PharmaSUG2011 - Paper CD21

Ensuring Consistent Data Mapping Across SDTM-based Studies – a Data Warehouse Approach

Annie Guo, ICON Clinical Research, North Wales, PA

ABSTRACT

SDTM is about standardization of clinical trials data. This paper presents a tool that helps ensure consistent data mapping across SDTM-based studies. The tool is comprised of a series of SAS® programs. The input to the SAS programs consists of three sources: annotated CRF, SDTM data set specifications, and SDTM SAS data sets. The SAS programs run across each study, and summarize the information from the input files. The output is a set of standardized SAS data sets per study that serve as a data warehouse storing the metadata and data contained in the SDTM data sets. This data warehouse approach allows for direct access and comparison among existing studies, bypassing the original sources, as well as providing a reference database useful for facilitating the programming of new studies.

INTRODUCTION

SDTM is about standardizing and normalizing clinical trial data. Therefore, consistent mapping from CRF raw data to SDTM is important especially for projects with multiple studies. However, discrepancies may arise as the number of studies increases. New studies may be assigned to new programmers who sometimes refer to only one or two previous studies that are thought to be most similar to the new ones, and miss out information in other existing studies. On the other hand, it is not realistic to expect anyone to look up the documentation and SDTM SAS data sets in all previous studies, which can be time consuming, in order to cover everything. Alternatively we may designate a lead programmer to oversee and ensure the consistency, but it is only efficient that everyone is on top of his/her own assignment and does not rely on another person to check the work. At the end of the day what we need is a tool that centralizes the metadata and data in our SDTM-based studies. Like a data warehouse, it stores our experiences with SDTM, and serves as a one stop source to look up anything we may need when developing new SDTM-based studies.

REQUIREMENTS

Requirements focus on two areas: data warehouse, and reporting of the data in the data warehouse.

Data Warehouse

On the data warehouse side, the three sources of SDTM metadata and data, i.e., annotated CRF (aCRF) in PDF, SDTM data set Specifications (Specs) in Excel and SDTM SAS data sets, must be all integrated into a set of SAS data sets across each study. The SAS data sets follow a uniform structure to store SDTM domain names, variable names, and variables values from the three sources. The uniform structure allows for listing or harmonization comparison across studies.

Report 1

List of CRFs and associated SDTM domain(s), with hypertext links to aCRF and Specifications

This is a high level overview of the association between CRF and SDTM. In many cases the name of the CRF determines the SDTM domain it is mapped to. For example, Concomitant Medications CRF goes to the published Concomitant Medications - CM domain. However, confusion may arise when it comes to custom domains. For example, Human Anti-Human Antibody Samples CRF, does it go to a custom domain, and if so, have we had one for that CRF? A list like Table 1 would provide clarity and answers to those questions.

The hypertext links to aCRF and Specs in Table 1 provides direct access to the SDTM document. It saves us time and we do not need to navigate through the folder structure on server in order to locate the file and then open the specific PDF page or Excel tab.

Table 1: Report requirement 1

Domain	SUPP--	CRF	Study #	aCRF Page # Link	Specs Link
Bi		Archived Tumor	Study1	aCRF #12	BI
BI		Human Anti-Human Antibody Samples	study1	aCRF #13	BI
BI		Human Anti-Human Antibody Samples	Study2	aCRF #21	BI
BI		Serum Ca125 Levels	Study2	aCRF #16	Bi
CM		Concomitant Medications	Study1	aCRF #39	CM
CM		Concomitant Procedures	Study1	aCRF #41	CM
CM		Concomitant Procedures	Study2	aCRF #52	CM
CM	SUPPCM	Prior And Concomitant Medications	Study2	aCRF #39	CM

Report 2

List of distribution of SDTM variables on CRF annotations, Specifications and SDTM data sets, including variable values if they are annotated on CRF

This structure of this list is one row per SDTM variable per variable value annotated on aCRF, such as the sample in Table 2. The focus is on the CRF annotations. The reason is, in general not all SDTM variables are annotated on aCRF. However, those variables or variable values annotated on aCRF must appear in the Specifications for the study. In addition, the Specifications and the SDTM data sets must have exactly the same SDTM variables within a study. This list would point out any deviation from those rules.

