Paper CT08

Patient Privacy – The Topmost Priority!

Shiva Gogu, Vita Data Sciences, Waltham, MA, USA

ABSTRACT

As part of a general trend to improve transparency and data sharing in clinical trials, healthcare and Sponsor companies are beginning to share their trial data with researchers to yield new opportunities for innovation in science to contribute to patient health. For them to share their data, protecting patient privacy and comply with legal governing laws throughout the entire process is critical. Therefore, implementation of data de-identification or anonymization is required. This paper will primarily focus on the guidelines, processes and best practices for achieving this. In addition, we will discuss the risks, benefits and challenges associated with data sharing and protection of patient's personally identifiable information.

INTRODUCTION

Unlike CSR (clinical study report) result disclosure, data disclosure has the high risk for patient privacy. Therefore, it is very important to protect patient privacy while disclosing the study data. The primary purpose of this paper is to focus on implementation of the data de-identification process with metadata approach using SAS®. Though the steps discussed here sounds familiar to most of the statistical analysis programmers, there are some key points to be noted while masking the data. Here we discuss about end to end processing of the data points as part of protecting Personally Identifiable Information of the patient. Following are the steps to be considered and followed to generate masking data. As you may know, there are multiple ways to create de-identified datasets but the process we discuss is the one which most of the sas programmers are familiar with and work on day-to-day basis.

METADATA DRIVEN APPROACH

The idea of having metadata driven approach is for the user to have thoroughly reviewed data, element by element across all domains. That way, there is very minimal to no chance of errors being taken place. Hence, it can be considered as the best practice. Having global metadata with action items defined for all elements upfront for any standard data helps in long run. Since this process of data de-identification is driven by metadata, it's better to have global metadata in place to efficiently implement across multiple studies. It is necessary to identify direct and indirect identifiers across all datasets to build global metadata, steps discussed below can be performed once global metadata is in place. It is up to individual preference on how to build the global metadata. However, let us take the below example for instance. Metadata driven process specifies action required for each element which serves as the input for programming purposes. Data need to be thoroughly reviewed in order to define an action item for each element.

Global metadata:

MEMNAME	NAME	TYPE	VARNUM	LENGTH	LABEL	ACTION	COMMENT
DM	STUDYID	CHAR	1	9	Study Identifier	NONE	CRF Collected
DM	DOMAIN	CHAR	2	2	Domain Abbreviation	NONE	CRF Collected
DM	USUBJID	CHAR	3	17	Unique Subject Identifier	MERGE	Randomly assigned patient number for De-Identity
DM	SITEID	CHAR	4	10	Study Site Identifier	MERGE	Randomly assigned site number for De-Identity
DM	AGE	NUM	5	8	Age	GROUP	>=90 and <90
AE	STUDYID	CHAR	1	9	Study Identifier	NONE	CRF Collected
AE	DOMAIN	CHAR	2	2	Domain Abbreviation	NONE	CRF Collected
AE	USUBJID	CHAR	3	17	Unique Subject Identifier	MERGE	Randomly assigned patient number for De-Identity
AE	AESEQ	NUM	4	8	Sequence Number	NONE	CRF Collected
AE	AETERM	CHAR	5	70	Reported Term for the Adverse Event	DROP	Verbatim information will be removed from deidentify dataset.
AE	AEDECOD	CHAR	6	40	Lowest Level Term	NONE	No action needed for dictionary terms
AE	AESTDTC	CHAR	7	20	Start Date/Time of Adverse Event	DATE	All dates will be dropped after deriving Relative Day (or) Offset dates are applied.

STEP 1 CREATE LOCAL METADATA

Local metadata is generated based on the source data available. The only difference between global and local metadata is not having the Action and Comment columns in Local metadata, rest of the content should exactly match between both the metadata. Study specific items may need to be modified in local metadata based on the inputs from the study team as appropriate.

