Paper DH06

A Standardized Approach to De-Identification

Benoit Vernay, Novartis, Basel, Switzerland
Ravi Yandamuri, MMS Holdings Inc., Canton, USA

ABSTRACT

Data transparency has become a popular topic in the Pharmaceutical industry over the last few years. De-identified clinical data is an important newdeliverable that requires resourcing and defined processes. This paper will share how Novartis is tackling the challenges around programming activities for standardized data de-identification as well as performing a 3-stage pilot with Privacy Analytics to investigate the risk of re-identification. The journey will start with a process overview and will then be followed by an indepth discussion of key topics : de-identification tool, standard de-identification metadata, validation approach, programming utilities and templates, risk of re-identification and finally the multi-sponsor environment where the data is shared externally. All the sections will be tied towards the goal of achieving speed and efficiency. We also hope to demystify a little this new realm and to encourage others to embrace the data sharing initiative.

INTRODUCTION

Data is the key asset in clinical Research and Development. This statement not only holds true in the Pharmaceutical industry but also for other stakeholder. Safety and efficacy data transparency via data sharing initiatives have a great potential to advance science, benefit patients and increase public trust in the industry. At the same time transparency carries risk for the subjects' privacy as there is a potential that individuals could be identified from the personal data that has been shared.. To mitigate this risk, data must be de-identified prior to any disclosure. Data de-identification is the process by which a dataset is derived in such a way that the data subject is no longer identifiable.

As of today, data de-identification while preserving data utility as much as possible is quite a challenge in our industry. Policies and guidances from regulators are still evolving. From a programmer perspective, much has still to be defined for both tools and processes.

The goal of this paper is to describe Novartis approach around programming activities. We will provide you with an overview of our process from original data in our system to de-identified data accessible to external parties. Then, we will discuss the key tools and components of the process. Together, we believe the tools and the process forms a standardized and efficient solution.

Whether you are getting started in data sharing or you are a pioneer. Whether you have a dedicated team in your organization or you work with an external partner. Whether you develop your own tools or you purchase amongst available ones on the market. Regardless of where you are in the efforts to develop your data sharing initiative, we hope to nurture knowledge and experience on data de-identification.

DE-IDENTIFICATION PROCESS

While defining a new process in the highly regulated Pharmaceutical industry, we have to think about how we will document the process with Standard Operating Procedures (SOP) and Working Practices (WP). Thus, you will need to assess which documents need to be updated and/or created. Novartis strived to defined a unique process, light and simple with the goal to automate as much as possible.

The idea is to bring efficiency by combining a consistent approach with standard components. Everything can be adapted and improved as our needs evolve. It is centered around two key components: a validated tool and metadata. These are integrated into an iterative approach. Each iteration attempts to process a set of clinical data using the validated tool. The metadata provides de-identification specifications to be followed by the tool. At the end of an iteration, the tool allows for the programmer to identify variables in the clinical data without associated specifications in the metadata. Thus, new specifications can be defined and added to existing metadata. The objective is dual: fully de-identify the set of data and iteratively build a standard metadata to support future de-identifications.

Thediagram below presents a process overview. The accompanying table provides a description of each numbered diagram shape and references to sections of this paper providing full details. Therefore you can link easily from the process overview to the remaining of the paper contents.

