PharmaSUG 2014 - Paper PO01

A User Friendly Tool to Facilitate the Data Integration Process

Yang Wang, Seattle Genetics, Inc., Bothell, WA
Boxun Zhang, Seattle Genetics, Inc., Bothell, WA

ABSTRACT

For any clinical program, meta-analysis, including Development Safety Update Report (DSUR), Integrated Summary of Safety (ISS), and Integrated Summary of Efficacy (ISE), integrating data is a challenging process and is often labor-intensive and error-prone. The challenges arise primarily due to the following reasons: (1) different study designs and data standards, (2) evolving industry standards including CDISC ®, MedDRA and WHODrug, and (3) different data collection methods across vendors. In order to ensure all the variables can be integrated for analysis, we built a user-friendly tool to dynamically display the differences in variable attributes and values across multiple clinical studies. This paper, with examples and programming components, shows how this tool can be used to facilitate data integration by identifying and resolving issues resulting from variable inconsistencies.

INTRODUCTION

Data integration is often needed in clinical trial studies. Integration across studies is usually done to pool variables for meta-analysis, while integration within a study is usually done to merge data (such as when demographic data is merged with labs and endpoint results). In this paper, we focus on pooling data from multiple studies for meta-analysis.

Meta-analysis usually focuses on efficacy and safety data. Integrated safety data is used to generate regulatory reports such as DSUR, Periodic Benefit-Risk Evaluation Report (PBRER), etc. Integrated efficacy data is used to generate ISE reports, which typically involves multiple subgroup analyses.

The integration process involves four elements as illustrated in Figure 1: Select data sources, compare data, resolve issues, and analyze. After datasets are integrated, new datasets can be added to the pooled datasets, and so a new round of integration starts.

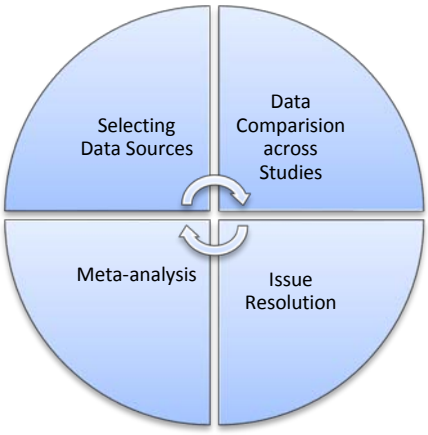


Figure 1. Four Elements in Data Integration Process

Clinical trial data are collected from multiple sources at different times. eCRF data are collected at investigator sites and entered by study coordinators whereas lab data are sent from different vendors. Data collected at different times may have different standards and conventions. Before multiple studies are integrated, the study design, data sources, and the data collected have to be fully understood, to determine whether each study can be integrated. Once the studies that will be integrated and analyzed are identified, critical variables for meta-analysis will be identified and compared. The key for data integration is to ensure pooled variables have the same definitions and attributes. Data attributes include variable length, variable type, label, and format. Data values also include units and coding. By checking the variable labels, variable values or ranges, we can usually conclude if the variable has the same definition. For example, if two variables have the same label of “gender” with values “M” and “F”, it is very likely these two variables are defined in the same way and can be stacked together using SAS “set” statement for meta-analysis. If a variable with different attributes or definitions is identified, it will be processed and harmonized before it can be

pooled with other studies to perform meta-analysis.

We developed a simple tool that can facilitate the integration process using SAS® and Visual Basic for Applications (VBA). This tool focuses on the data comparison element of the integration process. It facilitates easy comparison of variables across multiple studies and systematically identifies the variables that cannot be stacked simplifying variable standardization.

PROCESS FLOW FOR THE TOOL DEVELOPMENT

The following steps describe the process of the tool development as illustrated in Figure 2:

1. Set up a source information page that describes the locations of the data (we use an Excel sheet to store library names and the path/name of the source SAS Data files.)
2. Set up an interface in Excel using VBA to call SAS programs to generate outputs for data comparison.
3. Use the Excel interface to select datasets and variables of interest.
4. Display the data comparison results in Excel with color coded output to make the result easy to read. The attributes and values of each variable from multiple studies are displayed in separate Excel sheets. Any discrepancy from any study for each variable will be displayed in red. User can select specific studies and variables of interest.

