PharmaSUG China 2019 - Paper AD-010

Data De-identification Automation in SAS

Huan Lu, Sanofi

ABSTRACT

In 2013, Pharmaceutical Researchers and Manufacturers of America (PhRMA) and European Federation of Pharmaceutical Industries and Associations (EFPIA) achieved a commitment to enhance public health and medical and scientific knowledge and streamline regulatory compliance by facilitating the sharing and transparency of clinical trial information, it was the first time that data transparency was highlighted in public eye. Data in clinical trials submitted to Food and Drug Administration (FDA), European Medicines Agency (EMA) or national competent authorities of EU Member States shall be shared in order to fulfill the commitment by considering how de-identification and anonymization techniques can be applied to individual patient data (IPD). Given appropriate de-identification specification and plan, automation in data de-identification process becomes very much needed, DEID as data de-identification automation SAS macro package was created under such a background, which would be an ideal tool to de-identify data automatically.

INTRODUCTION

Given the fact that the world is escalating demands for clinical trial transparency, data and information sharing, and the participation of industry authorities for the responsible sharing of clinical trial data, data sharing policy developed and governance board established. Pharmaceutical companies on data sharing platform are seeking Standard Operation Procedure (SOP) and de-identification guidance and redaction for either internal or external sharing. As one of the most important part, developing programming tools to industrialize de-identification of clinical trial data is the key to increase automation. Figure 1 shows the development of data sharing and transparency activities.

BACKGROUND

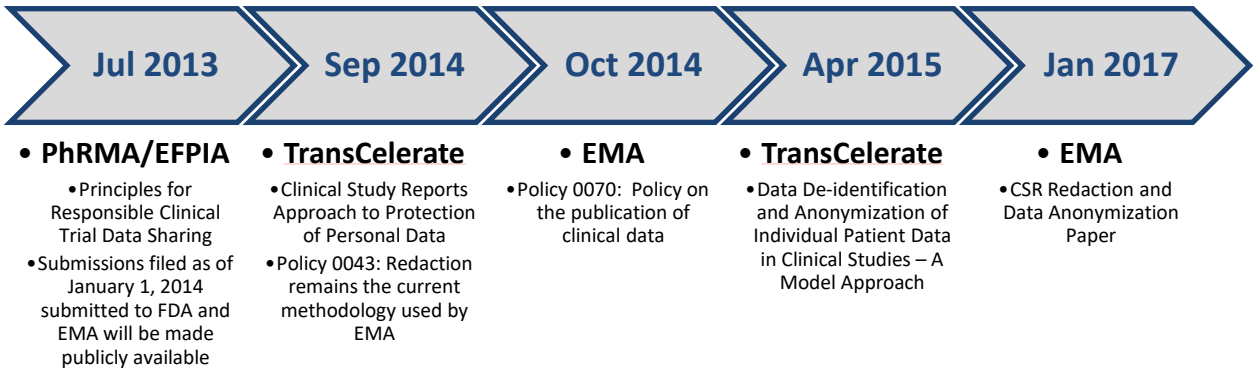

Figure 1. Development of Data Sharing and Transparency Activities

Since Pharmaceutical Researchers and Manufacturers of America (PhRMA) and European Federation of Pharmaceutical Industries and Associations (EFPIA) adopted principles for responsible clinical trial data sharing into effect as of January 1, 2014, there are five commitments were made.

- Enhancing Data Sharing with Researchers
- Enhancing Public Access to Clinical Study Information
- Sharing Results with Patients who Participate in Clinical Trials
- Certifying Procedures for Sharing Clinical Trial Information
- Reaffirming Commitments to Publish Clinical Trial Results

In order to fulfill the commitments, pharmaceutical companies are obliged to de-identify individual patient data (IPD), certain rules initialized by industry authorities that shall be applied to the de-identification process to protect privacy:

US HIPAA PRIVACY RULE

- Requires appropriate safeguards (Safe Harbor) to protect the privacy of personal health information
- Sets limits and conditions on the uses and disclosures that may be made of such information without patient authorization
- De-identification permits sharing without Protected Health Information (PHI)

HIPAA 18 Identifiers

- Names
- Geographic subdivisions smaller than state
- All elements of dates except year
- Telephone numbers
- Fax numbers
- E-mail addresses
- Social Security numbers
- Medical record numbers
- Health plan beneficiary numbers
- Account numbers
- Certificate/license numbers
- Vehicle identifiers/serial numbers
- Device identifiers/serial numbers
- URLs
- IP addresses
- Biometric identifiers
- Full face photographic images
- Any other unique identifying number or code