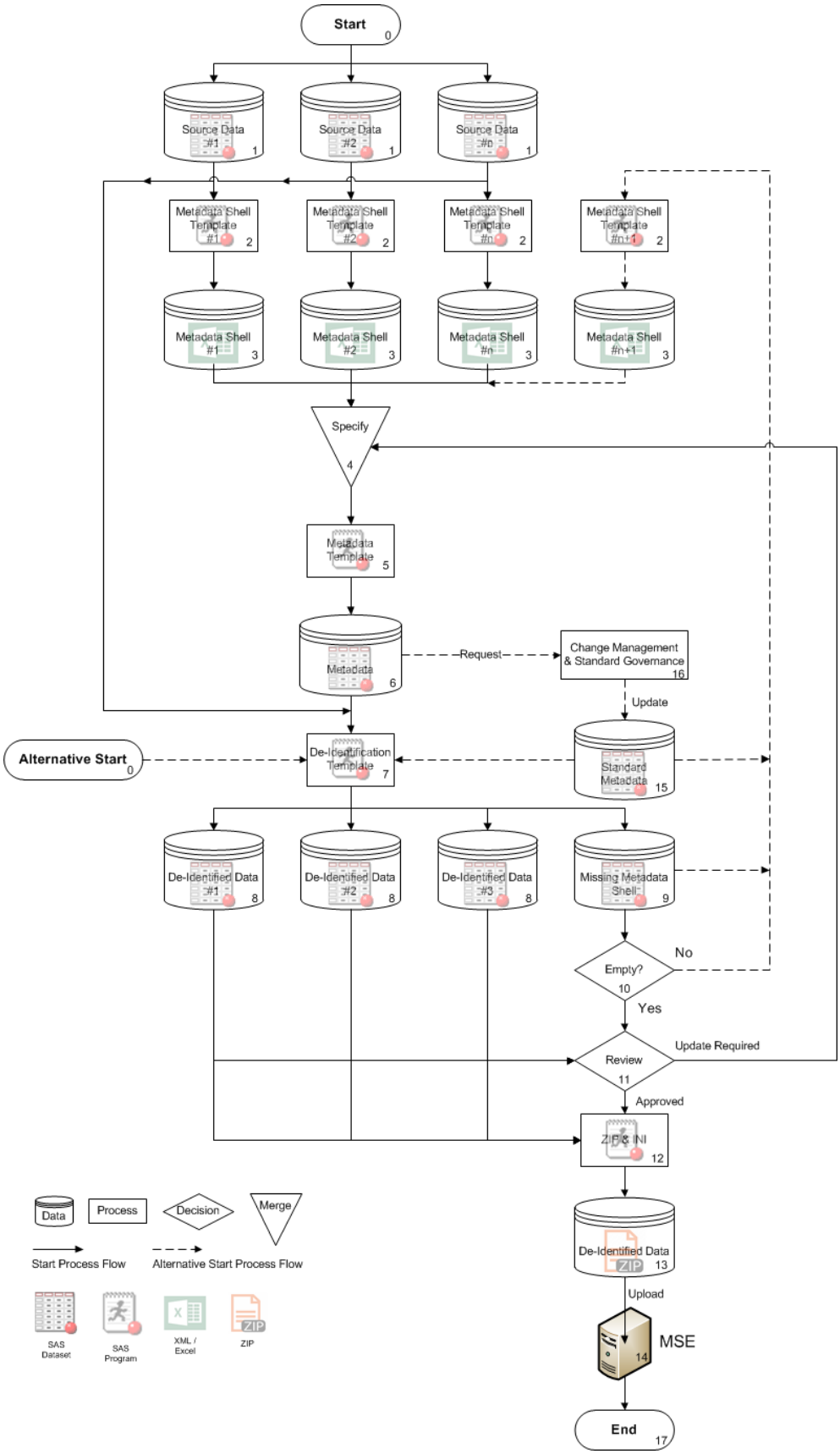

Figure 1: Process Flow Overview

Shape #	Attribute		Value	
1	Label	Format	Source Data	SAS Dataset
	Description		**Our process start with our source (i.e. not de-identified) data selected for disclosure.** The same datasets and files used for clinical reporting are read and processed for de-identification in the same Analytics Computing Environment. No copy or transfer is required. However, all of files created along the way (including de-identified datasets) are stored in a separated sub-folders structure. This structure is standardized i.e. it will be the same one every time. Access rights are managed consistently with the source data: whoever can access the source data can also access all files related to de-identification. Some scenarios require to process at the same time source data stored in separated folders. For example, a core study and its extensions are processed together to ensure consistent data de-identification.	
	Input	Output	N/A	N/A
	Refer to paper section about		N/A	
2	Label	Format	Metadata Shell Template	SAS Program
	Description		**Source data are read in order to create a metadata shell.** The objective is to scan through the entire data structure both horizontally and vertically. Horizontally, we fetch all variables attributes (name, Label / Format...). Wherever applicable, we also fetch attributes for a vertical data structure. For example typically, in data following a CDISC model, we look into SDTM TESTS and ADaM PARAMETERS values. All these attributes are outputted in an XML file readable with Excel. It is created using SAS ExcelXP tagset allowing us to interact on both the contents and the formatting from within the program. For examples, we adjust the column width to its contents, set autofilter on and freeze the column header row. All these minor formatting actions become time saver in a very repetitive process. The consistency in the output generated by the program also supports quality.	
	Input	Output	Source Data	Metadata Shell
	Refer to paper section about		- Programming Utilities and Templates	
3	Label	Format	Metadata Shell	XML / Excel
	Description		**The metadata shell is a standard XML file.** This XML file is also standard and always contains the same columns. Therefore, multiple files can easily be combined. Some columns are empty placeholders for de-identification specifications to be populated later in the process. That is why we call it a metadata shell.	
	Input	Output	N/A	N/A
	Refer to paper section about		- Standard Metadata	
4	Label	Format	Specify	N/A
	Description		**Metadata shells are combined and place holders are populated with de-identification specifications.** These specifications tell how to de-identified data: drop a variable, offset a date, set a value to missing... This manual step requires good knowledge and expertise on both clinical data and de-identification approach. Also, these specifications are not free text and rather standardized. Each column value must be within a pre-defined list or it must follow a pre-defined rule. As such, the resulting output metadata can be read, imported and processed further with minimal human Input / Output.	
	Input	Output	Metadata Shell	Metadata
	Refer to paper section about		- Standard Metadata	
5	Label	Format	Metadata	SAS Program
	Description		**The metadata XML file is converted into a SAS dataset.**	

