SCSUG2010

Facilitating Data Integration for Regulatory Submissions

John R. Gerlach, SAS / CDISC Analyst, Hamilton, NJ
John C. Bowen (Retired), Merck & Co., Rahway, NJ

ABSTRACT

The process of integrating data from multiple clinical trials for regulatory submissions poses many challenges, which is often labor intensive, as well as error prone. Even if the SAS data libraries are supposedly CDISC compliant, the process of integrating the data far exceeds a simple concatenation of data sets. Certainly, the integration process becomes even more intricate involving legacy studies that used different proprietary standards, as well as CDISC. Consequently, there are harmonization issues at both the metadata level and content level. More importantly, effective data integration is critical to the ISS / ISE analysis and the integrity of the submission. This paper explains a reporting tool, implemented as a SAS® macro, that facilitates the data integration process by comparing pair-wise similar data sets at the *metadata* and *content* level.

INTRODUCTION

The proposed SAS solution is strictly a reporting tool that facilitates data integration. It does not compare with data integration platforms like the SAS Enterprise Data Integration Server that can access various data sources and perform the gamut of ETL (Extract, Transform, and Load) processes. Also, data integration is not about using the COMPARE procedure, which analyzes the contents of two SAS data sets, such as matching variables having different values or, even, one data set having more observations (or variables) than the other. In fact, the terms *Base* and *Comparison* data sets associated with the COMPARE procedure are inappropriate, because no single data set determines necessarily what values are *correct*; that is, there is no Base data set per se. Instead, the objective is to integrate data sets, by harmonizing the data, not simply by comparing values. Yet, part of this process requires a comparison of one data set with another at the metadata, rather than just the content level. Fundamentally, data integration involves an identification of dissimilar data, such as an inconsistent collection of range values, terms, encodings with the intent to reconcile the disparity by consensus, thereby creating a homogenous set of data for analysis.

Consider an Integrated Summary of Safety (ISS) study consisting of ten individual studies. Also, let's assume that the individual studies are CDISC compliant or at least, a variant of the standard called Plus / Minus. Now, let's consider the Demographic (DM) domain, all ten of them. Which one should be used as a base data set? Does it matter? Perhaps not. Will the set of variables represent the union or intersection of the ten DM domains? The union of all variables might create a target data set that contains variables having mostly missing values because such variables existed only in one or two of the studies. Conversely, the intersection of variables might be too restrictive, causing a loss of information needed for the intended analysis. In practice, it is the clinical and analytical objectives of the study (i.e., to show safety and efficacy) that determines which variables are important.

Whether or not there is a base study, the process of integrating similar data sets from multiple studies requires a natural process for consolidating data in order to create a harmonious aggregate collection. Thus, reasonably, we begin with two data sets from which we incorporate the other similar data sets to the collection, until all the demographic data, for example, have been integrated. This pair-wise methodology ensures that all the data sets are integrated without the usual labor-intensive effort that creates numerous redundant reports.

Since there is no Base or Comparison data set *per se*, let’s agree on a more abstract naming convention for a pair of data sets – simply *Left* and *Right*, such that the Left data set acts as a pivotal data set from one study and the Right data set contributes more data. Thus, this process involves initially the integration of pair-wise similar data sets likely having the same name only, albeit residing in different data libraries. In fact, for this discussion, let's assume that the pair-wise data sets have the same name. Also, keep in mind that the Left data set might represent an intermediate aggregate collection of data, which is paired with the next contributing data set.

REPORT AND LAYOUT FEATURES

The proposed SAS solution generates two reports, that is, at the metadata-level and content-level. Below is the layout for the metadata-level report. Basically, it lists all the variables from the Left data set along with the data type, length, and label of each variable and, in juxtaposition, the metadata from the Right data set. In the event that the Right data set does not contain a respective variable, the information is left blank.

Comparison of the DM Data Set in the *Left* and *Right* Data Libraries
(Metadata Level)

	========== Left ==========			========== Right ==========		
Name	**Type**	**Length**	**Label**	**Type**	**Length**	**Label**
<Variable>	*N/C*	*n*	*<Description>*	*N/C*	*n*	*<Description>*

The content-level report is almost crude by appearance in that it contains simply the name of the variable along with as many as 30 **unique** values from the Left and Right data set. The values are ordered; however, missing values are always listed first. Moreover, for character variables having null values, the term "< Null >" is used.

