PharmaSUG China 2019 - Paper DM-045

A USER-CUSTOMIZED AND AUTOMATIC SDTM DATA CHECKING TOOL

Miao Yu, Beixin Lu, Xiao Fu, Zhijuan Yu, Zongfeng Lei, BeiGene Co., Ltd.

ABSTRACT

High quality of SDTM is essential to clinical trials and submission preparation. Current approaches of checking data issues need manual review and the issue summaries are often partial. In addition, current data checking tools are usually not convenient enough, which makes users difficult to learn and use. We need a standard and accessible method to maintain the data quality by finding the issues comprehensively and efficiently. In this paper, a data checking tool is proposed to fit the need. The tool consists of two parts. First is checking specification, an excel based user interface helps user to define the issues they are interested in checking. Second part is a set of backstage supportive function consists of VBA and SAS macros to generate and output issue summary. The biggest advantage of this checking tool is its ability to generate issue summary automatically with a simple click and it can prevent us from doing repeated works. We have tried this tool in several studies, and it is proved to be user-friendly and helpful according to feedbacks. The idea of replacing manual work with automatic tools can be adopted to and improve every aspect of our work in this industry. This paper gives a detailed description of developing and using this tool.

INTRODUCTION

In Pharmaceutical industry, data are collected in Case Report Forms (CRF) from different sites. Although, the data collected in CRFs and Study Data Tabulation Model (SDTM) need to follow the CDISC standards. The various sources of data entry may cause heterogeneity in data format and type. Also, heterogeneity and mess in data may cause trouble when developing SDTM and further impact Analysis Data Model (ADaM) datasets and Table/Listing/Figure (TLF). Therefore, it's especially important to keep the source datasets, SDTM, in high quality.

To maintain data quality, data issue checking is one of the most important methods. Although the Data Management team have already established validated checking system to clean the raw dataset, the programming team also needs to take responsibility to assure the data quality. So that data issue checking still cannot be skipped in programming process. Currently, we check issues by going through data manually or finding issues while building SDTM, ADaM or TLFs. Traditional way is time-consuming and finding results are partial. A more efficient and well-organized way is necessary to support programmers in data issue checking.

Now, we propose a data checking tool to provide convenience to our daily checking process. There are four main advantages of this tool:

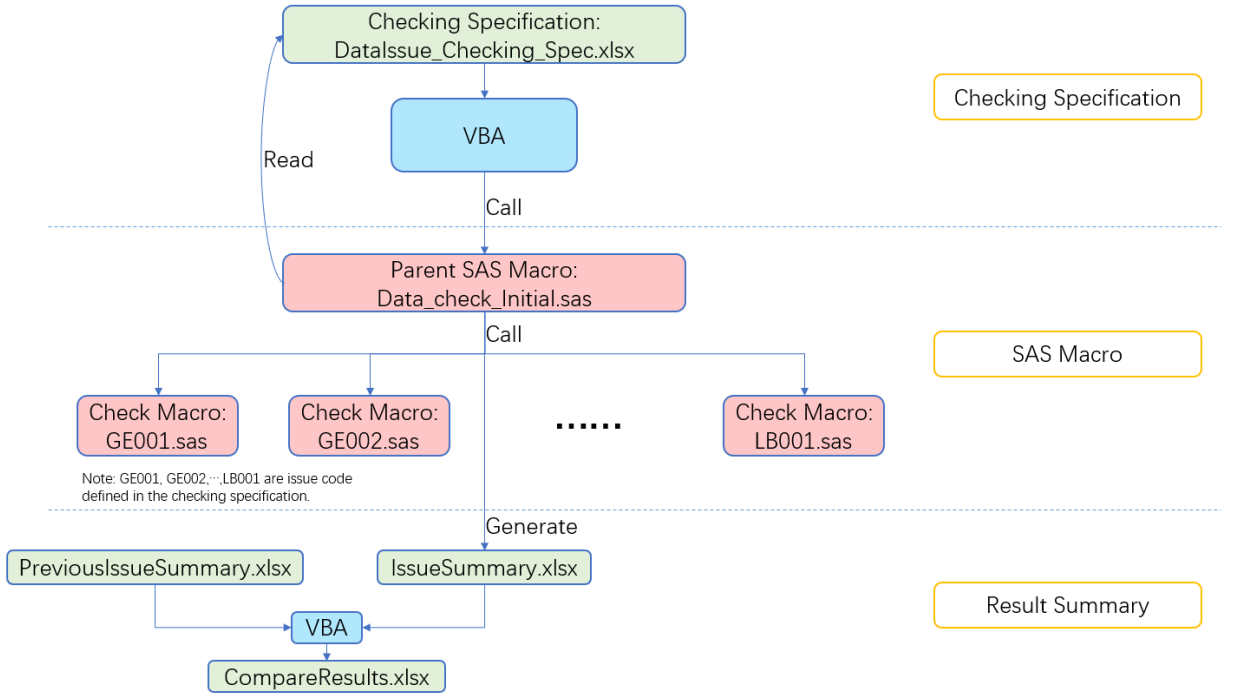
1. Easy to use with automatic generation of issue summary. There is no need to run programs by users.
2. The data issue library can be defined and expanded by user. Everyone can utilize the tool's platform and build their own checking by defining their own study-specific checking rules
3. The issue summary can be used directly to report issues to data management group or SDTM mapping vendor.
4. Different versions of comparison summaries can be compared automatically and comparation result will be output, which can aware us of newly arisen issues.