Shape #	Attribute		Value	
	Input	Output	Metadata	Metadata
	Refer to paper section about		- Programming Utilities and Templates	
6	Label	Format	Metadata	SAS Dataset
	Description		**The metadata is a standard SAS dataset.**	
	Input	Output	N/A	N/A
	Refer to paper section about		- Standard Metadata	
7	Label	Format	De-Identification	SAS Program
	Description		**Source data are de-identified.** This program reads all source data and the metadata file. It de-identifies data as per specifications contained in the metadata. Also, an extra SAS dataset is created containing any missing metadata. If a variable in the source data does not appear in the metadata, it is dropped by default and details are included in the missing metadata.	
	Input	Output	- Source Data - Metadata	- De-Identified Data - Missing Metadata Shell
	Refer to paper section about		- Programming Utilities and Templates - De-Identification Tool - Validation Approach	
8	Label	Format	De-Identified Data	SAS Dataset
	Description		**De-identified data are outputed in their respective sub-folders.**	
	Input	Output	N/A	N/A
	Refer to paper section about		N/A	
9	Label	Format	Missing Metadata Shell	SAS Dataset
	Description		**Missing metadata structure is similar to the metadata.** Missing de-identification specifications can be defined and they can be combined easily with the rest.	
	Input	Output	N/A	N/A
	Refer to paper section about		- Standard Metadata	
10	Label	Format	Empty?	N/A
	Description		**Is the missing metadata file empty?** If no, the extra dataset contains all details. At this point, the process allows to loop back and run through a 2nd iteration. Therefore, we can specify these missing pieces in light of the entire metadata rather than separately. If yes, then we can proceed to the next step.	
	Input	Output	Missing Metadata Shell	N/A
	Refer to paper section about		N/A	
11	Label	Format	Review	N/A
	Description		**De-identified data and metadata are reviewed for approval.**	
	Input	Output	- De-Identified Data - Metadata	N/A
	Refer to paper section about		- Validation Approach	
12	Label	Format	ZIP & INI	SAS Program

Shape #	Attribute		Value	
	Description		**Creation of a single ZIP file.** This program zips all de-identified files together in a pre-defined folder structure. An INI text file is written and included along in the ZIP archive. Both the folder structure and the INI are pre-requisites so that the ZIP file can be uploaded to the Environment where external researchers will be accessing de-identified data.	
	Input	Output	De-Identified Data	De-Identified Data
	Refer to paper section about		- Programming Utilities and Templates	
13	Label	Format	De-Identified Data	ZIP
	Description		**The ZIP archive stores all de-identified data in a standard folder structure. An INI file provide settings for the upload.**	
	Input	Output	N/A	N/A
	Refer to paper section about		N/A	
14	Label	Format	MSE	SAS Dataset
	Description		**The ZIP archive is uploaded to the Multi Sponsor Environment. Data files are automatically extracted based on the folder structure and the INI file.**	
	Input	Output	De-Identified Data	De-Identified Data
	Refer to paper section about		- Multi-Sponsor Environment	
15	Label	Format	Standard Metadata	SAS Dataset
	Description		**Standardization of our de-identification specifications.** As Novartis goes through multiple de-identification exercises, we are building up standard metadata. These metadata are re-usable from one study to another.	
	Input	Output	N/A	N/A
	Refer to paper section about		- Standard Metadata	
0	Label	Format	Alternative Start	N/A
	Description		**Leveraging the standard metadata.** Over the long term, the standard metadata will simplfy the de-identification process. So rather than starting from the source data to build specific metadata, our process allows an alternative start using the standard ones. In this alternative route, we go through a 1st iteration running the de-identification program with the standard metadata. Then any missing metadata is combined with the standard ones while going through a 2nd final iteration.	
	Input	Output	N/A	N/A
	Refer to paper section about		N/A	
16	Label	Format	Change Management & Standard Governance	N/A
	Description		**Maintaining our standard metadata.** While following the alternative start route, new de-identification specifications are created from missing metadata (during the 2nd iteration). We may decide to include this new pieces in our standard metadata. To maintain quality in our standards, such update follow a very well define change management process including standard governance.	
	Input	Output	Metadata	Standard Metadata
	Refer to paper section about		- Validation Approach	
17	Label	Format	End	N/A
	Description		**End of the process flow.**	

Shape #	Attribute		Value	
	Input	Output	N/A	N/A
	Refer to paper section about		N/A	

DE-IDENTIFICATION TOOL

The Novartis de-identification tool (de-id tool) consists of a standalone set of SAS macros. It is designed to allow data de-identification in accordance with the Safe Harbor Approach described in the US Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule. In this approach, eighteen data elements are identified as Personally Identifiable Information (PII) including Names, geographic subdivision information, dates etc…

The de-id tool is designed to apply five different methods of de-identification on the input data. They can be categorized into two types:

- Masking (e.g Subject identifiers, Dates)
- Removal (e.g Free text Verbatims, Investigator Names, Subject Initials)

Masking includes translation, date offset and age categorization methods. Removal includes dropping the data fields, or setting the values to null/missing. Each of these methods will be discussed in detail with examples.

As data passes through the de-id tool, one of the above methods will be applied to all the data fields which contain PII. A dictionary of data fields names with pre-defined methods of de-identification is leveraged by the tool to apply de-identification techniques on the data. This dictionary is referred to as de-identification metadata. Complex processing with robust algorithms are required to handle multiple datasets and achieve consistent and complete de-identification.

The de-id tool also has the ability to handle multiple input data libraries in a single run. This feature allows the tool to de-identify “basket” studies (i.e. same patient enrolled in different clinical trials) or core and extension studies together without losing the subject relationship across studies. For example, when de-identified together, subject 001 in the extension study will be de-identified to the same value as to what the same subject is de-identified to in the core study. In addition, the dates of a subject in both core and extension or basket studies, are offset by the same random number. This allows the data recipient to perform analysis on all the data together without the need to doing any pre-processing. However, it was noticed that, processing large volumes of data adds significantly to the processing time required to de-identify the data.

