PharmaSUG China 2018 – Paper 64

A Tool to Compare Different Data Transfers

Jun Wang, FMD K&L, Inc., Nanjing, China

ABSTRACT

For an ongoing study, especially for middle-large size studies, regular or irregular data reviewing is an import method to monitor the progress and quality of the study. In this paper, we develop a tool to compare the different data transfers, including added or deleted records, values updated for the existing records and no change of records.

INTRODUCTION

In clinical trial, for an ongoing study, many people, include medical, data management, statistician and so on are concerned about the progress and quality of the study. Regular or irregular data reviewing is an important method. But it will take a boundary of time and energy due to the huge data, especially for middle-large size studies. For different data transfers, the changes can be generalized into three parts: added or deleted records, values updated for the existing records and no change of records. Figure 1 shows a general sample data (only list one subject and some variables).

	SUBJID	PARAMCD	VISIT	VSDTC	AVAL
1	01-001	WEIGHT	SCREENING	2017-05-22	66.5
2	01-001	WEIGHT	RUN IN	2017-06-08	67.5
3	01-001	WEIGHT	FIRST DOSING	2017-07-22	67.5
4	01-001	WEIGHT	WEEK 1	2017-06-29	68
5	01-001	WEIGHT	WEEK 2	2017-07-07	68.5
6	01-001	WEIGHT	WEEK 3	2017-07-13	68
7	01-001	WEIGHT	WEEK 4	2017-07-20	70
8	01-001	WEIGHT	WEEK 8	2017-08-17	48
9	01-001	WEIGHT	WEEK 12	2017-09	.

	SUBJID	PARAMCD	VISIT	VSDTC	AVAL
1	01-001	WEIGHT	SCREENING	2017-05-22	66.5
2	01-001	WEIGHT	RUN IN	2017-06-08	67.5
3	01-001	WEIGHT	FIRST DOSING	2017-06-22	67.5
4	01-001	WEIGHT	WEEK 1	2017-06-29	68
5	01-001	WEIGHT	WEEK 2	2017-07-07	68.5
6	01-001	WEIGHT	WEEK 3	2017-07-13	68
7	01-001	WEIGHT	WEEK 4	2017-07-20	70
8	01-001	WEIGHT	WEEK 8	2017-08-17	68
9	01-001	WEIGHT	WEEK 12	2017-09-14	64.5
10	01-001	WEIGHT	WEEK 16	2017-10-12	65
11	01-001	WEIGHT	WEEK 20	2017-11-09	65
12	01-001	WEIGHT	WEEK 24	2017-12-07	65.5

Figure 1. Sample Data

As we see, there are some changes between these two. How to identify the changes clearly with shorter time? A utility tool will be discussed in following part.

PROGRAMMING PROCESS FLOW

The programming process flow will display in this part. It includes two.

1. Apply PROC COMPARE for all data sets. It will get the comparison results entirely.

2. For the exactly equal dataset, the comparison is finished. What we concerned about is the unequal data sets. It includes added or deleted records, value changed and no changes.

Figure 2 is the programming process flow.

