QC of the aCRF using SAS

By Senior Consultant Rune Østergaard Pedersen and Principal Consultant Niels Both
Outline

- The SDTM format
- The aCRF
- QC as a part of documentation for SDTM
  - ways to do QC
- Using the hierarchical structure of SDTM
- Examples using a sample aCRF
- Small examples on how to make the SAS-code
- Output tables
- Improvements
The SDTM format

- Has become the standard for submission of clinical trials.
- The SDTM data are divided into a number of datasets, each with a set of variables, and values within variables.
- The SDTM has a hierarchical structure, and this can be utilized when doing QC.

![SDTM Hierarchy Diagram]

- **Toplevel:**
  - Domain e.g. QS=QUESTIONAIRES

- **Intermediate level:**
  - QSCAT, QSSCAT, QSORRES etc.

- **Lower level:**
  - Value e.g. QSORRES when QSTESTCD = MMSEA1
The aCRF

- The raw CRF is annotated to make a link between the SDTM data and the questions on the CRF.
- Often previous trials are applied as templates for annotations.
- The annotations serve as input to the define.xml, and should be correct.
- When designing the aCRF a number of rules may be applied to link the data (domains in SDTM).
  - Colours of background in annotation boxes.
  - Standardized ways of writing notes, supplemental qualifiers, relations and variable names etc.
Macro-view of the ideal direct dataflow

Does not origin from aCRF, only cosmetic like sample dates

Iterative updating the aCRF and the SDTM until the final run
QC as a part of documentation for SDTM and aCRF

- Overview of the process

Input data source

SDTM framework

SDTM- QC

Data level (is raw data represented)

Internal SDTM inconsistency (Open CDISC validator)

aCRF versus SDTM and SDTM versus aCRF

Import FDF-file/SDTM and data preparation

Run SAS macros

List of files and problem

Correct aCRF and data

Manual QC of CRF paperform

Electronic QC raw data
<table>
<thead>
<tr>
<th>Level</th>
<th>Rule of inclusion</th>
<th>Rule of exclusion (data driven)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Domain</strong></td>
<td>Grouping using colors of domain variable</td>
<td></td>
</tr>
<tr>
<td></td>
<td>On each page the domain should be mentioned e.g. AE=ADVERSE EVENTS</td>
<td></td>
</tr>
<tr>
<td><strong>Variable</strong></td>
<td>The variable names are left of equal signs. In case of no equal sign the text is a variable. QNAM variables are separated with “-” and numbered successively like QNAM=AEACN1-AEACN6</td>
<td>Origin in the data that is not CRF e.g. EDC, could be adjudication data.</td>
</tr>
<tr>
<td><strong>Value</strong></td>
<td>Separation of values is done with “,”. Certain words are used to identify variables. For example LBORRES when LBTESTCD=TOT_CARH. The “when” and “=” are identifiers.</td>
<td></td>
</tr>
<tr>
<td><strong>Notes and other labels</strong></td>
<td>None</td>
<td>Notes are always initiated with NOTE:, this allows to remove the notes from the SAS dataset originating from the FDF-file. “Not Submitted” labels are excluded.</td>
</tr>
<tr>
<td><strong>Replicated forms (paper forms)</strong></td>
<td>Replications of the same forms on the aCRF are written AS PAGE x (where x indicates a number).</td>
<td></td>
</tr>
<tr>
<td><strong>RELREC</strong></td>
<td>The domains in RELREC are identified on the aCRF as RELREC AE,DS. In case of more than two domains the list is expanded as RELREC ZZ,YY,XX</td>
<td>None</td>
</tr>
</tbody>
</table>
Identifying potential errors

SDTM library - SDTM data

SASHELP library - vcolumn data

aCRF

FDF-file

Common dataset

Exclude the common (no errors)

Finding the values only in SDTM

Combined dataset with errors

Remove potential false positives by search in metadata

Finding the values only in aCRF

Excel report for documentation, including comments still relevant from previous reports
The SAS code starts by checking on the toplevel, moving to the intermediate level and finally the low level. Each level may be exemplified by a pseudocode. In general, terms a logical condition can be made in three loops:

1) %If not %sysfunc(exist(sdtm.var)) %then  %do;
   %end
2) %else %do; /Intermediate level*/
   /*Making a macro variable containing the intermediate level variable*/
   /*If macro variable not in common dataset then output*/
   3) /*else do*/
   /*Testing a variable on the low level*/
   /*If low level variable do not exist then output error in report*/
   /*end*/
/end*/
%end
Testing the toplevel:
We make count if the domain exist on the aCRF

%if not %sysfunc(exist(sdtm.RELREC)) %then
  %do;

    data RELREConCRF_notDATA;
    LENGTH DOMAIN $ 20 problem $ 500 checkdesc $200;
    domain="RELREC";
    problem="Missing SDTM-RELREC domain";
    checkno=5;
    checkdesc="Find if RELREC domain is missing from SDTM data";
    run;