MASKING TECHNIQUES

Data translation is based on a hexadecimal algorithm. The de-id tool applies this technique in a way that preserves the ability to combine datasets and perform analysis. The key takeaway here is that the de-id tool translates values consistently across all de-identified datasets of the study. Aditionally, this technique also applies consistent translation of values across multiple studies. This is especially useful to maintain subject relationships when a subject participates in more than one study. One may encounter this scenario frequently when de-identifying core and extension data of a study.

Figure - 2 below shows an example translation of a subject identifier value, when de-identified separately vs when de-identified together. You may notice, when de-identified separately, the same subject 12345 is translated to two different values in output de-identified data – nn1AF, nn9CG. In this case, the subject relationships cannot be established between the two studies as the translated values are different. However, when de-identified together, the subject is translated to the same value in both the studies. This allows the data recipient to perform meaningful analysis on core and extension studies together.

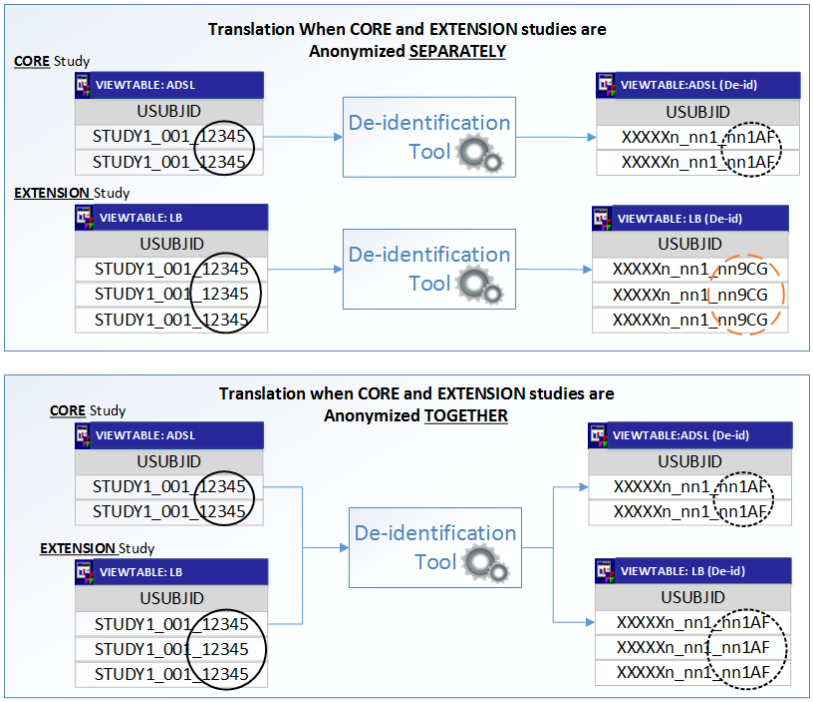

Figure 2: Examples of translation

Date offsetting probably entails the most complex processing within the macro code of de-id tool as it poses many challenges. To enable the maximum utility of the data post de-identification, all full dates will be provided with a random offset per subject. Any partial dates however will be set to null to prevent re-identification. Date components are de-identified to the same values as would their corresponding full date fields.

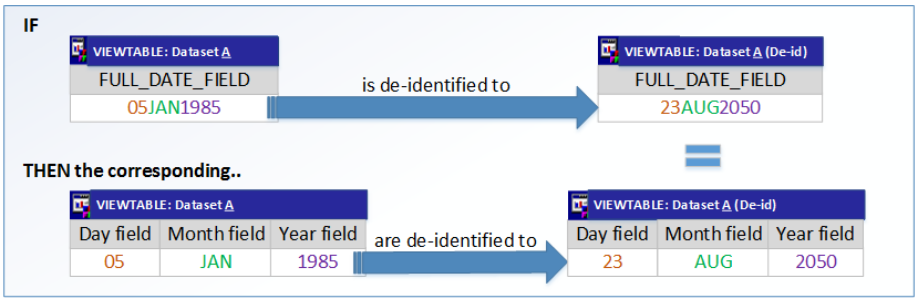

Figure 3: Example of date component offsetting

The last masking technique, age categorization, is designed to group age values above 90 to reduce the risk of patient re-identification. Additionally, the de-id tool allows the user to specify an interval by which all the age values less than 90 can be categorized to. The de-id tool is designed to perform this efficiently. This was found to significantly reduce the risk of re-identifying a patient from the dataset. Refer to Risk Based Approach section for details about risk of re-identification.