Table 2: Report requirement 2

			Study 1			Study 2		
Domain	Variable	Value (if present on aCRF)	aCRF	Specs	Dataset	aCRF	Specs	Dataset
AE	AEREL		√	√	√	√	√	√
CM	CMDOSE			√	√	√	√	√
CM	CMDOSTXT		√	√	√		√	√
EG	EGORRES	ABNORMAL	√	√	√			
EG	EGORRES	NORMAL	√	√	√	√	√	
EG	EGORRES	ABNORMAL CLINICAL SIGNIFICANCE				√	√	√
EG	EGORRES	ABNORMAL NOT CLINICAL SIGNIFICANCE				√	√	√

This list summarizes CRF annotations by SDTM domain and variable, and gives us an idea about the data collected on CRF, without having to open and look at the aCRF files. For example, in Table 2, both studies collect adverse event casualty, AEREL. It appears that there is no annotation for AEREL variable values, so most likely the values are according to pre-printed text on CRF. For CM domain, Study 2 has CMDOSE annotated on CRF, so this is probably a numeric data field to collect medication dose. Study 1 has CMDOSTXT annotated instead, so we can guess the data field on CRF collects character data for not only medication dose but also medication unit or other information.

This list also helps identify inconsistency among studies. For example, EGORRES in Table 2, the variable values for abnormal ECG test results are different according to the CRF annotations. Without actually opening the aCRF to verify, we may guess Study 1 collects ECG test result as either Abnormal or Normal, but Study 2 also asks if an abnormal test result is clinically significant. Another possible cause for the difference would be inconsistent mapping between the two studies. In other words, both studies collect Clinical Significance, but Study 1 has that piece of information mapped to SUPPEG, and EGORRES is set to ABNORMAL regardless of Clinical Significance.

Report 3

List of all variables values in SDTM SAS data sets, cross referencing controlled terminology terms in Specifications

The structure of the list is one row per combination of SDTM variable name, variable value, and variable

label from SDTM SAS data sets and/or Specifications across studies. For those variables subject to controlled terminology, they are cross-checked to show if the variable labels and values are consistent with the Specifications.

Table 3: Report requirement 3

Domain	Variable	Value	Label	Study 1 Dataset	Study 1 Specs	Study 2 Dataset	Study 2 Specs	Study 3 Dataset	Study 3 Spec
AE	AEREL	MULTIPLE	Causality			√	√		√
AE	AEREL	NOT RELATED	Causality		√			√	
AE	AEREL	POSSIBLY RELATED	Causality		√				
AE	AEREL	RELATED	Causality	√	√			√	

The purpose of this list is to show variable value mapping across studies. For example, in Table 3, variable AEREL, there are the values MULTIPLE, NOT RELATED, POSSIBLY RELATED and RELATED from the three studies. Breakdown by study, Study 1 has all the values but MULTIPLE according data set Specifications, and only RELATED actually collected on CRF and stored in the AE data set. For Study 2, the only value on AEREL is MULTIPLE. Though it looks like a little different between Study 1 and Study 2, it could be because the two studies following different versions of SDTM IG. So overall there seems to be no discrepancy.

This list can also identify possible data issues in the SDTM SAS data sets. For example, in Table 3, Study 3 seems to have an error. The values on AEREL in AE data set are NOT RELATED and RELATED, but the Specifications file has MULTIPLE as the only controlled terminology term for AEREL.

Report 4

List of paired variable values for --TESTCD and --TEST variables, and paired QNAM and QLABEL in SDTM SAS data sets, cross referencing value level metadata in Specifications

Test codes and test names in Findings domains, and QNAM and QLABEL in supplemental qualifiers are one-to-one relation. They may be used consistently across studies unless there are study or sponsor specific requirements.

The structure of this list is one row per paired variable values from SDTM SAS data sets and/or the value level metadata in Specifications. Since it displays all the possible combinations, it is straightforward for anyone to look up what we have had on --TESTCD / --TEST and QNAM / QLABEL, and to make a good judgment if sticking to the existing convention or creating new ones.

Table 4: Report requirement 4

Domain	Variable	--TESTCD / QNAM	--TEST / QLABEL	Study 1 Dataset	Study 1 Specs	Study 2 Dataset	Study 2 Specs	Study 3 Dataset	Study 3 Specs
EG	EGTESTCD	INTP	ECG Interpretation	√	√	√	√	√	
EG	EGTESTCD	INTRP	ECG Interpretation	√	√	√	√		√
EG	EGTESTCD	PR	PR Interval	√	√	√	√	√	√
SUPPEG	QNAM	EGCLSIG	ECG Clinical Significance	√	√	√	√		
SUPPEG	QNAM	EGCLSP	ECG Clinical Significance Specify	** Absence **	√	√	√		
SUPPEG	QNAM	EGCS	ECG Clinical Significance					√	√

This list is also a tool for identifying differences across studies or validating SDTM SAS data sets against Specifications within a study. For example, the red text in Table 4, Study 3 has used the value EGCS for the variable EGTESTCD, as opposed to the value EGCLSIG used by the other two studies. Another problem in Study 3 is it has the value INTP stored in the SDTM SAS data set, but the Specification file has INTRP as the controlled terminology term for the variable EGTESTCD.