%end;
Testing inside the low level/intermediate loops

Data mytest; /*Datacleaning step*/
Set rawdata;
Where myvar1="wish" and myvar2 ="" etc;
Run;

Proc Sql noprint; /*Testing step*/
Create table problems as
Select b.var as var 1 "Problem" as problem
From mytest where (a.key not in data_on_crf); quit;
### Examples using a sample aCRF from CDISC

**IE=Inclusion/Exclusion**

<table>
<thead>
<tr>
<th>CDISC Study</th>
<th>VISIT</th>
<th>Screening</th>
</tr>
</thead>
<tbody>
<tr>
<td>CDISC01</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>IEDTC</td>
</tr>
<tr>
<td></td>
<td>Assessment Date:</td>
<td></td>
</tr>
</tbody>
</table>

#### ELIGIBILITY CRITERIA

**INCLUSION CRITERIA**

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Response</th>
<th>Status</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Is age 18 - 85.</td>
<td>Yes</td>
<td>No</td>
<td>Will give false positive</td>
</tr>
<tr>
<td>2. Has Xyz disease of at least 10 weeks duration confirmed by biopsy</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Did not respond to a standard course of medication ABC.</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **IEORRES when IETESTCD = INCL01**
- **IEORRES when IETESTCD = INCL02**
- **IEORRES when IETESTCD = INCL03**

**Wrong colour**
To help the reader we would add: NOTE: RACE, when more than one selected, RACE=MULTIPLE and individual responses are RACE1, RACE2, etc. in SUPPDM

Loop is made for qnam

False positive error, only white, black or african american and Asian exist.
Missing domain annotation on top of continuing pages will give multiple errors in the program

Should be MHCAT=MEDICAL

<table>
<thead>
<tr>
<th>CDISC Study CDISC01</th>
<th>SCREENING</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MHDTC</td>
</tr>
<tr>
<td>Assessment Date:</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>MEDICAL AND SURGICAL HISTORY</th>
<th>MHCAT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Does the subject have any significant medical or surgical history? [NOT SUBMITTED]</td>
<td></td>
</tr>
<tr>
<td>Yes, list the condition(s) below</td>
<td>No</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>MHTERM</th>
<th>Year</th>
<th>“✓” if RESOLVED</th>
<th>“✓” if ONGOING</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MHSTDTC</td>
<td>MHENRF = BEFORE</td>
<td>MHENRF = DURING/AFTER</td>
</tr>
</tbody>
</table>
False error due to noone using the supp-domain

<table>
<thead>
<tr>
<th>EGTESTCD = INTP</th>
</tr>
</thead>
<tbody>
<tr>
<td>OVERALL INTERPRETATION (Please check one):</td>
</tr>
<tr>
<td>1 = ☐ Normal (do not comment) <strong>EGORRES = NORMAL</strong></td>
</tr>
<tr>
<td>2 = ☐ Abnormal, not clinically significant (do not comment) <strong>EGORRES = ABNORMAL</strong></td>
</tr>
<tr>
<td>3 = ☐ Abnormal, clinically significant. Specify and comment: <strong>EGORRES = ABNORMAL</strong></td>
</tr>
<tr>
<td>Comments [char(200)]</td>
</tr>
<tr>
<td><strong>EGCLSIG=N in SUPPEG</strong></td>
</tr>
<tr>
<td><strong>EGCLSP in SUPPEG</strong></td>
</tr>
</tbody>
</table>

Potential error in EGCLSP needs investigation in raw data.
### DS=Disposition

<table>
<thead>
<tr>
<th>CDISC</th>
<th>Assessment Date: <em><strong><strong>/</strong></strong></em>/_____</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study CDISC01</td>
<td>[NOT SUBMITTED]</td>
</tr>
</tbody>
</table>

#### TERMINATION

Did patient complete the study? 

- [ ] Yes
- [x] No

**DSDECOD** / **DSTERM = COMPLETED when Yes**

If patient did not complete the study, indicate the date of termination and check one primary reason to indicate why:

- **Date of Termination:** _____/_____/_____  **DSSTDTC**
- **DSDECOD**
  - Patient did not meet Inclusion/Exclusion Criteria at Screening or baseline (specify):  **DSTERM**
- **DSDECOD** Discontinued due to lack of Therapeutic Response
- **DSDECOD** Discontinued due to Adverse Event

Adverse Event No. ________(Enter the number from the ADVERSE EVENTS Form)

**Missing RELREC: AE, DS**
# Example of a report (modified from Excel)

<table>
<thead>
<tr>
<th>Check No</th>
<th>Check Description</th>
<th>Category</th>
<th>Problem</th>
<th>Annotation</th>
<th>Domain</th>
<th>Page</th>
<th>Color</th>
<th>Variable</th>
<th>Value</th>
<th>MetaData</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Find if a variable has been annotated on a page, but with no matching domain annotation. This could be caused by color differences.</td>
<td>aCRF Annotation Problem</td>
<td>No domain annotated on aCRF</td>
<td>IEDTC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>A color difference has been removed</td>
</tr>
<tr>
<td>2</td>
<td>Find if a variable with value (eg XXTESTCD=TES T) is annotated on the aCRF, but that value does not exist in the SDTM data</td>
<td>IETESTCD present but no value of INCL01 exists in data</td>
<td>IEORRES when IETESTCD=INCL01 exists in data</td>
<td>IE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>The inclusion criteria was never violated, hence the value is not in SDTM</td>
</tr>
</tbody>
</table>
Example of a report continued.

