PhUSE 2011

Paper CD04

Ensuring Compliant and Consistent Data Mappings for SDTM-based Studies – an ICON Approach

Jennie Mc Guirk, ICON, Dublin, Ireland
Steven Thacker, ICON, Marlow, UK

ABSTRACT

This paper presents the various tools that are leveraged, in order to ensure compliant and consistent data mappings, across Study Data Tabulation Model (SDTM) based studies within ICON Clinical Research.

An increasing number of companies are looking to convert legacy studies of various ages to fully CDISC compliant packages ahead of submission, therefore it is essential to have the means available to ensure that all of the studies are harmonized and compliant in an efficient and automated way to realize what can be tight timelines, utilizing limited resource.

In addition to standard checks such as OpenCDISC and WebSDM, ICON Clinical Research's compliance and consistency toolkit also comprises of a series of checks based around a data warehouse structure that captures information from the annotated CRF, metadata and the transformed data itself. These checks and reports can be run within an individual study, or between/across a series of studies to ensure that data is fully CDISC compliant and harmonized. The input to these checks consists of three main sources: the annotated CRF, the SDTM dataset specifications, and SDTM SAS® datasets. The output is a group of reports used to identify inconsistencies, and other potential issues for resolution.

This paper is an introduction to the ICON Clinical Research tools developed within SAS, the reports produced and how they are interpreted to aid harmonization. Anonymized case study examples will be included to ground the presentation in a practical setting.

INTRODUCTION

SDTM is about standardizing and normalizing clinical trial data. Therefore, consistent mapping of Clinical raw data to SDTM is important especially for projects with multiple studies, and in particular for projects where integrated analysis may be needed.

However, differences can arise as the number of studies increases, either due to human error or difference in interpretation of the implementation guides.

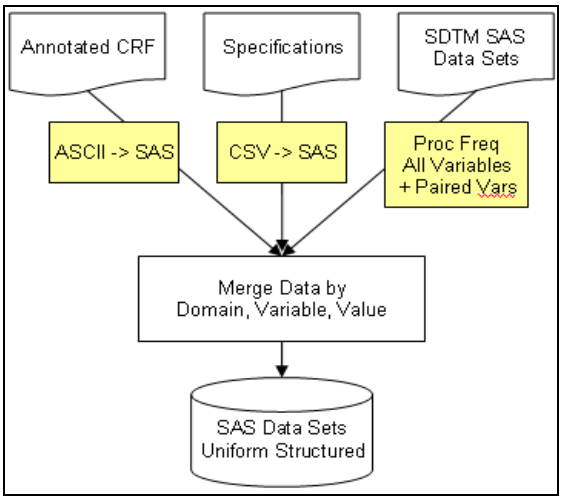
To increase across study comparisons, ICON has a tool that centralizes the metadata and data in our SDTM-based studies. Like a data warehouse, it stores our experience with SDTM, and is a one stop source to look up anything we may need when developing new SDTM-based studies.

BACKGROUND

Base SAS is the programming environment for both the data warehouse and the reports.

ICONs' tool imports both the CRF annotations, and the SDTM dataset Specifications into SAS, creating the metadata that will be used as a basis for comparison. The metadata from the SDTM SAS datasets is readily available. We now have metadata from the three key components of our submission package, which can be merged into a series of SAS datasets and compared. This process flow is represented within Process Flow 1 below.

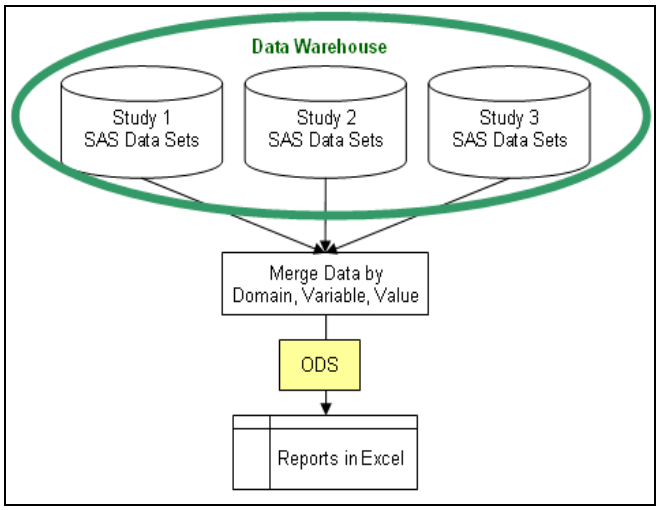
Process Flow 1



Within an individual study, we can now run reports to determine if we have any discrepancies or inconsistencies between the Annotated CRF, SDTM specifications and/or SDTM datasets.

More importantly, due to the now uniform structure of the output created by the ICON tool, the metadata from each study can also be merged and compared across studies, to create reports as shown in Process Flow 2.

Process Flow 2



Reports are created using SAS ODS. Final reports are in Excel files, so the reviewer of the reports can take advantage of the AutoFilter functionality and drill down on potential issues and/or inconsistencies.

1. SETUP

The setup of the ICON Data Warehouse tool is described for each of the inputs.