US COMMON RULE / FDA REGULATIONS

- De-identification permits data sharing in absence of explicit consent for secondary research

EU DATA PROTECTION DIRECTIVE

- Anonymization permits data sharing for secondary research
- Cross-border transfer of records achievable with binding corporate rules

OBJECTIVE

The objectives of de-identification is that de-identification is needed when the data to be used for the purpose beyond the contents of patient informed consent. Before sharing the data no matter internally or externally, the pharmaceutical companies need to protect patients' privacy by minimizing the risk for every single patient to be re-identified. Ideally, for anonymization, the risks of re-identification should be zero, but theoretically, for de-identification, such risks cannot be zero. Figure 2 shows the approaches from transparency to de-identification

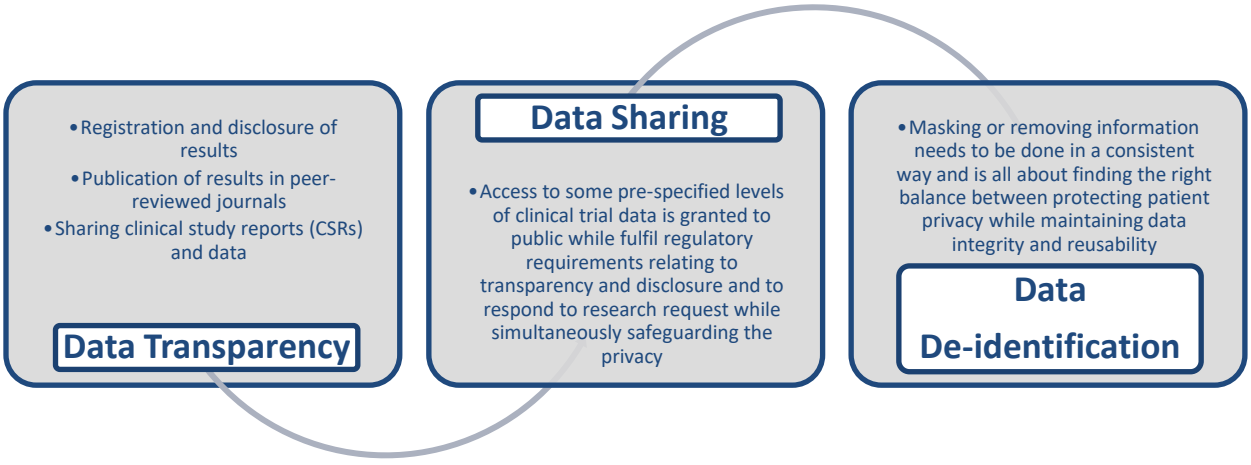

Figure 2. Approaches from Data Transparency to Data de-identification

CHALLENGE

Known need of data de-identification and given defined process and guideline, looking at pharmaceutical companies, at programming level, there are three main challenges that they need to face.

STANDARDIZATION

Automation tool must be universally adapted to submission data sets. The tool itself was developed based on Clinical Data Interchange Standards Consortium (CDISC) standards in order to adapt the tool to follow the majority scenarios. Parameters and options are also designed in tool to make sure that flexibility exists in order to fit the tool to minority events on study level. The key to the high level automation in tool is whether majority scenarios have been covered or not.

This challenge lies in how can we systematically and accurately define the definition of majority scenarios mentioned above. The sufficiency of de-identification is measured by the risk of re-identification mostly, the measurement of such risk is related to the sample size. Speaking of majority scenarios, there certainly is no majority sample size, for instance, comparatively significant small sample size in rare diseases or pediatric studies.

This would be an only one out of hundreds of challenges in standardization.

OVER REDACTION

Over redaction can be the side effect of over standardization, given inappropriately overwhelming rules for covering majority scenarios, there would be risk of damaging data utility under constraint of maximize standardization.

Extreme values in age shall be removed for the reason that those values are determined as outliers, for example, age, height or weight. Based on different algorithms such as confidence interval of distribution, percentile or an ad-hoc number, the outcomes can be significantly different. Over redaction would likely happen if the inappropriate algorithm was applied, that is, non-outliers has been determined as outliers and got excluded from the data sets. Privacy and confidentiality is successfully protected in this way, however, there might not be much information left in data sets, standardization in this way has taken its toll on data utility.

For data de-identification, as much and impact as possible information should be shared to public, even though the information has been de-identified. Pharmaceutical companies need to seek a way to de-identify data with minimum works without contradicting the laws.

