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. VTESTCODE="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

<table>
<thead>
<tr>
<th>Column</th>
<th>Label</th>
<th>Format</th>
<th>SDTM Domain</th>
<th>Other Details</th>
<th>Where Condition</th>
<th>Assigned</th>
<th>CORB Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>NAME</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AGE</td>
<td>Age (Years)</td>
<td>CRF</td>
<td>6 DM AGE</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ALC_AMT</td>
<td>Alcohol Amount</td>
<td>CRF</td>
<td>6 SU SUBDOSE</td>
<td></td>
<td>SUBTR^&quot;ALCOHOL&quot;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ALC_UNIT</td>
<td>Alcohol Amount Unit</td>
<td>$</td>
<td>6 SU SUBDOSE</td>
<td></td>
<td>SUBTR^&quot;ALCOHOL&quot;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ALC_USR</td>
<td>Alcohol User</td>
<td>YN</td>
<td>6 SU SUCCUR</td>
<td></td>
<td>SUBTR^&quot;ALCOHOL&quot;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A_No</td>
<td>Volunteer A Number</td>
<td></td>
<td>6 DM ANUMBER</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DOB</td>
<td>Date of Birth</td>
<td>DATE</td>
<td>6 DM BIRTHDC</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DTSTOP</td>
<td>Date Stopped Smoking</td>
<td>DATE</td>
<td>6 SU SUBENDC</td>
<td></td>
<td>SUBTR^&quot;TOBACCO&quot;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DTSTOPPT</td>
<td>Date Stopped Smoking (text)</td>
<td>$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ENTRYCOM</td>
<td></td>
<td>300</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RACE</td>
<td>Race</td>
<td></td>
<td>6 DM RACE</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Notes:*  
- Age expressed in AGEL. May be derived from INFTOTDC and BIRTHDC, but BIRTHDC may not be available in all cases due to subject source.  
- Amount of SUBTR† consumed. Not populated if SUBDOSE is populated. Units for SUBDOSE, SUBSTOT, or SUBSTSTOT, Examples: UNITS, CIGARETTE EQUIVALENTS, GRAMS.  
- When the use of specific substances is solicited, SUCCUR is used to indicate whether or not (YN) a particular pre-specified substance was used. Values are not for substances not specifically solicited.

*Column Notes:*  
- A: Memo Name  
- B: Name  
- C: Type  
- D: Length  
- E: Label  
- F: Format  
- G: SDTM Domain  
- H: Other Details  
- I: Where Condition  
- J: Assigned  
- K: CORB Notes  
- L: SDTM Variable Label  
- M: SDTM Variable  
- N: SUPP QNAM  
- O: SUPP QLABEL  
- P: SDTM Variable Label  
- Q: SDTM Variable  
- R: CORB Notes
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

Variable list conditional on domain

SDTMIG Label and Notes are shown
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
<table>
<thead>
<tr>
<th>Module</th>
<th>Variable</th>
<th>Label</th>
<th>Format_Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw_eDT</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Source_Columns</td>
<td>Table</td>
<td>Column</td>
<td></td>
</tr>
<tr>
<td>Source_Codelists</td>
<td>XmlCodelist</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Source_Values</td>
<td>Table</td>
<td>Column</td>
<td>WhereClause</td>
</tr>
<tr>
<td>SDTM_CDISC_IG</td>
<td>Table</td>
<td>Column</td>
<td></td>
</tr>
<tr>
<td>Annotation_Text</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Raw_CRF_Annotations</td>
<td>CRF_Page</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Postion_CRF_Page</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Module</td>
<td>Variable</td>
<td>Label</td>
<td></td>
</tr>
<tr>
<td>WhereClause</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not_Submit</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Raw_to_SDTM</td>
<td>Module</td>
<td>Variable</td>
<td>RawWhereClause</td>
</tr>
<tr>
<td>Source_Values</td>
<td>Table</td>
<td>Column</td>
<td>WhereClause</td>
</tr>
<tr>
<td>Type</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Label</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
• 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.
Define paths to study metadata.

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.
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 SDTM highlight 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 SDTM.

• 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