For example, the values for the AESER (Serious AE) variable might be listed; whereas, values for the AEREL (AE Related to Study Drug) variable might look quite different, as follows:

Comparison of the AE Data Set in the *Left* and *Right* Data Libraries
(Content Level)

Variable	***Left***	***Right***
AESER	N	N
	Y	Y
AEREL	< Null >	DEFINITELY RELATED
	N	NOT RELATED
	Y	POSSIBLY RELATED
		PROBABLY RELATED
		UNLIKELY RELATED

Notice (especially obvious for the variable AEREL) that the listed values from the Left and Right data sets are independent of each other. The objective is to readily discern the compatibility of the data sources with respect to a common variable. Clearly, AEREL poses an integration issue that requires data mapping. Also, this report lends itself to be expanded with a comments column that could be a mechanism for capturing the agreed upon data mappings, recoding or reformatting.

THE SAS SOLUTION

The reporting utility consists of a single SAS macro that contains three positional parameters and one keyword parameter, as follows:

- Left — Pivotal Data Library
- Right — Contributing Data Library
- DSN — Common-named (existing) SAS Data Sets
- HTML=Y — ODS HTML output, as well as regular output

As mentioned earlier, the so-called base data library may be a misnomer in the context of data integration; however, it is often used as the "standard" to which contributing data sets conform. And, even though the pair-wise data sets should exist, the SAS macro aborts with an error message if both do not exist. The HTML keyword parameter generates more aesthetically pleasing reports that can be viewed by team members via a browser.

Consider the following invocations of the macro **%data_integrate**.

```
%data_integrate(study101,   study201,  AE,   HTML=N) ;

%data_integrate(study101,   study201,  DM)           ;

%data_integrate(study101,   study201,  ADSL)         ;

%data_integrate(study101,   study201,  QQQ)          ;
```

All invocations attempt to produce metadata-level and content-level reports on the pair-wise data sets stored in STUDY101 and STUDY201, the Left and Right data libraries, respectively. The first invocation wants standard output only, no HTML document, which is the default, and the last invocation specifies a non-existing data set, which promptly generates an error message, and then aborts.

At the onset of analyzing two data sets, it is necessary to make sure that both exist in their respective data libraries. The macro contains the following code that creates two macro variables, *&leftdsn* and *&rightdsn*, which a %IF statement uses in order to decide whether to proceed or abort.

```
proc sql noprint;
   select count(*) into :leftdsn
      from dictionary.tables
       where libname eq "%upcase(&left.)"  and memname eq "%upcase(&dsn.)";
quit;

proc sql noprint;
   select count(*) into :rightdsn
      from dictionary.tables
       where libname eq "%upcase(&right.)" and memname eq "%upcase(&dsn.)";
quit;
```

Given that both data sets exist, the utility proceeds to generate first the Content-level report consisting of the metadata, specifically: variable name, data type, length, and label. Again, using Dictionary tables, the SQL procedure accomplishes this task easily. A Data step performs a match-merge of the metadata from both data sets and the Report procedure generates the desired Content-level report, as illustrated by the following code.

```
proc sql noprint;
   create table left as
   select upcase(name) as name, upcase(type) as type1,
      length as len1, label as lab1
      from dictionary.columns
       where libname eq "%upcase(&left.)" and memname eq "%upcase(&dsn.)"
       order by name;
quit;
proc sql noprint;
   create table right as
   select upcase(name) as name, upcase(type) as type2,
      length as len2, label as lab2
      from dictionary.columns
       where libname eq "%upcase(&right.)" and memname eq "%upcase(&dsn.)"
      order by name;
quit;
data rep;
   merge left(in=left) right(in=right);
       by name;
   if left
      then do;
         if right and lab2 eq ' ' then lab2 = lab1;
         output;
         end;
```

```
run;
proc report data=rep nowindows headline headskip;
    columns name ("= %upcase(&left.) ="  type1 len1 lab1)
                 ("= %upcase(&right.) =" type2 len2 lab2);
   define name    / display     width=8                          'Name';
   define type1   / display     width=4                          'Type';
   define len1    / display     width=6   format=3.   center     'Length';
   define lab1    / display     width=40                         'Label';
   define type2   / display     width=4                          'Type';
   define len2    / display     width=6   format=3.  center      'Length';
   define lab2    / display     width=40                         'Label';
run;
```

Consider the following abridged (i.e. labels not shown completely) Metadata-level report for the Demography (DM) data set. This report clearly shows that both data sets are very compatible. One would think that the task of integrating these data sets would require little more than the APPEND procedure, for example. But, are these data sets copasetic at the content-level?