RISK ASSESSMENT

The performance of the de-identification tool, on study level, is evaluated by the risk assessment report, which relies on the certification from external expertise. Given the fact that data de-identification is still a relatively new concept, there is no industry standard to either process or assess the de-identified data. Pharmaceutical companies are slowly getting there with standards and rules provided by TransCelerate BioPharma Inc.

Internally, considering the expense, budget and also the availability of external expertise, pharmaceutical companies are more cling to their own expertise to evaluate the de-identified data sets and tools, even though for internal data sharing, creditable and effective evaluation on de-identification requires at least two different teams, one is working on de-identification, and the other on is working on re-identification, in order to validate the efficiency of de-identification and avoid any potential misconducts. Data sharing requestor would in this way receive de-identified data among different pharmaceutical companies based on different standards and rules, which would be a headache for requestor to tell whether the data is ready to share or not, there will never be a final greenlight in this case for data sharing safely and effectively.

Externally, different standards and rules for de-identification are rarely shared between companies in global pharmaceutical industry, de-identified data is less likely to be easy to access for the researchers, let alone for the public. When a pharmaceutical company answer to the industry authorities with the de-identified data, there is no globally agreed standard or rule the company needs to follow, they can simply answer the requestor with the data, perhaps along with the self-developed de-identification specification and certificate

mentioned above. No wonder that for data de-identification, globally agreed assessment standard is very much needed in the near future in order to let all pharmaceutical companies speak the same language in this field.

DEVELOPMENT

As Sanofi received enormous requests from different industry authorities for data sharing, we developed internally a system to handle the situation.

STANDARD OPERATION PROCESS

Figure 3 shows the data de-identification standard operation process in Sanofi.

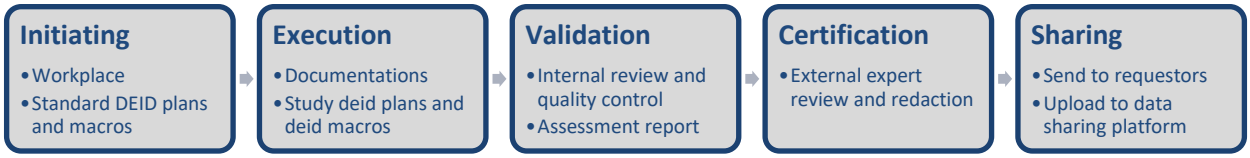

Figure 3. Data De-identification Standard Operation Process (SOP)

Data de-identification flows from end to end requiring close cooperation within multiple teams, Statistician and project lead would take the most of responsibility of initiating the project and determination of research methods and rationales. Involves an external expert with appropriate knowledge of and experience with generally accepted statistical and scientific principles and methods for rendering information not individually identifiable, who would also need to document the methods and results of the analysis that justify such a determination in the assessment report along with the certificate.

For this paper, the step of execution with automated programming tool would be addressed mostly.

SOLUTION

Data de-identification automation tool (DEID) provides SAS® macros procedure to de-identify sensitive information within SDTM and ADaM, which is designed and maintained on Sanofi internal programming environment. Figure 4 shows the logo of data de-identification automation tool.

Figure 4. Logo of Data De-identification Automation Tool

According to US HIPAA Privacy Rules and industry general rules, DEID would be able to protect privacy, provide de-identification process tracking records and provide final de-identified data sets.

- De-identified study IDs and site IDs
- Randomized subject IDs and randomized unique subject IDs
- De-identified outliers
- Criterial groups
- Remove identifiers, quasi-identifiers
- Modification list
- Final de-identification reports

Given defined study de-identification plan based on standard plan, simply modifying the parameters and options accordingly in DEID macros would de-identify automatically. Table 1 shows the module of DEID.

Module	Macro Name	Objectives
Initialization	%di_init	Define input/output/temporary study, folder and data names Create global variable list among all datasets, dataset list with unique subject ID, common variable list, outlier list, modification tracking list
Randomization	%di_dataprep	Define ADSL dataset Create global variable/dataset list, randomized IDs Update global common variables list and modification tracking list Save all global datasets in temporary path
Criterion	%di_criterion	Define detail criterions
Outlier	%di_calc	Define de-identified variable and desired variable name, lower/upper threshold and desired label name, dataset desired/ exception list Create de-identified variables and labels Update global outliers list and modification tracking list
Date	%di_date	Define ADSL dataset and dataset desired/exception list Remove dates Calculate relative day Update modification tracking list
Formatting	%di_format	Create global variable formatting list Rename all randomized and de-identified variables Output original values in each corresponding dataset Update modification tracking list
Deletion	%di_remove	Define variable desired/exception list, dataset desired/exception list Remove variable or dataset accordingly Update modification tracking list
Reporting	%di_report	Define dataset desired/exception list Create final de-identified dataset list, final de-identified datasets Update modification tracking list
Restoration	%di_restore	Define dataset desired/exception list Remove global datasets and restore original datasets layout

Table 1. DEID Module

DEID macros are mutually dependent and correspondingly independent, each of macro should be executed chronologically, and initialization and randomization macros must be executed first. Before actually execute DEID on certain study, reading related documentations is strongly recommended.