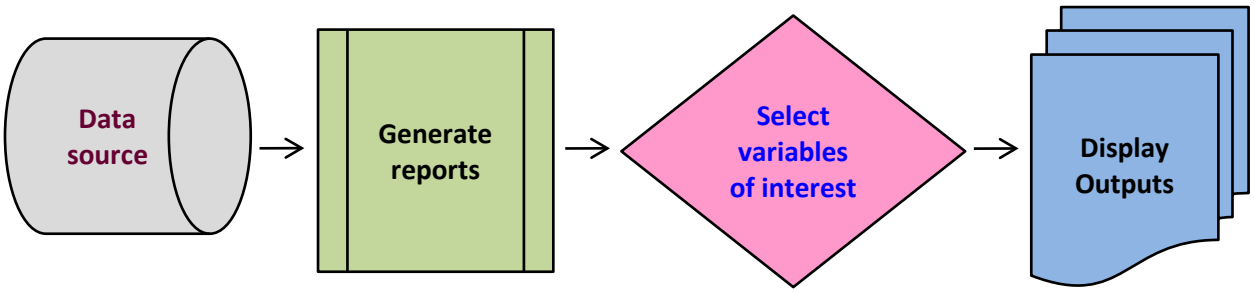


Figure 2. Process flow for the tool development

Data sources are SAS datasets. Reports are generated in SAS and displayed in Excel. Excel is also used as an interface for data source selection and output display. This is done by SAS and VBA. The SAS code is called from VBA scripts. The connection code is shown below.

```
'--> SAS IOM Workspace Connection Declarations
Public obWSM As New SASWorkspaceManager.WorkspaceManager
Public obWS As SAS.Workspace
Public obWSflag As Integer
Sub fetchFromSAS()
'--> ADO Database Connection Declarations
Dim obConn As New ADODB.Connection
Dim obRS As New ADODB.Recordset
Dim errorstring As String
Dim SAScode As String

'--> Create a Local SAS Workspace (One-Time Processing)
If obWSflag = 0 Then
Set obWS = obWSM.Workspaces.CreateWorkspaceByServer("", VisibilityProcess, Nothing,
"", "", errorstring)
obWSflag = 1
End If

'--> Set Path Vars
sPath = ActiveWorkbook.path
SAS_DataPath = VBAProject.UserForm1.Text_Browse.Value

'--> Call SAS code which is stored in string and declare variables
obWS.languageservice.submit SAScode
SASlog = obWS.languageservice.FlushLog(50000)
```

```
'--> First set a string which contains the path to the file you want to create.
'--> This example creates one and stores it in the root directory
MyFile = sPath & "/SASlog.log"

'--> Set and open file for output
fnum = FreeFile
Open MyFile For Output As fnum
Print #fnum, SASlog
Close #fnum
```

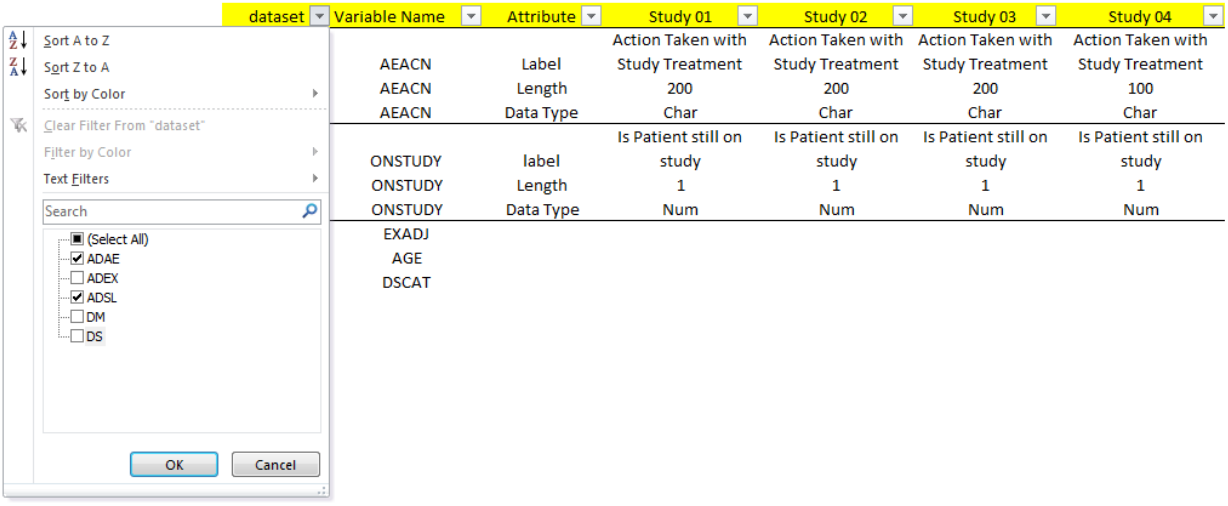
PROCESS FLOW OF HOW TO USE THE TOOL DATA SELECTION

Data selection interface is an Excel file. It has two steps:

1. User puts data source paths in the interface and all the datasets that exist in the source paths will be compared.
2. In the results page, the user will select the studies, datasets, and variables of interest for display purpose. Figure 3 and Figure 4 show the data selection interfaces in Excel.

