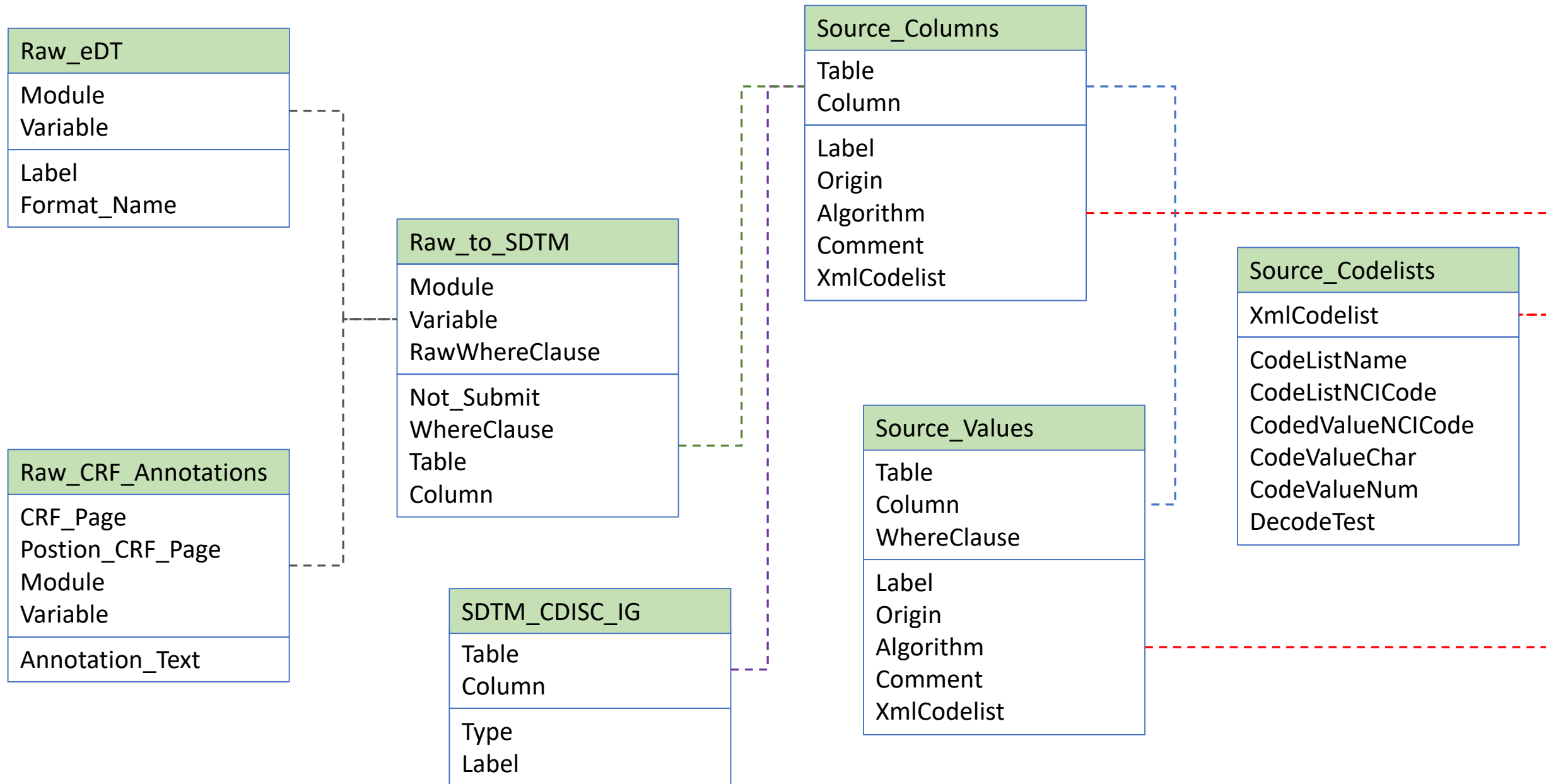
- Forces all raw data to be considered
- Quick to produce.
- Requires programmer to determine which other SDTM variables are needed
- Allows one person to go through all raw data and make mapping decisions.
- Does not contain all codelist information

# What is SAS Clinical Standards Toolkit

- A set of metadata definitions and macros to aid reporting clinical trials in SAS
- These are free from SAS
- Cover a range of reporting activities e.g. define.xml generation, TFL definition, helpful utility macros (e.g. Replacing Extended ASCII Characters in a SAS Data Set)
- Supports CDISC standards and regulatory requirements for defines

# Trial of SAS CST

- Focussed on mapping SDTM metadata to SAS CST metadata so as to allow the use of SAS CST macros to generate define.xml
- Mapped to the six standard SAS CST metadata datasets
  - SOURCE\_STUDY – one record containing study description
  - SOURCE\_TABLES – one record per SDTM dataset
  - SOURCE\_COLUMNS – one record per variable in each SDTM
  - SOURCE\_VALUES – value level information required for variables
  - SOURCE\_CODELISTS – one record per codelist item
  - SOURCE\_DOCUMENTS – document link (e.g. CRF page) info for each variable referencing a document