Create Local metadata:

```
proc sql noprint;
  create table loc_mdata as
  select *
  from dictionary.columns
  where LIBNAME = "SDTM";
quit;
```

Combine Local metadata with Global metadata:

```
proc sql noprint;
  create table deid_mdata as
  select a.*, b.action, b.comment
  from work.loc_mdata a left join global.mdata b
  on a.memname = b.memname and a.name = b.name;
quit;
```

STEP 2 REVIEW ACTION ITEMS FOR EACH ELEMENT

Once the action items are imported into metadata, action item specified for each element needs review in local metadata. Study specific elements may need to be updated based on the following categories:

1. Direct identifiers (usubjid, sex, dob, race, ethnic, etc.)
2. Indirect identifiers (age, dates)
3. Outliers (extreme values such as abnormal lab results)
4. Comments (elements that contains investigator comment)

Action item for each element can be classified as following:

1. None (No action needed)
2. Merge (Get de-identified information)
3. Group (Element will be grouped to protect subject PII as per HIPPA rules: If age is greater than 89 then "90+".)
4. Date (All dates will be pushed 3 to 6 months forward (or) dates will be dropped after deriving relative day)
5. Missing (Set all values to missing)
6. Drop (Remove elements from dataset)

STEP 3 CREATE MACROS FOR EACH ACTION ITEM

e.g., For Missing:

```
%macro _std_didmiss(ds,var) ;
%let dsid = %sysfunc(open(&ds));
  %if &dsid > 0 %then %do;
    %let varnum = %sysfunc(varnum(&dsid,&var));
     %if &varnum > 0 %then %do;
      %let vartype = %upcase(%sysfunc(vartype(&dsid,&varnum)));
       %if &vartype = C %then %let vartype = $;
      %if &vartype = N %then %let vartype = N;
     %let varlen  = %sysfunc(varlen(&dsid,&varnum));
    %end;
   %let numobs = %sysfunc(attrn(&dsid,nobs));
  %let dsid = %sysfunc(close(&dsid));
 %end;

%if &vartype eq $ %then
%do;
data &ds;
   set &ds;
   &var="";
run;
%end;

%if &vartype eq N %then
%do;
data &ds;
   set &ds;
   &var=.;
run;
%end;
%mend _std_didmiss;
```

e.g., For Drop:

```
%macro _std_diddrop(ds, var);
%let dsid = %sysfunc(open(&ds));
%if &dsid > 0 %then
%do;
      %let varnum = %sysfunc(varnum(&dsid,&var));
      %let  dsid  =  %sysfunc(close(&dsid));
%end;
%if &varnum > 0 %then
%do;
       data &ds ;
            set  &ds ;
            drop &var. ;
       run;
%end;
%mend _std_diddrop;
```

e.g., For Merge:

```
%macro _std_didmerge(ds,var);
%let dsid = %sysfunc(open(&ds));
 %if &dsid > 0 %then %do;
   %let varnum = %sysfunc(varnum(&dsid,&var));
    %if &varnum > 0 %then %do;
proc sql noprint;
  create table &ds. as
  select a.*, b.d&var. as &var. /*Could be USUBJID, SUBJID, SITEID in some cases PRVSUBJID */
/* DEID dataset having randomly generated id's for USUBJID, SUBJID, SITEID and PRVSUBJID (if applicable) */
  from &ds. (rename=(&var. = _&var.)) a left join misc.deid b
  on a.&var. = b.&var.;
quit;
    %end;
 %end;
%mend _std_didmerge;
```

STEP 4 PROCESS EACH DOMAIN, ELEMENT BY ELEMENT FROM LOCAL METADATA

Both step 4 and step 5 requires quite a bit of coding and it is difficult to place the entire code here. The purpose of this step is to execute the action items by calling independent utility macros for a domain as defined for each element in metadata.

```
/*Sample Macro Call:*/
%did_exec(domn=ae, ddset=%str(dsdtm.ae), odset=%str(sdtm.ae), byvar=%str(usubjid aedecod aestdtc aeendtc);
```

STEP 5 COMPARE DE-IDENTIFIED DATA WITH ACTUAL DATA TO ENSURE QUALITY

Following macro generates differences between actual data and de-identified data in excel with color coding. Helps to identify the changes performed to the data element by element across all domains.

```
/*Sample Macro Call:*/
ods listing;
%let outexcel = /..../diff.xls;
ods tagsets.excelxp file="&outexcel" style=statistical;
%did_compare(dom=ae, ddset=%str(dsdtm.ae), odset=%str(sdtm.ae), byvar=%str(usubjid aedecod ), labl = %str(Adverse
Events));
ods tagsets.excelxp close;
ods listing close;
```

Sample Output:

SITEID	INVID	INVNAME	USUBJID	AESTDTC	AEENDTC	AETERM	AEDECOD	De-idenfitied?	Change
1011	412		ABCD_2343	17MAR2020	18MAR2020		Abdominal pain upper	Modified	# Value change in (SITEID) from 2125 to 1011 # Value change in (INVID) from 1254 to 412 # Value change in (AESTDTC) from 18DEC2019 to 17MAR2020 # Value change in (AEENDTC) from 19DEC2019 to 18MAR2020
1011	412		ABCD_2344	12MAR2020	14MAR2020		Headache	Modified	# Value change in (SITEID) from 2125 to 1011 # Value change in (INVID) from 1254 to 412 # Value change in (AESTDTC) from 13MAR2019 to 12MAR2020 # Value change in (AEENDTC) from 15MAR2019 to 14MAR2020
1011	412		ABCD_2345	07MAR2020	07MAR2020		Eye pain	Modified	# Value change in (SITEID) from 2125 to 1011 # Value change in (INVID) from 1254 to 412 # Value change in (AESTDTC) from 08DEC2019 to 07MAR2020 # Value change in (AEENDTC) from 08DEC2019 to 07MAR2020

WHAT DOES GUIDELINE SAY

There are two methods to achieve data de-identification in accordance with the HIPAA Privacy Rule.

1. **Expert Determination -** Very small risk that anticipated recipient could identify individual:
   - A person with appropriate knowledge of and experience with generally accepted statistical and scientific principles and methods for rendering information not individually identifiable:
     - Applying such principles and methods, determines that the risk is very small that the information could be used, alone or in combination with other reasonably available information, by an anticipated recipient to identify an individual who is a subject of the information; and
     - Documents the methods and results of the analysis that justify such determination.

2. **Safe Harbor –** No actual knowledge residual information can identify individual:
   - The following 18 identifiers of the individual or of relatives, employers, or household members of the individual, are removed:
     - Names
     - All geographic information including street address, city, county, precinct, ZIP code, and their equivalent geocodes, except for the initial three digits of the ZIP code if, according to the current publicly available data from the Bureau of the Census
     - All elements of dates (except year) for dates that are directly related to an individual, including birth date, admission date, discharge date, death date, and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older
     - Telephone numbers
     - Vehicle identifiers and serial numbers, including license plate numbers
     - Fax numbers
     - Device identifiers and serial numbers
     - Email addresses
     - Web Universal Resource Locators (URLs)
     - Social security numbers
     - Internet Protocol (IP) addresses
     - Medical record numbers
     - Biometric identifiers, including finger and voice prints
     - Health plan beneficiary numbers
     - Full-face photographs and any comparable images
     - Account numbers
     - Any other unique identifying number, characteristic, or code
     - Certificate/license numbers

IMPORTANT POINTS TO BE CONSIDERED

- All the above points considered to be removed except dates as they will be managed to replace all dates with new dates (or) replace all dates with study days relative to reference date based on inputs from statistician.
- Any geographic information that will not be removed should be decided by statistician by ensuring the information is not compromising on patient privacy.
- Subject, Sponsor and Site identifiers need to be recoded in a similar fashion to have randomly assigned values in case if siteid needs for any statistical purpose.
- Remove all free text verbatim terms (AE's, Medications, Medical history), comments, names including investigator names.
- All standard dictionary coded terms will be retained.
- Genetic data will be not be shared at all
- It is recommended to set comment variables across all datasets to missing as it is a free text field and may contain sensitive information of the subject.
- CO dataset must be deleted as it contains free text fields and may serve as a direct identifier for Patient sensitive information.
- Extreme values for duration variables (--DUR) may be considered for low frequency and collapsed into categories
- Operational variables (--LOT, --XFN) with low data utility must be removed as it may hold PII

CHALLENGES

It is always challenging to deal with the legacy data and data collected in non-standard format as it is time consuming to trace and identify the purpose of the elements and to implement standard macros. From programming perspective, it is recommended to use individual programs when the data is not in standard format.

DATA DE-IDENTIFICATION VS DATA ANONYMIZATION:

- De-identification of data refers to the process of removing or obscuring any personally identifiable information from individual records in a way that minimizes the risk of unintended disclosure of the identity of individuals and information about them.
- Anonymization of data refers to the process of data de-identification that produces data where individual records cannot be linked back to an original as they do not include the required translation variables to do so.

