PharmaSUG 2018 - Paper AD-16

An Efficient Tool for Clinical Data Check

Chao Su, Merck & Co., Inc., Rahway, NJ
Shunbing Zhao, Merck & Co., Inc., Rahway, NJ
Cynthia He, Merck & Co., Inc., Rahway, NJ

ABSTRACT

High-quality data in clinical trials is essential for compliance with Good Clinical Practice (GCP) and regulatory requirements. If data issues are observed during statistical analysis, the interpretation of study results is impacted. As such, statistical programmers may be asked to perform data validity checks and detect potential data issues across multiple datasets during the analysis and reporting (A&R) processes. The focus of this paper is NOT on finding data issues prospectively, but on how to accurately and efficiently track findings during the course of statistical analysis and programming processes. In this paper, some routinely performed data checks are discussed briefly. The results of the data checks are generated and summarized through a macro consisting of two sub-macros. The macro combines the results into one Excel file and compares the current check results with the ones in the previous version automatically. The changes between two versions are marked in different colors so that the reviewers can identify pending issues quickly. This tool provides real time documentation that can track data issues related to the A&R process accurately and efficiently.

INTRODUCTION

In the pharmaceutical industry, statistical analysis and reports are based on clinical data which must be of high quality. During the trial, the data management (DM) team applies many edit checks to clean the data. In practice, most of the checks are univariable and within the same domain. As a result, discrepancies across multiple variables and mismatches among different datasets might exist after edit checks are complete with DM tools. Additional data issues may be identified using SAS programs for analyses and reporting.

Besides identifying data issues, reporting and tracking data issues are also critical. Both the clinical team and the statisticians need to review the data issue report and include comments and feedback before the report with issue findings from SAS programs is sent to the DM. The DM will correct or query the issues according to the report. Excel is commonly used among professionals and has the flexibility required to integrate different comments into the tracking report. Therefore, it is used here as a tool to bring all the addressed data issues and feedback together. All the data issues captured by SAS programs from different sources are integrated into a single Excel workbook with multiple worksheets.

Although Excel is flexible and easy to use, one needs to put a great deal of time-consuming manual effort on adjustments like column formatting to produce a tracking report with a user friendly interface. , The level of effort is incompatible with frequent reports. A macro developed to automatically output the tracking spreadsheet with a customized interface is necessary. Historical information like comments from reviewers and the date on which the issue was found are kept in the report. Updates of issue contents are highlighted with different colors so that it is very easy for the reviewer to track activities among different roles and determine issue status.

DATA CHECK PROCEDURES

The purpose of the data check here is to record and track data issues found during statistical analysis and reporting. It works as a supplement instead of a replacement to the data edit checks done by DM. The first step of the data check is to create the data check specification. The programmers will consult with statisticians and the clinical team to draft the data checks specification for a given project. DM will review the data check specification to avoid duplicate work by the DM edit check tool. After the specification is ready, the study programmers will follow the specification to identify issues. Different SAS datasets are generated for the various issue categories. These SAS datasets work as the input source for the developed macro to create a single Excel file with one worksheet for each issue category. Another worksheet with a status summary of the data issues is also provided by the macro. If a previous version of issue spreadsheet exists, the macro can compare the new spreadsheet with the old one. The comments from the previous version of the spreadsheet will be carried over to the new report if the issue still exists there. Different colors of cells are used to distinguish whether the issue was reported previously, or if it is a new one. This will facilitate reviewers to track the issues status easily and efficiently.

DATA CHECK ITEMS

It is important to check the data in advance so that there is sufficient time to query and correct data issues before the database is locked. Some examples of items to check are listed as below.

1. Discrepancies of randomized subjects:
   (1) Subject with randomization number but without reference date
   (2) Subject with randomization number but without randomization date
   (3) Subject randomized but with a record of screen failure in the Disposition dataset.