Comparison of the DM Data Set in the *Left* and *Right* Data Libraries
(Metadata Level)

	========== Left ==========			========== Right ==========		
Name	**Type**	**Length**	**Label**	**Type**	**Length**	**Label**
AGE	NUM	8	Age in AGEU at …	NUM	8	Age in AGEU at …
AGEU	CHAR	5	Age Units	CHAR	5	Age Units
ARM	CHAR	10	Description of …	CHAR	10	Description of …
ARMCD	CHAR	10	Planned Arm Code	CHAR	10	Planned Arm Code
BRTHDTC	CHAR	10	Date of Birth	CHAR	10	Date of Birth
COUNTRY	CHAR	3	Country	CHAR	3	Country
DOMAIN	CHAR	2	Domain Abbreviation	CHAR	2	Domain Abbreviation
RACE	CHAR	10	Race	CHAR	10	Race
RFENDTC	CHAR	20	Subject Reference End …	CHAR	20	Subject Reference End …
RFSTDTC	CHAR	20	Subject Reference Start …	CHAR	20	Subject Reference Start …
SEX	CHAR	6	Sex	CHAR	6	Sex
SITEID	CHAR	8	Study Site Identifier	CHAR	8	Study Site Identifier
STUDYID	CHAR	20	Study Identifier	CHAR	20	Study Identifier
SUBJID	CHAR	10	Subject Identifier …	CHAR	10	Subject Identifier …
USUBJID	CHAR	15	Unique Subject Identifier	CHAR	15	Unique Subject Identifier

Consider another metadata-level report involving adverse events. The variables AEENRF and AESDTH do not even exist in the Right data set. The AEREL (causality) variable found in the Left data set contains only one byte (Y/N); whereas, the Right data set stores twenty bytes of information, probably a more descriptive response (e.g. Definitely Related). Clearly, there's a harmonization issue here. Finally, there may be concern for the variables AEOUT and AETERM whose length is considerably shorter in the contributing Right data set with respect to losing information.

Comparison of the AE Data Set in the *Left* and *Right* Data Libraries
(Metadata Level)

	========== Left ==========			========== Right ==========		
Name	**Type**	**Length**	**Label**	**Type**	**Length**	**Label**
AEACN	CHAR	100	Action Taken with …	CHAR	100	Action Taken with …
AEBODSYS	CHAR	100	Body System or Organ Class	CHAR	100	Body System or Organ Class
AEDECOD	CHAR	100	Dictionary-Derived Term	CHAR	100	Dictionary-Derived Term
AEENDTC	CHAR	20	End Date/Time of Adverse …	CHAR	20	End Date/Time of Adverse …

AEENDY	NUM	8	Study Date of End of Event		NUM	8	Study Day of End of Event
*** AEENRF**	**CHAR**	**16**	**End Relative to Reference …**		**.**		
AEHLGT	CHAR	200	MedDRA Highest Level …		CHAR	200	MedDRA Highest Level …
*** AEOUT**	**CHAR**	**50**	**AE Outcome**	**CHAR**	**25**	**Outcome of Adverse Event**	
*** AEREL**	**CHAR**	**1**	**Causality**	**CHAR**	**20**	**Causality**	
*** AESDTH**	**CHAR**	**1**	**Results in Death**		**.**		
AESEQ	NUM	8	Sequence Number	NUM	8	Sequence Number	
AESER	CHAR	1	Serious Event	CHAR	1	Serious Event	
AESEV	CHAR	20	Severity	CHAR	20	Severity	
AESTDTC	CHAR	20	Start Date/Time of …	CHAR	20	Start Date/Time of …	
AESTDY	NUM	8	Study Day of Start of Event	NUM	8	Study Date of Start of Event	
*** AETERM**	**CHAR**	**200**	**Reported Term for the …**	**CHAR**	**100**	**Reported Term for the …**	
DOMAIN	CHAR	2	Domain Abbreviation	CHAR	2	Domain Abbreviation	
STUDYID	CHAR	20	Study Identifier	CHAR	20	Study Identifier	
USUBJID	CHAR	15	Unique Subject Identifier	CHAR	15	Unique Subject Identifier	