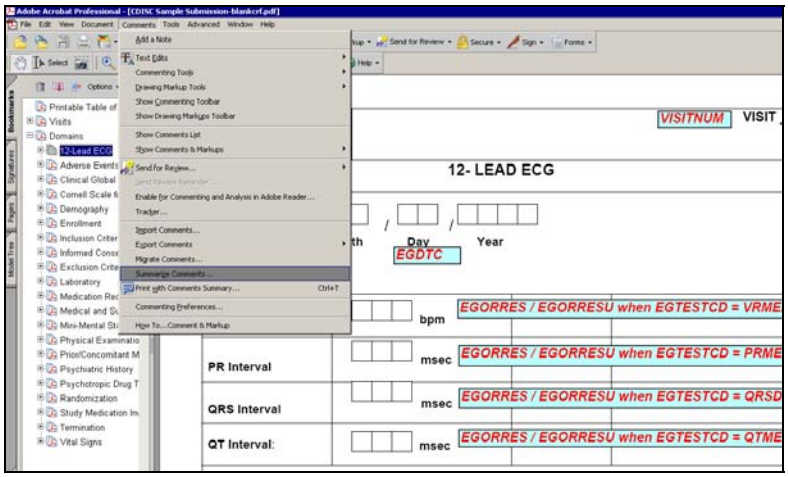
1.1 Annotated CRF
1.2 SDTM/CRT Specifications
1.3 SDTM/CRT SAS Datasets

1.1 ANNOTATED CRF

The annotations in the CRF are created as Comments in the PDF file. ICON follows the SDTM Submission Guidelines for consistent format and layout.

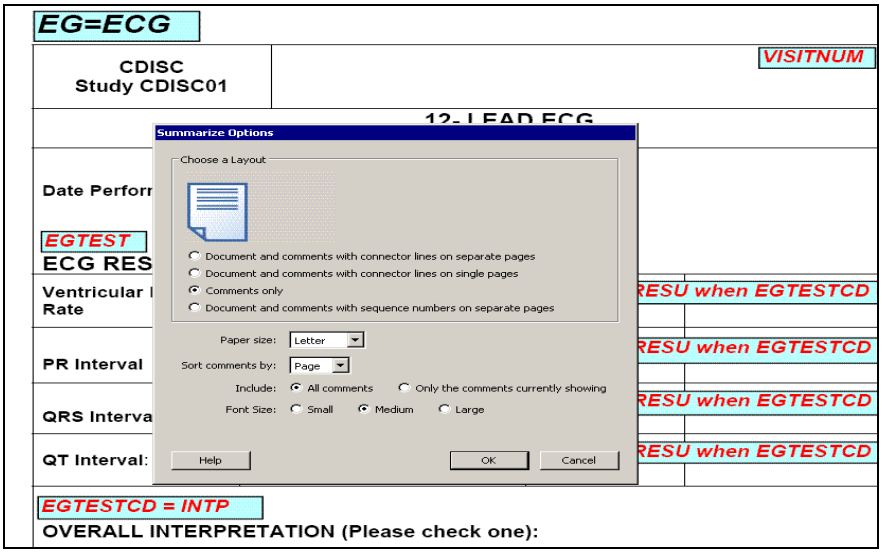
To import the annotations to SAS, we first save the annotations in ASCII file.
In Adobe Acrobat, in the menu bar click Comments -> Summarize Comments as shown in Screen 1.1.1

Screen 1.1.1



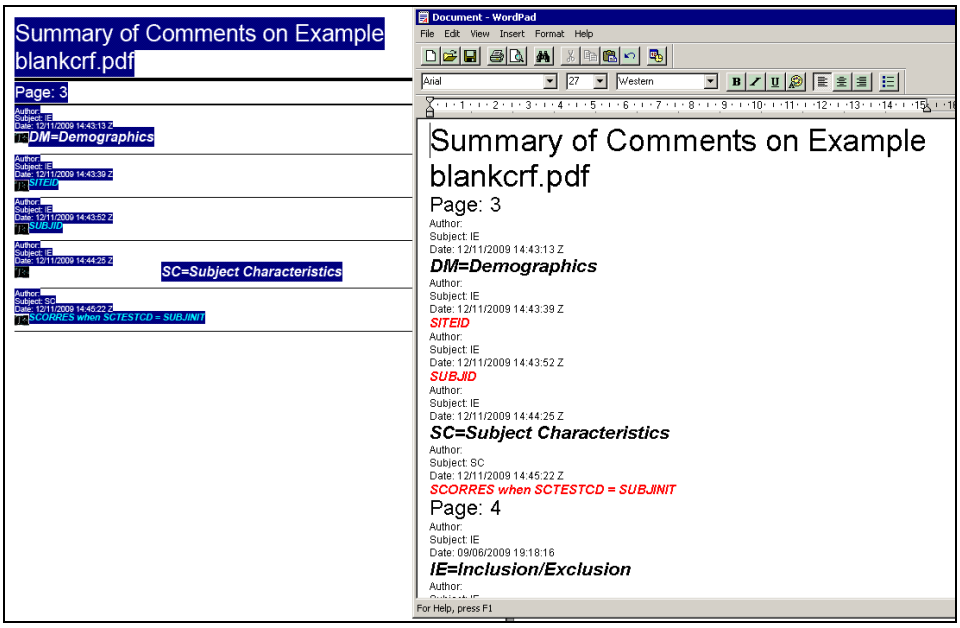
Select Comments only, as shown in Screen 1.1.2. This extracts the annotation text to a separate window.

Screen 1.1.2



Copy all the text and paste to a text editor, as shown in Screen 1.1.3.

Screen 1.1.3



The annotations are saved to an ASCII file (e.g. aCRF_comments.txt). The annotations can now be read to SAS by using a simple data step.

Note that from Screen 1.1.3, only the comments in red text are the annotations. Because the annotations follow certain rules in terms of format, we can now scan the imported text, and extract the information about SDTM domains, the variable names, values and also the CRF page number. Screen 1.1.4 shows sample SAS code to process the imported text.

