PharmaSUG 2022 - Paper QT-135

Data Masking

Sumit Pratap Pradhan, Navneet Agnihotri, and Rachel Brown, Syneos Health®

ABSTRACT

When Clinical trials are conducted in double blinded manner and continuous analysis is needed by both blinded and unblinded team, data should be handled with special attention. Ideally unblinded data should not be seen by blinded team until Database Lock is completed, but sometimes analyzing data is necessary by blinded team to take important decisions for the study. Here Data Masking comes into the role to help sharing data with blinded team without violating any rule. In the process of data masking, original subject Id of a patient will be replaced by a dummy Id.

Since the study is ongoing, analysis must be performed on individual data cuts. Another important requirement is that keep the dummy Id assigned consistent across all data cuts. This is particularly helpful in doing comparison between different data cuts. Example – Consider 1st data cut has 2 subjects, subject 101-33 is assigned dummy id DUM-01 and subject 101-45 is assigned dummy id DUM-02.

Consider 2nd data cut has 3 subjects (1 additional subject 102-65 as compared to 1st data cut). Only new subject 102-65 will be assigned new dummy id as DUM-03 and there will not be any change in dummy id assigned to old subjects (subject 101-33 and 101-45).

In this paper, details of achieving this result will be explained.

INTRODUCTION

Consider a study where multiple vendors are involved which are providing data. Since the study is blinded and ongoing, blinded sponsor team is not allowed to see original subject id but for review purpose both blinded and unblinded sponsor team needs to look at data. To encounter this scenario, unblinded team will mask original subject id which will facilitate seeing data by both blinded and unblinded team without violating any rule. For doing this, two points will be considered:

1) Original Subject id will be masked to some randomized dummy id
2) Since this is being done for ongoing study, new subjects will be added in subsequent transfers. Assigned randomized dummy id will be consistent across multiple transfers.

   Below is original data having subject details. See Figure 1

	SITEID	SUBJID	USUBJID
1	7654	001	7654-001
2	7654	001	7654-001
3	7654	001	7654-001
4	7654	011	7654-011
5	7654	011	7654-011
6	7654	011	7654-011
7	7654	014	7654-014
8	7654	014	7654-014
9	7654	014	7654-014

Figure 1. Input Dataset

MASKING ALGORITHM

1) Order will be assigned to each subject in input dataset. See last variable in below figure. See Figure 2

	SITEID	SUBJID	USUBJID	ord_var
1	7654	001	7654-001	1
2	7654	001	7654-001	1
3	7654	001	7654-001	1
4	7654	011	7654-011	2
5	7654	011	7654-011	2
6	7654	011	7654-011	2
7	7654	014	7654-014	3
8	7654	014	7654-014	3
9	7654	014	7654-014	3

Figure 2. Populate Order for each Subject

2) PROC PLAN will be used to generate random sequence for total number of subjects present in input dataset.

   a. Below SQL procedure will be used to count total number of subjects.

```
proc sql;
      select max(ORD_VAR) into :ALL_N from _SOURCE1;
quit;
```

   b. PROC PLAN will generate random sequence using SEED.

      i. SEED=779 helps to generate random sequence

      ii. &ALL_N macro variable contains total number of subjects

```
*** Dummy subject id ***;
PROC PLAN SEED=779;
      FACTORS DUMSUBJ_=&ALL_N. random /noprint;
      OUTPUT OUT=TEMP DUMSUBJ_ random;
run;
```

Below is random sequence generated. See Figure 3.

	DUMSUBJ_
1	25
2	19
3	26
4	34
5	24
6	22
7	30
8	31
9	14
10	8

Figure 3. Random sequence generated by Proc Plan

3) Dummy id is created using random sequence generated by PROC PLAN (See Figure 3). Below SAS code is used. Output is shown in Figure 4.

```
data TEMP1;
   set TEMP;
   length DUMSUBJ $100;
   if DUMSUBJ_ < 10 then UMSUBJ="DUM00"||strip(put(DUMSUBJ_,best.));
   else if DUMSUBJ_ < 100 then DUMSUBJ="DUM0"||strip(put(DUMSUBJ_,best.));
   else DUMSUBJ="DUM-"||strip(put(DUMSUBJ_,best.));
   ORD_VAR=_N_;
run;
```

	DUMSUBJ_	DUMSUBJ	ORD_VAR
1	25	DUM-025	1
2	19	DUM-019	2
3	26	DUM-026	3
4	34	DUM-034	4
5	24	DUM-024	5
6	22	DUM-022	6
7	30	DUM-030	7
8	31	DUM-031	8
9	14	DUM-014	9
10	8	DUM-008	10

Figure 4. Creation of dummy id

4) Above Produced dataset (See Figure 4) will be used to merged with source dataset to generate masked dataset.