For the content-level report, the SAS macro identifies character and numeric variables based on the Left data set, first processing character variables, the numeric variables, according to the following algorithm:

- Identify character variables, if any.
- For each variable
  - Obtain unique values found in the Left data set.
  - Determine the data type of the respective variable in the Right data set.
  - Obtain unique values, keeping 30 observations only, storing them as character values, regardless of its data type.
  - Perform a 1-1 merge on the Left and Right data sets containing the unique values.
    - Assign the text '< Null >' for missing values (blanks only).
  - Append this data set representing the i*th* variable to the reporting data set.
- Produce the report representing all character variables.

Assigning the "< Null >" text needs to be done only at the first iteration of the merge. Why? Because the juxtaposed values are disjoint (recall the variable AEREL); that is, the Left or Right data set lists fewer values than the other; thus, the Data step that performs the one-to-one merge contains the code below. Not surprisingly, this algorithm applies to the numeric variables, as well.

```
if _n_ eq 1
   then do;
      if &left. eq ''
         then &left.  = '< Null >';
      if &right. eq ''
         then &right. = '< Null >';
      end;
```

DATA INTEGRATION ISSUES

Consider now various issues that have been identified by the proposed SAS solution during the data integration process. For this exercise, assume that CDISC data libraries are being integrated. Also for convenience, the issues, denoted by variables, are listed in alphabetical order. Again, keep in mind that the context of each issue is *either* two studies *or* the aggregate collection along with the next study, which are called *Left* and *Right* studies. Each issue contains a brief discussion followed by a list of values in juxtaposition, that is, the content-level report, as needed.

AEOUT (Outcome of Adverse Event) in the AE domain – Notice that the Right study represents a subset of values, except for the value ONGOING whose value should change accordingly by the time of database lock. Hence, there should be no harmonization issue here.

Variable	***Left Study***	***Right Study***
AEOUT	FATAL	FATAL
	RESOLVED	ONGOING

	RESOLVED WITH SEQUELAE	RESOLVED
	UNKNOWN	RESOLVED WITH SEQUELAE
	UNRESOLVED	

AEREL (Relationship to Study Drug) in the AE domain – In the Left study, the variable AEREL contains dichotomous values (Yes / No); whereas, the Right study indicates five unique descriptive values, which is more commonly found in CDISC domains. Hence, in this case, it is likely that the *Y* and *N* values would be mapped to *Definitely Related* and *Not Related*, respectively. Keep in mind that the point is not the mapping per se, which is typically a clinical issue. Instead, this discussion explains a method for identifying these issues efficiently as part of the integration process.

AEREL	N	Definitely Related
	Y	Not Related
		Possibly Related
		Probably Related
		Unlikely Related

AESDTH (Adverse event resulting in death) in the AE domain – Imagine that this variable exists in all the studies, except one. However, fortunately that particular study does have the variable AEOUT from which you can impute the value for AESDTH when integrating the AE domain. Thus, here the data integration process requires a rule. Notice that this issue is found in the metadata report, not at the content level since the variable does not exist in the Right (contributing) data set. The CDISC standard stipulates that this variable typically contains CDISC Control Terminology: Y, N, or Null.

AESDTH → Imputed from AEOUT (FATAL).

AESEV (Severity / Intensity of Adverse Event) in the AE domain – The following report indicates a possible issue in the Left study; that is, the UNKNOWN and Null values. It may be convenient to recode the Null values to UNKNOWN. Notice, however, this issue has nothing to do with the Right study. On the other hand, the Right study has a harmonization issue; that is, the values in the Right study are stored in mixed case, unlike the Left study. Consequently, these values should be converted into upper case. Also, the value LIFE THREATENING may be a harmonization issue, perhaps extraneous, depending on the analysis plan.

