PhUSE 2015

Paper DH09

The rocky road to Data Transparency

Poritosh Modak, Shafi Consultancy, Bangladesh
Rafi Rahi, Shafi Consultancy Limited, U.K

ABSTRACT

To make the clinical trial data transparent is a great initiative for global research. However, implementing the various guidelines for data transparency is a rocky path. Whether it is trying to automate the different processes, or working out how to both identify and deal with low frequency events, this was not a smooth journey. Free text and key dates are as ever problems that need the manual intervention, but are there alternatives? This paper will highlight some of the difficulties that organisations will need to overcome when trying to implement the data transparency rules, and suggest some techniques that may make the road a little smoother.

INTRODUCTION

This paper describes the discrepancies that arise during the implementation of data transparency techniques on clinical trial data and possible solutions to resolve these issues. To accurately de-identify patients' data gives rise to generating a duplication of patients' identification, which in turn also causes incorrect linking among patients' ID, site ID and other laboratory ID. Identifying and dealing with the low frequency data dynamically using a standard method is more difficult and risky. De-identification of Date and/or Date-Time throughout the clinical trial database leads to a challenge of keeping the original format and shifting them in a consistent way. To replace free-text and Personally identifiable information (PII) with approved text from those concerned is a problematic task because they need the involvement of a manual and automatic system. To eliminate these discrepancies, a process has been developed. This process uses some standard methods in identifying low frequency data, advanced computing for offset dates, defensive coding techniques for random number generation and linking of ID variables, as well as user-friendly techniques of replacing PII and free text with the approved text. Other transparency rules are implemented in an efficient manner to avoid all probable discrepancies.

CONCEPT OF DATA TRANSPARENCY

The technique of data transparency is a concept which was mainly invented to share clinical trial data with other organizations for research purpose. Due to the advent of modern electronic computing devices, the concept of data transparency has become easier and more accessible to a wide variety of organisations.

The demand of sharing data to different research organisations and to export multi-sponsor system is increasing day by day. Now the question arises about the patients' data confidentiality which is one of the commitments made to ensure that any data related to the patient is not disclosed to others. The purpose of data transparency is to completely de-identify clinical study data at patient-level and study-level, but in a consistent way so that these can be used for further meaningful research and analysis without compromising patients' data confidentiality.

WHAT ARE THE STANDARDS FOR DATA TRANSPARENCY?

Different organisations have different standards on data transparency techniques to strictly maintain trial patients' confidentiality when they start conducting their clinical trial research. However, the demand to establish a set of common standards lead to defining standards for CDISC data models, starting with SDTM which ensures that all possible de-identification takes place securely for patients' data at the different stages listed below, in order to maintain the patients' confidentiality.

1. Recode Subject ID
2. Recode ID variable
3. Offset dates
4. Remove
5. Provide coded content of the variable if free-text
6. Keep
7. Elevate to continent
8. Derive Age
9. Aggregate Age
10. Scan and only redact values with personal information
11. Review and only redact values with personal information
12. Remove dataset

EXAMPLE SECNARIOS OF PROBLEMATIC ISSUES

A research organisation started thinking to share their trial data to other volunteer research organisations for further research and analysis. They called for a meeting and asked some statistical programmers to develop a process where a number of clinical trial datasets can be used as input data for analysis and the process will de-identify all

possible patients' identification. The organisation requires that the important guidelines include, but are not limited to, de-identifying the subject ID to prevent the subject from further identification, remove any free-texts available in the dataset, identify low frequency variables, shift dates in a consistent way so that it should not affect any further analysis but cannot be used to identify any subject event or elevate geographical information.

Organisations involved with the data transparency process should be cautious when de-identifying patients' data is performed automatically using a subject identifier in SAS® software. Study site numbers and corresponding variables need to be taken care of during this automatic subject ID generation as failure to do so may cause incorrect linking between sites and patient ID.

Implementation of date shifting involves both character and numeric dates, which may sometimes be partial. The biggest issue in this part is to correctly identify date variables and appropriate format of those variables in order to display the dates and/or times correctly. Along with these partial dates must be identified and shifted according to the standard imputing procedures. At the end the process involves formatting the transformed data back into their original formats.

Once the programmers in the organisation started developing the process, there were some questions regarding how the process is ensuring that the data is perfectly de-identified and this led programmers to investigate the process thoroughly. Was it a smooth journey to implement the Standard Guidelines? What were the crucial factors that should be taken care of when implementing rules from the Standard Guidelines?