   Below SAS code is used. Output is shown in Figure 5.

```
data OUTDS;
   length SUBJID USUBJID $100;
   merge _SOURCE1 TEMP1;
   by ord_var;
   drop DUMSUBJ_ ord_var DUMSUBJ;
   SUBJID_OLD=SUBJID;
   USUBJID_OLD=USUBJID;
    SITEID_OLD=SITEID;
   SUBJID=DUMSUBJ;
   if SITEID eq '7654' then SITEID='DUMS99';
   USUBJID=strip(SITEID)||'-'||strip(SUBJID);
run;
```

	SUBJID	USUBJID	SITEID	SUBJID_OLD	USUBJID_OLD	SITEID_OLD
1	DUM-025	DUMS99-DUM-025	DUMS99	001	7654-001	7654
2	DUM-025	DUMS99-DUM-025	DUMS99	001	7654-001	7654
3	DUM-025	DUMS99-DUM-025	DUMS99	001	7654-001	7654
4	DUM-019	DUMS99-DUM-019	DUMS99	011	7654-011	7654
5	DUM-019	DUMS99-DUM-019	DUMS99	011	7654-011	7654
6	DUM-019	DUMS99-DUM-019	DUMS99	011	7654-011	7654
7	DUM-026	DUMS99-DUM-026	DUMS99	014	7654-014	7654
8	DUM-026	DUMS99-DUM-026	DUMS99	014	7654-014	7654
9	DUM-026	DUMS99-DUM-026	DUMS99	014	7654-014	7654

Figure 5. Dummy id against original Subject id

5) Consider masked data is requested for new data cut which includes 2 new subjects. SAS programming will be used using PROC PLAN to generate masked dataset in a manner so that assignment of dummy id remains same for old subjects and new dummy id is assigned to new subjects only. Below figure shows new subjects in yellow color (See Figure 6). New subjects 022 and 031 are assigned dummy id 034 and 022 respectively whereas old subjects 001,011 and 014 are assigned same old dummy id (Ref. to Figure 5).

	SUBJID	USUBJID	SITEID	SUBJID_OLD	USUBJID_OLD	SITEID_OLD
2	DUM-025	DUMS99-DUM-025	DUMS99	001	7654-001	7654
3	DUM-025	DUMS99-DUM-025	DUMS99	001	7654-001	7654
4	DUM-025	DUMS99-DUM-025	DUMS99	001	7654-001	7654
5	DUM-019	DUMS99-DUM-019	DUMS99	011	7654-011	7654
6	DUM-019	DUMS99-DUM-019	DUMS99	011	7654-011	7654
7	DUM-019	DUMS99-DUM-019	DUMS99	011	7654-011	7654
8	DUM-026	DUMS99-DUM-026	DUMS99	014	7654-014	7654
9	DUM-026	DUMS99-DUM-026	DUMS99	014	7654-014	7654
10	DUM-026	DUMS99-DUM-026	DUMS99	014	7654-014	7654
11	DUM-034	DUMS99-DUM-034	DUMS99	022	7654-022	7654
12	DUM-034	DUMS99-DUM-034	DUMS99	022	7654-022	7654
13	DUM-034	DUMS99-DUM-034	DUMS99	022	7654-022	7654
14	DUM-022	DUMS99-DUM-022	DUMS99	031	7654-031	7654
15	DUM-022	DUMS99-DUM-022	DUMS99	031	7654-031	7654
16	DUM-022	DUMS99-DUM-022	DUMS99	031	7654-031	7654

Figure 6. Masked data with new subjects

CONCLUSION

Data masking is useful SAS programming technique which can be used on multiple occasions in clinical domain.

REFERENCES

1. SAS Documentation

   https://go.documentation.sas.com/doc/en/pgmsascdc/9.4_3.4/statug/statug_plan_syntax01.htm

CONTACT INFORMATION

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

Sumit Pratap Pradhan
Syneos Health, Principal Statistical Programmer
Building No. 14, Tower B, DLF Cyber City, Gurgaon - 122002, Haryana, India
E-mail: sumit.pradhan@syneoshealth.com
LinkedIn: https://www.linkedin.com/in/sumit-pradhan-71133345/

Navneet Agnihotri
Syneos Health, Senior Statistical Programmer
Building No. 14, Tower B, DLF Cyber City, Gurgaon - 122002, Haryana, India
E-mail: navneet.agnihotri@syneoshealth.com
LinkedIn: https://www.linkedin.com/in/navneet-agnihotri-78540553/

Rachel Brown
Syneos Health, Manager, Statistical Programming
Regional | TX, United States
E-mail: rachel.brown@syneoshealth.com