2. Discrepancies of death subjects:
   (1) Subject with a death record in the Death dataset but not in the Disposition dataset
   (2) Subject with a death record in the Disposition dataset but not in the Death dataset
   (3) Subject with different death dates between the Death dataset and the Disposition dataset

3. Discrepancies of administered drug:
   (1) Administered drug dose amount mismatched with the dispensed dose amount
   (2) Administered drug dose start date behind stop date
   (3) Administered drug dose with overlapped date

4. Discrepancies of administered drug and disposition datasets:
   (1) Administered drug dose date behind treatment or study discontinued date
   (2) Gap between last dose date and treatment discontinued date
   (3) Treatment discontinued date behind study discontinued date
   (4) Subject with Administered drug but not randomized

INDIVIDUAL WORKSHEET BUILDING

After the data check specification is ready, a SAS program is developed to generate SAS datasets according to different data issue categories. All the SAS datasets are exported into one excel file with one worksheet for each data issue category. The SAS code (key part) is shown below:

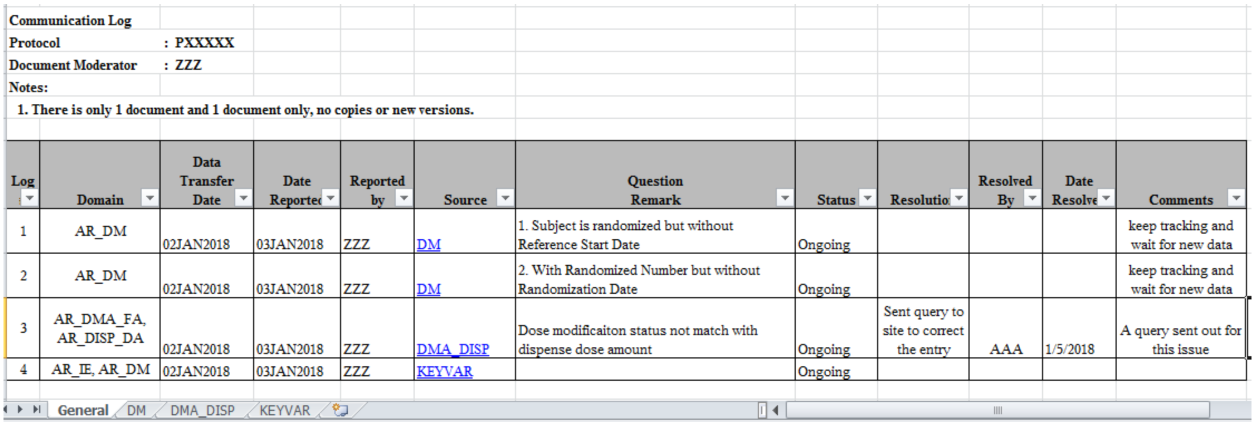
```
*=-------------- Generate Excel file with multiple worksheets -------*;

%macro ExcelGen(outfile=%str(date_issue_check), outsheet=, indata=,
tab=, desc=);
*--- Check whether dataset is empty ----;
 %let Exist = No;
 %let NumObs = 0;

 %if %sysfunc(exist(&InData)) %then %let Exist = Yes;

 %if &Exist = Yes %then %do;
    %let DSNId  = %sysfunc(open(&InData));
    %let DSObs  = %sysfunc(attrn(&DSNId.,nobs));
    %let rc     = %sysfunc(close(&DSNId.));
    %let NumObs = &DSObs.;
 %end;

%if &numobs > 0 %then %do;
    proc sql;
      create table &indata._ as
        select *, "" as DM_Comments length 200 from
            &indata. ;
    quit;

    proc export data=&indata._
      outfile="&path.\Data Issues.xls"
      dbms= excelcs REPLACE;
      sheet="&outsheet";        *--- The name of checked issue type ---;
    run;
%end;
%else %do;
  %put There is no record at dataset &indata. ;
%end;
%mend ExcelGen;
```

According to the data check specification, this macro will be called repeatedly to generate an individual worksheet when a dataset is created and checked. All the worksheets are saved into one excel file for further use.

WORKSHEET FORMAT AND SUMMARY BUILDING

The excel file created by the macro ExcelGen contains all detailed information about the data issues, but has not yet been formatted. A sub macro is developed to format the cells at each worksheet such as defining the width of columns, setting filter, and page orientation.

In order to give reviewers an overview of all data issues, a new worksheet “general” is created to summarize the data issues from all worksheets, as shown in Fig. 1, where the column “domain” is the raw dataset names; the column “Source” is the hyperlink to the individual worksheet; the column “Question Remark” is the description of the data issue. A column “Comments” is added to adopt comments from reviewers. Additional notes could be added to the “General” worksheet as shown on the top of Fig. 1.



Communication Log
Protocol : PXXXXX
Document Moderator : ZZZ
Notes:
1. There is only 1 document and 1 document only, no copies or new versions.