Screen 1.1.4

```
filename acrf aCRF_comments.txt';
data work.aCRF;
    infile aCRF lrecl = 2000 dsd firstobs = 1 missover;
    attrib text length = $1000;
    input text;
    length Domain $10 Variable $40 Value $200 Page 8;
    ** More statements here...;
    if index(text,'Author:') = 0;
    if index(text,' in ') > 0 and index(text,' = ') > 0 then do;
         domain = scan(text,5,' ');
         Variable = scan(text,1,' ');
         value = scan(text,3,' ');
    end;
    ** More statements here...;
run;
```

The ICON tool now outputs a SAS dataset. An example is shown in Screen 1.1.5.

Screen 1.1.5

VIEWTABLE: Work.Acrf

	Domain	Variable	Value	Page	CRF	text
1	EG	EGTESTCD	INTP	12	12-Lead ECG	EGTESTCD = INTP
2	EG	EGORRES	NORMAL	12	12-Lead ECG	EGORRES = NORMAL
3	EG	EGORRES	ABNORMAL	12	12-Lead ECG	EGORRES = ABNORMAL
4	SUPPEG	EGCLSIG	N	12	12-Lead ECG	EGCLSIG = N in SUPPEG
5	EG	EGORRES	ABNORMAL	12	12-Lead ECG	EGORRES = ABNORMAL
6	SUPPEG	EGCLSIG	Y	12	12-Lead ECG	EGCLSIG = Y in SUPPEG
7	SUPPEG	EGCLSP		12	12-Lead ECG	EGCLSP in SUPPEG

1.2 SDTM/CRT SPECIFICATIONS

SDTM dataset Specifications are prepared and saved as individual CSV files per domain. Each CSV file includes the SDTM domain name, variable names, variable labels, type, length, controlled terminology terms and other required metadata. Screen 1.2.1 is an example for EG domain and its supplemental qualifier SUPPEG.

Screen 1.2.1

Dataset	Variable Name	Label	Type	Length	Controlled Terminology
EG	STUDYID	Study Identifier	Char	15	
EG	DOMAIN	Domain Abbreviation	Char	2	EG
EG	USUBJID	Unique Subject Identifier	Char	25	
EG	EGSEQ	Sequence Number	Num	8	
EG	EGTESTCD	ECG Test or Examination Short Name	Char	8	INTP; PR
EG	EGTEST	ECG Test or Examination Name	Char	40	ECG Interpretation; PR Interval
EG	EGORRES	Result or Finding in Original Units	Char	200	NORMAL; ABNORMAL
SUPPEG	EGCLSIG	Clinically Significant	Char	1	Y
SUPPEG	EGCLSP	Clinically Significant Specify	Char	200	

Using a SAS data step, combined with a SAS Macro, the tool loops through each domain and reads in the individual CSV files into SAS. Screen 1.2.2 is the sample code.

Screen 1.2.2

```
%macro loopcsv;
    ** Create macro variable &maxdsn: total # of CSV files;
    ** Create macro variable &&dsname&didx: name of each CSV file;
    ** Loop through individual CSV files;
    %do didx=1 %to &maxdsn;
        filename incsv "&&dsname&didx..csv";
        data &&dsname&didx;
            length Domain $10 Variable $40 Label $200 Type $40
                   Length 8 Terms $2000;
            infile incsv dlm="," dsd missover lrecl=10000 firstobs=2;
            input Domain $ Variable $ Label $ Type $ Length Terms $;
        run;
    %end;
%mend;
```

The ICON tool now outputs a SAS dataset. An example is shown in Screen 1.2.3.

Screen 1.2.3

TABLE: Work.Specs_terms

Domain	Variable	Value	Label
EG	EGTESTCD	INTP	ECG Test or Examination Short Name
EG	EGTESTCD	PR	ECG Test or Examination Short Name
EG	EGTEST	ECG Interpretation	ECG Test or Examination Name
EG	EGTEST	PR Interval	ECG Test or Examination Name
EG	EGORRES	NORMAL	Result or Finding in Original Units
EG	EGORRES	ABNORMAL	Result or Finding in Original Units
SUPPEG	QNAM	EGCLSIG	Qualifier Variable Name
SUPPEG	QLABEL	Clinically Significant	Qualifier Variable Label

Two other output datasets are also created (shown in Screen 1.2.4 and 1.2.5).

The first in Screen 1.2.4, is the controlled terminology terms extracted from the SDTM Specifications.
The column Label is the SDTM variable label.

Screen 1.2.4

TABLE: Work.Specs_terms

Domain	Variable	Value	Label
EG	EGTESTCD	INTP	ECG Test or Examination Short Name
EG	EGTESTCD	PR	ECG Test or Examination Short Name
EG	EGTEST	ECG Interpretation	ECG Test or Examination Name
EG	EGTEST	PR Interval	ECG Test or Examination Name
EG	EGORRES	NORMAL	Result or Finding in Original Units
EG	EGORRES	ABNORMAL	Result or Finding in Original Units
SUPPEG	QNAM	EGCLSIG	Qualifier Variable Name
SUPPEG	QLABEL	Clinically Significant	Qualifier Variable Label

The second in Screen 1.2.5, is the value level metadata extracted from the SDTM Specifications.
The column Variable is the variable name from the Specification, i.e., EGTESTCD and QNAM.
The column Value stores the values of the two target variables, i.e., INTP and PR, and EGCLSIG, respectively.
The column Label is the corresponding test name from EGTEST, and the qualifier variable label from QLABEL i.e. where EGTESTCD = 'INTP' then EGTEST = 'ECG Interpretation'.