Study	Data Source	Generate Report
Study 01	C:\Desktop\study 01	
Study 02	H:\My Documents\study 02	
Study 03	K:\Data\study 03	
Study 04	K:\Data\regulatory\study 04	
more...		

Figure 3. Excel interface that will be used to select studies and data sources.



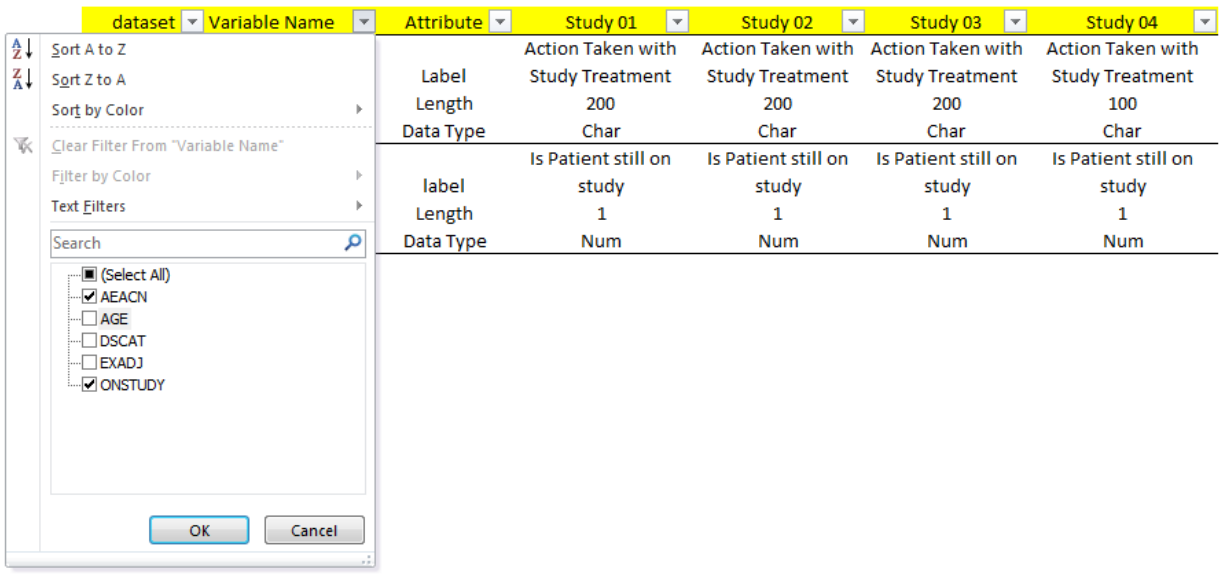


Figure 4. Excel interface to select studies and variables of interest.

RESULTS DISPLAY

Data comparison includes two aspects: attributes and values. They are displayed on two Excel sheets illustrated in Figure 5 and Figure 6.

Dataset	Variable Name	Attribute	Study 01	Study 02	Study 03	Study 04
ADAE	AEACN	Label	Action Taken with Study Treatment	Action Taken with Study Treatment	Action Taken with Study Treatment	Action Taken with Study Treatment
ADAE	AEACN	Length	200	200	200	100
ADAE	AEACN	type	Char	Char	Char	Char
ADSL	SAFFL	Label	Safety Population Flag	Safety Population Flag	Safety Population Flag	Safety Population Flag
ADSL	SAFFL	Length	1	1	1	1
ADSL	SAFFL	type	Char	Char	Char	Num
ADSL	onstudy	label	Is Patient still on study	Is Patient still on study	Is Patient still on study	Is Patient still on study
ADSL	onstudy	length	1	1	1	1
ADSL	onstudy	type	Num	Num	Num	Num

Figure 5. Variable attributes are displayed in an Excel sheet. The red text brings attention to the difference in variable length across studies for AEACN and the difference in variable type across studies for SAFFL.