For supplemental qualifier, if a combination of QNAM and QLABEL is specified in Specifications but missing from SDTM SAS data sets, the list displays ** Absence **, for example, in Table 4 the paired values EGCLSP and ECG Clinical Significance Specify. The reason is, not all QNAM and QLABEL values defined in Specifications are required to appear in SDTM SAS data sets. If the CRF data field for that QNAM is completely blank in the raw data, the QNAM / QLABEL is not included in the SUPP-- data set.

Report 5

Extended from Report 4, list of paired variable values on --TESTCD / --TEST plus --CAT / --STRESU, and paired QNAM / QLABEL plus --QORIG / --QEVEL in SDTM SAS data sets, cross referencing value level metadata in Specifications

Table 5: Report requirement 5

Domain	Variable	--TESTCD / QNAM	--TEST / QLABEL	--CAT / QORIG	--STRESU / QEVAL	Study 1 Dataset	Study 1 Specs	Study 2 Dataset	Study 2 Specs	Study 3 Dataset	Study 3 Specs
LB	LBTESTCD	ALB	Albumin	CHEMISTRY	g/L	√	√	√	√		√
LB	LBTESTCD	BASO	Basophils	HEMATOLOGY	x10^3/uL	√	√	√	√	√	√
LB	LBTESTCD	BUN	Blood Urea Nitrogen	CHEMISTRY	mmol/L	√	√			√	√
LB	LBTESTCD	BUN	Blood Urea Nitrogen	CHEMISTRY	mg/dL					√	
SUPPEG	QNAM	EGCLSIG	ECG Clinical Significance	CRF	INVESTIGATOR	√	√	√	√		
SUPPEG	QNAM	EGCLSIG	ECG Clinical Significance	CRF	SPONSOR					√	√

This list is to drill down Report 4, with the addition of the category variable --CAT and the standard unit variable --STRESU for SDTM Findings domains, and the qualifier variable origin QORIG and evaluator QEVAL for supplemental qualifier data sets. The structure of the list is one row per combination of the respective 4 variables.

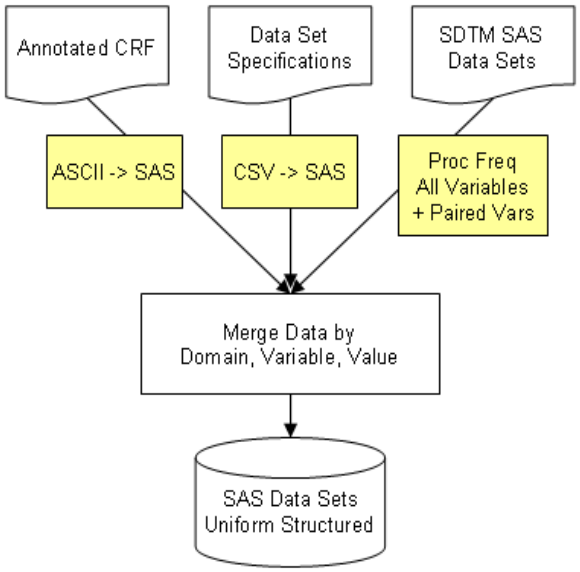
This list displays detailed information about the paired variables. For example, in Table 5, the data related to LBTESTCD is fairly consistent except for Study 3; Study 3 has multiple standard units mmol/L and mg/dL for the lab test Blood Urea Nitrogen, but only mmol/L is present in the value level metadata for that test in Specifications. This implies possible programming issue in Study 3.

For supplemental qualifier data sets, QORIG and QEVAL can be standardized unless there are study or sponsor specific requirements. For example, if QEVAL = INVESTIGATOR has been used as the default mapping, we may stick to it rather than using a different value such as SPONSOR, for example, Study 3 in Table 5.