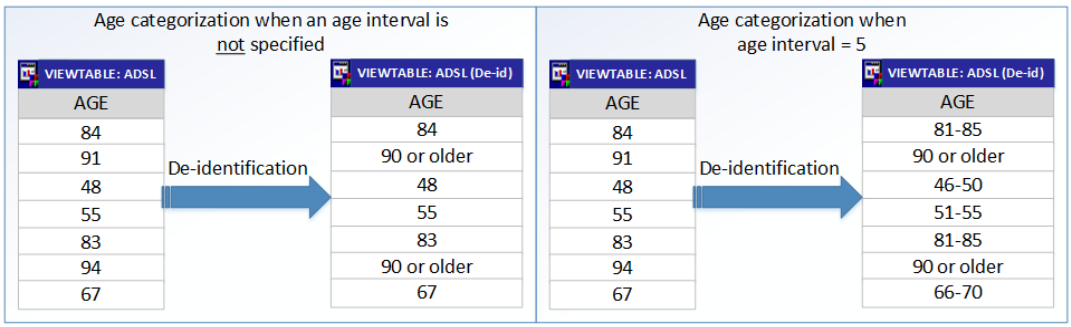

Figure 4: Examples of age categorization

REMOVAL TECHNIQUES

The removal techniques are comparatively more straight forward than the masking techniques. Variables can be dropped from the de-identified dataset. Alternatively, variables can be retained but all data values set to missing.

DATA STRUCTURE

In clinical healthcare data we quite often encounter vertical data structure such as in CDISC findings domains or in supplemental qualifier. The de-id tool is designed to handle such data along with conventional horizontal data structure. The vertical data structure poses unique challenges when only the results of certain parameters/assessments may contain PII. The results of all remaining parameters/assessments may not contain any PII. With this in mind, the tool is equipped with the ability to de-identify only the results of selected parameters.

METADATA

STANDARD METADATA (VS) METADATA

Standard metadata refers to the mode of de-identification for a specific variable that will be repeated across clinical trials. Standard metadata is data model specific: one set is built for CDISC or legacy models. The methods of de-identification which are populated in this file follow Novartis Global Data De-identification Standards (a set of de-identification standards established based on the HIPAA Privacy rules). For CDISC data model, the PhUSE De-identification Standards for SDTM can be leveraged as a starting point to build on.

STANDARD METADATA IN THE DE-IDENTIFICATION PROCESS

The standard metadata is built iteratively either after de-identifying each study, or in batch mode after certain number of studies are de-identified. The dotted line in figure 1 shows the process flow of how the standard metadata is built from the de-identification of a study. The standard metadata is the input for the de-identification tool to populate the methods of de-identification for study variables. Figure 5 shows a simplified process map of how Standard metadata are built.

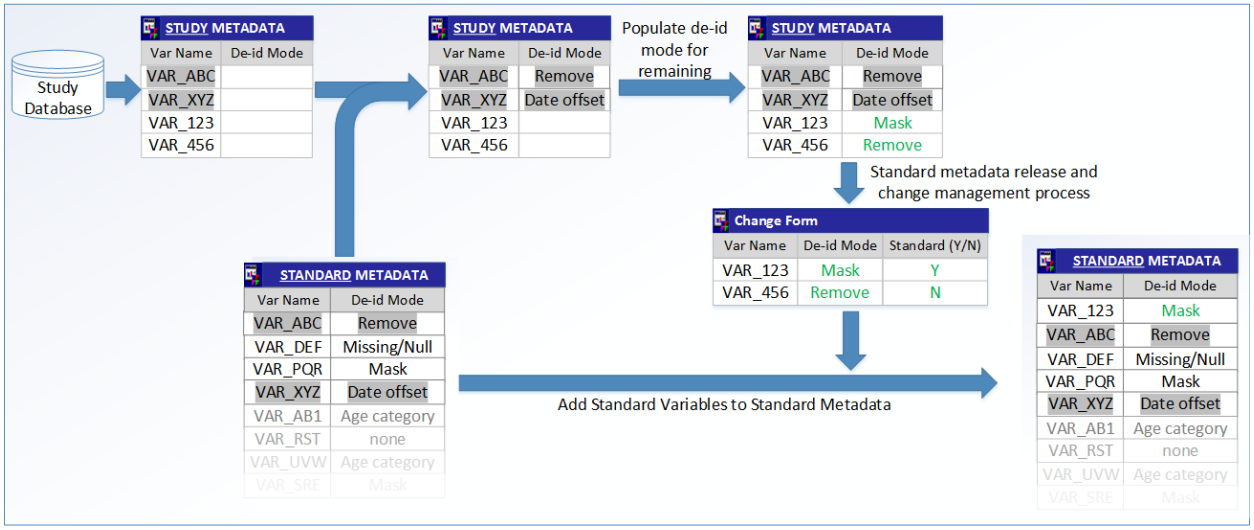

Figure 5: Standard Metadata update process

Changesto standard metadata are controlled by a validated change management process. This process includes a standards governance of the de-identification method defined for each data field (refer to the Validation Approach section). Approved changes are applied to the standard metadata dataset by a validated SAS macro. This SAS macro checks the existing metadata against the requested changes and generates a summary report detailing the changes. The macro is designed to performthe following actions on the standard metadata:

- Add a new record to existing standard metadata,
- Update the method of de-identification applied to an existing record,
- Retire a record from the existing standard metadata.

The macro also maintains full history and traceability of our metadata.

VALIDATION APPROACH

Because of its non-reproducible nature, validation of de-identified data through double programming is impossible. Therefore Novartis ensures quality in three stages: validation of de-id tool, Change Management & Standard Governance, and a review of the de-identified data.