The tool is established on VBA and SAS language. And the tool will be described in detail in the following sections.

TOOL BUILDING

TOOL STRUCTURE

Here is the structure of the data checking tool：



Display 1. Data Checking Tool Structure

This data checking tool can be divided into 3 parts: Checking Specification, SAS Macro and Result Summary. The organization and mechanism of the tool are described as follows.

Checking Specification

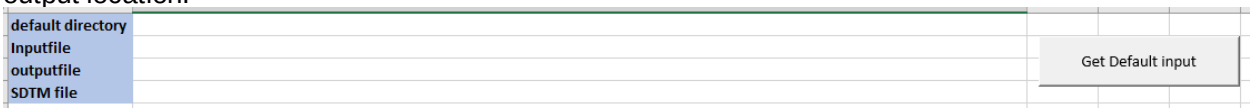
As shown in Display 1. Data Checking Tool Structure, our tool is managed by an excel based user interface, Checking Specification. This excel file can help us read in required inputs, user customizations and call backstage SAS macros.

Input Panel

In company, we usually have multiple studies ongoing. To apply this tool to the SDTM data of specific study, required inputs as the specific SDTM directory and I/O directory are needed. To extract those directories, Users need to copy the Checking_Specification.xlsx to study folder and click “Get default input” in the tool as shown in Display 2. All the directories can also be modified manually after extraction. To obtain the data path, the user will need to copy the Checking_Specification.xlsx to the study folder. As shown in Display 2, clicking “Get default input” in the tool can give us the extracted folder paths. All the directories can also be modified manually after extraction.

We use VBA to identity the study and data path by extracting the directory of copied Checking_Specification.xlsx, which will be read by SAS macros to identify the necessary file path and

output location.