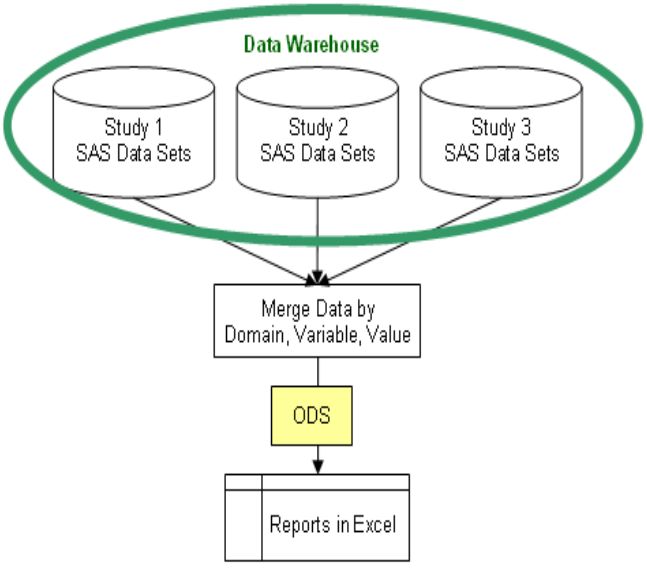
DESIGN

Base SAS is the programming environment for both the data warehouse and the reports. It is used to import the CRF annotations, and the metadata in Specifications into SAS. For SDTM SAS data sets across each study, as illustrated in Flowchart 1, Proc Freq is used to summarize the values from individual variables and also paired variables. Then the data from the three sources are merged appropriately by SDTM domain, variable and value, and saved in a set of SAS data sets.

Flowchart 1: Data processes across each study



Flowchart 2: Reports across multiple studies



The output SAS data sets are merged across studies, and reports created as illustrated in Flowchart 2. Reports are created with SAS ODS. Final reports are in Excel to take advantage of its AutoFilter tool.

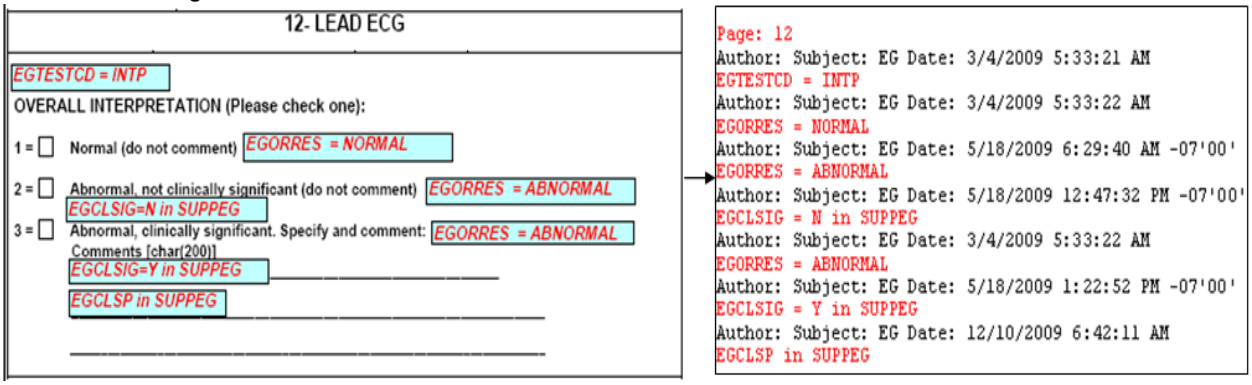
IMPLEMENTATION OF DATA WAREHOUSE

Annotated CRF → Data Warehouse

The annotations on CRF are created as Comments in the PDF file. They have consistent format and layout by following SDTM Submission Guidelines. For example, in Table 6, the annotation for a test code can be *EGTESTCD = INTP*, and for a test result *EGORRES = NORMAL*. QVAL for supplemental qualifier is annotated as, for example, *EGCLSIG = N in SUPPEG*.

To import the annotations to SAS, first save the annotations in ASCII file. In Adobe Acrobat, in the menu bar, click Comments and Summarize Comments…, and then click Comments Only. This extracts the annotation text to a separate window. Copy all the text and paste to a text editor and save them as ASCII file. For example, in Table 6 the annotations from the PDF file for the 12-LEAD ECG CRF can be saved as the text in ASCII file on the right.

Table 6: Converting annotations from PDF to ASCII file



Once the annotations are in ASCII file, they can be read into SAS with Data step. Note in Table 6, only the text in red is CRF annotations, plus Page 12 that is the page number from the PDF file and is part of the PDF Comments. The text is structured since we follow certain rules when creating the annotations. Therefore we can scan the imported text and extract specification information about SDTM domains, variable names, and variable values. Table 7 is the sample SAS code to process the text. The output SAS data set from the sample code is as in Table 8.

