PharmaSUG 2020 - Paper SA-013

Diaries and Questionnaires: Challenges and Solutions

Marina Komaroff, Sandeep Byreddy, Noven Pharmaceuticals Inc., Jersey City, NJ

ABSTRACT

In PharmaSUG2019 conference opening session, there was a question about the most unfavorable data set for programmers and statisticians to work with. Diaries and questionnaires (QS) was named among the first five! The rationale was the complexity of QS: too many items to work with, and hard to compare responses across the time points, within and between subjects.

Clustering and/or categorization of diaries' items using clinical judgement is known approach that helps with analyses. However, to compare the responses of the questions across multiple time points still requires deep understanding of the research question and strong programming skills.

The goal of this paper is to convert diaries-haters to diaries-lovers and explain how appropriate algorithm should be developed and programmed. As example, the research question was to find a fraud in filling up the diaries and check out if subjects repeat the same responses (possibly randomly changing a couple of points) across different time points of the study. The authors suggest an algorithm and provide SAS® Macro to answer this research question; yet, this program can be easily adapted for other needs.

INTRODUCTION

Modern technology came to the clinical trials. Data are collected through tablets, slates, and different technological devices. It means that the information from questionnaires and diaries that ends up in the QS data sets consists of thousands of questions with corresponding answers. Once programmers create data set for analysis (ADQS data set), everything is already structured and standardized. Even if we work with Real World Data (RWD), the data set for analyses should be created the same way.

The hardship for analyses is coming when we need manipulation of multiple questions/answers, and compare the answers within and between the visits for the same question. Such tasks require optimization of the complex algorithms.

This paper provides SAS MACRO that was developed by the authors with a goal to easily compare the answers within and between the visits. The subject and/or clinical site with the suspected fraud for repeated answers can be flagged. It means that if the answers are same throughout the trial, or having minimum, like one, or two random different answer, they are suspected for fabrication. The MACRO can be expanded and easily adapted for different tasks that arises within and after the trials. In addition, it may serve as quality control of your data.

DESCRIPTION

Let's consider the imaginary company X that develops a drug for osteoarthritis (OA) pain relief. The Western Ontario and McMaster Universities Osteoarthritis (WOMAC) Index Likert (LK) version 3.1 is the most recent version of this instrument for the assessment of OA pain in hip or knee. The WOMAC Osteoarthritis Index is widely used to measure pain, stiffness, and physical function in subjects with OA pain. For simplicity, in this paper, only WOMAC pain will be considered.

The WOMAC Pain scale consists of 5 questions about how difficult for the subject to perform following daily activities: Q1: Walking, Q2: Stair climbing, Q3: Nocturnal, Q4: Rest, Q5: Weight bearing. Each question should be answered on a scale of 0 to 4, with 0 being no difficulty and 4 being extreme difficulty: 0=None, 1=Slight, 2=Moderate, 3=Very, 4=Extremely difficult. The final ADQS data set will capture this information in the standardized format (**Output 1**).

Output 1: Snapshot of ADQC Data Set

SUBJID	TRTA	TRTAN	AVISIT	AVISITN	PARAM	PARAMCD	AVAL
01001	Placebo	2	BASELINE	1	Pain Walking on Flat Surface Last 24 HRS	Q01	2
01001	Placebo	2	VISIT 2	2	Pain Walking on Flat Surface Last 24 HRS	Q01	1
01001	Placebo	2	VISIT 3	3	Pain Walking on Flat Surface Last 24 HRS	Q01	1
01001	Placebo	2	VISIT 4/ET	4	Pain Walking on Flat Surface Last 24 HRS	Q01	1
01001	Placebo	2	BASELINE	1	Pain climbing stairs	Q02	3
01001	Placebo	2	VISIT 2	2	Pain climbing stairs	Q02	2
01001	Placebo	2	VISIT 3	3	Pain climbing stairs	Q02	1
01001	Placebo	2	VISIT 4/ET	4	Pain climbing stairs	Q02	1
…..							
01001	Placebo	2	BASELINE	1	Standing – Pain	Q05	3
01001	Placebo	2	VISIT 2	2	Standing – Pain	Q05	1
01001	Placebo	2	VISIT 3	3	Standing – Pain	Q05	1
01001	Placebo	2	VISIT 4/ET	4	Standing – Pain	Q05	1

The biostatistician suspects that some subjects were not compliant and fabricated the same answers for some or all questions across all visits. Thus, the programmer was asked to: (1) flag the questions that had the same answers for all visits; (2) flag the questions with answers randomly differ in 1 or 2 visits; (3) flag the subjects that had the same answers for all questions at every visit.