- SAS code that creates the SAS CST metadata datasets from
  - RAW to SDTM mapping
  - Additional variable metadata information. Covering assigned and derived variables.
  - SDTM datasets themselves – list of variables, lists of codelist values
  - SDTMIG information
  - CDISC Controlled terminology – NCI codes
- Then use SAS CST supplied driver program to run SAS CST macros to create define.xml

Define the standards you are using. Not just CDISC.

```
*****;
* create_definexml_from_source.sas *;
* *;
* Sample driver program to perform a primary Toolkit action, in this case, *;
* creating the Define-XML data sets from the CDISC standard model (SDTM, ADaM or *;
* SEND source metadata) in the WORK library, and then create a CDISC-DEFINE-XML *;
* V2.0.0 define.xml file. *;
* A call to a standard-specific Define-XML macro is required later in this code. *;
* *;
* Sample driver program to perform a primary Toolkit action, in this case, *;
* creating the Define-XML data sets from the CDISC standard model (SDTM, ADaM or *;
* SEND source metadata). *;
* A call to a standard-specific Define-XML macro is required later in this code. *;
* *;
* Assumptions: *;
* The SASReferences file must exist, and must be identified in the *;
* call to cstutil_processsetup if it is not work.sasreferences. *;
* *;
version 1.7 *;
* *;
The following statements may require information from the user *;
*****;

_cstStandard=CDISC-DEFINE-XML;
_cstStandardVersion=2.0.0; * <----- User sets the Define-XML version *;

%let _cstSrcType to _crtdds, _saslib or _define;
_cstSrcType=_crtdds;

%let _cstTrgStandard=CDISC-SDTM; * <----- User sets to standard of the source study *;
%let _cstTrgStandardVersion=3.1.2; * <----- User sets to standard version of the source study *;
%let _cstDefineFile=define-sdtm-3.1.2.xml;

%*let _cstTrgStandard=CDISC-ADAM; * <----- User sets to standard of the source study *;
%*let _cstTrgStandardVersion=2.1; * <----- User sets to standard version of the source study *;
%*let _cstDefineFile=define-adam-2.1.xml;

%* Subfolder with the SAS Source Metadata data sets;
%let _cstSrcMetaFolder=%lowercase(&_cstTrgStandard)-&_cstTrgStandardVersion/metadata;

*****;
* The following code sets (at a minimum) the studyrootpath and studyoutputpath. These are *;
* used to make the driver programs portable across platforms and allow the code to be run with *;
* minimal modification. These macro variables by default point to locations within the *;
* cstSampleLibrary, set during install but modifiable thereafter. The cstSampleLibrary is assumed *;
* to allow write operations by this driver module. *;
*****;

%cst_setStandardProperties(_cstStandard=CST-FRAMEWORK,_cstSubType=initialize);
%cstutil_setstudyroot;
```

Define paths to study metadata.

```
%cst_setStandardProperties(_cstStandard=CST-FRAMEWORK,_cstSubType=initialize);
options mprint mlogic;
%cstutil_setcstsroot;
data _null_;
  call symput('studyRootPath',cats("P:\xxx\Data\Sandbox"));
  call symput('studyOutputPath',cats("P:\xxx\Output\Sandbox"));
run;
%let workPath=%sysfunc(pathname(work));
```

```
*****;
* One strategy to defining the required library and file metadata for a CST process *;
* is to optionally build SASReferences in the WORK library. An example of how to do *;
* this follows. *;
* *;
* The call to cstutil_processsetup below tells CST how SASReferences will be provided *;
* and referenced. If SASReferences is built in work, the call to cstutil_processsetup *;
* may, assuming all defaults, be as simple as: *;
* %cstutil_processsetup() *;
*****;
```

```
%let _cstSetupSrc=SASREFERENCES;
options mprint mlogic;
%cst_createdsfromtemplate(_cstStandard=CST-FRAMEWORK,_cstType=control,_cstSubType=reference,_cstOutputDS=work.sasreferences);
```