DE-ID TOOL VALIDATION

A stringent validation process was performed prior to the production release of the de-id tool to ensure that final product met the de-identification requirements. The Novartis Quality group recommended that the de-identification tool be validated following a V-model shown in figure 6. This validation process involved testing the tool in the following three areas: Installation Qualification (IQ), Operational Qualification (OQ) and Performance Qualification (PQ). Rigorous testing with multiple test cases/scenarios is performed to ensure robustness and quality of the tool. The validation procedure involved extensive documentation at each step of the process, along with review and approval for audit readiness. This exercise required a great amount of collaboration between Quality compliance, Quality management and IT support teams.

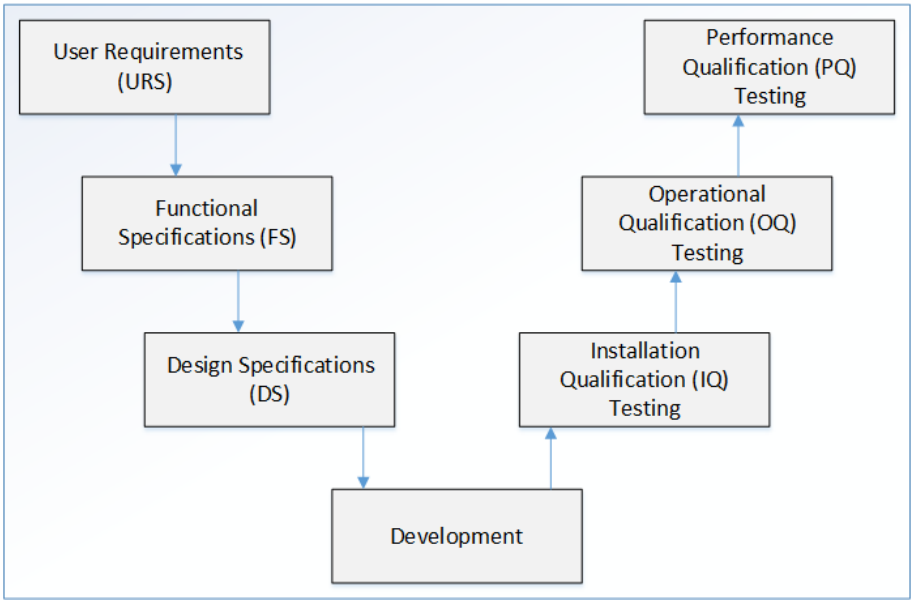

Figure 6: De-identification tool Validation model

CHANGE MANAGEMENT & STANDARD GOVERNANCE

A change management process including standard governance is used to maintain our standard metadata. It is integrated into the overall de-identification process. The standard governance involves a group of individuals who are subject matter experts in the data de-identification area. They establish rules that help in distinguishing between standard and non standard variables.

REVIEW

The third stage for ensuring quality – Data review, is also integrated into the de-identification process. The review step is designed to be performed by two groups:

- De-identification team: Individual who runs the de-id tool to de-identify the data.
- Study team: At least one Statistician and one Programmer with a knowledge of the therapeutic area or who supported the reporting of the data.

The study teams remain responsible for the review and approval of de-identification metadata and de-identified data. However, as the de-id tool is strictly validated to apply the definitions, the teams mainly focus on reviewing the de-identification metadata. At times, the study teams provide valuable insights into the type of data that may or may not be de-identified to fit the analysis needs. These insights and decisions certainly ensure quality and data utility. The data review performed by the de-identification team ensures that all personally identifiable information is de-identified.

PROGRAMMING UTILITIES AND TEMPLATES

A set of programming utilities and templates to improve efficiency in the process were built. The utilities brings simplicity and consistency. They are paired with template programs providing flexibility to the programmer to address study specific needs.

UTILITIES

The programming utilitites are SAS macros that mostly support streamlining the data de-identification process and/or creation of metadata file. The utility programs do not effect the data itself. As these utility programs are designed to

carryout isolated tasks in the process, it is agreed that these can be validated as separate macro programs and not as part of the de-id tool. Multiple programming utility macros were created and each of these are discussed in sections below.

%DS2XML

This utility macro creates an xml file designed to be opened with Excel from a dataset.

Key features: The de-identification process involves review of the de-identification metadata file at multiple points in the process. Therefore, this macro is very often used to generate an excel version of the metadata file to allow the reviewer to filter the file and add comments as needed. The macro employs ExcelXP tagsets to define a consistent style to the output Excel file.

%RULE2DEFDATA

The %rule2defdata macro is designed to populate modes and other attributes in metadata dataset as per Novartis Global Data De-identification Standards.

Key features: In data de-identification process, two different catagories of rules are defined.

- Rules to define and distinguish standard vs non-standard variables. Refer to Metadata section for details on this.
- Rules defined for modes and attributes of metadata variables as per Novartis Global Data De-identification Standards.

As shown in figure 7, the macro employs an excel file which contains SAS code for the rules defined. The macro applies these SAS codes to populate appropriate de-identification mode values that should be specified for successful de-identification.

NOTE: The rules defined here are also reviewed and approved by the standards governance.

Rules are processed in sequence as per a sequence number (not shown in picture here) defined in the excel file. In one macro run, multiple rules can be processed for one variable. However, the result will be of that of the last processed rule.