1. DIFFERENT RANDOM ALPHANUMERIC DATA GENERATION

The rocky path begins when the process is about to be accomplished dynamically throughout the trial database. This de-identified alphanumeric data will be frequently used in replacing original alphanumeric data. Here the consistency matters because the random numbers used in one patient's detailed data are inherently related to their other resultant data. Misplacement of random numbers affects all types of analysis. At times the randomly generated numbers may become similar to the original numbers and the generation of duplicate random numbers is a common scenario. Study site number and corresponding variables need to be taken care of during this automatic Subject ID generation as failure to do so may cause incorrect linking between sites and patient ID.

	STUD	USUBJID	DOMAI	SUBJID	SITEID	SEX	REGION	BRTHDTC	RFICDTC
1	studyx	studyx_74191	DM	74191	Site1	female	Europe	1946-04-18	2010-01-01
2	studyx	studyx_74191	DM	74191	Site2	Male	Europe	1947-02-04	2010-09-17
3	studyx	studyx_76361	DM	76361	Site3	female	North America	1946-09-10	2010-01-01
4	studyx	studyx_76361	DM	76361	Site6	Male	North America	1953-02-17	2010-01-21
5	studyx	studyx_77269	DM	77269	Site5	Male	North America	1947-08-16	2010-01-20
6	studyx	studyx_77269	DM	77269	Site4	female	North America	1969-02-13	2010-01-01

Figure 1: Duplicate of Subject ID and Incorrect linking among other related variables.

2. CONSISTENCY OF OFFSET DATES

This is one of the most challenging and advanced computing part of the system. In the aforementioned scenario, a researcher performed conventional offset date techniques (a fixed number for all dates for all subjects) in their trial data very carefully. The mistake resulting by the system during this process is the actual treatment end date being just prior to the consent date, and not only that, the partial character dates and its equivalent numeric dates are not quite consistent either. The researcher found that issues kept coming one after another. Afterwards the researcher noted down some findings based on the errors they encountered. Implementation of the date shifting part involved both character and numeric dates which were occasionally partial dates. The biggest issue at this stage of the process is to correctly identify date variables and appropriate format used in those variables to display the dates and/or times. Along with this, partial dates are also required to be identified at their place and need to be shifted. At the end of this 'consistent offset dates process' the process also includes the issue that the dates need to be back in original format after transformation.

Is there any way to resolve these issues which requires only input datasets along with some predefined date? Yes. In order to escape from such nightmare, we came up with a dynamic process which will be explained later in this paper.

3. IDENTIFYING LOW FREQUENCY DATA

Another challenge came up when dealing with Low frequency data. These should be notified to those concerned prior to exporting data. Determining the efficient method to identify them is a bit risky when the database contains millions of pieces of information. A little mistake in identifying Low frequency data might result in the governing board of clinical trial being penalised. Here the researcher preferred the dynamic method to find them and further replace them with approved data. Outliers' data may disclose the subject-related information, so it is another crucial part to identify and deal with them without changing the actual significance of the values.

4. IDENTIFYING FREE-TEXT AND PII

Dynamic de-identification of PII and Free-text data is obsolete. It needs manual intervention to determine the free-text containing data, otherwise what the system returns could be risky and most of the unrelated but important texts for analysis would be de-identified which is undesirable. Is there an easier way to identify free-text and PII so that only the texts containing details of confidentiality will be identified, and will be replaced with the approved text? Manually identifying these data and sending these to the concerned parties for their approval may help to get rid of this nightmare, although using special software can reduce some of the burden.

5. DYNAMIC ALLOCATION OF REGION TO ELEVATE THE COUNTRY

Developing an automatic process to perform the implementations of this guideline is an efficient approach. The automated process is preferable to a manual process as this is easier and saves time in man power, thus greatly reducing the total estimate of time spent on this task.

WHAT ARE THE SOLUTIONS?

To resolve these issues a programming process has been developed in which all of these aforementioned difficulties are considered to make the path little smoother for end users. The full process is divided into several sections to make the process more manageable, according to the types of the guidelines of data transparency.

