The rocky road to Data Transparency

Shafi Consultancy Ltd.

Overview

- Introduction to Data Transparency
- Standards used
- Issues found during implementation
- Solutions
- Example macro
- Summary

Shafi Consultancy Ltd.

Introduction to Data Transparency

- Data Transparency is a process to de-identify patients' data without compromising **patients' data confidentiality**
- Different organisations have taken different approach to de- identify trial data
- PhUSE has released the first version of a guideline for data transparency to clarify some issues and provide suggestions on implementation. This can be downloaded from PhUSE website.

Shafi Consultancy Ltd.

Standards for Data Transparency

Key PhUSE De-identification standard for STDM 3.2

- Offset dates
- Provide coded content of the variable if free-text
- Elevate country to continent
- Aggregate Age
- Scan free-text and only redact values with personal information

Issues found during implementation

- Duplication in generating random subject 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	Site3	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

Shafi Consultancy Ltd.

Issues found during implementation

- Inconsistency of Offset date

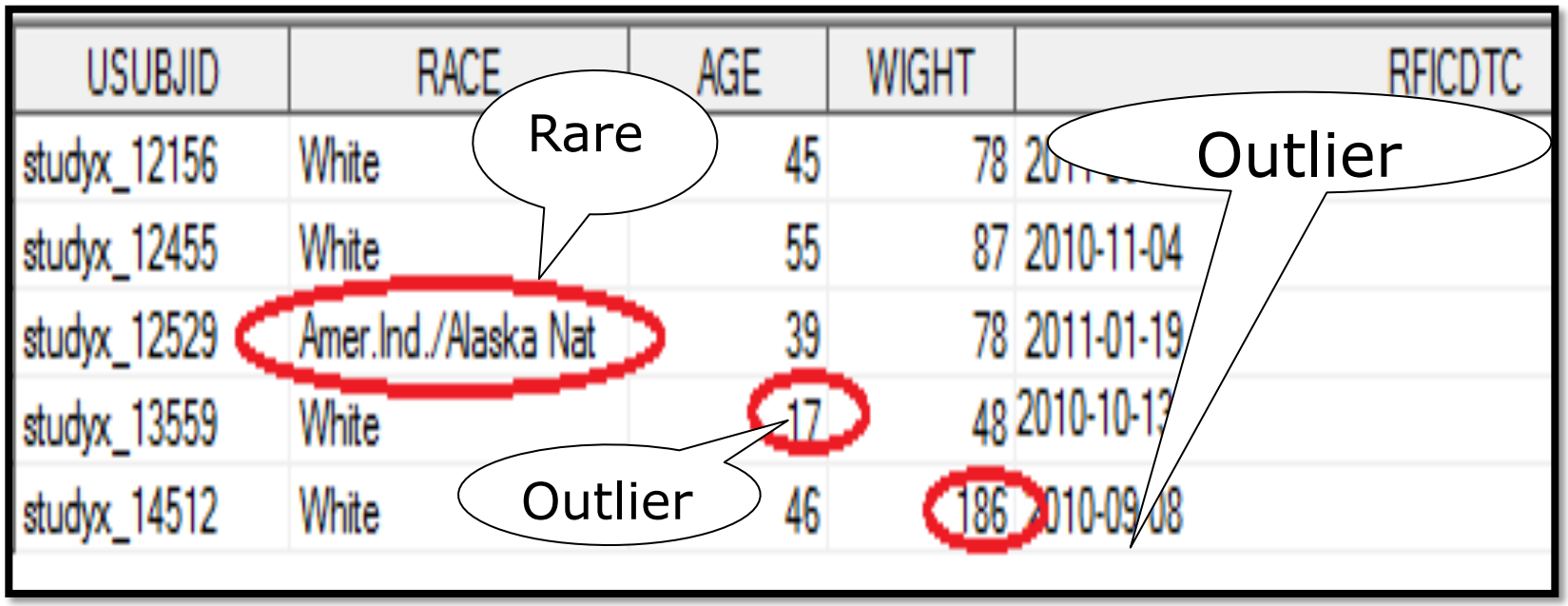
  Historical AEs can shift to post consent date

USUBJID	BRTHDTC	RFICDTC	RFSTDTC	AEONDTC
studyx_12156	1954-01-27	2011-06-02	2011-06-22	20JAN2011
studyx_12455	1962-12-26	2010-11-04	2009-11-10	04FEB2009
studyx_12529	1934-02-21	2011-01-19	2011-02-10	20MAR2014
studyx_13559	1946-05-24	2010-10-13	2011-01-10	20JUN2012
studyx_14512	1972-06-26	2010-09-08	2010-10-22	13JAN2010

Shafi Consultancy Ltd.

Issues found during implementation

- Identifying Low Frequency Data



Shafi Consultancy Ltd.

