PharmaSUG 2017 - Paper DA04

Data Change Report

Eric Kammer, Novartis Pharmaceuticals Corporation, East Hanover, NJ

ABSTRACT

This paper shows a data change concept for comparing data between data milestones for clinical studies that was developed using SAS©. SAS© allows the data changes to be colored in GREEN in an output Excel© file indicating what data has changed which can be new, deleted, or changed records. While this program was developed for a Unix© system it is applicable to any environment since it uses a metadata concept. Data can then be compared to see if any significant changes occurred between milestones that could affect safety or any other important decisions applicable with analysis, monitoring, or clinical assessment.

INTRODUCTION

It is important to ensure that data has not changed significantly from one milestone to another that would impact the submission or a safety review of the project. This is especially true for oncology studies where the patients have difficult diseases and the drugs have challenging safety profiles. A data change report is required which gives clinicians and data managers the ability to review any changes to previously collected data after a milestone.

WHAT CHANGES ARE WE LOOKING FOR?

Metadata and Data Changes

What changes are we looking for? Data changes can include new/inserted, deleted, or changed records and fields. The data structure may change, including format, length, and the addition/removal of records. The data change report that was developed was in a SAS© programming environment using a SAS© Proc Compare on datasets. Many people have performed a SAS© Proc Compare, but this does not properly account for metadata changes and is often unreadable by end users.

The two compared datasets, part1 and part2 are split from the metadata in preparation for the compare. The proc compare in combination with collected variables from the CRF can determine what changes have taken place from the previous data. Sashelp.volumn and proc contents can tell if the correct variable names are collected between the two datasets. This is important because if too many variables have been added between milestones because of amendments the comparison might not be valid any longer.

Other considerations for development.

The output should display one record per row, rather than one field from a record per row. Reviewers must be able to interpret and easily summarize the changes to the data, by including: flags to identify the type of change, coloring to indicate which fields and records were changed and both records should be visible side-by-side in full detail. Note if there are too many changes, it will be difficult to review and summarize the changes in an efficient manner. However, this would suggest the formal statistical analysis should be revisited. Also, if you want to see what changes occurred over a 1 year period of time with an ongoing study this might not be ideal so need to restrict it to a shorter period of time otherwise there will be too many changes of data items making the output results not usable. Sometimes data is collected in a vertical manner that needs to be transposed such as laboratory data. This poses additional issues as if you are looking only at selected lab tests these might appear missing but in actuality they might not have been collected at that visit. So if you are looking for WBC, RBC and these are missing it might be best not to remove this record as a lab date might have been changed for another set of labs and it will show that your changes went from missing data to missing data which is not really a change because it is not applicable to your current subset of data.

Sources of Data need to be determined

What is needed? The sources of data must be identified (Oracle© data or SAS©). The report can be set up in two ways. First, the metadata report shows a change for any data elements within the dataset. Second, which seems most preferable, the data changes for pre-specified fields of interest. Another option is that the user must be able to run the report on demand using current data, previous snapshots, or milestones otherwise there would be reliance on a programmer.

Metadata vs Fixed Variables

There are two considerations for determining how to compare the data. One, either a metadata change report that allows users to select the domains of interest for a complete comparison and the report would read the data and determine what fields/data may have changed. This is a great concept but sometimes the users do not want to see all data elements that have changed in data. Two, a semi fixed approach seems most preferable for the data changes for pre-specified fields of interest with fields entered at the site.

KEY FIELDS

Reliable comparison requires key fields that are part of the CRF that will not allowed to be changed by the site or the data managers. As an example, with concomitant medications if data is entered with a record number on the CRF this would be the key field. If there were 10 medications and record 6 was deleted then this would mean record number 1-5,7-10 would be present. Since 6 was no longer valid it will no longer be present and appear as a deleted record. A formid or specific CRF page label is very important not just center, subject, and visit (e.g. Docnum for OC©).

User Driven System

There are many ways to allow an on-demand system to work and for this report JReview© was chosen which allowed users to either keep the defaults given by domain or change depending on their needs and then running SAS© within a Unix© environment. However, this report can easily just be run using SAS© or another user driven system that executes the SAS© environment.

Also, another very important feature is to allow user friendly messages generated as an example stating the comparison date was not found or if the date are incorrectly entered, or variables that might have been added or deleted due to an amendment.

Here is an example of date that was chosen and is not available as a SAS© dataset.

15APR2015 was missing in the compare dataset so the report cannot process.

What is in the output?

The output includes NEW, DELETED, or CHANGES to records.

1. NEW means the record did not exist previously (using primary keys). The primary key is very important as without a valid primary key there is no way to determine accurately the changes that can occur between the two compared datasets.
2. DELETED is a record that existed before but does not exist now (again using primary keys), and
3. CHANGES are any change based on the key fields identified for monitoring for an example AE Report Term and Start Date could be changed. The corresponding records from each dataset are always outputted based on the milestone dates entered by user.