EXAMPLE

The ADQS data set has to be transposed to the horizontal format (**Output 2)** with one record per question and all answers by time point (V1=visit 1, V2=visit 2, V3=visit 3 and V4= visit 4) to be prepared for analyses.

```
proc transpose data=QS out=myds(drop=_NAME_ _LABEL_) prefix=V;
    by subjid trtan paramcd;
    var aval;
    id avisitn;
run;
```

Output 2: The Snapshot of ADQC Data Set by Visits in Horizontal Format

SUBJID	TRTAN	PARAMCD	V1	V2	V3	V4
01001	2	Q01	2	1	1	1
01001	2	Q02	3	2	1	1
01001	2	Q03	2	1	0	0
01001	2	Q04	2	1	1	0
01001	2	Q05	3	1	1	1
…….						
01010	2	Q01	2	1	1	1
01010	2	Q02	3	2	1	1
01010	2	Q03	2	1	0	0
01010	2	Q04	2	1	1	0
01010	2	Q05	3	1	1	1

The simplified for learning purposes SAS code for this example with 5 Questions and answers at Visit1 through Visit 4 is presented in a few steps. The complete MACRO for any number of visits and questions is provided in MACRO section.

Step 1: Initiation

```
data QSout;
    set myds;
    array avisit {4} V1-V4;      /*array for answers by Visit 1 through 4*/
    array diffs1 {4} d1_1 d1_2 d1_3 d1_4; /*diff: answer in V1 and others*/
    array diffs2 {4} d2_1 d2_2 d2_3 d2_4; /*diff: answer in V2 and others*/
    array diffs3 {4} d3_1 d3_2 d3_3 d3_4; /*diff: answer in V3 and others*/
    array diffs4 {4} d4_1 d4_2 d4_3 d4_4; /*diff: answer in V4 and others*/
```

NOTE: The first step is to set up *avisit* array to capture the *aval* answers at each Visit, and *diffsx* arrays to capture differences in values at Vx and others, where x is 1, 2, 3 and 4. Those arrays can be set up as _TEMPORARY_ arrays.

For simplicity, next steps explain logic of the algorithm for Visit 1 (V1).

Step 2: Find the differences with answers for V1 and others and capture in diffs1 array

```
fl_same1=0;
do i=1 to 4;   /*where 4 is the number of visits*/
if avisit{1}>.z then
  do; if avisit{i}>.z then diffs1{i}=avisit{1}-avisit{i};
    if diffs1{i}=0 then fl_same1=fl_same1+1;
    if diffs1{i}>0 then do;
      if except1="" then except1=trim(left(put(i,8.)));
      else except1=trim(left(except1))||","||trim(left(put(i,8.)));
    end;
  end;
end;
```

NOTE: Similarly, the differences between answers for Vx and others (where x=2,3 and 4) should be calculated into *diffsx* arrays. The character variable *exceptx* captures the Visit number where mismatch was identified.

Step 3: Flag if answers in the visits are the same as in V1

```
if fl_same1=4 then COM=trim(left(PARAMCD))||":"||" ALL ANSWERS THE SAME";
```

Step 4: Flag if 1 answer in the visits is different and others are the same as in V1

```
if fl_same1=3 then COM=trim(left(PARAMCD))||":"||" SAME V1(except
#"||trim(left(except1))||")";
```

Step 5: Flag if 2 answers in the visits are different and others are the same as in V1

```
If fl_same1=2 then COM=trim(left(PARAMCD))||":"||" SAME V1(except
#"||trim(left(except1))||")";
```

NOTE: Similarly, the same flags are created for Vx as fl_samex (where x =2, 3, and 4). The character variable *exceptx* is used for setting up the COM variable.