Issues found during implementation

- Identifying Free-text and Personally Identifying Information (PII)

USUBJID	VISIT	LBNM	LBDESC	LBGCOM
studyx_12156	visit1	URIC	Uric acid	trea wsa oto ole
studyx_12156	visit2	CHOL	Cholesterol, total	Nrt govd
studyx_12156	visit3	MON	Monocyte, absol.	Vesblt under-process
studyx_12156	visit4	BASA	Baso, absol.	Cmount was too minimum
studyx_12156	visit5	AMYL	Amylase	Rvbvdlt ngt gbod

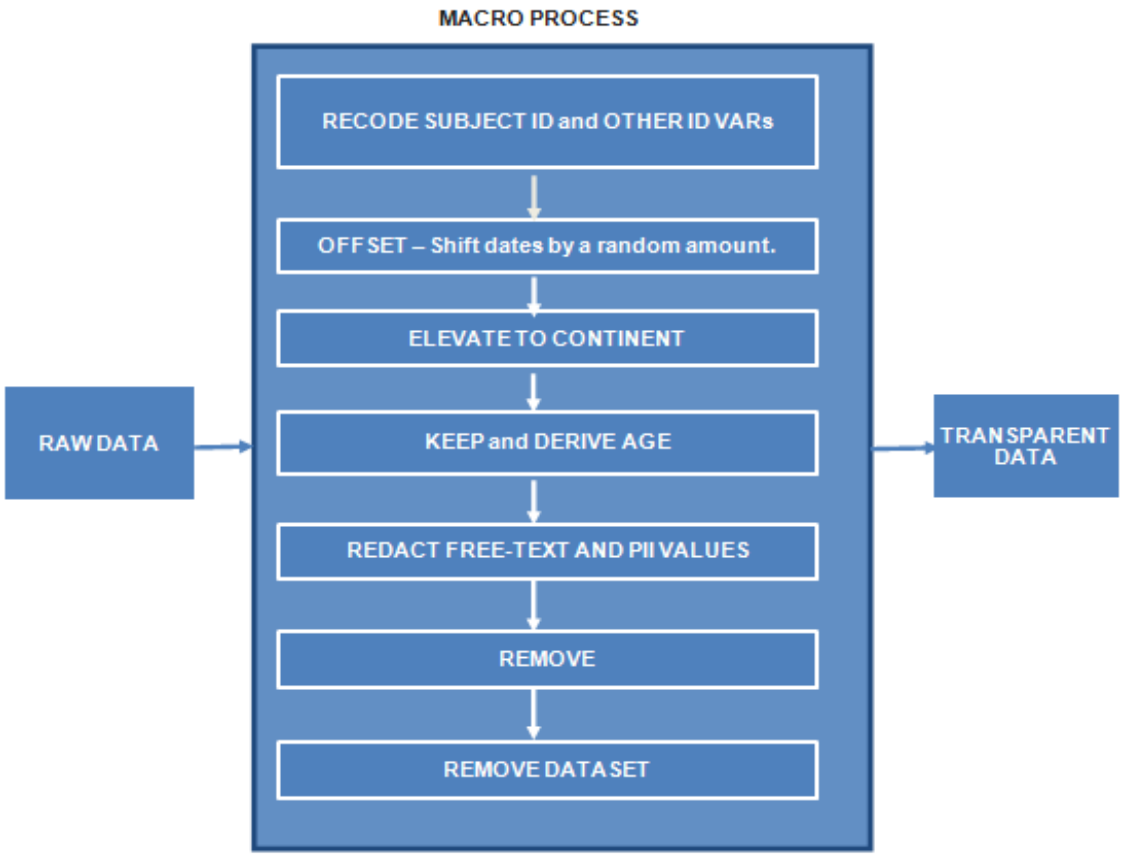
Solutions

- Develop a step by step process to ensure issues are identified and resolved
- A macro program for each de-identification standards
- Automated and manual approaches, both techniques need to be considered
- Collaboration among teams

Shafi Consultancy Ltd.

Complete process of Data Transparency

- Full process has been implemented individually based on standard rules



Shafi Consultancy Ltd.

Example of Macro to implement Data Transparency Guidelines

```
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;
```

Each macro Input is the previous macro output

Shafi Consultancy Ltd.

Conclusion

- During **OFFSET** dates
  - Flag partial dates
  - Check historical dates
- **Low frequency** data
  - Check cross tabulation, not just single values
  - Check grouped data to ensure consistency
- **FREE-TEXT AND PII**
  - Use automatic scanning tools
  - Perform manual review

Shafi Consultancy Ltd.

Conclusion

- Standards help to anonymize the data without compromising patients' data confidentiality
- A thorough process needs to be in place to ensure the efficient and correct implementation of the standards
- A macro for each standard/rule can smooth the process and minimise updates if new data introduces new problems

Questions or Comments?

Shafi Chowdhury

shafi@shaficonsultancy.com

www.shaficonsultancy.com

Shafi Consultancy Ltd.