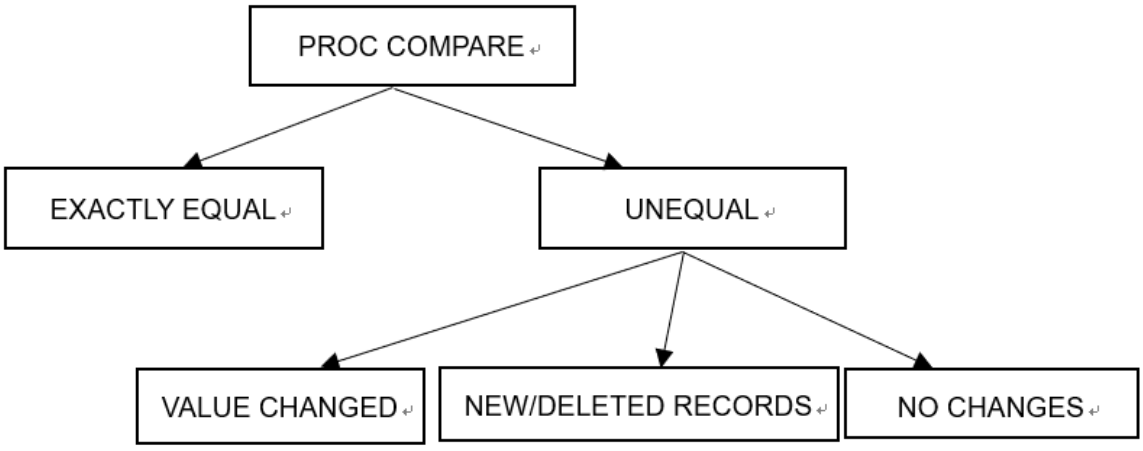

Figure 2. Programming Process Flow

STEP 1

Apply PROC COMPARE for all data sets, it will get the comparison result entirely.

```
Filename filelist pipe "dir ""&logpath"" /b ";

DATA filelist;
infile filelist length=reclen;
input filename $varying1024. reclen ;
if index(lowcase(filename), ".sas7bdat");
memname=scan(filename,1);
Run;

DATA _null_;
set filelist end=eof;
call symputx(compress("data"||put(_N_,best.)),filename);
if eof then call symputx("num",_N_);
Run;
```

Firstly, get all data sets in the folder. Then read each data sets name "***memname***" into a macro variable ***data*** and use macro variable ***num*** to represent the total number of data sets. After that, get all the data sets prepared and using PROC COMPARE, related code is listed as follow:

```
%do i=1 %to &num;

Proc compare data=old.&&data&i compare=new.&&data&i outbase outcomp out=diff&i
outnoequal listall criterion=10e-9 maxprint=(50,32767);
Run;

DATA _null_;
set diff&i end=eof;
if eof then call symput("mvar",trim(left(put(_n_,8.))));
Run;

%if &mvar=0 %then %do;
Proc delete data=diff&i;
Run;
%end;
```

```
%end;
```

Here &logpath is the path of your data sets stored. We can save the result as PDF if necessary. At the same time, it necessary to output the unequal data set and named as diff&i. Then check the observation in diff&i. It will be deleted if there is no observation (exactly equal) in diff&i. Figure 3 is the result of unequal.

	TYPE	_OBS_	SUBJID	PARAMCD	VISIT	VSDTC	AVAL
1	BASE	3	01-001	WEIGHT	FIRST DOSING	2017-07-22	67.5
2	COMPARE	3	01-001	WEIGHT	FIRST DOSING	2017-06-22	67.5
3	BASE	8	01-001	WEIGHT	WEEK 8	2017-08-17	48
4	COMPARE	8	01-001	WEIGHT	WEEK 8	2017-08-17	68
5	BASE	9	01-001	WEIGHT	WEEK 12	2017-09	.
6	COMPARE	9	01-001	WEIGHT	WEEK 12	2017-09-14	64.5
7	COMPARE	10	01-001	WEIGHT	WEEK 16	2017-10-12	65
8	COMPARE	11	01-001	WEIGHT	WEEK 20	2017-11-09	65
9	COMPARE	12	01-001	WEIGHT	WEEK 24	2017-12-07	65.5

Figure 3. PROC COMPARE Result (Unequal)

STEP 2

After PROC COMPARE, the overall comparison result will display. For the unequal data sets, it's time consuming to check the unequal records, especially for middle-large size and multiple variables. We need to develop a new tool to identify the changes clearly. It includes four steps.

1. Find the common variables. For SDTM and ADaM datasets, we can identify the common variables by classes.

```
/* Find common variables*/
Proc contents data=new.&&data&i out=contents noprint;
Run;

DATA common;
set contents;
where name in('SUBJID','VISIT') or find(name,'DECOD') or find(name,'TRT') or
find(name,'TESTCD') or find(name,'PARAM') or find(name,'QNAM');
Run;

DATA _null_;
set common end=eof;
call symputx(compress("common"||put(_N_,best.)),name);
if eof then call symputx("count",_N_);
Run;

DATA common1;
length common common_ $200;
retain common common_;
common='';common_='';
%do i=1 %to &count;
if not missing(common) then common=strip(common)||' '||"&&common&i.";
else common="&&common&i.";
if not missing(common_) then common_=strip(common_)||','||"'&&common&i.'";
else common_="'&&common&i.'";
%end;
Run;

DATA _null_;
set common1;
call symputx("common1",common);
call symputx("common1_",common_);
Run;
```

Macro variable ***common1*** list all common variables and ***common1***_ is the character of them.