Screen 1.2.5

TABLE: Work.Specs_terms

Domain	Variable	Value	Label
EG	EGTESTCD	INTP	ECG Interpretation
EG	EGTESTCD	PR	PR Interval
SUPPEG	QNAM	EGCLSIG	Clinically Significant

1.3. SDTM/CRT SAS DATASETS

Proc Freq and Merge statements in SAS are used to process the data from SDTM SAS datasets.

Screen 1.3.1 is sample SDTM SAS dataset for the EG domain.

Screen 1.3.1

TABLE: ECG Test Results

STUDYID	DOMAIN	USUBJID	EGSEQ	EGTESTCD	EGTEST	EGORRES	EGSTAT
STUDY1	EG	STUDYID-10011001	1	INTP	ECG Interpretation	ABNORMAL	
STUDY1	EG	STUDYID-10011001	2	PR	PR Interval	120	
STUDY1	EG	STUDYID-10011001	3	INTP	ECG Interpretation	NORMAL	
STUDY1	EG	STUDYID-10011001	4	PR	PR Interval	142	
STUDY1	EG	STUDYID-10011001	5	EGALL	ECG Data		NOT DONE
STUDY1	EG	STUDYID-10011002	1	INTP	ECG Interpretation	ABNORMAL	

Screen 1.3.2 is the supplemental qualifier SUPPEG dataset.

Screen 1.3.2

TABLE: Supplemental EG

STUDYI	USUBJID	RDOMAIN	IDVAR	IDVARVAL	QNAM	QLABEL	QVAL	QORIG	QEVAL
STUDY1	STUDYID-10011001	EG	EGSEQ	1	EGCLSIG	Clinically Significant	N	CRF	INVESTIGATOR
STUDY1	STUDYID-10011002	EG	EGSEQ	1	EGCLSIG	Clinically Significant	N	CRF	INVESTIGATOR

Screen 1.3.3 is sample SAS code, used to process the SDTM SAS datasets. Proc Freq combined with data step in a Macro loop is run for each variable in the SDTM SAS dataset.

Screen 1.3.3

```
%macro loopsdtm;
    ** Set macro variable &maxdsn: total # of SDTM data sets;
    ** Set macro variable &&dsname&i: each SDTM data set name;
    ** Loop through all SDTM data sets;
    %do i=1 %to &maxdsn;
        ** Set macro variable &maxvarn:
           total # of variables in this SDTM data set;
        ** Set macro variable &&varname&j:
           each variable name in this SDTM data set;
        ** Loop through all variables in this SDTM data set;
        %do j=1 %to &maxvarn;
            ** Frequency of this variable;
            proc freq data=sdtmsas.&&dsname&i noprint;
                tables &&varname&j/out =freq_&&dsname&i._&&varname&j;
            run;
            ** Update variable attributes;
            data freq_&&dsname&i._&&varname&j
                                (keep=domain variable value);
                length domain $10 varname $200 value $200;
                set freq_&&dsname&i._&&varname&j;
                domain = "&&dsname&i";
                variable = "&&varname&j";
                value = &&varname&j;
                ** Adjust Value if SDTM variable type is Num;
            run;
            ** Combined frequency from all variables;
            data freqall;
                set freqall freq_&&dsname&i._&&varname&j ;
            run;
        %end;
    %end;
%mend;
```

The output dataset from the sample code is shown in Screen 1.3.4.
The column Value displays all the variable values in the SDTM SAS datasets.
The column Label is the variable labels in the SDTM SAS datasets.

Screen 1.3.4

TABLE: Work.Sdtm_single

Domain	Variable	Value	Label
EG	STUDYID	STUDYID1	Study Identifier
EG	DOMAIN	EG	Domain Abbreviation
EG	USUBJID	STUDY1-10011001	Unique Subject Identifier
EG	USUBJID	STUDY1-10011002	Unique Subject Identifier
EG	EGSEQ	1	Sequence Number
EG	EGSEQ	2	Sequence Number
EG	EGSEQ	3	Sequence Number
EG	EGSEQ	4	Sequence Number
EG	EGSEQ	5	Sequence Number
EG	EGTESTCD	EGALL	ECG Test or Examination Short Name
EG	EGTESTCD	INTP	ECG Test or Examination Short Name
EG	EGTESTCD	PR	ECG Test or Examination Short Name
EG	EGTEST	ECG Data	ECG Test or Examination Name
EG	EGTEST	ECG Interpretation	ECG Test or Examination Name
EG	EGTEST	PR Interval	ECG Test or Examination Name
EG	EGORRES	120	Result or Finding in Original Units
EG	EGORRES	142	Result or Finding in Original Units
EG	EGORRES	ABNORMAL	Result or Finding in Original Units
EG	EGORRES	NORMAL	Result or Finding in Original Units
EG	EGSTAT	NOT DONE	Completion Status
EG	VISITNUM	0	Visit Number
EG	VISITNUM	1	Visit Number
EG	VISITNUM	2	Visit Number
EG	VISIT	Screening	Visit Name
EG	VISIT	Visit 1	Visit Name
EG	VISIT	Visit 2	Visit Name

The output datasets from Proc Freq for paired variables in the EG and SUPPEG datasets are shown in Screen 1.3.5.
The column Variable is the target variable name, i.e., EGTESTCD or QNAM, in the EG and SUPPEG datasets.
The column Value is the values on the two target variables.
The column Label is the corresponding values from EGTEST and QLABEL, respectively.