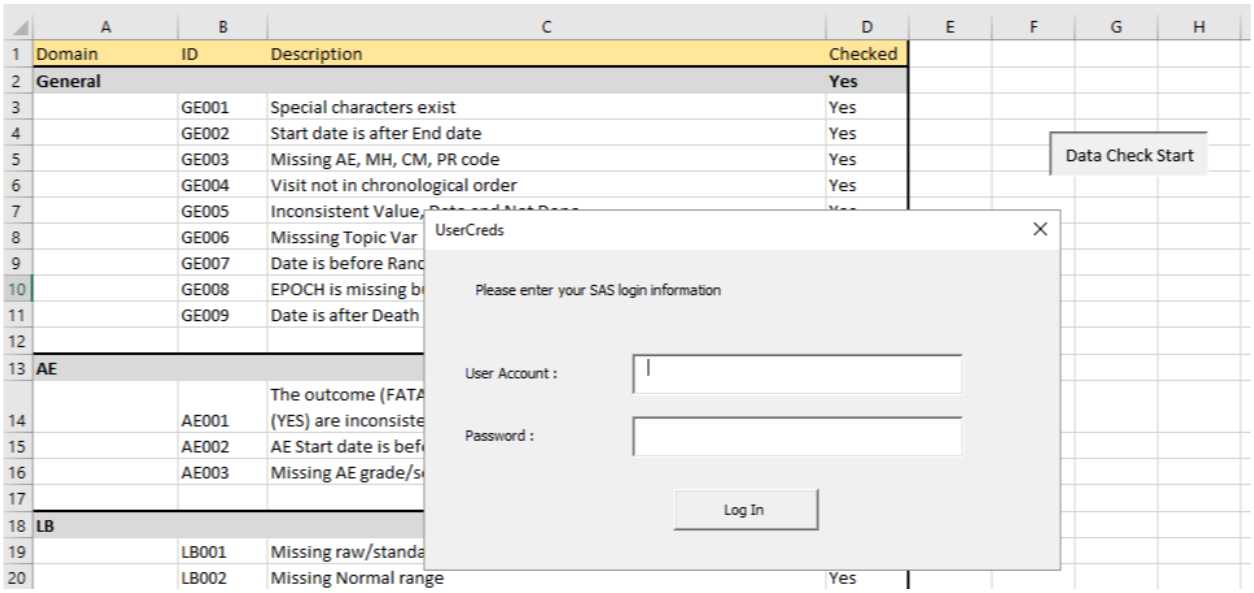
Display 2. Checking Specification: get default location

Checking Content Panel

	A	B	C	D	E	F	G	H
1	Domain	ID	Description	Checked				
2	**General**			**Yes**				
3		GE001	Special characters exist	Yes				
4		GE002	Start date is after End date	Yes				
5		GE003	Missing AE, MH, CM, PR code	Yes			Data Check Start	
6		GE004	Visit not in chronological order	Yes				
7		GE005	Inconsistent Value, Date and Not Done	Yes				
8		GE006	Misssing Topic Var	Yes				
9		GE007	Date is before Random date	Yes				
10		GE008	EPOCH is missing but with non-missing date	Yes				
11		GE009	Date is after Death Date	Yes				
12								
13	**AE**			**Yes**				
14		AE001	The outcome (FATAL), AE grade (GRADE 5) and resulting in death (YES) are inconsistent.	Yes				
15		AE002	AE Start date is before the 1st dose date	Yes				
16		AE003	Missing AE grade/severity	Yes				
17								
18	**LB**			**Yes**				
19		LB001	Missing raw/standard unit	Yes				
20		LB002	Missing Normal range	Yes				
21		LB003	Potential Outliers	Yes				
22								
23	**DD**			**Yes**				
24		DD001	Partial or missing Death date	Yes				
25		DD002	Death reason is unknown	Yes				
26								
27	**DM**			**Yes**				
28		DM001	Missing SEX	Yes				
29								
30	**DS**			**Yes**				
31		DS001	Multiple study discontinuation records exist per one patient	Yes				