Dataset	Variable Name	Value	Study 01	Study 02	Study 03	Study 04
ADSL	TRT1P	1.0 mg/kg			x	
ADSL	TRT1P	2.0 mg/kg				x
ADSL	TRT1P	VI	x	x		
ADSL	TRT1P	(MISSING)			x	
ADSL	SAFFL	Y	x	x	x	x
ADSL	SAFFL	N	x	x	x	x
ADSL	onstudy	1	x	x		
ADSL	onstudy	0	x			
ADSL	onstudy	(MISSING)		x	x	

Figure 6. Variable values are displayed in an Excel sheet. Variable TRT1P has values of "VI" in study 01 and 02, "1.0mg/kg" and missing values in study 03, and "2.0 mg/kg" in study 04. Variable SAFFL is consistently set to "N' or "Y" in all studies. Variable ONSTUDY is not available in study 04, has missing values in study 03, "1" or missing values in study 02, and "1" or "0" in study 01.

EXAMPLES OF HOW THE RESULTS FROM THIS TOOL HELP DATA INTEGRATION

VARIABLE LABEL AND LENGTH DIFFERENCES

When variables with the same name but different lengths are stacked by SAS, the following warning message is generated: "*WARNING: Multiple lengths were specified for the variable SAFFL by input data set(s).This may cause truncation of data*". When multiple studies are involved, without a tool, each study must be checked before the variable with the different length can be identified. With this integration tool, we can easily identify *all* the studies that contain variables with different lengths and modify them before stacking them together. The example below shows variable DTHAE has a different length in studies 03 and 04. It is noticeable that variable labels are different as well. This variable is defined differently among studies. Use of this tool will allow these discrepancies to be rectified prior to stacking the data.

Dataset	Variable Name	Attribute	Study 01	Study 02	Study 03	Study 04
ADAE	DTHAE	Label	Death due to AE	Death due to AE	Primary Cause of Death (Preferred term)	Death caused by this AE
ADAE	DTHAE	Length	20	20	200	12

Figure 7. Example of Variable Label and Length Differences

VARIABLE TYPE DIFFERENCES

When variables with different data types are stacked, an error message is generated and data cannot be pooled. With this tool, it is very easy to identify the studies that contain the discrepant variables and harmonize the data types prior to integration. An example can be found in Figure 5 where variable name SAFFL has a different data type in study 04.

VARIABLES WITH DIFFERENT DEFINITIONS AND VALUES

Variables with different values from different studies usually indicate those variables that are defined differently in each study. These variables need to be identified and issues have to be resolved before data can be analyzed. The example below shows variable DTHAE is defined differently in study 03 when compared with studies 01 and 02. Additionally, this variable is not available in study 04. Similar observations can be found in variable DTHREL.

Dataset	Variable Name	Value	Study 01	Study 02	Study 03	Study 04
ADAE	DTHAE	AE1			x	
ADAE	DTHAE	AE2			x	
ADAE	DTHAE	AE3			x	
ADAE	DTHAE	AE4			x	
ADAE	DTHAE	AE5			x	
ADAE	DTHAE	N	x	x		
ADAE	DTHAE	AE6			x	
ADAE	DTHAE	AE7			x	
ADAE	DTHAE	Y	x	x		
ADAE	DTHAE	(MISSING)			x	
ADAE	DTHREL	N	x	x	x	
ADAE	DTHREL	U			x	
ADAE	DTHREL	Y	x	x	x	
ADAE	DTHREL	(MISSING)	x	x		

Figure 8. Variable value and variable definitions are different.

CONCLUSION

This paper introduced a user-friendly tool that compares and summarizes datasets in a simple output so the data can be harmonized prior to integration. This tool is powerful because an unlimited number of studies and variables can be compared. Yet the tool is also flexible and easy to set up. All the sources are set up in an Excel interface, so it is very generic, easy to maintain and can be used in any data source structure folders. It identifies the variables that differ in attributes and values, to ensure the integrated datasets' variables have the same definitions. We recommend using this tool for any integration of data for meta-analysis, or to QC already integrated data before final analysis.

ACKNOWLEDGMENTS

We thank Norm Fox, Shawn Hopkins for their assistance of VBA scripts, Rajeev Karanam and Shefalica Chand for their valuable suggestions and comments.

CONTACT INFORMATION

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

Name: Yang Wang
Enterprise: Seattle Genetics, Inc.
Address: 21823 30th Drive Southeast
City, State ZIP: Bothell, WA 98021
E-mail: yawang@seagen.com

Name: Boxun Zhang
Enterprise: Seattle Genetics, Inc.
Address: 21823 30th Drive Southeast
City, State ZIP: Bothell, WA 98021
E-mail: bzhang@seagen.com