Table 7: Sample SAS code to import ASCII file to SAS

```
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;
```

Table 8: CRF annotations saved in SAS data set

TABLE: Work.Acrf

Domain	Variable	Value	Page	CRF	text
EG	EGTESTCD	INTP	12	12-LEAD ECG	EGTESTCD = INTP
EG	EGORRES	NORMAL	12	12-LEAD ECG	EGORRES = NORMAL
EG	EGORRES	ABNORMAL	12	12-LEAD ECG	EGORRES = ABNORMAL
SUPPEG	EGCLSIG	N	12	12-LEAD ECG	EGCLSIG = N in SUPPEG
EG	EGORRES	ABNORMAL	12	12-LEAD ECG	EGORRES = ABNORMAL
SUPPEG	EGCLSIG	Y	12	12-LEAD ECG	EGCLSIG = Y in SUPPEG
SUPPEG	EGCLSP		12	12-LEAD ECG	EGCLSP in SUPPEG

Data Set Specifications → Data Warehouse

SDTM data set Specifications sample are saved in CSV files. They include SDTM domain names, variable names, variable labels, type, length, controlled terminology terms and other required metadata. Table 9 is an example for EG domain and its supplemental qualifier SUPPEG.

Table 9: Sample data set Specifications in CSV file for EG and SUPPEG

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	

Data step combined with SAS Macro is used to loop through all domains and read individual CSV files into SAS. Table 10 is the sample code. Table 11 is the output SAS data set, and it matches the Specifications in Table 9.

Table 10: Sample SAS code to import Specifications CSV files to SAS

```
%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;
```

Table 11: SDTM data set Specifications saved in SAS data set

TABLE: Work.Eg

Domain	Variable	Label	Type	Length	Terms
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; N
SUPPEG	EGCLSP	Clinically Significant Specify	Char	200	

Two other output data sets are created as shown in Table 12 and Table 13. In Table 12 the column Value is the controlled terminology terms extracted from the Specifications, and the column Label is the SDTM variable labels. Table 13 is the value level metadata extracted from the Specifications, where the column Variable is the target variable names from the Specifications, i.e., EGTESTCD and QNAM, and the column Value stores the values of the two target variables, i.e., INTP and PR, and EGCLSIG, respectively. The column Label in Table 13 is the corresponding test names from EGTEST, i.e., ECG Interpretation and PR Interval, and the qualifier variable label

from QLABEL, i.e., Clinically Significant.

Table 12: Controlled terminology terms from Specifications saved in SAS data set

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

Table 13: Value level metadata from Specifications saved in SAS data set

TABLE: Work.Specs_terms

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

SDTM SAS Data Sets → Data Warehouse

Proc Freq and Merge statement are used extensively to process the data from SDTM SAS data sets. Table 14 is sample SDTM SAS data set for EG domain, and Table 15 the supplemental qualifier SUPPEG data set.

Table 14: SDTM SAS data set for EG domain

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	

Table 15: SDTM SAS data set for SUPPEG supplemental qualifier

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

Table 16 is the sample SAS code to summarize the SDTM SAS data sets. Proc Freq combined with Data step in Macro loop is run for each variable in the SDTM SAS data set. The output data set based on the EG and SUPPEG data sets is in Table 17, where the column Value displays all the variable values in the SDTM SAS data sets. The column Label is the variable labels in the SDTM SAS data sets.

The output data sets from Proc Freq for the paired variables in the EG and SUPPEG data sets are as in Table 18. The column Variable is the target variable names, i.e., EGTESTCD or QNAM, from the EG data set and SUPPEG data set, respectively. The column Value is the values of the two target variables, e.g., EGALL, INTP and PR for EGTESTCD, and EGCLSIG for QNAM. The column Label is the corresponding values from EGTEST and QLABEL. That is, ECG Data, ECG Interpretation and PR Interval that are associated with the three ECG test codes, and Clinically Significant that is associated with the qualifier variable EGCLSIG.

Table 16: Sample code to summarize SDTM SAS data sets

```
%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;
```

Table 17: SDTM variables values summarized and saved in SAS data set

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

Table 18: Paired SDTM variable values summarized and saved in SAS data set

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

IMPLEMENTATION OF REPORTS

Reports are created with SAS ODS. Table 19 is sample code. Note that with ODS the output file type is XML. XML file can be open with Excel and saved as a new Excel file.