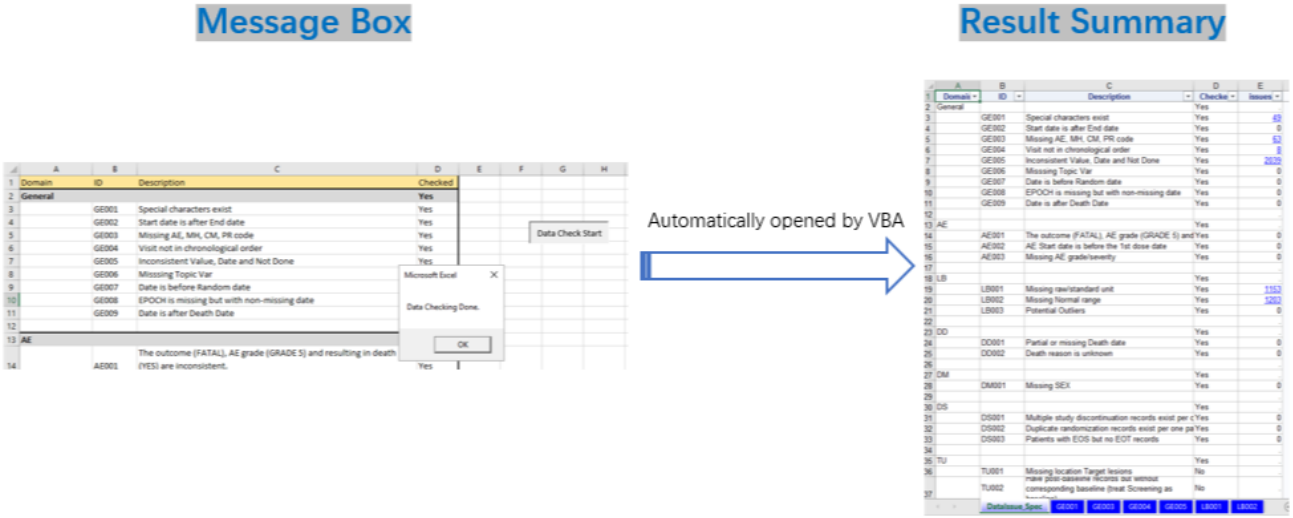
Display 3 Checking Specification: specify checking issues

Display 3 is the main panel in our Checking_Specification.xlsx file. It is the issue checking library ordered by SDTM domain, where “General” means the issue is checked across all the domains, for example. All the issues are assigned with codes like “GE001”, indicating issue number 1 in general domain, which simplify the way we represent a data issue. In Description Column, the issues are described in detail. The ‘Checked’ column with value “Yes”/ “No” defines which domain and specific issue need to be checked. If it is selected as “No”, the checking will not be implemented by not calling its corresponding checking macro. After customizing and saving the excel file, the checking summary will be one click away. By clicking “Data Check Start” button, the backstage VBA program will connect to SAS server and call the SAS macros to generate the data checking issue summary, as shown in Display 4.



Display 4. Connecting to SAS server

When the check is completed, there will be a pop-up message box notifying us that the summary result is ready. By clicking the message box, the result summary will open automatically, saving us time from finding and opening the summary manually.



There is also a domain called “User_Define” to allow the user to customize this checking tool, given the facts that the user may want to explore more possibilities. Therefore, we offer a ‘User_Define’ panel so that they can write their own checking macros.

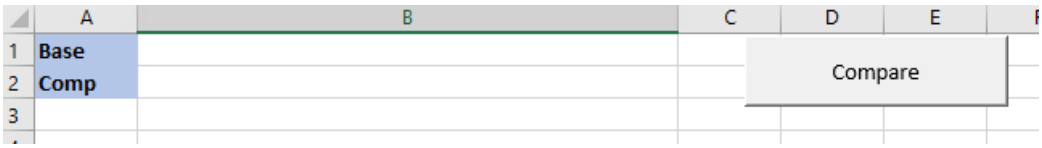
User_Define			Yes
		1.Please assign a code for your checking following "UD00X" format and put it in column B. 2. Add detailed description of your checking in column C. 3. Put your check macro in input directory as in sheet "Input". (The macro should follow some specific rules so that it can be read by the program, and the macro name should be in **lowcase**. The detailed instruction will be provided along with the macro template. Please refer to the template **ud00x.sas** at X:\Biometrics\04-Innovation\03-Incubator\SH_data_checking\v2.0.)	
	UD00X	4. Assign "Yes" in column "D", save it and run.	Yes

Display 5. Checking Specification: User_Define domain

Compare Panel

Besides the issue summary of current SDTM, we are also interested if the issues are new or stale in previous data. This panel helps us to determine if the two checking results are the same. The comparation results are shown in the next section.

After filling out the two summary file names of the checking results that you want to compare and click "Compare" button, we will get a newly generated file containing the comparation results.