Screen 1.3.5

TABLE: Work.Sdtm_paired

Domain	Variable	Value	Label
EG	EGTESTCD	EGALL	ECG Data
EG	EGTESTCD	INTP	ECG Interpretation
EG	EGTESTCD	PR	PR Interval
SUPPEG	QNAM	EGCLSIG	Clinically Significant

2. IMPLEMENTATION

Once the setup of the tool has been completed, we now have all metadata required for the comparison reports. The metadata, as described in Section 1, is the input to the reporting tool.

The reports are created using SAS ODS. Screen 3.1 shows sample code.

Screen 3.1

```
ods tagsets.ExcelXP path="&REPPDIR" file='Report1.xml'
style=XLsansPrinter;
    ods tagsets.ExcelXP options(embedded_titles='yes'
                                embedded_footnotes='yes'
                                sheet_name='Summary'
                                absolute_column_width='9');
    title Report1;
    footnote;
    proc print data=unique_domains noobs label split='*';
    run;
ods tagsets.ExcelXP close;
```

Using ODS the output is XML file type. XML file can be open with Excel and saved as a new Excel file.

The Reports are the output from the data warehouse tool. The reports can now summarize all information in a user friendly format for review.

3. OUTPUT

There are 7 main reports created.

Report 1: Summary of SDTM and CRF Variables by Study.
Report 2: Distribution of SDTM Variables
Report 3: SDTM Variable Values
Report 4: SDTM Paired Variable Values
Report 5: SDTM Extended Paired Variable Values
Report 6: SDTM Nested Paired Variable Values
Report 7: CRF Page Number Comparisons

3.1 REPORT 1 - SUMMARY OF SDTM AND CRF VARIABLES BY STUDY

Report 1 serves as a summary or index for all domains for each study. The structure of this report is one row per SDTM domain per CRF per study. Screen 3.1.1 below is an example report.

Primary Purpose:
We can review this report to identify potential inconsistencies across studies on use of domain name. This is particularly important for custom domains.

We can see from the example in Screen 3.1.1 that the AE and SUPPAE domain has been used for all 3 studies. However, we can also identify that there is a conflict between the use of the custom ZO domain when comparing studies 1 and 2 (Efficacy Examination), with Study 3 (Safety Examination).

Screen 3.1.1

SDTM Domains in aCRF and Specs
Studies : Study 1, Study 2, Study 3

SDTM_Domain	SDTM_SUPPxx	aCRF_Page_Header	aCRF_Page	Study Number	Protocol_Number	acrf_link	spec_link
AE	SUPPAE	Adverse Events	89	Study 1	Study 1	aCRF #89	AE
AE	SUPPAE	Adverse Events	55	Study 2	Study 2	aCRF #55	AE
AE	SUPPAE	Adverse Events	49	Study 3	Study 3	aCRF #49	AE
ZO	SUPPZO	Efficacy Examination	28	Study 1	Study 1	aCRF #28	ZO
ZO	SUPPZO	Efficacy Examination	29	Study 2	Study 2	aCRF #29	ZO
ZO	SUPPZO	Safety Examination	30	Study 3	Study 3	aCRF #30	ZO

Secondary Purpose:

The last two columns on the report are useful reviewing tools, as they serve as hyperlinks to the input documentation. These hyperlinks are useful when reviewing the other 6 reports.

The aCRF Link and Specs Link, are embedded with the Excel function HYPERLINK in each cell.

For example, the function for aCRF page 89 is =HYPERLINK("D0008041-aCRF.pdf#89", "aCRF #89"), where the first argument in the function is the aCRF file name plus the page number, and the second argument is the label for the hypertext link.

Similarly, the function for the AE and ZO tab in the SDTM Specifications file is =HYPERLINK("[8888041-CRT-Specification.xls]AE!A1", "AE"), where the first argument is the SDTM Specification Excel file name plus the tab name AE (or ZO), and the second argument is the text for the hypertext link.

3.2 REPORT 2 - DISTRIBUTION OF SDTM VARIABLES

Report 2 serves as a summary of the domain and variable distribution between the 3 key sources as shown in Screen 3.2.1 – the Annotated CRF (shown in the aCRF column), the SDTM Specifications (shown in the Specs column) and the SDTM Dataset (shown in the CRT column).

This report can be generated for an individual study, or for multiple studies as shown in Screen 3.2.1.

Primary Purpose:
Within an individual study, we can review this report to identify inconsistencies across the 3 key sources. It is essential that all dataset variables are defined within the SDTM Specifications, as this forms the basis of the define.xml file. It is also essential that all annotations from the CRF, are defined within the SDTM Specifications, and also present in the SDTM dataset.

For example, in the report displayed in Screen 3.2.1, we can see that AEBODSYS is not annotated for Study 1, 2 or 3. The reviewer can confirm the report and determine if this is expected. In this case we would deem this as acceptable, since this coding variable annotation is not required. However, we can also see that there is an issue with Study 2, since the CRF has been annotated with AEHOS instead of the correct AEHOSP variable. The reviewer can then correct the annotated CRF.

Screen 3.2.1

Variables and variable values in aCRF cross-checking Specs and CRT
Studies : Study 1, Study 2, Study 3

Domain	Variable	_Study 1_aCRF_	_Study 1_spec_	_Study 1_crt_	_Study 2_aCRF_	_Study 2_spec_	_Study 2_crt_	_Study 3_aCRF_	_Study 3_spec_	_Study 3_crt_
AE	AEBODSYS		Y	Y		Y	Y		Y	Y
AE	AESHOS				Y					
AE	AESHOSP	Y	Y	Y		Y	Y	Y	Y	Y

Secondary Purpose 1:
Another powerful use of this report is consistency in naming conventions within the supplemental domains across studies.