Table 19: Sample SAS code to create reports

```
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;
```

Reports are the output from the data warehouse and the tool for programmers or anyone who wishes to know about our SDTM-based studies without having to access the annotated CRFs, data set Specifications, or SDTM SAS data sets. Given all the detailed information from the three sources is captured in the data warehouse, with the exception of patient level association, the purposes and the uses of the reports can be unlimited, whether it is for knowledge transfer, comparison across studies or validation within a study. The following describes the five types of reports according to the Requirements section in this paper.

Report 1: CRF–SDTM Mapping Overview

In Table 20, this is a list of the SDTM SAS data sets and the corresponding CRFs where the raw data are collected. This CRF - SDTM association serves as a reference for new studies to ensure consistent data mapping. The area in yellow in Table 20 shows the various forms we have mapped to the custom domain BI - Biomarkers, and the area in green the forms mapped to the published domain CM - Concomitant Medications, including the supplemental qualifier SUPPCM consistently with data coming from the Prior Cancer Therapy CRF and the Prior Radiation Therapy CRF.

The structure of this list is one row per SDTM domain per CRF per study. It may look like displaying a lot of duplicate domains and CRF names. However, with AutoFilter in Excel, we can easily drill down to display only what we are interested in. For example, with AutoFilter by Study # = 8888/033, it displays only the unique combination of SDTM data sets and CRFs for that study, as in the bottom of Table 20.

The last two columns on this report, aCRF Link and Specs Link, are embedded with the Excel function HYPERLINK in each cell. For example, the function for Study 8888041 aCRF page 27 is =HYPERLINK("8888041-aCRF.pdf#27", "aCRF #27"), where the first argument in the function is the aCRF file name plus the page number, and the second argument is the label for this hypertext link. Similarly, the function for the BI tab in the Specifications Excel file is =HYPERLINK("[8888041-CRT-Specification.xls]BI!A1", "BI"), where the first argument in the function is the Excel file name plus the tab name BI, and the second argument is the text for this hypertext link.

Report 2: Distribution of SDTM Variables

The sample report in Table 21 compares between studies. The cells in yellow highlight shows the variables CMDOSE, CMDOSFRQ, CMDOSTXT and CMDOSU are included in both studies, except for CMDOSTXT that is not annotated on CRF. The variable EXCAT appears in study 8888/043 only, and the value “STUDY DRUG” is annotated on CRF. For EX domain, the cells in green highlight show that some permissible variables are present in SDTM SAS data sets and annotated on aCRF, e.g., EXADJ in study 8888/043, and EXDOSFRQ and EXDOSTOT in study 8888/062. Overall, the SDTM SAS data sets match aCRF and Specification within each study.

Report 3: SDTM Variable Values vs. Controlled Terminology Terms

The sample report in Table 22 compares three studies on the variable EGORRES. First of all, the variable label for EGORRES is identical among SDTM data sets and Specifications for all studies. Next, in terms of the variable values of EGORRES, this report can identify even slight differences. For example, the three studies all collect abnormal test results being clinical significance or not. However, due to different pre-printed text on CRF used as the controlled terminology terms, the values for EGORRES are different. There are “ABNORMAL, CLINICALLY SIGNIFICANT” vs. “CLINICALLY SIGNIFICANT FINDINGS”, and “ABNORMAL, NON-CLINICALLY SIGNIFICANT” vs. “ABNORMAL, NOT CLINICALLY SIGNIFICANT”.

This sample report also points out a data issue in study 8888/047. Comparing with the controlled terminology terms for the study, the text in red, “SIGNIFICAN”, is likely a data truncation when in SDTM SAS data set for EG.

When reviewing this list we need to be aware of any free text collected on CRF. For example, EGORRES may contain numeric test results. In Table 22, there are the values 996, 997 and 998 for EGORRES. They are from specific ECG tests such as RR Interval, and are not subject to controlled terminology.