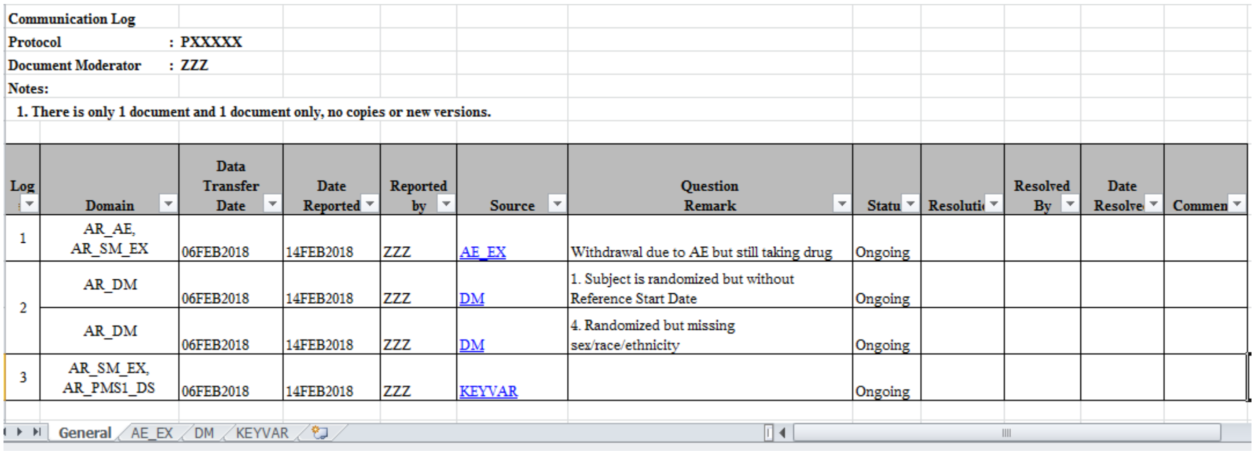
Log	Domain	Data Transfer Date	Date Reportec	Reported by	Source	Question Remark	Status	Resolutio	Resolved By	Date Resolve	Comments
1	AR_DM	02JAN2018	03JAN2018	ZZZ	DM	1. Subject is randomized but without Reference Start Date	Ongoing				keep tracking and wait for new data
2	AR_DM	02JAN2018	03JAN2018	ZZZ	DM	2. With Randomized Number but without Randomization Date	Ongoing				keep tracking and wait for new data
3	AR_DMA_FA, AR_DISP_DA	02JAN2018	03JAN2018	ZZZ	DMA_DISP	Dose modificaiton status not match with dispense dose amount	Ongoing	Sent query to site to correct the entry	AAA	1/5/2018	A query sent out for this issue
4	AR_IE, AR_DM	02JAN2018	03JAN2018	ZZZ	KEYVAR		Ongoing				

Fig. 1 Summary table of previous data issue tracking

COMPARISON OF TRACKING SHEETS

For a long running trial, not only is a data issue report needed once, but also a cumulative data issue sheet is required to track the data issues and their status. A sub- macro “edit0check0compare_excel” (key part is shown below) is developed to create the cumulative data issue spreadsheet, which will compare the current spreadsheet shown in Fig. 2 with the previous version (Fig. 1) to track the issue status, as marked with different colors. The summary table of the output is shown in Fig. 3. In this sub-macro, yellow means new data issues; red means data issues exist but have occurred before; green means this kind of data issue occurred before but is resolved now. Comparing Fig.1 and Fig.2, the data issue at source AE_EX is a new one. Therefore, it is marked with yellow. The issue at source DMA_DISP, existed in Fig.1 but not in Fig.2. As a result, it is marked with green to show that it is solved at in the new data source. The issues existing in both Fig.1 and Fig.2 are marked with red.

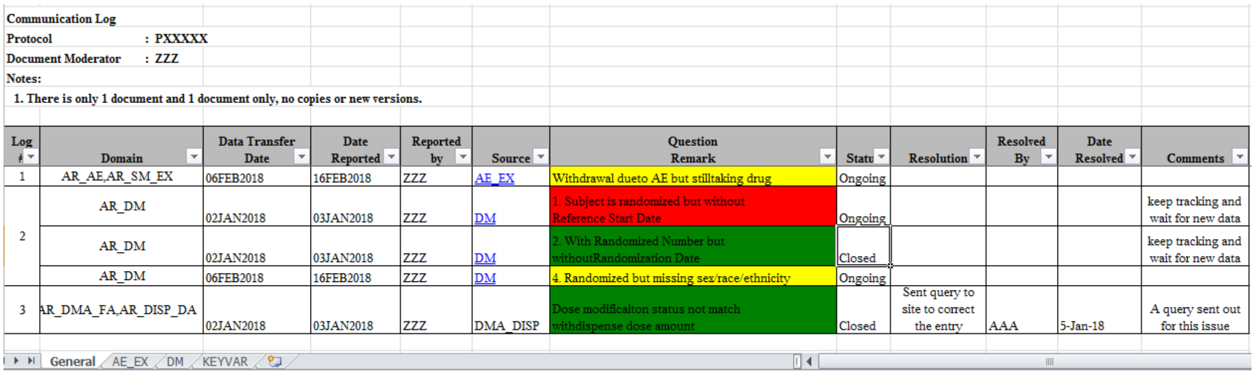
Fig. 4 is a screen shot of an individual worksheet in the cumulative report.