Display 6. Checking Specification: compare checking results

SAS macro

Individual checking macros

The data issues are various and difficult to be covered thoroughly. To enable the tool to have expandable library, we use separate macro for individual data issue. Therefore, to add a new issue checking, we only need to develop its own checking macro test it in different studies and then add it to the Checking Specification.

Parent macro to execute the checking

Besides the individual checking macros, we also developed a parent macro. The parent macro will execute individual checking macros according to the input information defined by the Checking Specification. The parent macro will do the following work and the description of the attached codes are highlighted in green:

1. Read the Checking Specification.
    1.1 Read the data and the file path specified in "Input sheet" in specification as shown in Display 2.
    1.2 Read the checking specification's ID and the "Checked" column, which will determine if a macro will be called or not.

    Here are related codes, consists of VBA and SAS languages.

```
/*************************************************************************************************************/
VBA:

' 1. Connect to sas server
Set obSAS = obObjectFactory.CreateObjectByServer("sasserver1", True, observer, Username, Password)
' 2. Include parent SAS macro
obSAS.LanguageService.Submit ("options source2; %include '/YOURPATH/data_check.sas'; ")
' 3. Get the location of Checking Specification, Issue Summary and SDTM datasets
Dim outfile As String
Dim sdtmpath As String
Dim inpath As String
inpath = ThisWorkbook.Worksheets("Input").Cells(2, 2).Value
outpath = ThisWorkbook.Worksheets("Input").Cells(3, 2).Value
sdtmpath = ThisWorkbook.Worksheets("Input").Cells(4, 2).Value
' 4. call the data_check macro
obSAS.LanguageService.Submit ("options source2;
%data_check(infile=" & inpath & ", outfile = " & outpath & ", sdtm= " & sdtmpath & "); ")
/*************************************************************************************************************/


/*************************************************************************************************************/
SAS:

%macro data_check(infile=, outfile=,sdtm=);
/*1. Set up SDTM and import Checking Specification*/
libname sdtm "&sdtm.";
libname general xlsx "&infile.";
/*2. Identify issues to be checked*/
data general;
  set general.dataissue_spec;
  id=strip(id);
run;

**summary issues need to check;
data check;
  set general;
  if ^missing(ID) and upcase(checked)="YES";
run;
/*************************************************************************************************************/
```

2. Call the checking macros.

Call the checking macros indicated by the “Checked” column.

```
/*************************************************************************************************************/
SAS:

/*Call the corresponding macros of data issues*/
data _null_ ;
  set check;
  include="%nrstr(%include 'YOURPATH"||strip(ID)||".sas'; )";
  call execute(include);
  call_macro='%nrstr(%'||strip(ID)||');';
  call execute(call_macro);
run;
/*************************************************************************************************************/
```

3. Output the checking result to excel file.

   Using PROC REPORT can easily generate excel file to restore these checking issues. The filter and header format also be added during the PROC REPORT, which can help Data Management team to better understanding the issues listed in the result summary.