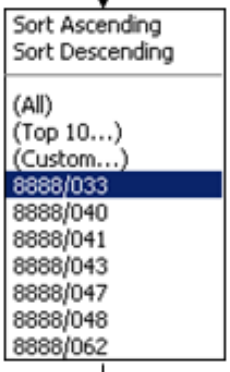
Table 20: Report 1, summary of SDTM – CRF mapping

=HYPERLINK("[8888041-CRT-Specification.xls]BI!A1", "BI")

=HYPERLINK("8888041-aCRF.pdf#27", "aCRF #27")

SDTM Domain	SUPP--	aCRF Title	Study #	aCRF Link	Specs Link
BI		Human Anti-Human Antibody Samples	8888/041	aCRF #27	BI
BI		Investigational Biomarkers	8888/041	aCRF #34	BI
BI		Investigational Biomarkers - Archived Tumor	8888/043	aCRF #25	BI
BI		Investigational Biomarkers - Vegf	8888/043	aCRF #43	BI
BI		Serum Apoptosis Biomarkers	8888/033	aCRF #49	BI
BI		Serum Cea Biomarkers	8888/033	aCRF #52	BI
BI		Snap Frozen Tumor Sample	8888/041	aCRF #20	BI
CM		Concomitant Medications	8888/062	aCRF #39	CM
CM		Concomitant Procedures	8888/033	aCRF #41	CM
CM		Concomitant Procedures	8888/040	aCRF #52	CM
CM		Non Drug Treatment Procedures	8888/047	aCRF #52	CM
CM		Non Drug Treatment Procedures	8888/048	aCRF #47	CM
CM		Non-Drug Treatment/Procedures	8888/043	aCRF #53	CM
CM	SUPPCM	Prior Cancer Therapy	8888/033	aCRF #11	CM
CM	SUPPCM	Prior Cancer Therapy	8888/040	aCRF #11	CM
CM	SUPPCM	Prior Radiation Therapy	8888/033	aCRF #13	CM
CM	SUPPCM	Prior Radiation Therapy	8888/040	aCRF #12	CM

AutoFilter by **Study #**



SDTM Domain	SUPP--	aCRF Title	Study #	aCRF Link	Specs Link
BI		Serum Apoptosis Biomarkers	8888/033	aCRF #49	BI
BI		Serum Cea Biomarkers	8888/033	aCRF #52	BI
CM		Concomitant Procedures	8888/033	aCRF #41	CM
CM	SUPPCM	Prior Cancer Therapy	8888/033	aCRF #11	CM
CM	SUPPCM	Prior Radiation Therapy	8888/033	aCRF #13	CM

Report 4: SDTM Paired Variable Values vs. Value Level Metadata

The sample report in Table 23 displays the combination of QNAM and QLABEL in the supplemental qualifier SUPPEX from three studies. This can serve as a reference for new studies to carry over the same mapping.

This list shows that the data in SUPPEX data sets match the value level metadata in Specifications. It also shows consistent mapping across the three studies, for example, QNAM = DOSEINTE and QLABEL = Dose Interrupted in both studies 8888/043 and 8888/047, and QNAM = DOSERED and QLABEL = Dose Reduced in studies 8888/043 and 8888/062.

The sample report also identifies qualifier variables that are defined in Specifications but not present in SDTM SAS data sets, for example, Study 8888/062, the paired QNAM = AUCDOSE and QLABEL= Target AUC Dose. This might be simply because of no raw data collected for the data field Target AUC Dose on CRF. However, it could be a programming issue that has dropped data points when transferring the raw CRF data to SDTM. So, it is worth checking the raw CRF data to verify if this data field is indeed blank.

Table 21: Report 2, distribution of SDTM variables in aCRF, Specifications and SDTM SAS data sets

			8888/043			8888/062		
Domain	Variable	Values (if present on aCRF)	aCRF	Specs	Dataset	aCRF	Specs	Dataset
CM	CMDOSE		Y	Y	Y	Y	Y	Y
CM	CMDOSFRQ		Y	Y	Y	Y	Y	Y
CM	CMDOSTXT			Y	Y		Y	Y
CM	CMDOSU		Y	Y	Y	Y	Y	Y
EX	EXCAT	STUDY DRUG	Y	Y	Y			
EX	EXADJ		Y	Y	Y			
EX	EXDOSE		Y	Y	Y		Y	Y
EX	EXDOSFRM			Y	Y		Y	Y
EX	EXDOSFRQ						Y	Y
EX	EXDOSTOT					Y	Y	Y
EX	EXDOSTXT			Y	Y		Y	Y
EX	EXDOSU		Y	Y	Y	Y	Y	Y
EX	EXROUTE			Y	Y		Y	Y