Figure 5 shows the before and after comparison throughout DEID data de-identification process.

Center ID	Investigator ID	Investigator Name	SUBJID	USUBJID	AGE
012345	531441	Dr. Bryce	6	EFC12345-001-0006	25
012345	531441	Dr. Bryce	9	EFC12345-001-0009	31
456789	046656	Dr. Joshua	19	EFC12345-013-0019	86
456789	046656	Dr. Joshua	91	EFC12345-013-0091	28

AESTDTC	AEENDTC
2018-10-14	2018-10-16
2017-12-04	2017-12-06
2018-05-10	2018-05-24
2018-02-09	2018-02-13

Death Related Variable	Free Text Variable
Recovered	Diarrhea
Recovered	Diarrhea
Recovered	Pain
Recovered	Pain

Randomized Center ID	Randomized Investigator ID	Remove
315462	684654	
315462	684654	
791203	013516	
791203	013516	

Randomized SUBJID	Randomized USUBJID	AGE	AGEC	Create Relative Start Day	Create Relative End Day
8344	EFCXXXXX-999-8344	25	[20,30)	45	47
6346	EFCXXXXX-999-6346	31	[30,40)	50	52
0546	EFCXXXXX-998-0546	.	>40	30	44
8470	EFCXXXXX-998-8470	28	[20,39)	10	14

Remove Remove

Figure 5. DEID data de-identification process demonstration example

De-identification of large scale data sets is still risky even though with the help automation tool, approach based on differential privacy can be a new way to re-identify information, which is rising up recently and rapidly. Incorporating some noise in data sets would be able to strengthen the data and prevent information

leaking. Since both topics are still comparatively new, voices from both sides should be raised loud enough to be heard by the world. As we process in data de-identification, pharmaceutical companies should continue to standardize the process and support it to become more efficient and meaningful.

CONCLUSION

The goal of data transparency is to tell the stories behind pills, every single patient is titled to know the details in the tiny little pill he or she swallows in. The reason why we did not fully achieve data transparency is that we barely pay attention or have doubt in the credibility of the medicine we take, we always have unshakable faith and trust in pharmaceutical companies. However, clinical trials come with uncertainties and potential risks for sure as always, once pharmaceutical companies have successfully conquered all the difficulties and got approved by the authorities, evidences shall be conditionally shared in order to support the faith and trust of public. In other words, conditionally sharing complied with legal governing laws requires data de-identification to protect patient confidentiality, so that pharmaceutical companies would be able to share data with reasonable incentives. Automation in SAS to de-identify data would continue to support this goal with enormous efficiency and minimum effort.

REFERENCES

"Protection of Personal Data in Clinical Documents – A Model Approach." Available at http://www.transceleratebiopharmainc.com/wp-content/uploads/2017/02/Protection-of-Personal-Data-in-Clinical-Documents.pdf.

"De-identification and Anonymization of Individual Patient Data in Clinical Studies". Available at http://www.transceleratebiopharmainc.com/wp-content/uploads/2015/04/TransCelerate-De-identification-and-Anonymization-of-Individual-Patient-Data-in-Clinical-Studies-V2.0.pdf.

"CDISC Standards in the Clinical Research Process". Available at https://www.cdisc.org/standards.

"TransCelerate BioPharma Inc.". Available at https://transceleratebiopharmainc.com/.

ACKNOWLEDGMENTS

I would like to acknowledge Jianfeng Ye for developing profoundly this project at early stage. Furthermore, I want to thank Veronique Poinsot, for being incredibly supportive and actively leading this project all the way along. Additionally, thanks to my parents, my families and friends, and all the colleagues in Sanofi.

RECOMMENDED READING

- *Dwork, C and Roth, A. 2014. "The Algorithmic Foundations of Differential Privacy". Foundations and Trends® in Theoretical Computer Science, Vol. 9, Nos. 3–4 (2014) 211–407SAS® For Dummies®*

CONTACT INFORMATION

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

Huan Lu
Sanofi
Chengdu Yintai Centre Tower 3, Tianfu Dadao Beiduan 1199, Wuhou District
Chengdu, Sichuan 610041
bryce.lu@sanofi.com