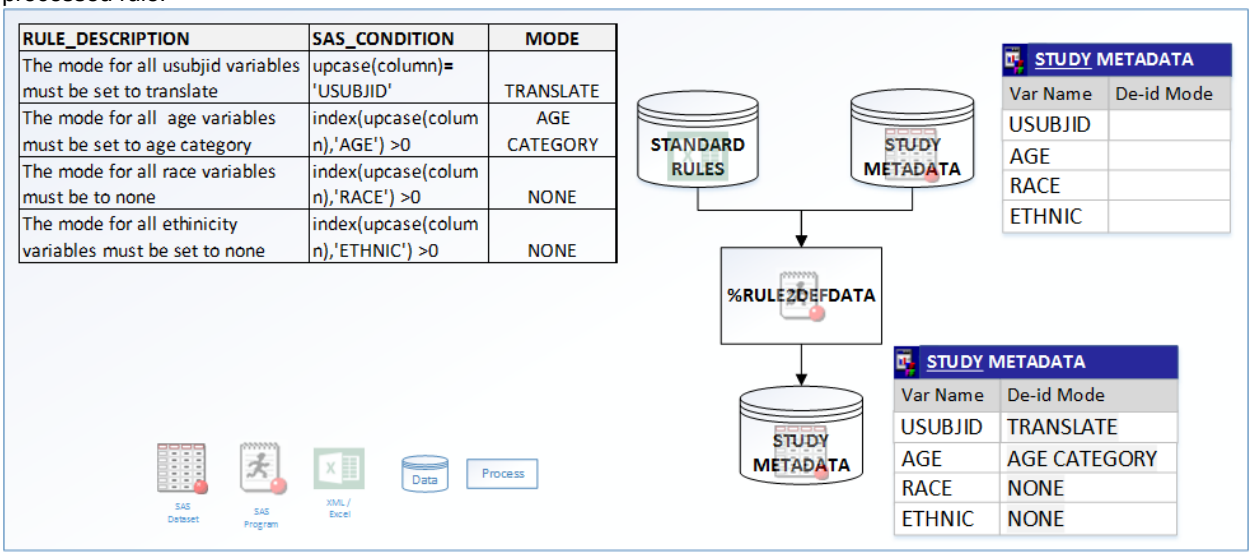

Figure 7: Description of how standard rules are leveraged to populate modes in metadata

%GPS2SSR

The %gps2ssr macro is designed to programmatically create a zip file of de-identified datasets, in a way that will allow upload to the Multi-Sponsor Environment (MSE).

Key features: After successful de-identification, the de-identified datasets will need to be uploaded to the MSE to allow controlled sharing with third parties. However, the MSE only accepts zip files which also must contain an initialization file. The %gps2ssr macro is designed to programmatically create this zip file of datasets from multiple libraries, along with the initialization file.

TEMPLATES

On a high level, data de-identification is a repetitive process for each study/group of studies. However, each study is different and at times require a custom solution. Therefore, template programs were written to allow the programmers to start with an initial code also offering freedom for customization to fit individual study needs. They also permit integration of consistent checks like comparing lists of datasets to ensure none have been missed in the process. The 3 templates described below significantly improved our efficiency.

METADATA SHELL TEMPLATE PROGRAM

To create a metadata shell of the study variables, into which the methods of de-identification and other attributes will be populated.

Key features: As in shape #2 of figure 1, this program is employed in the beginning of de-identification process. This program takes the original data that is to be de-identified as its input. In addition to its primary purpose of creating a metadata shell, it is also designed to perform the following:

- Compares the variables in study metadata with standard metadata. If the same variable exists in standard metadata, the corresponding methods of de-identification are assumed.
- Also, pre-populates methods of de-identification based on standard rules that are defined by Standard governace. This is performed by calling the %rule2defdata utility macro.
- Create rows in metadata file to allow de-identification of vertical data.
- Check that all the datasets of the input library are specified in the metadata dataset.

METADATA TEMPLATE PROGRAM

As shown in shape #5 of figure 1, this template program creates a metadata dataset for use as input to the de-id tool.

Key features: The metadata template creates a metadata dataset from the metadata excel file created and filled out previously. It also allows the user to keep only the required columns for the de-identification macro.

DE-IDENTIFICATION TEMPLATE PROGRAM

The primary purpose of this program is to call the de-idetification macro.

Key features: In addition to calling de-identification macro, this program also allows the user to perform any pre and post processing that might be required based on the study needs. For example:

- To de-identify datasets from multiple data libraries. This is especially required in case of core and extension studies.
- Exclude/include datasets which are not required to be shared.
- This template also allows the user to delete any empty datasets within the de-identified data.

RISK-BASED APPROACH

SAFE HARBOR METHOD (VS) EXPERT DETERMINATION METHOD

In recent years, many major pharmaceutical companies are looking into performing data de-identification in one the following two methods – 'Safe harbor approach' and/or 'Expert determination approach'. As discussed briefly within the De-identification tool section, the safe harbor approach is focused on de-identifying the data elements which fall within at least one of the 18 PII described in HIPAA Privacy rule.