AESEV	<Null>	Mild
	LIFE THREATENING	Moderate
	MILD	Severe
	MODERATE	
	SEVERE	
	UNKNOWN	

AESTDY (Study Day of Start of Adverse Event) in the AE domain – With continuous data such that the values aren't related per se, it takes familiarity with the respective study designs to make sense of the values collected across studies. The Left data set contains null values, as well as negative numbers; whereas, the Right data set contains whole numbers ranging from 0 to 30 (even though it is arguable that there should not be a study day zero). Regardless, these are not necessarily data integration issues. However, it may be necessary to know how to group adverse events or related data by study day, which is likely an analysis or reporting question. Finally, notice that the proposed SAS solution converts numeric variables in character format for reporting, which explains the order.

AESTDY	. (Null)	0
	1	1
	2	2
	3	:
	< More values >	30
	Negative values	
	< More values >	
	20	

AETERM (Reported Term for the Adverse Event) in the AE domain – The Right study shows that the variable contains at least one instance of a null value, which is inappropriate for a Required CDISC variable. Perhaps the Right study is ongoing such that it contains values that require follow-up at the study site. Also noteworthy, the

variables AEDECOD (Dictionary Derived Term) and AEBODSYS (Body System or Organ Class) should have null values as well, since these variables were imputed from the MedDRA dictionary based on the imputed Preferred Term. A similar situation can occur for the Concomitant Medications (CM) domain. The CMTRT (Reported Name of Drug, Med, or Therapy) variable might contain null values; whereupon, the CMDECOD (Standardized Medication Name) and CMCLAS (Medication Class) must contain null values, as well, again, since these values were imputed from the WHO dictionary.

ARMCD (Treatment Arm) in the DM domain – There are two issues in the Demographic domain. First, the Right study has at least one instance of a missing value, which is unacceptable for a Required CDISC variable. The other situation concerns the values PROD_NAME and CMPD_NAME, which actually represents the same treatment group, even though the Left study identifies the study drug using its product name and the Right study uses a code name. Obviously, the name of the study drug must be consistent for integrated reporting.

ARMCD	PROD_NAME	<Null>
	PLACEBO	CMPD_NAME
		PLACEBO

CMROUTE (Route of Administration) in the CM domain – This integration issue concerns the harmonization of values for the administration of a treatment by intravenous. Certainly, the values in the Left study seem more varied and detailed in contrast to the Right study. More likely, the Left study represents the collection of several studies, already, which itself contains harmonization issues (i.e. the values I/V versus Intravenous). Thus, the inclusion of additional studies affords a retrospective understanding of the integration process, even exposing issues that were overlooked. Regardless, the protocol, analysis plan, and case report forms from the respective studies must be consulted in order to explain these gradations (I/V, IV/PO, etc.) and to harmonize the data appropriately for the purpose of the integrated study. Finally, this particular example clearly demonstrates that data integration is more than just a concatenation of similar data sets.

CMROUTE	I/V	IV (INTRAVENOUS)
	IV / PO	
	IVPO	
	Intravenous	
	Intravenous direct	
	Intravenous injection	

COUNTRY in the DM domain – This variable is not really an issue per se. That is, the variable utilizes the ISO 3166 standard correctly. Also, null values are acceptable. However, the SAS solution revealed an important matter concerning another variable (e.g. REGION), which is not defined in the CDISC standard.

COUNTRY	ARG	< Null >
	DNK	CAN
	IND	CZE

Consider another issue where both studies use the ISO 3166 standard; however, the Left study uses the 2-byte version and the Right study uses the 3-byte version. Also, notice that the contributing Right data set is a superset of the Left data set.

COUNTRY	US	USA
		ENG
		ITA

DOMAIN (Domain Abbreviation) in the DM domain – The Right study represents an embarrassing oversight for this ubiquitous Required variable. Fortunately, the proposed SAS solution recognizes such oversights as a harmonization issue, which can easily be resolved.

DOMAIN	DM	< Null >
		DM

EXDOSE (Dose per administration) in the EX domain – In the Left study the EXDOSE variable is numeric in compliance with the CDISC standard. However, in the Right study, this variable is character such that the values are left justified, which is not CDISC compliant. For whatever reason, perhaps a failure in the CDISC conversion process, this variable must be corrected. Obviously, the metadata report indicates the conflict with respect to data

type. Also, in the content report, the values are right-justified for the numeric variable and left-justified for the character variable. (Results are not shown.)