Note: While Study, Subject, and Visit can be unique it does not take into account the CRF page or CRF Formid where the data was entered. Most programmers do not understand how data is entered at the site as their focus is analysis.

An example of primary keys that shows date would not be ideal

Subject Identifier
Visit Number
Document Number
Visit Name
Vendor Identification for Local Lab
Date of Specimen Collection-Local Lab

In this case lab date could change so it is best to use
Subject Identifier
Visit Number
Document Number
Visit Name
Vendor Identification for Local Lab

Key fields cannot change they are based on the CRF and are usually preprinted.

SO HOW DO WE COLOR?

So how do we color? Set up a flag variable for each field and use an appropriate color. You need to prepare your data ahead of time and then just continue the process. The Proc Compare using character converted fields gives Xs indicated changed data.

Figure 1. Example of Proc Compare with key field X and a comparison variable y

x	y
First term bad	X
2nd term good	

Figure 2. Example of the Coloring with Changed Records

Data diferences between dataset ae120 30MAY2016 and dataset ae130 10MAR2017 23MAR17

ver	Subject Identifier for	Reported Term for the Adverse Event	Dictionary-Derived Term	Serious Event	Causality
30MAY2016	9999992	rupture left acute otitis media		N	NO
10MAR2017	9999992	left acute otitis media	Otitis media acute	N	NO
30MAY2016	9993002	skin infection	SKIN INFECTION	N	NO
10MAR2017	9993002	skin infection - fungal	Fungal skin infection	N	NO
30MAY2016	9993002	appetiate change		N	YES, BOTH AND/OR INDISTINGUISHABLE
10MAR2017	9993002	appetiate decrease	Decreased appetite	N	YES, BOTH AND/OR INDISTINGUISHABLE
30MAY2016	9993004	Left thigh pain	PAIN IN EXTREMITY	N	NO
10MAR2017	9993004	left thigh pain	Pain in extremity	N	NO
30MAY2016	9997001	hypogammaglobulinemia	HYPOGAMMAGLOBULINAEMIA	N	YES, INVESTIGATIONAL TREATMENT
10MAR2017	9997001	Intermittent hypogammaglobulinemia	Hypogammaglobulinaemia	N	YES, INVESTIGATIONAL TREATMENT
30MAY2016	9997003	Hypogammaglobulinemia	HYPOGAMMAGLOBULINAEMIA	N	YES, INVESTIGATIONAL TREATMENT
10MAR2017	9997003	Intermittent Hypogammaglobulinemia	Hypogammaglobulinaemia	N	YES, BOTH AND/OR INDISTINGUISHABLE

Here is an example of the code for comparison and coloring:

```
title "Comparison of 01JAN2017 vs 02FEB2017";
%let key=key;
%let keeps=y z;

/* Compare the changed records */
proc compare data=one
  compare=two out=compare(rename=(y=yx z=zx)) noprint;
 by &key;
  var &keeps;
run;

/* and the user name */
%let user=EricKammer;

/* set up the Excel template */
/* note the color field based on the compare indicator of X */
/* yx to show the approach */
  ODS TAGSETS.EXCELXP
FILE="c:\temp\data_change1_&sysdate9._&user..xls"
style=Styles.Meadow
OPTIONS (sheet_name = "Compare" autofit_height="yes"
absolute_column_width="15" embedded_titles='yes'
center_horizontal ='yes' MERGE_TITLES_FOOTNOTES='no' WRAPTEXT='no');

/* tag the color with compute */
proc report data=all2 nowindows missing;
column  ver &key &keeps;
 compute y;
  if index(y,'X') then
   call define(_col_,"style","style={background=light green}");
 endcomp;
 compute z;
  if index(z,'X') then
   call define(_col_,"style","style={background=light green}");
 endcomp;
```

```
format z y $X.;
ods listing close;
run;
ods tagsets.excelxp close;
```

Here is an example of changed data highlighted in color so that a user in Excel© can easily filter and see which fields have changed from Date1 to Date2:

Figure 3. Comparison of 01JAN2017 vs 02FEB2017

VER	*Key*	*y*	*z*
01JAN2017	*1*	*1*	*32*
02FEB2017	*1*	*1*	*40*
01JAN2017	*2*	*2*	*33*
02FEB2017	*2*	*5*	*33*
Del	*3*	*99*	*31*
New	*4*	*3*	*37*

CONCLUSION

In the dinosaur era users would get a data dump from SAS©, Jreview© , Excel©, or other Oracle© databases and check manually any data that was updated. This report meets the needs to compare two requested milestone dates and gives the users a readily available output in Excel© with color using SAS© showing changes that have occurred. Also, meets the needs for an “on-demand” environment where the user can run the report whenever they want from SAS© on any date that occurred during a study.

CONTACT INFORMATION

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

Name: Eric Kammer
Enterprise: Novartis Pharmaceuticals
Address: One Health Plaza
City, State ZIP: East Hanover, NJ 07936
E-mail: eric.kammer@novartis.com