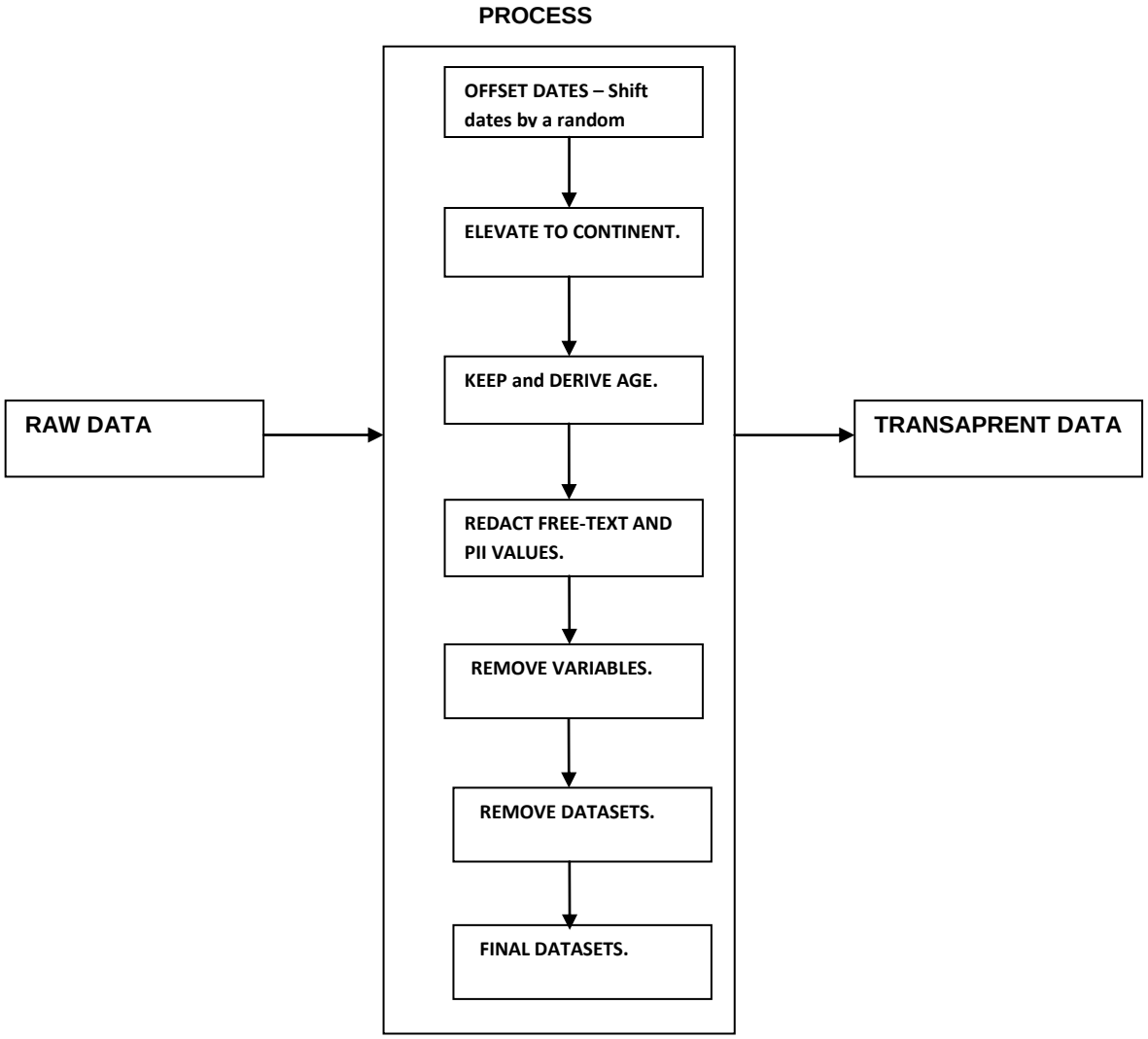

Figure 2: Full Process of Data Transparency

During the development of this process all interrelations between subject identifier and other variables have been carefully assessed and considered. A set of macros has been designed to process a group of guidelines to confirm the de-identification of subject information in their several stages. Each of the macro uses the input of the preceding macro output except the first guideline/rule, so there are no chances that an observation misses any of the guidelines.

1. DIFFERENT RANDOM ALPHANUMERIC DATA GENERATION

To recode subject ID and other subject-related ID variables the macro uses randomly generated characters and numbers to replace the original values with newly generated random codes throughout the clinical trial database. This

is done by considering the interrelation with other study-related variables like site, which ensures the consistency of the data, and checking that the same ID values in different datasets have the same random code values. Also the generation of duplicate numbers and the repetition of original numbers when producing the random numbers are taken care of by using defensive programming. The following approaches have been adopted for de-identification of the Patients, Sites and Laboratory ID.

Approaches taken for making transparent of subject ID

a) Read in input datasets.
b) Gather all unique subjects that exist throughout the database.
c) Generate respective random alphanumeric number.
d) Check for duplicate generation of value.
e) Check if original ID re-generated in any of the steps.
f) Replace all original ID with random alphanumeric number.
g) Carry these random alphanumeric numbers for later processing.
h) Send datasets to partial output library.