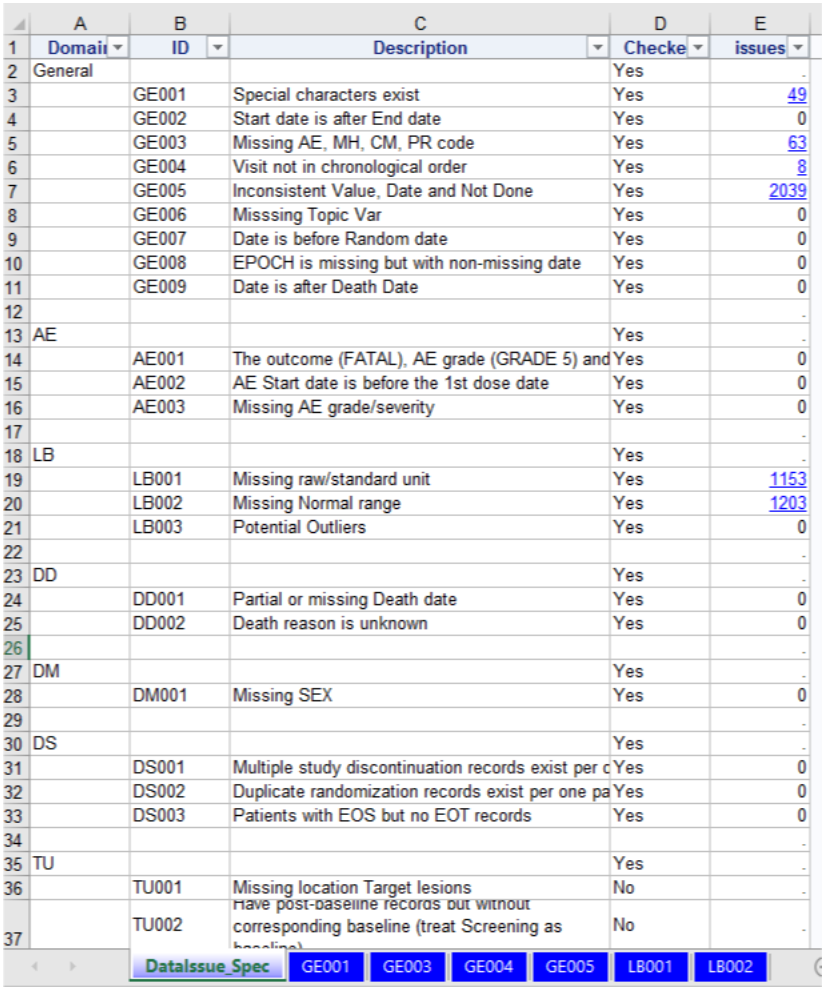
4. Compare data checking results.

   Comparing different version of data checking results is supported by VBA. The example code is shown as follows.

```
/*************************************************************************************************************/
VBA:

' 1. Compare the number of sheets
        Dim nbasesheet, ncompsheet As Integer
        Dim allcompsheet, allissue As String

        For ncompsheet = 1 To compwb.Sheets.Count
            allcompsheet = allcompsheet & "|" & compwb.Sheets(ncompsheet).Name
        Next ncompsheet

        For nbasesheet = 1 To basewb.Sheets.Count

            If InStr(allcompsheet, basewb.Sheets(nbasesheet).Name) = 0 Then
                '3=Red , 4=green,5=blue,6=yellow
                'flag for New issue sheet
                basewb.Sheets(nbasesheet).Tab.ColorIndex = 6
            End If
            If InStr(allcompsheet, basewb.Sheets(nbasesheet).Name) > 0 Then
                allissue = allissue & """,""" & basewb.Sheets(nbasesheet).Name
            End If

        Next nbasesheet
' 2. Compare the navigation page
    allissue = allissue & ""","""

    Dim issue As Variant
    Dim nissue As Integer
    issue = Split(allissue, ",")

    For nissue = 0 To UBound(issue)

        If issue(nissue) <> """" And issue(nissue) <> "" Then

            Dim theIssue
            theIssue = Replace(issue(nissue), """", "")

            Call copy(theIssue)
            Call detail_compare(theIssue)
            Call clear

        End If

    Next nissue
/*************************************************************************************************************/
```

Result summary

Navigation Page



	A	B	C	D	E
1	Domain	ID	Description	Checke	issues
2	General			Yes	
3		GE001	Special characters exist	Yes	49
4		GE002	Start date is after End date	Yes	0
5		GE003	Missing AE, MH, CM, PR code	Yes	63
6		GE004	Visit not in chronological order	Yes	8
7		GE005	Inconsistent Value, Date and Not Done	Yes	2039
8		GE006	Misssing Topic Var	Yes	0
9		GE007	Date is before Random date	Yes	0
10		GE008	EPOCH is missing but with non-missing date	Yes	0
11		GE009	Date is after Death Date	Yes	0
12					
13	AE			Yes	
14		AE001	The outcome (FATAL), AE grade (GRADE 5) and	Yes	0
15		AE002	AE Start date is before the 1st dose date	Yes	0
16		AE003	Missing AE grade/severity	Yes	0
17					
18	LB			Yes	
19		LB001	Missing raw/standard unit	Yes	1153
20		LB002	Missing Normal range	Yes	1203
21		LB003	Potential Outliers	Yes	0
22					
23	DD			Yes	
24		DD001	Partial or missing Death date	Yes	0
25		DD002	Death reason is unknown	Yes	0
26					
27	DM			Yes	
28		DM001	Missing SEX	Yes	0
29					
30	DS			Yes	
31		DS001	Multiple study discontinuation records exist per c	Yes	0
32		DS002	Duplicate randomization records exist per one pa	Yes	0
33		DS003	Patients with EOS but no EOT records	Yes	0
34					
35	TU			Yes	
36		TU001	Missing location Target lesions	No	
37		TU002	Have post-baseline records but without corresponding baseline (treat Screening as	No	