For example, in the report displayed in Screen 3.2.2, we can see that Study 1 and 2 have followed the same naming conventions for the MedDRA coding of the Adverse Event i.e. AEHLT and AEHLCD - High Level Term and High Level Code, but Study 3 has used a different naming convention i.e. AEHLTERM and AEHLCODE. In this case the reviewer can confirm the report and determine if this is expected or correct the SDTM Specification and SDTM Datasets for Study 3 to ensure consistency with the other 2 studies.

Screen 3.2.2

Variables and variable values in aCRF cross-checking Specs and CRT
Studies : Study 1, Study 2, Study 3

Domain	Variable	_Study 1_aCRF_	_Study 1_spec_	_Study 1_crt_	_Study 2_aCRF_	_Study 2_spec_	_Study 2_crt_	_Study 3_aCRF_	_Study 3_spec_	_Study 3_crt_
SUPPAE	AEHLT		Y	Y		Y	Y			
SUPPAE	AEHLTERM								Y	Y
SUPPAE	AEHLCD		Y	Y		Y	Y			
SUPPAE	AEHLCODE								Y	Y
SUPPAE	AELLT		Y	Y		Y	Y			
SUPPAE	AELLTERM								Y	Y
SUPPAE	AELLCD		Y	Y		Y	Y			
SUPPAE	AELLCODE								Y	Y

3.3 REPORT 3: SDTM VARIABLE VALUES

Report 3 serves as a summary of the domain and variable values between the SDTM Specifications (shown in the Specs column) and the SDTM Dataset (shown in the CRT column).

Primary Purpose:
Report 3 has the functionality to display the Variable value. This is useful to identify if the same variable has differing values across studies.

For example, in the report displayed in Screen 3.3.1, we can see that for Study1 and 3, EGORRES has the value 'ABNORMAL, NOT CLINICALLY SIGNIFICANT', whereas in Study 2 EGORRES has the value 'ABNORMAL, NON-CLINICALLY SIGNIFICANT'.

Screen 3.3.1

Variables and variable values in aCRF cross-checking Specs and CRT
Studies : Study 1, Study 2, Study 3

Domain	Variable	Value	_Study 1_crt_	_Study 1_spec_	_Study 2_crt_	_Study 2_spec_	_Study 3_crt_	_Study 3_spec_
EG	EGORRES	ABNORMAL, NOT CLINICALLY SIGNIFICANT	Y	Y			Y	Y
EG	EGORRES	ABNORMAL, NON-CLINICALLY SIGNIFICANT			Y	Y		

3.4 REPORT 4: SDTM PAIRED VARIABLE VALUES

Report 4 serves as a summary of the domain label versus the Controlled Terminology List, Screen 3.4.1 shows the domain variable name, the variable value (in the Value column) and the value of its paired variable in the Label column).

Primary Purpose:
This report is most useful looking when looking for inconsistencies across multiple studies. The report lists the variable value and also the value of its paired variable.

The example in Screen 3.4.1, shows that for the EG domain variable EGTESTCD, Study 1 and 3 have a 1-1 relationship between the EGTESTCD for 'MRHYABN' and the EGTEST value 'Morphological / Rhythm Abnormality', whereas Study 2 has deviated from this and used the EGTEST value of 'Morphological / Rhythm Abnormalities'.

Screen 3.4.1

Paired variable values in CRT, cross-checking Specs
Studies : Study 1, Study 2, Study 3

Domain	Variable	Value	Label	_Study1_ crt_	_Study 1_spec_	_Study 2_crt_	_Study 2_spec	_Study 3_crt_	_Study 3_spec
EG	EGTESTCD	EGSA	Sinus Arrhythmia	Y	Y	Y	Y	Y	Y
EG	EGTESTCD	EGSBR	Sinus Bradycardia	Y	Y	Y	Y	Y	Y
EG	EGTESTCD	EGSPB	Supraventricular Premature Beats					Y	Y
EG	EGTESTCD	EGSRNAB	Sinus Rhythm Assessment			Y	Y		
EG	EGTESTCD	EGSTNAB	ST Segment Assessment	Y	Y				
EG	EGTESTCD	EGTACH	Sinus Tachycardia	Y	Y	Y	Y	Y	Y
EG	EGTESTCD	EGTWF	T Wave Findings	Y	Y				
EG	EGTESTCD	EGTWNAB	T Wave Assessment			Y	Y	Y	Y
EG	EGTESTCD	EGUWN	U Wave Present	Y	Y				
EG	EGTESTCD	EGUWNAB	U Wave Assessment			Y	Y	Y	Y
EG	EGTESTCD	EGVPB	Ventricular Premature Beats	Y	Y	Y	Y	Y	Y
EG	EGTESTCD	HR	Heart Rate	Y	Y	Y	Y	Y	Y
EG	EGTESTCD	INTP	ECG Interpretation	Y	Y	Y	Y	Y	Y
EG	EGTESTCD	MRHYABN	Morphological / Rhythm Abnormalities			Y	Y		
EG	EGTESTCD	MRHYABN	Morphological / Rhythm Abnormality	Y	Y			Y	Y
EG	EGTESTCD	PR	PR Interval	Y	Y	Y	Y	Y	Y
EG	EGTESTCD	PRMEAN	Mean PR Interval	Y	Y	Y	Y	Y	Y
EG	EGTESTCD	QRS	QRS Interval	Y	Y	Y	Y	Y	Y

3.5 REPORT 5: SDTM EXTENDED PAIRED VARIABLE VALUES

Report 5 serves a similar purpose to Report 4, however it has the capability to extend the paired variables within a domain, and compare these across studies.