1. If the flag for difference between any Visit with other is equal to the number of Visits, then all answers for particular Question are the same
2. If the flag for difference between any Visit with other is equal to the number of Visits - 1, then all answers for particular Question are the same except one Visit
3. If the flag for difference between any Visit with other is equal to the number of Visits - 2, then all answers for particular Question are the same except for two Visits
4. The variable COM captures at what Visits the answers were the same, and which one or two were different.

SAS MACRO

```
%MACRO QS(inds=QS2,   /* this is the name of original data set*/
          outds=QS3,        /* this is the name of output data set*/
          nv= ,             /* this is the number of Visits*/
          nq=         /* this is the number of Questions*/
          );

data &outds(drop=i);
 set &inds;
```

```
 format com $200. except1-except&nv $20.;
 array avisit {&nv} V1-V&nv;
 %do D=1 %TO &nv;
   array diffs&D {&nv} d&D._1 - d&D._&nv;
 %end;

 /* find the differences with answers at all Visits */
 com=" "; /*initiate */
 %do kk=1 %to &nv;
   fl_same&kk=0;
   do i=1 to &nv;
    if avisit{&kk}>.z then
       do; if avisit{i}>.z then diffs&kk {i}= avisit{&kk}-avisit{i};
         if diffs&kk {i}= 0 then fl_same&kk=fl_same&kk + 1;
       if (diffs&&kk{i}> .z) and (diffs&&kk{i} ne 0) then do;
            if except&kk="" then except&kk=trim(left(put(i,8.)));
            else
except&kk=trim(left(except&kk))||","||trim(left(put(i,8.)));
       end;
     end;
   end;

/* set up Flags */
 if fl_same&kk=&nv then
COM=left(trim(PARAMCD))||":"||"ALL ANSWERS THE SAME";
  if (fl_same&KK =%eval(&nv-1)) or (fl_same&KK =%eval(&nv-2)) then
do; if com ="" then COM=left(trim(PARAMCD))||":"||" SAME V&kk(except #"||
trim(left(except&kk))||")";
    else COM=left(trim(com))||":"||" SAME V&kk(except
#"||trim(left(except&kk))||")";
end;
%end;
run;
%MEND QS;
/*Call the Macro for Simulated Data set QS2 with 5 Questions and 4 Visits*/
%QS(inds=QS2, outds=QS3, nv=4 , nq=5);
```

Having this data set, it is possible to request an output with the most problematic Questions or Subjects using PROC FREQ (see **Output 3**, and **Output 4**):

Output 3: The Most Problematic Questions

PARAMCD	Com	NumberOfQ
Q01	Q01: SAME V2(except #1): SAME V3(except #1): SAME V4(except #1)	12
Q02	Q02: SAME V2(except #1): SAME V3(except #1): SAME V4(except #1)	6
Q03	Q03: SAME V2(except #1): SAME V3(except #1): SAME V4(except #1)	6
Q04	Q04: SAME V2(except #1): SAME V3(except #1): SAME V4(except #1)	9
Q05	Q05: SAME V2(except #1): SAME V3(except #1): SAME V4(except #1)	11
Q05	Q05: SAME V3(except #1, 2): SAME V4(except #1, 2)	6

Output 4: The List of Subjects Having Questions with the Same Answers, or with Difference in one or two Answers

SUBJID	PARAMCD	V1	V2	V3	V4	Com
01001	Q01	2	1	1	1	Q01: SAME V2(except #1): SAME V3(except #1): SAME V4(except #1)
01001	Q02	3	2	1	1	Q02: SAME V3(except #1, 2): SAME V4(except #1, 2)
01001	Q03	2	1	0	0	Q03: SAME V3(except #1, 2): SAME V4(except #1, 2)
01001	Q04	2	1	1	0	Q04: SAME V2(except #1, 4): SAME V3(except #1, 4)
01001	Q05	3	1	1	1	Q05: SAME V2(except #1): SAME V3(except #1): SAME V4(except #1)
….						
01010	Q01	**2**	**2**	**2**	**2**	**Q01:ALL ANSWERS THE SAME**
01010	Q02	3	3	3	2	Q02: SAME V1(except #4): SAME V2(except #4): SAME V3(except #4)
01010	Q03	**2**	**2**	**2**	**2**	**Q03:ALL ANSWERS THE SAME**
01010	Q04	3	2	2	2	Q04: SAME V2(except #1): SAME V3(except #1): SAME V4(except #1)
01010	Q05	3	2	3	2	Q05: SAME V1(except #2, 4): SAME V2(except #1, 3): SAME V3(except #2, 4): SAME V4(except #1, 3)

The snapshot of output data set QS3 is presented in **Output 5**.