2. Merge the new and old data sets by common variables. Check the total number of the observation. It will be flagged as 'NEW' if added records and flagged as 'DELETE' if deleted records.

```
/* FLAG='NEW' and FLAG='DELETED' */
if not a and b then FLAG='NEW';

if a and not b then FLAG='DELETE';
```

3. Check the records in common and rename the variables except common variables. Rename with prefix 'OLD' for the old and 'NEW' for the new. Apply the loop to compare all the variables between the 'OLD' variables and 'NEW' variables. It will be flagged as 'NO CHANGES' if all the variables are exactly equal. And it will display the details if the value in 'OLD' and 'NEW' is different.

```
/*FLAG='NO CHANGES' and FLAG=details*/
length FLAG $50 TERM $200;
retain CHANGE TERM;
CHANGE=0;
TERM='';
%do i=1 %to &num1;
if NEW_&&cname&i=OLD_&&cname&i then CHANGE=CHANGE;
else if NEW_&&cname&i^=OLD_&&cname&i then CHANGE=CHANGE+1;

if NEW_&&cname&i^=OLD_&&cname&i and not missing(TERM) then
TERM=strip(term)||' , '||"&&cname&i."||':'||strip(OLD_&&cname&i)||' to
'||strip(NEW_&&cname&i);
else if NEW_&&cname&i^=OLD_&&cname&i then
TERM="&&cname&i."||':'||strip(OLD_&&cname&i)||' to '||strip(NEW_&&cname&i);
%end;
if CHANGE=0 then FLAG='No CHANGES';
else if CHANGE>0 then FLAG='HAVE CHANGED';
```

Here ***num1*** and ***cname*** are macro variables. And ***num1*** represents the total number of variables except common variables and ***cname*** is the name of variables. Variable CHANGE represents the total number of changes and TERM shows the details.

4. Set the three datasets together and all the changes will display in the final data set.

RESULT

Apply the utility tool described above, it will get the comparison result and list in Figure 4.

	SUBJID	PARAMCD	VISIT	FLAG	CHANGE	TERM
1	01-001	WEIGHT	SCREENING	NO CHANGES	0	
2	01-001	WEIGHT	RUN IN	NO CHANGES	0	
3	01-001	WEIGHT	FIRST DOSING	HAVE CHANGED	1	VSDTC:2017-07-22 to 2017-06-22
4	01-001	WEIGHT	WEEK 1	NO CHANGES	0	
5	01-001	WEIGHT	WEEK 2	NO CHANGES	0	
6	01-001	WEIGHT	WEEK 3	NO CHANGES	0	
7	01-001	WEIGHT	WEEK 4	NO CHANGES	0	
8	01-001	WEIGHT	WEEK 8	HAVE CHANGED	1	AVAL:48 to 68
9	01-001	WEIGHT	WEEK 12	HAVE CHANGED	2	AVAL:. to 64.5 , VSDTC:2017-09 to 2017-09-14
10	01-001	WEIGHT	WEEK 16	NEW	1	
11	01-001	WEIGHT	WEEK 20	NEW	1	
12	01-001	WEIGHT	WEEK 24	NEW	1	

Figure 4. Details of Unequal

From Figure 5, it is clear to obtain the details of these two, even though there are multiple variables changes.

CONCLUSION

Apply the utility tool to compare data transfer, it will identify the details in shorter time and clearly. It's not only about the data transfer, but also check the data issue. But there is a restriction for the tool. It's better to keep the structure of the two datasets the same, especially the variables type.

RECOMMENDED READING

- Base SAS® Procedures Guide
- SAS® For Dummies®

CONTACT INFORMATION

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

Name: Jun Wang
Enterprise: FMD K&L, Inc.
Address: 4F，Building 4，Nanan Ruizhi, No.99 Shengtai Road, Jiangning District
City, State ZIP: Nanjing, 210000
E-mail: jun.wang@fountain-med.com