Primary Purpose:
This report is most useful when looking for inconsistencies across multiple studies. The report lists the variable value and also the value of its paired variable(s).

The example in Screen 3.5.1, shows that for the LB domain variable LBTESTCD, Study 1 and 3 have a 1-1 relationship between the LBTESTCD for 'ALB' and the LBTEST value 'Albumin'.

Screen 3.5.1

Extension 1 paired variable values in CRT, cross-checking Specs
Studies : Study 1, Study 2, Study 3

Domain	Variable	Value	Label	STRESU/QEVAL	CAT/QORIG	_Study1_crt_	_Study 1_spec_	_Study 2_crt_	_Study 2_spec_	_Study 3_crt_	_Study 3_spec_
LB	LBTESTCD	ALB	Albumin	g/L	CHEMISTRY	Y	Y			Y	Y
LB	LBTESTCD	ALB	Albumin	g/dL	CHEMISTRY			Y	Y		
LB	LBTESTCD	ALP	Phosphatase	IU/L	CHEMISTRY	Y	Y	Y	Y	Y	Y
LB	LBTESTCD	ALT	Aminotransferase	IU/L	CHEMISTRY	Y	Y	Y	Y	Y	Y
LB	LBTESTCD	AST	Aminotransferase	IU/L	CHEMISTRY		Y	Y	Y	Y	Y
LB	LBTESTCD	AST	Aminotransferase	U/L	CHEMISTRY	Y					
LB	LBTESTCD	BASO	Basophils	10^9/L	HEMATOLOGY				Y		
LB	LBTESTCD	BASO	Basophils	x10^9/L	HEMATOLOGY	Y	Y	Y		Y	Y

The extended paired variable that is also looked at in this report, is the LBSTRESU variable, where we can see that Studies 1 and 3 have listed the LBSTRESU standard unit for Albumin as 'g/L', whereas Study 2 has deviated from this and used the LBSTRESU value of 'g/dL'.

Another issue we see in the report, is that the standard unit for LBTESTCD = 'AST' is the same for all 3 studies within the SDTM dataset, but we have an issue with Study 1 SDTM specification. We see the same issue for Study 2 for Basophils. We can also confirm from this report that for the Lab tests listed, all LBCAT values are consistent across all 3 studies.

Secondary Purpose:
This tool also has the capability to be extended beyond the above paired variables, to also include additional paired variables.

The example in Screen 3.5.2, is a similar report to 3.5.1, but has the additional LBORRES variable listed.

We can see that Study 2 has an additional LBORRESU (mmol/L) that has not been converted in programming, and hence has a missing LBSTRESU.

Screen 3.5.2

Extension 1 paired variable values in CRT, cross-checking Specs
Studies : Study 1, Study 2, Study 3

Domain	Variable	Value	Label	ORRESU	STRESU/QEVAL	CAT/QORIG	_Study1_crt_	_Study 1_spec_	_Study 2_crt_	_Study 2_spec_	_Study 3_crt_	_Study 3_spec_
LB	LBTESTCD	ALB	Albumin	g/L	g/dL	CHEMISTRY	Y	Y			Y	Y
LB	LBTESTCD	ALB	Albumin	g/dL	g/dL	CHEMISTRY	Y	Y			Y	Y
LB	LBTESTCD	ALB	Albumin	mmol/L		CHEMISTRY			Y	Y		

Issues such as this in 3.5.2, can arise when programs or macros are being utilized across studies, where an unexpected value has appeared for one study.

3.6 REPORT 6: SDTM NESTED PAIRED VARIABLE VALUES

Report 6 serves a similar purpose to Report 4 and 5, but instead of looking at paired values, Report 6 looks at nested values.

Primary Purpose:
This report is most useful when looking for inconsistencies across multiple studies. The report lists the variable value and also the value of its nested variable(s).

The example in Screen 3.6.1, shows that for the LB domain variable LBTESTCD, Studies 1, 2 and 3 have a 1-1 relationship between the LBTESTCD for 'AMPHET' and the LBTEST value 'Amphetamine'. The first nested value is also the same for all 3 studies, where the Lab Category LBCAT is 'ABUSE'. However, we see that for Study 3 the second nested value for the Specimen Type LBSPEC is different for Study 3, and has the value 'BLOOD' instead of 'URINE'. This may be expected as the specimen type may be study specific for this group of studies, or it may be an issue with the transformation programming.

Screen 3.6.1

Nested paired variable values in CRT, cross-checking Specs
Studies : Study 1, Study 2, Study 3

Domain	Variable	Value	Label	NestVar1	NestVar2	_Study1_crt_	_Study 1_spec_	_Study 2_crt_	_Study 2_spec_	_Study 3_crt_	_Study 3_spec_
LB	LBTESTCD	AMPHET	Amphetamine	ABUSE	BLOOD					Y	Y
LB	LBTESTCD	AMPHET	Amphetamine	ABUSE	URINE	Y	Y	Y	Y		
LB	LBTESTCD	BILI	Bilirubin	URINALYSIS	URINE					Y	Y
LB	LBTESTCD	BILI	Bilirubin	CHEMISTRY	BLOOD	Y	Y	Y	Y	Y	Y
LB	LBTESTCD	HBSAB	Hepatitis B Virus Surface Antibody	VIROLOGY						Y	Y
LB	LBTESTCD	HBSAB	Hepatitis B Virus Surface Antibody	VIROLOGY	BLOOD	Y	Y	Y	Y		
LB	LBTESTCD	HBSAG	Hepatitis B Virus Surface Antigen	VIROLOGY						Y	Y
LB	LBTESTCD	HBSAG	Hepatitis B Virus Surface Antigen	VIROLOGY	BLOOD	Y	Y	Y	Y		
LB	LBTESTCD	HCG	Choriogonadotropin Beta	CHEMISTRY	BLOOD					Y	Y
LB	LBTESTCD	HCG	Choriogonadotropin Beta	PREGNANCY	BLOOD	Y	Y	Y	Y		