RE-IDENTIFICATION BY CODE

Often, de-identified data may need to be re-identified (rendered distinguishable) for various purposes including to enhance research activities. For these situations, a special code or pseudonym may be assigned to individual records that meet the following criteria:

- The code or pseudonym may not be derived from other related information about the individual. The most common de-identification techniques include the use of a one-way cryptographic function known as hashing, or a random number generator
- Only authorized parties should know or have access to the re-identification method.
- The re-identification method is documented and include, at a minimum, the following security controls:
  - Physical, technical, and administrative safeguards to protect the index of codes used to re-identify the data
  - Physical and/or logical separation of storage of the de-identified data and the index of codes
  - Documented retention period for the index of codes
  - Documented steward responsible for safeguarding the index of codes
  - Documented list of individuals authorized to access the index of codes
  - Description of re-identification purpose, frequency and duration (e.g., will the data remain de-identified for the duration of the study, or will it be re-identified periodically, by whom, and for what reasons?)

DATA BEFORE AND AFTER DE-IDENTIFICATION

Before:

SITEID	INVID	INVNAME	USUBJID	AGE	AESTDTC	AEENDTC	AETERM	AEDECOD
2125	1254	James	ABCD_2343	65	18Dec2019	19Dec2019	Stomach ache	Abdominal pain upper
2125	1254	James	ABCD_2344	78	13Dec2019	15Dec2019	Worsening of headache	Headache
2125	1254	James	ABCD_2345	84	08Dec2019	08Dec2019	Right eye pain	Eye pain
4254	1651	Samuel	ABCD_2346	91	09Dec2019	28Dec2019	Common cold	Nasopharyngitis
4254	1651	Samuel	ABCD_2347	55	04Dec2019		Fever	Pyrexia
4254	1651	Samuel	ABCD_2348	95	15Nov2019	08Dec2019	Knee pain	Arthralgia

After:

SITEID	INVID	INVNAME	USUBJID	AGEGRP	AESTDTC	AEENDTC	AETERM	AEDECOD
1011	412		ABCD_51285	<90	17Mar2020	18Mar2020		Abdominal pain upper
1011	412		ABCD_62289	<90	12Mar2020	14Mar2020		Headache
1011	412		ABCD_73590	<90	07Mar2020	07Mar2020		Eye pain
1114	515		ABCD_84615	>=90	08Mar2020	27Mar2020		Nasopharyngitis
1114	515		ABCD_98845	<90	03Mar2020			Pyrexia
1114	515		ABCD_99586	>=90	13Feb2020	07Mar2020		Arthralgia

- Site, Investigator and Subject identifiers are re-assigned with random values
- Age values have been categorized per HIPAA guidelines
- AETERM is set to missing as it is a free text field and may contain sensitive information that reveals patient identity
- AEDECOD is retained as it is dictionary term

CONCLUSION

In conclusion, protection of patient privacy is equally important as patient safety in clinical trials. Hence, implementing efficient steps while adhering to the applicable guidelines is important before sharing the trial data. Process discussed in this paper is one of the multiple ways to de-identify the data in order to protect patient personally identifiable information.

REFERENCES

1. Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule: (https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html.)
2. A Visual Guide To Practical Data De-Identification: (https://fpf.org/2016/04/25/a-visual-guide-to-practical-data-de-identification/)
3. De-identification of Clinical Trials Data Demystified: (https://pdfs.semanticscholar.org/446e/52a0c4cd850e4931fff2180746f87e610342.pdf?_ga=2.167361145.613123977.1580228432-1887479253.1580228432)

ACKNOWLEDGMENTS

I would like to acknowledge my wonderful colleagues who helped me finalize this paper by providing constructive feedback and encouragement along the way.

RECOMMENDED READING

[1] Phuse De-Identification Standards:
https://www.phusewiki.org/wiki/index.php?title=PhUSE_De-Identification_Standards

[2] De-Identification Rules defined for SDTM IG 3.2:
https://www.phuse.eu/data-transparency-download

[3] SAS Macros for Data De-Identification:
https://www.projectdatasphere.org/projectdatasphere/html/resources/PDF/DEIDENTIFICATION_SAS_MACROS

CONTACT INFORMATION

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

Author Name: Shiva Gogu
Company: Vita Data Sciences
Address: 281 Winter St, Suite 100
City / Postcode: Waltham, MA 02451
Work Phone: +1 (781) 833-0256
Email: sgogu@vitadatasciences.com
Web: https://vitadatasciences.com/