DataIssue_Spec | GE001 | GE003 | GE004 | GE005 | LB001 | LB002

Display 7. Navigation Page

As shown in Display 7. Navigation Page Display 7, the Navigation Page has the same layout as Checking Specification (Display 1) with new column “issues” indication the number of issues we found. By clicking the number, it will direct you to the corresponding sheet, which contains the listing of the data issues.

Issue listing

Take GE001 as an example: Special Character Exists

A	B	C	D	E	F	G
domain	usubjid	general_seq (Sequence)	raw_value (Original value in SDTM)	special_var (Vars contain non-printable)	non_print (non printable special characters)	ASCII (ASCII characters for non printable special characters)
LB	BGB-2019-999-00-000	LBSEQ = 42	JIAMUSI CENTRALÂ HOSPITAL	LBNAM	Â	byte(194)
LB	BGB-2019-999-00-001	LBSEQ = 43	JIAMUSI CENTRALÂ HOSPITAL	LBNAM	Â	byte(194)
LB	BGB-2019-999-00-002	LBSEQ = 44	JIAMUSI CENTRALÂ HOSPITAL	LBNAM	Â	byte(194)

Display 8. Issue listing

As shown in Display 8, we can see the reason of issue and easily go back to the original data. In addition, the label of each variable is also displayed in the column to help to communicate with Data Management team smoothly.

Compare Results

	A	B	C	D	E	F
1	Domai	ID	Description	Checke	issues	Status
2	General			Yes	.	
3		GE001	Special characters exist	Yes	49	New
4		GE002	Start date is after End date	Yes	0	Existing
5		GE003	Missing AE, MH, CM, PR code	Yes	63	New
6		GE004	Visit not in chronological order	Yes	8	Existing
7		GE005	Inconsistent Value, Date and Not Done	Yes	2039	Existing
8		GE006	Misssing Topic Var	Yes	0	Existing
9		GE007	Date is before Random date	Yes	0	Existing
10		GE008	EPOCH is missing but with non-missing date	Yes	0	Existing
11		GE009	Date is after Death Date	Yes	0	Existing
12					.	
13	AE			Yes	.	
14		AE001	The outcome (FATAL), AE grade (GRADE 5) and	Yes	0	Existing
15		AE002	AE Start date is before the 1st dose date	Yes	0	Existing
16		AE003	Missing AE grade/severity	Yes	0	Existing
17					.	
18	LB			Yes	.	
19		LB001	Missing raw/standard unit	Yes	1153	Existing
20		LB002	Missing Normal range	Yes	1203	Existing
21		LB003	Potential Outliers	Yes	0	Existing
22					.	
23	DD			Yes	.	
24		DD001	Partial or missing Death date	Yes	0	Existing
25		DD002	Death reason is unknown	Yes	0	Existing
26					.	
27	DM			Yes	.	
28		DM001	Missing SEX	Yes	0	Existing
29					.	
30	DS			Yes	.	
31		DS001	Multiple study discontinuation records exist per c	Yes	0	Existing
32		DS002	Duplicate randomization records exist per one pa	Yes	0	Existing
33		DS003	Patients with EOS but no EOT records	Yes	0	Existing
34					.	
35	TU			Yes	.	
36		TU001	Missing location Target lesions	No	.	Existing
37		TU002	Have post-baseline records but without corresponding baseline (treat Screening as	No	.	Existing

DataIssue_Spec | GE001 | GE003 | GE004 | GE005 | LB001 | LB002

Display 9. Compare results

As shown in Display 9, we can see that GE001: Special character exist is labeled as ‘New’ while others are ‘Existing’, indicating that there are new issues only of GE001. This function can save your time from searching for discriminating newly discovered issues from existing ones. Here is the detailed new issue in GE001.