The expert determinarion method takes many other different factors into consideration. In this approach, the residual risk of re-identifying a subject from the de-identified data is quantified and compared against a risk threshold. Other key factors include – Context of data release and security controls, data sharing and confidentiality agreements with data recepients, sensitive terms, uniqueness etc. The EMA policy–0070 also emphasized on performing a risk of re-identification for public release of redacted CSR's and data in general.

Even though risk-based de-identification methods have been in use for some time for de-identifying health data, they are new to the pharmaceutical industry, and specifically in the context of sharing clinical trial data. It is necessary for industry to develop expertise in this area and to customize these methods for clinical trial data. This learning and expertise development process means that the pace to readily integrate these techniques into our process has been slow. Although Novartis has started this effort already and we hope to report on this in the future.

DESCRIPTION OF PILOT

The current data de-identification process follows the safe harbor approach. In addition to this, the risk-based approach is also under investigation. This approach is not integrated into our process shown in figure 1, but is discussed here based on experience with a pilot performed outside the current de-identification process. The end goal of the pilot is to maximize data utility whilst keeping the risk of re-identifying a subject less than the risk threshold. Multiple iterations of risk assessments are performed for each study. In each iteration the identifier fields in the data are adjusted to bring the risk values close to the risk threshold. Possible further de-identification steps are implemented to lower the risk if needed.

KEY LEARNING

The risk of re-identification involves complex statistical methods and principles in order to account for multiple factors that contribute to the risk both intrinsic and extrinsic of the data. Therefore, input from statistical personnel or subject matter experts with deep understanding of requirements in this area is key to implement these techniques. The expert determination method is the only quantitative measure balancing data utility and subjects' privacy protection. The evolution of de-identification practices is intended to give a defensible argument that we are protecting privacy and reducing the company`s risk exposure and liability.

MULTI-SPONSOR ENVIRONMENT

The solution used to share their de-identified data is left to the discretion of sponsors. Novartis is an active user of the MSE solution, used by multiple sponsors to share their de-identified data with external researchers in a controlled and uniformed platform. The controls in place prevents the external researcher from downloading the clinical data. Other features of MSE include analytic and reporting tools (SAS and R), access rights management, secure file transfer and management of upload/download requests from users.

IN OUR PROCESS

The Programming Group and the Clinical Disclosure Office share responsibilities in the management of this environment. Programmers are responsible for the data de-identication and upload of all files to the system. The Clinical Disclosure Office is responsible for granting access to data in the MSE along with implementing a data sharing agreement with external parties. Such controls contributes towards lowering the risk of re-identification. In the data de-identification process, these roles and responsibilities are to be outlined and detailed in the relevant SOP and WP.

FUTURE

The recent guidelines in Part A of EMA policy-0070 recommended a public release of redacted CSR. Similarly, the part B of this policy on clinincal trial data de-identification may recommend either Public, semi-public or controlled data sharing. Therefore, it is highly recommended to remain open to adapt to changing regulations and guidances.

CONCLUSION

You now have a good understanding of the different pieces in Novartis data de-identification process as well as how they are implemented. We hope we have helped you to design or improve your own approach within your organization.

Such process should always go through continuous improvement. It needs to evolve to adapt to a changing environment. In the near future, we believe new regulations will have significant impact. Technological solutions will be created or updated to support compliance to these same regulations. More standards will be created, refined and shared as stakeholders collaborate and publish in this emerging landscape.

REFERENCES AND RECOMMENDED READING

1. Health Insurance Portability and Accountability Act (HIPAA) Resources
   http://privacyruleandresearch.nih.gov/
2. European Medicines Agency policy on publication of clinical data for medicinal products for human use
   http://www.ema.europa.eu/docs/en_GB/document_library/Other/2014/10/WC500174796.pdf
3. External guidance on the implementation of the European Medicines Agency policy on the publication of clinical data for medicinal products for human use
   http://www.ema.europa.eu/docs/en_GB/document_library/Regulatory_and_procedural_guideline/2016/03/WC500202621.pdf
4. Clinical Study Data Request
   https://www.clinicalstudydatarequest.com/
5. Novartis Global Clinical Data De-identification Standards
   https://www.clinicalstudydatarequest.com/Documents/Novartis-Global-Clinical%20-Data-Anonymization-Standards%20(v2).pdf
6. Pharmaceutical Users Software Exchange - Data Transparency
   http://www.phuse.eu/Data_Transparency.aspx
7. PhUSE De-identification Working Group: Providing identification Standards to CDISC Data Models
   http://www.phusewiki.org/docs/Conference%202015%20DH%20Papers/DH01.pdf
8. TransCelerate Biopharma Inc.
   http://www.transceleratebiopharmainc.com/wp-content/uploads/2015/04/Data-Anonymization-Paper-FINAL-5.18.15.pdf

ACKNOWLEDGMENTS

We would like to thank Chris Hurley (MMS Holdings Inc.) for his input and for presenting this paper in Barelona. We would like to thank as well Gregory Pinault, Guillaume Breton, Janice Branson, Joseph Rowley, Jacques Lanoue (Novartis), Stephen Korte, Geordan Chester and Khaled El Emam (Privacy Analytics) for their review.

CONTACT INFORMATION

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

Benoit Vernay
Novartis
benoit.vernay@novartis.com

Ravi Yandamuri
MMS Holdings Inc.
ryandamuri@mmsholdings.com