Output 5: The Snapshot of ADQS Data Set by Visits and Calculated Flags

SUBJID	PARAMCD	V1	V2	V3	V4	d1_1	d1_2	d1_3	d1_4	d2_1	d2_2	d2_3	d2_4	d3_1	d3_2	d3_3	d3_4	fl_same1	fl_same2	fl_same3	fl_same4	com
01001	Q01	**2**	**1**	**1**	**1**	0	1	1	1	-1	0	0	0	-1	0	0	0	**1**	**3**	**3**	**3**	Q01: SAME V2(except #1): SAME V3(except #1): SAME V4(except #1)
01001	Q02	**3**	**2**	**1**	**1**	0	1	2	2	-1	0	1	1	-2	-1	0	0	**1**	**1**	**2**	**2**	Q02: SAME V3(except #1, 2): SAME V4(except #1, 2)
01001	Q03	**2**	**1**	**0**	**0**	0	1	2	2	-1	0	1	1	-2	-1	0	0	**1**	**1**	**2**	**2**	Q03: SAME V3(except #1, 2): SAME V4(except #1, 2)
01001	Q04	**2**	**1**	**1**	**0**	0	1	1	2	-1	0	0	1	-1	0	0	1	**1**	**2**	**2**	**1**	Q04: SAME V2(except #1, 4): SAME V3(except #1, 4)
01001	Q05	**3**	**1**	**1**	**1**	0	2	2	2	-2	0	0	0	-2	0	0	0	**1**	**3**	**3**	**3**	Q05: SAME V2(except #1): SAME V3(except #1): SAME V4(except #1)
……																						
01010	Q01	**2**	**2**	**2**	**2**	0	0	0	0	0	0	0	0	0	0	0	0	**4**	**4**	**4**	**4**	**Q01:ALL ANSWERS THE SAME**
01010	Q02	**3**	**3**	**3**	**2**	0	0	0	1	0	0	0	1	0	0	0	1	**3**	**3**	**3**	**1**	Q02: SAME V1(except #4): SAME V2(except #4): SAME V3(except #4)
01010	Q03	**2**	**2**	**2**	**2**	0	0	0	0	0	0	0	0	0	0	0	0	**4**	**4**	**4**	**4**	**Q03:ALL ANSWERS THE SAME**
01010	Q04	**3**	**2**	**2**	**2**	0	1	1	1	-1	0	0	0	-1	0	0	0	**1**	**3**	**3**	**3**	Q04: SAME V2(except #1): SAME V3(except #1): SAME V4(except #1)
01010	Q05	**3**	**2**	**3**	**2**	0	1	0	1	-1	0	-1	0	0	1	0	1	**2**	**2**	**2**	**2**	Q05: SAME V1(except #2, 4): SAME V2(except #1, 3): SAME V3(except #2, 4): SAME V4(except #1, 3)

CONCLUSION

This paper clarified the logic how to identify questions or subjects with the same (or possibly randomly changing a point or couple of points) responses across different time points of the study. The authors suggested an algorithm and provided a SAS® Macro to identify the most problematic questions or subjects; yet, this program can be easily adapted for other needs. We really hope that this paper convert diaries-haters to diaries-lovers and bring confidence to many programmers and statisticians.

CONTACT INFORMATION

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

Marina Komaroff, Dr.P.H.
Senior Director-Biostatistics, Product Development
Noven Pharmaceuticals, Inc.
mkomaroff@noven.com

Sandeep Byreddy, M.S.
Manager-Biostatistics, Product Development
Noven Pharmaceuticals, Inc.
sbyreddy@noven.com