RACE in the DM domain – The Left study seems clear; whereas, the Right study might require a consolidation of the values OTHER and BLACK or, perhaps, the issue might be with the values NON-WHITE and BLACK. As always, it depends how these items were defined in the respective Case Report Forms. As always, it depends on the clinical factors and the statistical objectives of the study. The null values in both studies are not an issue.

RACE	< Null >	< Null >
	NON-WHITE	BLACK
	WHITE	OTHER
		WHITE

RFENDTC (Subject Reference End Date / Time) in the DM domain – Here's an interesting observation concerning the ISO 8601 standard for date / time variables. This variable contains null values in both studies, which is acceptable for screen failures, however, not acceptable for randomized subjects (Note: the CDISC Core status of this variable is both Required and Expected). However, the Left study includes both date and time values; whereas, the Right study contains dates only. Is it an integration issue? It depends on how, or if, this data will be used for analysis purposes

RFENDTC	< Null >	< Null >
	2002-01-17T17:00	2008-07-16
	2002-01-23T12:30	2008-07-18
	2002-01-30T03:30	2008-07-21
	2002-01-30T16:51	2008-07-22
	2002-01-31T08:00	2008-07-31
	2002-02-12T14:10	2008-08-11

SEX in the DM domain – The Left study appears complete, as well as compliant to the CDISC standard. However, the Right study contains null values, which would be recoded to U for Unknown. Notice that the AEREL example used the Right study to determine the assignment; whereas, the SEX variable relied on the Left study.

SEX	M	<Null>
	F	M
	U	F

SITEID (Site Identifier) in the DM domain – The Right study indicates null values for this Required variable. Fortunately, this study affords the opportunity to impute these values from the USUBJID variable, which should have been done already at the individual study level. Nonetheless, this issue poses no harmonization issue.

SITEID	101	< Null >
	201	501
	301	701
	401	

STUDYID (Study Identifier) in the DM domain (*for example*) – Imagine that the first four studies have been integrated such that target data library contains the aggregated collection of integrated data. However, in the fifth study, the STUDYID variable contains at least one missing value for a Required variable, which poses a data management issue. Otherwise, in the context of this variable, the inclusion of this domain requires little more than a concatenation of data sets.

STUDYID	GAST_01	<Null>
	GAST_02	GAST_05
	GAST_03	
	GAST_04	

Besides these several (of many) data issues discussed here, there are other issues worth noting, which are outside the scope of the SAS utility. For example, consider Control Terminology such as the MedDRA and WHO dictionaries. What if the studies represent a span of several years such that more than one version of these dictionaries were used. Certainly, the dictionary values need to be consistent or leveled for the analysis. Without a

central database utility to do such leveling (which is often the case with integrating other organization's disparate data), it poses a huge effort far beyond the recoding of AE Severity or gender that uses different code lists.

CONCLUSION

As pharmaceutical companies continue to expand its pipeline through acquisitions, data integration methods and tools have become more than just a convenience. Even with the advent of CDISC standards, the consolidation of multiple clinical studies for ISS / ISE submissions poses real challenges that must be addressed in a more dirigible manner such that it becomes a part of the IT landscape.

The proposed SAS solution demonstrates a reliable method by which data integration issues can be easily discerned. Moreover, this method can prevent the probable need to interrupt the analysis downstream, thereby expediting the submission. Finally, this utility can be applied outside the scope of this paper, such as using a pre-defined database consisting of a superset of metadata to use as the first (i.e., Left) study for comparison.

REFERENCES

Stander, Jeff; "For Base SAS Users: Welcome to SAS Data Integration!." *Proceedings of the SAS Global Forum Conference, 2009.*

ACKNOWLEDGEMENTS

The authors greatly appreciate the *Global Scientific Programming* team at Merck & Company for their participation in the development of this paper.

CONTACT INFORMATION

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

Name:	John R. Gerlach	John C. Bowen
Enterprise:	SAS / CDISC Analyst	Independent Consultant
City, State ZIP:	Hamilton, NJ	Rahway, NJ
Work Phone:	609-672-5034	
Fax:		
E-mail:	jrgerlach@optonline.net	jhnbwn@gmail.com

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.