Approaches taken for making transparent of ID variables

a) Read in datasets from prior partial output library.
b) Fetch all ID variables from macro parameter.
c) Identifying the value pattern of individuals' variable.
d) Generate random alphanumeric number according to their value appearance.
e) Check for duplicate generation of value.
f) Check if original ID re-generated in any of the steps.
g) Replace all original ID with random alphanumeric number.
h) Carry these random alphanumeric numbers for later processing.
i) Send datasets to partial output library.

2. CONSISTENCY OF OFFSET DATES

For offset dates the process used advanced computing method. The system needs only trial initiation date as input and all other processing is done inside the macro. The system first identifies all numeric and characters date variables, as well as all partial dates and times. Then the process also keeps track of original attributes of the identified date variables. After this the macro identifies all date values under a certain subject and shifts them in a consistent way by considering the trial initiation date. All Dates and Date/time (both numeric and character) values are shifted by a calculated day difference which is determined from subtraction of a trial initiation date and subject consent date. Shifted dates will be saved with original format by using the original variable attribute.

Approaches taken for OFFSET DATES

a) Get all datasets from Reference input Library to temporary work library.
b) Pick all the numeric and character date or date-time variables dynamically.
c) Calculate difference of dates.
d) Find partial dates and flag them.
e) Process all character dates and make them ready to be shifted.
f) Subtract that amount with original dates to shift the dates in a consistent way per subject.
g) Convert all shifted character dates attributes as they were in original dataset.
h) Send datasets to partial output library.

3. IDENTIFYING LOW FREQUENCY DATA

For identification of Low frequency data the system uses 99th percentiles of Proc Univariate for numeric data and for identifying character low frequency data the process uses 10 times lower procedures i.e. calculate average percentage of all categories ,divide average percentage by calculated percentage ,finally identify low frequency data which is greater than 10. Thus the low frequency data are separated/ removed to prevent identification of sites and subject. These low frequency data are stored for further investigation.

Approaches taken for LOW FREQUENCY DATA

a) Get all datasets from Reference input Library to temporary work library.
b) List of all possible variables that may contain LOW Frequency data.
c) Identify numeric and character values containing variables.
d) Apply 99th percentiles of PROC UNIVARIATE method to identify numeric LOW FREQUENCY data.
e) Apply 10 times lower procedure to identify character LOW FREQUENCY data.
f) Send these identified variables to a spreadsheet file for reviewing by authority.
g) Replace the respective values as authority approved.
h) Send to partial output library.

4. IDENTIFYING FREE-TEXT AND PII

A list of free-text variables is used as process input but later steps of the process are done automatically. It sends all free-text values in a spreadsheet file for seeking approval value from concern under redacted text column, so any parts requires to be replaced has to put in that redacted column and later on these original values are replaced with that approved redacted column value automatically.

Approaches taken for making transparent of FREE-TEXT AND PII

a) Get all datasets from Reference input Library to temporary work library.
b) Identify all possible variables name from specification.
c) Produce a spreadsheet file containing all Free Text variables list along with domain name, original value.
d) Use an automated tool to identify initial PII that should be redacted.
e) Manually find these free Text values and define some recoded values under “Redacted” column.
f) This spreadsheet file along with Redacted values will be sent for approval.
g) If the redacted values are approved then again the process will replace original values with these redacted values.
h) Send datasets to partial output library.

Rest of the guidelines such as elevate to continent, age categorization are implemented automatically by the developed process just assigning the required format.

This process is checked and validated by independent programming, frequency testing, and manual checking for attributes and other things with respect to the specification of original data.