Table 22: Report 3, variable values in SDTM SAS data sets vs. Specifications

				8888/043		8888/047		8888/062	
Domain	Variable	Value	Label	Datase	Specs	Datase	Specs	Datase	Specs
EG	EGORRES	996	Result or Finding in Original Units					Y	
EG	EGORRES	997	Result or Finding in Original Units					Y	
EG	EGORRES	998	Result or Finding in Original Units					Y	
EG	EGORRES	ABNORMAL, CLINICALLY SIGNIFICANT	Result or Finding in Original Units		Y		Y		
EG	EGORRES	ABNORMAL, NON-CLINICALLY SIGNIFICAN	Result or Finding in Original Units			Y			
EG	EGORRES	ABNORMAL, NON-CLINICALLY SIGNIFICANT	Result or Finding in Original Units	Y	Y		Y		
EG	EGORRES	ABNORMAL, NOT CLINICALLY SIGNIFICANT	Result or Finding in Original Units					Y	Y
EG	EGORRES	CLINICALLY SIGNIFICANT FINDINGS	Result or Finding in Original Units						Y
EG	EGORRES	NORMAL	Result or Finding in Original Units	Y	Y	Y	Y	Y	Y

Table 23: Report 4, paired variable values in SDTM SAS data sets vs. Specifications

				8888/043		8888/047		8888/062	
Domain	Variable	Value	Label	Dataset	Specs	Dataset	Specs	Dataset	Specs
SUPPEX	QNAM	AUCDOSE	Target AUC Dose					** Absence **	Y
SUPPEX	QNAM	DOSCHINT	Dose Change/Interruption					Y	Y
SUPPEX	QNAM	DOSEDILU	Dilution			Y	Y		
SUPPEX	QNAM	DOSEHELD	Dose Held			Y	Y		
SUPPEX	QNAM	DOSEINTE	Dose Interupted	Y	Y	Y	Y		
SUPPEX	QNAM	DOSELVL	Dose Level	Y	Y				
SUPPEX	QNAM	DOSEMISS	Did the Subject Miss Any Doses?					Y	Y
SUPPEX	QNAM	DOSERED	Dose Reduced	Y	Y			Y	Y
SUPPEX	QNAM	DOSEREDT	Date of Dose Reduction	Y	Y				

Report 5: SDTM Extended Paired Variable Values vs. Value Level Metadata

The sample report in Table 24 displays the variable value level data associated with the HCG test in LB data sets from three studies. Though the values on LBTESTCD and LBTEST are consistent, there are differences on LBCAT and LBSTRESU. In studies 8888/043 and 8888/047, the values on LBCAT are PRENANCY TEST and PREGNANCY, respectively, and there is no standard unit associated. This implies that the test result must have been Positive or Negative. For study 8888/062, the HCG test is in the category of CHEMISTRY, and has the standard unit IU/L. So, there are two findings by reviewing this report. First of all, we may want to decide on the LBCAT value being either PREGNANCY TEST or PREGNANCY for the purposes of consistent data mapping. Secondly, when the HCG test is in the category of Chemistry, it shall be associated with the same standard unit, in this case, IU/L.

Table 24: Report 5, extended paired variable values in SDTM SAS data sets vs. Specifications

Domain	Variable	--TESTCD / QNAM	--TEST / QLABEL	--CAT / QORIG	--STRESU / QEVAL	8888/043		8888/047		8888/062	
						Dataset	Specs	Dataset	Specs	Dataset	Specs
LB	LBTESTCD	HCG	Choriogonadotropin Beta	PREGNANCY				Y	Y		
LB	LBTESTCD	HCG	Choriogonadotropin Beta	PREGNANCY TEST		Y	Y				
LB	LBTESTCD	HCG	Choriogonadotropin Beta	CHEMISTRY	IU/L					Y	Y

DISCUSSION

Various types and uses of reports can be created out of the data warehouse. For example, we can add VISIT and VISITNUM to the list of paired variables in Report 4, or the sub-category variable --SCAT and original test unit variable --ORRESU to Report 5.

Because of the standardized structure of the SAS data sets in the data warehouse, program templates and macros are developed for programmers to use and create custom reports for comparison or validation purposes. For example, based on Report 4, we can display rows only when there are multiple --TEST values corresponding to the same --TESTCD, which indicates data issues. Another modification is to add WHERE clause to any of the Reports to list a subset of SDTM variables of interest.

ACKNOWLEDGMENTS

The author would like to extend a special thanks to Robert Stemplinger for his review of this paper, as well as his ongoing support for progress, creativity and innovation.

CONTACT INFORMATION

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

Name: Annie Guo
Affiliation: ICON Clinical Research
Address: North Wales, CA, USA
Work Phone: 215-616-6597
Fax: 215-240-7595
E-mail: annie.guo@iconplc.com
Web: www.iconplc.com