In addition, when we look at Screen 3.6.1, the report shows that for the LB domain variable LBTESTCD, Studies 1, 2 and 3 have a 1-1 relationship between the LBTESTCD for 'BILI' and the LBTEST value 'Bilirubin'. Here we see that both nested values (LBCAT and LBSPEC) are different for study 3. Again, this may be expected.

More interestingly, the report also shows that for the LB domain variable LBTESTCD Study 1, 2 and 3 have a 1-1 relationship between the LBTESTCD for 'HBSAG' and the LBTEST value 'Hepatitis B Virus Surface Antigen'.

However, we can see that the second nested variable (LBSPEC) has not been populated for Study 3. This may be since the specimen type is unknown for Study 3, or it may again be an issue with the transformation programming.

3.7 REPORT 7: CRF PAGE NUMBER COMPARISONS

Report 7 is one of the most useful reports, which eliminates the need for manually checking the CRF page references between the annotated CRF and SDTM Specifications.

Primary Purpose:
This simple, but effective report compares the CRF annotation page number for each domain variable, with the CRF page number (or Origin) as described in the SDTM Specifications.

The report will identify where the Origins are different. An example of this is shown in Screen 3.7.1 below. This report would be run on an individual study basis.

Screen 3.7.1

Comparison of Variable Origin
Study : Study 1

Domain	Variable	aCRF Origin	Spec Origin	Different Origin
LB	LBCAT	CRF Pages 12, 13, 14, 15, 16	CRF Pages 13, 14, 15, 16, 17	Y
LB	LBORRES	CRF Pages 12, 13, 14, 15, 16	CRF Pages 13, 14, 15, 16, 17	Y
LB	LBTESTCD	CRF Pages 12, 13, 14, 15, 16	CRF Pages 13, 14, 15, 16, 17	Y
QSCG	QSMETHOD	CRF Pages 20, 21, 43, 61, 79	CRF Pages 20, 43, 61, 79	Y
SUPPDM	RACEOTH	CRF Page 8	Assigned	Y
SUPPPE	PEOTHER		CRF Page 10	Y

In the above example shown in Screen 3.7.1, the report has identified a number of issues.

For example, in the LB domain, the LBCAT variable has been annotated on the CRF pages 12, 13, 14, 15, and 16. However in the SDTM specification it gives the Origin as CRF pages 13, 14, 15, 16 and 17 instead.

Another example we see is that the Demographic supplemental dataset SUPPDM, has been annotated on the CRF Page 8, but the Origin in the SDTM specification has indicated that this variable is 'Assigned'.

A final example we see, is that the Physical Exam supplemental dataset SUPPPE, has not been annotated on the CRF, but the Origin in the SDTM specification has indicated that this variable should have been annotated on the CRF Page 10.

4. FURTHER ENHANCEMENTS

ICON has found the Data Warehouse tool very useful and efficient for highlighting the inconsistencies described in this paper. However, we do recognize that further development and enhancements could be implemented.

For example, the tool currently reports all observations from the CRF, SDTM specifications and SDTM datasets. ICON is looking to enhance the reports with flags or auto filters to aid the reviewer to subset the report only to records where inconsistency issues appear.

ICON also has a separate tool that produces OpenCDISC style reports with regards to variable attribute harmonization/inconsistencies across studies. This tool currently reports inconsistencies in variable attributes such as variable label, type, length and format. This report could be added to the Data Warehouse tool.

CONCLUSION

ICON has developed an effective tool that greatly facilitates the cross checking of the 3 key submission sources that make up the define.xml – the annotated CRF, the SDTM specifications and the SDTM datasets, thus enhancing and speeding up the harmonisation process and reducing incidences of human error in what has been a very manual task to date.

In addition, when consistency is critical, such as when an integrated or pooled analysis is required, Reports 4, 5 and 6 prove to give an effective way of catching major consistency issues.

It may be that some studies are just always going to be unique or individual .and ICON is OK with that too!

ACKNOWLEDGMENTS

The authors would like to extend a special thanks to Annie Guo, who is the sole developer of the tool described in this paper. Annie continues to maintain and add to the tools and reports, and is considered an ICON leader in creativity and innovation.

The authors would also like to extend a special thanks to Shaun Maraj and Robert Stemplinger for review, input and support.

REFERENCES

http://pharmasug.org/
Title: Ensuring Consistent Data Mapping Across SDTM-based Studies - a Data Warehouse Approach
Author: Annie Guo

CONTACT INFORMATION

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

Author Name: Jennie Mc Guirk
Company: ICON PLC
Email: jennie.mcguirk@iconplc.com
Web: http://www.iconplc.com/

Author Name: Steven Thacker
Company: ICON PLC
Email: Steven.Thacker@iconplc.com
Web: http://www.iconplc.com/


ICON
A Symbol of Excellence


Clinical Research