IMPLEMENTATION OF GUIDELINES USING MACROS

The whole guidelines implementation process has been completed by designing some macros. For the first rule implementation the process requires the input library name of raw data. Later on the process uses output library of the preceding process as input library. The entire individual library contains de-identification data that it processed and up to their preceding used library data.

```
1  *Recode subject id ,input library is raw_data database;
2  %recode_sub(inlibref=raw_data,outlibref=re_su_id,
3          subjid=subjid,subref= sub_file,seed=1);
4
5  *Recode idvars that placed in idvars parameter,input library is re_su_id database;
6  %re_id ( inlibref=RE_SU_ID,outlibref=re_id_va,
7           idvars=  SITEID INVID CMSPID EXREFID DSREFID DSSPID DVREFID DVSPID);
8
9  *Offset required inputs are consent date and dataset name, study initiazation date;
10 %offset(inlibref =re_id_va, outlibref=offset,subjid=usubjid,study=studyid,seed=0,
11          debug =0,stdindt=01JAN2010,condt=rficdtc,conds=dm );
12
13 *Elevate country to region needs only key dataset name;
14 %elevate_cntry(inlibref=OFFSET,outlibref=elevate, keydata=dm);
15
16 *Free text , Redact and scan values with PII values,validated EQ No for sending data to spreadsheet;
17 %free_text(srclib=keep,outlib=free_txt,validated=n);
18 *validated EQ Yes to replace the original value with approved redacted value;
19 %free_text(srclib=keep,outlib=free_txt,validated=y);
20
21 *Remove macro only needs the list of Drop variables list.;
22 %remove(inlibref=scan, outlibref=remove, dropvars=invnam brthdtc erm dsterm ;
23
24 * Remove dataset needs only the list of dataset;
25 PROC DATASETS LIB=anodata MEMTYPE=DATA NOLIST;
26   DELETE co;
27 RUN;
28 QUIT;
```

Figure 3: An example macro code for complete de-identification.

EXAMPLE OF INPUT RAW DATA AND TRANSPARENT DATA

Source Library				Transparent Library				
Original Dataset name	Variable name	Variable Attributes	Original value of the Variable	Transparent Dataset name	Variable name	Variable Attributes	De-Identified value of the Variable	Rules Implemented
DM	SUBJID	Length : 20	10001 …… ……..	DM	SUBJID	Same as original	32144 …… ……..	(Recode subject ID)
	SITEID	…	SG1001		SITEID	Same as original	PQ2011	(Recode Id Variables)
	COUNTRY	…	UK		REGION	Same as original	Europe	(Elevate to continent)
	RFICDTC	…	10-03-11		RFICDTC	Same as original	02-10-10	(Offset dates)
	BRTHDTC	…	01-02-1987		DROPPED	…	…	(Remove)
	AGE	…	70		AGECATDI	Categorized to character.	"<=89"	(Derive Age)
AE	AELOC	…	BONE	AE	AELOC	Same as original	Gone	(Free-text)
CM	CMREASND	…	Subject vomiting after taken SG220mg dose.	CM	CMREASND	Same as original	Subject relaxed after taken SG100mg dose.	(Scan and redact PII)
CO	…	…	…	CO	Dropped Dataset	…	Dropped Dataset	(Remove dataset)

ROLES AND LIMITATIONS OF THE MACRO

ROLES

1) Macro will deal with the whole SDTM trial database.
2) Macro parameter name may carry the significance of its role.
3) Any dataset opened to facilitate the execution is to be closed.
4) Each macro will perform each designed rule.
5) Care should be taken in selecting the variables for different rules.
6) Each macro designed for different rule is to de-identify only the values.
7) Macro will not have any impact in attributes of the variable throughout the database unless having any special instruction.
8) Macro will not have any impact in number of data files stored in the source library.
9) Macro will process all sorts of date covariates dynamically by itself.
10) Macro will de-identify the data considering the ability to merge and analysis.
11) Subject will be replaced by random number if only subject identifier exists.

LIMITATIONS

1) Any temporary dataset created within macro can have variable name up to 32 characters.
2) Global macro names created within a macro must have specific prefix related to that macro.
3) Temporary datasets created within a macro must have a specific prefix related to that macro.
4) No SAS functions should be used as variable or dataset names within a macro.

CONCLUSION

In this paper we explained the issues that were faced in implementing the guidelines of Data transparency and proposed solutions for them. The process and method used in the system are validated by individual programming and manual checking. The individual macro to implement each rule makes the data de-identification process easier, consistent and user friendly. Using individual macro for individual rule also makes it easy to maintain and update these macros if there is any change in the de-identification guidelines for any individual rule.

REFERENCES

1. PhUSE De-Identification Standard for SDTM 3.2 available at - http://www.phuse.eu/Data_Transparency.aspx

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Please contact the authors at:

Poritosh Modak
Shafi Consultancy Bangladesh
93, Kajolshah
Sylhet - 3100
Bangladesh
Mobile:+880 173 0049 205
Email: poritosh@shaficonsultancy.com
Web: www.shaficonsultancy.com

Rafi Rahi
Shafi Consultancy Limited
Regus House
Highbridge
Oxford Road
Uxbridge
Middlesex
UB8 1HR
United Kingdom
Tel: +44 (0)1895 876 533
Email: rafi@shaficonsultancy.com
Web: www.shaficonsultancy.com