	A	B	C	D	E	F	G	H
1	domain	usubjid	general_seq (Sequence)	raw_value (Original value in SDTM)	special_var (Vars contain non-printable)	non_print (non printable special characters)	ASCII (ASCII characters for non printable special characters)	Status
2	LB	BGB-A317-999-00-00	LBSEQ = 214	ERDOS HOSPITAL OF TRADITIONAL CHINESE MEDICINE	LBNAM		byte(9)	New
3	LB	BGB-A317-999-00-01	LBSEQ = 215	ERDOS HOSPITAL OF TRADITIONAL CHINESE MEDICINE	LBNAM		byte(9)	Existing
4	LB	BGB-A317-999-00-02	LBSEQ = 216	ERDOS HOSPITAL OF TRADITIONAL CHINESE MEDICINE	LBNAM		byte(9)	Existing

As for GE003, it is marked by yellow and its status is “New”, which means GE003 type of issues are newly arisen with no common issues in previous check version. Here is the detailed new issue in GE003.

	A	B	C	D	E	F	G	H
1	DOMAIN (Domain Abbreviation)	USUBJID (Unique Subject Identifier)	SEQ (Sequence Number)	TERM (Reported Term for the Adverse Event)	LLTCD (Lowest Level Term Code)	PTCD (Preferred Term Code)	HLTCD (High Level Term Code)	Status
2	CM	BGB-A317-999-00-00	59	METHIMIDAZOLE TABLET	.	.	.	New
3	CM	BGB-A317-999-00-01	56	RABEPRAZOLE SODIUM FOR INJECTION	.	.	.	New
4	MH	BGB-A317-999-00-02	1	SQUAMOUS NSCLC	.	.	.	New
5	MH	BGB-A317-999-00-03	1	SCLC	.	.	.	New
6	MH	BGB-A317-999-00-04	1	SQUAMOUS NSCLC	.	.	.	New

CONCLUSION

To ensure high quality of SDTM datasets, lots of efforts and time have been taken during our daily work. It is essential for us to make this process efficient. This data checking tool has been used in several internal studies and proved to be a very good practice. Based on the current tool, we can improve it by adding more checking macros and adapt it to raw data, ADaM data checking. In addition, we can seek opportunity to impact other aspects of our work by developing more and more automatic tools.

ACKNOWLEDGMENTS

Sincerely thanks to our data checking group and our programming team for their help and support.

RECOMMENDED READING

- *Base SAS® Procedures Guide*
- *SAS® For Dummies®*
- *Accessing SAS® Code via Visual Basic for Applications*
- *Jennifer Davies, Z, Inc, Silver Spring, MD, Paper 306-2013 SAS Global Forum 2013*
- *Not Just for Scheduling: Doing More with SAS® Enterprise Guide® Automation, Chris Hemedinger, SAS Institute Inc, Cary, NC, Paper 298-2012 SAS Global Forum 2012*
- *Using VBA to Debug and Run SAS® Program Interactively, Run Batch Jobs, Automate Output, and Build Applications*
- *SAS Enterprise Guide Scripts "http://support.sas.com/documentation/onlinedoc/guide/"*

CONTACT INFORMATION

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

<Miao Yu>
<BeiGene Inc.>
<miao.yu@beigene.com >

<Beixin Lu>
<BeiGene Inc.>
<beixin.lu@beigene.com >

<Xiao Fu>
<BeiGene Inc.>
<xiao.fu@beigene.com >

<Zhijuan Yu>
<BeiGene Inc.>
<zhijuan.yu@beigene.com >

<Zongfeng Lei>
<BeiGene Inc.>
<Zongfeng.lei@beigene.com >