Communication Log
Protocol : PXXXXX
Document Moderator : ZZZ
Notes:
1. There is only 1 document and 1 document only, no copies or new versions.

Log #	Domain	Data Transfer Date	Date Reported	Reported by	Source	Question Remark	Statu	Resoluti	Resolved By	Date Resolve	Commen
1	AR_AE, AR_SM_EX	06FEB2018	14FEB2018	ZZZ	AE_EX	Withdrawal due to AE but still taking drug	Ongoing				
2	AR_DM	06FEB2018	14FEB2018	ZZZ	DM	1. Subject is randomized but without Reference Start Date	Ongoing				
	AR_DM	06FEB2018	14FEB2018	ZZZ	DM	4. Randomized but missing sex/race/ethnicity	Ongoing				
3	AR_SM_EX, AR_PMS1_DS	06FEB2018	14FEB2018	ZZZ	KEYVAR		Ongoing				

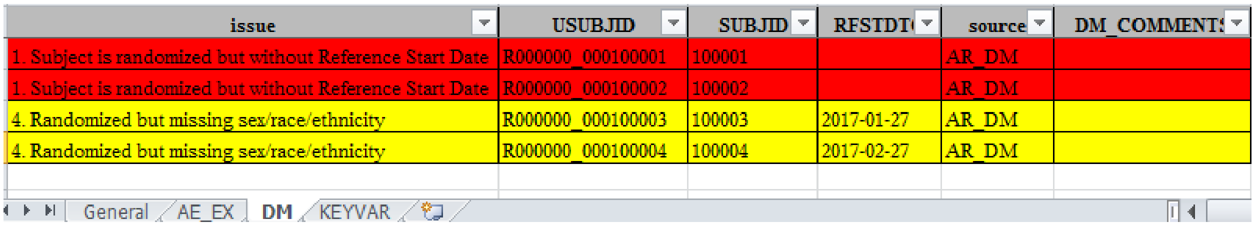
Fig. 2 Summary table of current data issue tracking



Communication Log
Protocol : PXXXXX
Document Moderator : ZZZ
Notes:
1. There is only 1 document and 1 document only, no copies or new versions.

Log #	Domain	Data Transfer Date	Date Reported	Reported by	Source	Question Remark	Statu	Resolution	Resolved By	Date Resolved	Comments
1	AR_AE,AR_SM_EX	06FEB2018	16FEB2018	ZZZ	AE_EX	Withdrawal dueto AE but stilltaking drug	Ongoing				
2	AR_DM	02JAN2018	03JAN2018	ZZZ	DM	1. Subject is randomized but without Reference Start Date	Ongoing				keep tracking and wait for new data
	AR_DM	02JAN2018	03JAN2018	ZZZ	DM	2. With Randomized Number but withoutRandomization Date	Closed				keep tracking and wait for new data
	AR_DM	06FEB2018	16FEB2018	ZZZ	DM	4. Randomized but missing sex/race/ethnicity	Ongoing				
3	AR_DMA_FA,AR_DISP_DA	02JAN2018	03JAN2018	ZZZ	DMA_DISP	Dose modificaiton status not match withdispense dose amount	Closed	Sent query to site to correct the entry	AAA	5-Jan-18	A query sent out for this issue

Fig. 3 Summary table of cumulative data issue tracking



issue	USUBJID	SUBJID	RFSTDT	source	DM_COMMENT
1. Subject is randomized but without Reference Start Date	R000000_000100001	100001		AR_DM	
1. Subject is randomized but without Reference Start Date	R000000_000100002	100002		AR_DM	
4. Randomized but missing sex/race/ethnicity	R000000_000100003	100003	2017-01-27	AR_DM	
4. Randomized but missing sex/race/ethnicity	R000000_000100004	100004	2017-02-27	AR_DM	