<table>
<thead>
<tr>
<th>Check No</th>
<th>Check Description</th>
<th>Category</th>
<th>Problem</th>
<th>Annotation</th>
<th>Domain</th>
<th>Page</th>
<th>Colour</th>
<th>Variable</th>
<th>Value</th>
<th>Metadatates</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>Find if a supplemental qualifier variable has been annotated on the aCRF, but is not present in data</td>
<td>Annotated on aCRF, but not in SDTM data</td>
<td>QNAM annotated on aCRF, but not present in supp qual data</td>
<td>RACE1 - RACE5 in SUPPD M</td>
<td>SUPP DM</td>
<td>6</td>
<td></td>
<td>QNAM</td>
<td>RA CE 4</td>
<td></td>
<td>The Native Hawaiian or Other Pacific Islander is never found in the raw data</td>
</tr>
<tr>
<td>4</td>
<td>Find if a variable has been annotated on a page, but with no matching domain annotation. This could be caused by color differences.</td>
<td></td>
<td>aCRF Annotation Problem</td>
<td>No domain annotated on aCRF</td>
<td>MHDT C</td>
<td>7</td>
<td>C[0.75 1 1]</td>
<td></td>
<td></td>
<td></td>
<td>The error is due to missing domain annotation</td>
</tr>
</tbody>
</table>
## Example of a report continued.

<table>
<thead>
<tr>
<th>Check No</th>
<th>Check Description</th>
<th>Category</th>
<th>Problem</th>
<th>Annotation</th>
<th>Domain</th>
<th>Page</th>
<th>Colour</th>
<th>Variable</th>
<th>Value</th>
<th>Metadatest</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>Find if a RELREC domain combination that exists in the RELREC SDTM domain has not been annotated on the aCRF</td>
<td>In SDTM data, but not annotated on aCRF</td>
<td>No annotation on aCRF for domains linked in RELREC</td>
<td>RELREC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>The annotation RELREC: AE.DS has been added to the aCRF</td>
</tr>
<tr>
<td>6</td>
<td>Find if a supplemental qualifier variable has been annotated on the aCRF, but is not present in data</td>
<td>Annotated on aCRF, but not in SDTM data</td>
<td>QNAM annotated on aCRF, but not present in supp qual data</td>
<td>EGCLSP</td>
<td>SUPPEG</td>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>After a look in raw data it looks like no one entered a comment, hence OK</td>
</tr>
</tbody>
</table>
Explanations of the output variables

- **Check No** – Refers to the list of checks used by the SAS macros

- **Check Description** – A short description of the problem/error. For example: Find if a variable has been annotated on a page, but with no matching domain annotation. This could be caused by colour differences.

- **Category** – The program uses one of the categories:
  1) aCRF Annotation Problem.
  2) Annotated on aCRF, but not in SDTM data.
  3) In SDTM data, but not annotated on aCRF.

- **Problem** – this is a more fine description of the problem. For example: No domain annotated on aCRF or QNAM annotated on aCRF, but not present in supp qual data.

- **Page** – the pagenumber of the aCRF that the annatation originates from
• **Annotation** – This is a variable containing the original annotation on the aCRF. Is only relevant if the problem is missing data for annotations present on the aCRF.

• **Domain** – the two letter abbreviation or supplemental qualifier abbreviation. E.g. LB and SUPPLB that the data/annotation is linking to.

• **Colour** – this is only relevant for colour issues on the aCRF and refers to the colour code that are under investigation.

• **Variable** – the variable in SDTM. However, in supplemental qualifier it may refer to a qnam.

• **Value** – this is a value of a SDTM variable e.g. a testcd for lbtestcd.

• **Comment** – these are comments used to justify the error, and may be applied for documentation of the QC-process of SDTM. Comments not changed, but still relevant are saved from one iteration of the program to the next.
Conclusion

- The QC program has proven as an important tool, enabling identification of many errors that cannot easily be found manually.
- The complexity of the dataflow often gives a number of false positives, which makes manual QC needed.
- The QC program can easily be applied as a part of the documentation for authorities.
- The way of creating annotation that can be applied to QC on PC have the positive side-effect of being more systematic.
- New employees often benefit from the errors detected by the PC, making their SDTM learning curve less steep.