proc sql;

```
insert into work.sasreferences
values ("CST-FRAMEWORK" "1.2" "messages" "" "messages" "libref" "input" "dataset" "N" "" ""
values ("&_cstStandard" "&_cstStandardVersion" "messages" "" "crtmsg" "libref" "input" "dataset" "N" "" ""
values ("&_cstStandard" "&_cstStandardVersion" "autocall" "" "auto1" "fileref" "input" "folder" "N" "" ""
values ("&_cstStandard" "&_cstStandardVersion" "control" "reference" "control" "libref" "both" "dataset" "Y" "" "&work
values ("&_cstStandard" "&_cstStandardVersion" "properties" "initialize" "inprop" "fileref" "input" "file" "N" "" ""
values ("&_cstStandard" "&_cstStandardVersion" "results" "results" "results" "libref" "output" "dataset" "Y" "" "&stud
values ("&_cstStandard" "&_cstStandardVersion" "sourcedata" "" "srcdata" "libref" "output" "folder" "Y" "" "&work
values ("&_cstTrgStandard" "&_cstTrgStandardVersion" "studymetadata" "study" "sampdata" "libref" "input" "dataset" "N" "" "&stud
values ("&_cstTrgStandard" "&_cstTrgStandardVersion" "studymetadata" "table" "sampdata" "libref" "input" "dataset" "N" "" "&stud
values ("&_cstTrgStandard" "&_cstTrgStandardVersion" "studymetadata" "column" "sampdata" "libref" "input" "dataset" "N" "" "&stud
values ("&_cstTrgStandard" "&_cstTrgStandardVersion" "studymetadata" "codelist" "sampdata" "libref" "input" "dataset" "N" "" "&stud
values ("&_cstTrgStandard" "&_cstTrgStandardVersion" "studymetadata" "value" "sampdata" "libref" "input" "dataset" "N" "" "&stud
values ("&_cstTrgStandard" "&_cstTrgStandardVersion" "studymetadata" "document" "sampdata" "libref" "input" "dataset" "N" "" "&stud
values ("&_cstStandard" "&_cstStandardVersion" "externalxml" "xml" "extxml" "fileref" "output" "file" "Y" "" "&stu
values ("&_cstStandard" "&_cstStandardVersion" "referencexml" "stylesheet" "xslt01" "fileref" "output" "file" "Y" "" "&stu
```

```
;
quit;
```

Set up work.references dataset that has necessary info for rest of code to run. User interventions stops here.

Process the study metadata to datasets better structured for define generation.

Create the define.  
Validation steps and steps to create output format other than xml (e.g. HTML) follow

```
else
    call execute("%sysfunc(tranwrd(options %cmpres(&_cstDebugOptions), %str( ), %str( no)));");
run;

*****;
* The following macro (cstutil_processsetup) utilizes the following parameters:      *;
*                                                                                      *;
* _cstSASReferencesSource - Setup should be based upon what initial source?          *;
*   Values: SASREFERENCES (default) or RESULTS data set. If RESULTS:                *;
*   (1) no other parameters are required and setup responsibility is passed to the   *;
*       cstutil_reportsetup macro                                                    *;
*   (2) the results data set name must be passed to cstutil_reportsetup as          *;
*       libref.memname                                                               *;
*                                                                                      *;
* _cstSASReferencesLocation - The path (folder location) of the sasreferences data set *;
*                               (default is the path to the WORK library)              *;
*                                                                                      *;
* _cstSASReferencesName - The name of the sasreferences data set                     *;
*                               (default is sasreferences)                          *;
*****;

%cstutil_processsetup();

*****;
* Run the standard-specific Define-XML macros.                                     *;
*****;

%define_sourcetodefine(
    _cstOutLib=srcdata,
    _cstSourceStudy=sampdata.source_study,
    _cstSourceTables=sampdata.source_tables,
    _cstSourceColumns=sampdata.source_columns,
    _cstSourceCodeLists=sampdata.source_codelists,
    _cstSourceDocuments=sampdata.source_documents,
    _cstSourceValues=sampdata.source_values,
    _cstFullModel=N,
    _cstLang=en
);

%define_write(
    _cstCreateDisplayStyleSheet=1,
    _cstHeaderComment=%str( Produced from SAS data using the SAS Clinical Standards Toolkit &_cstVersion )
);
```

# Findings

- Relatively easy to move mapping code from study to study
- Getting SAS CST define macro to run took some doing but it was more the environment than the CST macros. Once running it has been fine and not had to delve into SAS CST macro code.
- SAS CST macros appear robust with error tracking etc. Documentation and layout takes a while to get used to.
- Having metadata generated from both the specs and the actual SDTMs highlighted differences between code and spec
- Having metadata in SAS datasets made for easy transfer through to code for automatic generation of SDTM CRF annotations. Also useful when manually generating ADaM specs.

# Producing the Define.xml early

- In these studies the define.xml was run and created several times during prod coding of SDTMs.
- Relatively good quality define.xml available near to the start of QC programming
- Was intending that the define.xml would act as replacement for the Standard Phastar SDTM specs. In reality this did not really happen because programming was too far down the track by the time this was produced.

# Findings

- Define.xml needs to be produced before programming starts.
- Requires additional metadata files to identify required variables and some codelist information
- Retain production of define based on actual SDTM data as this highlights differences to spec and allows programming lead to quickly see what is actually going on in all datasets
- SAS CST is a good tool for define generation
- SAS CST metadata structure is clear and easy to use

# Next steps

- Combine with information coming from raw annotated CRF to produce define and CST metadata (see John McDade's poster)
- Generation of SAS SDTM code from CST metadata and raw to SDTM mapping information.

# Questions and Feedback