Fig. 4 Data issue at worksheet DM

```
*-------------------------Key code of  cumulative summary table -------------------------;

data general;
     merge old_general(in=a rename = (col3 = ocol3 col4 = ocol4 col8 =
ocol8))
             new_general(in=b );
        by _col6 _col7;
```

```
if a and b then     mssg='both';
          if a and not b then  mssg='old';
          if b and not a then  mssg='new';
          col7=compress(col7,,'kw');
          if ocol8 ne '' then col8 = ocol8;
          if ocol3 ne '' then col3 = ocol3;
          if ocol4 ne '' then col4 = ocol4;
       drop ocol3 ocol4 ocol8;
run;

proc report data = general  nowd spanrows
     style(header)={font_weight=bold font_size=10pt just=center
protectspecialchars=off borderstyle=solid bordercolor=black}
     style(column)={borderstyle=solid bordercolor=black};
        column col1-col12 mssg;
        define col1 /order order= data  'Log #' style(column)={vjust=c
just=c};
        define col2 /display 'Domain' style(column)={vjust=c just=c};;
        define col3 /display 'Data Transfer Date';
        define col4 /display 'Date Reported';
        define col5 /display 'Reported by';
        define col6 /display style(column)={url=$urlfmt. } 'Source';
        define col7 /display 'Question / Remark';
        define col8 /display 'Status';
        define col9 /display 'Resolution';
        define col10 /display 'Resolved By';
        define col11 /display 'Date Resolved';
        define col12 /display 'Comments';
        define mssg /noprint ;

        compute COL6;
             if upcase(strip(COL6)) in (&tablist) then do;
                call define(_col_,'style','style={color=blue
                             textdecoration=underline}')  ;
              end;
        endcomp;
        compute COL1;
             if upcase(strip(COL1)) eq 'LOG #' then do;
                call define(_row_,'style','style={background=grey
                            font_weight=bold}')  ;
             end;
             if upcase(strip(COL1)) =0 then do;
                 call define(_col_,'style','style={color=white }')  ;
             end;
         endcomp;

         compute mssg;
             if findw(upcase(mssg), "NEW") gt 0 then do;
                call
define('col7','style','style={background=yellow}')  ;
             end;

             else if findw(upcase(mssg), "BOTH") gt 0 then do;
                call define('col7','style','style={background=red}')  ;
              end;
```

```
  else if findw(upcase(mssg), "OLD") gt 0 then do;
                 call
define('col7','style','style={background=green}')  ;
              end;
           endcomp;
        run;
```

CONCLUSION

High quality clinical data are essential for statistical analysis and reporting. The macro code discussed in this article provides an efficient way to track and document data issues in addition to formal data review completed by Data Management. The automatic output saves programmers significant time in the generation of issue reports. The macro adopts comments from reviewers and marks issue statuses with different colors, which provides the reviewers a useful tool to track and monitor data issues. The method helps the whole study team to identify and resolve the potential data issues in a timely and efficient manner which supports a high quality statistical analysis.

REFERENCES

Jeff Xia, Lugang Larry Xie, Shunbing Zhao, “A Vivid and Efficient Way to Highlight Changes in SAS Dataset Comparison”, *PharmaSUG 2017, Paper AD04.*

Niraj J. Pandya, Vinodh Paida, “Data Edit-checks Integration using ODS Tagset”, *PharmaSUG 2011, Paper DM03.*

ACKNOWLEDGEMENT

The authors would like to thank Changhong Shi for her great support and valuable input of this paper.

CONTACT INFORMATION

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

Name: Chao Su
Enterprise: Merck
Address: 126 E. Lincoln Avenue
City, State ZIP: Rahway, NJ 07065-4607
Work Phone: 732-594-6459
E-mail: chao.su@merck.com
Web: www.merck.com

Name: Shunbing Zhao
Enterprise: Merck
Address: 126 E. Lincoln Avenue
City, State ZIP: Rahway, NJ 07065-4607
Work Phone: 732-594-3976
E-mail: shunbing.zhao@merck.com
Web: www.merck.com

Name: Cynthia He
Enterprise: Merck
Address: 126 E. Lincoln Avenue
City, State ZIP: Rahway, NJ 07065-4607
Work Phone: 732-594-3876
E-mail: cynthia.he@merck.com
Web: www.merck.com