PharmaSUG 2015 - Paper HA03

Now You See It, Now You Don't -- Using SAS to De-Identify Data to Support Clinical Trial Data Transparency

Dave Handelsman, d-Wise, Research Triangle Park, North Carolina

ABSTRACT

Clinical trial data transparency initiatives will only be successful if the data being shared is properly de-identified in order to protect patient confidentiality and comply with national regulations, while still supporting investigation and analysis. In this emerging area, however, the rules regarding clinical trial data de-identification can be confusing, open to interpretation and difficult to understand.

At a basic level, de-identification means that someone accessing trial data should not be able to match an individual patient's data to a real-life individual. This means not only obfuscating patient ID information (patient number and site number, for example), but also masking all dates, eliminating references to sensitive terms like "HIV", and a wide variety of additional, and often confusing, rules. All of these data modifications must be done in such a way that the clinical trial data can still be successfully analyzed on its own, or when combined with additional trial data. To further complicate matters, this additional trial data may frequently be provided by multiple biopharmaceutical companies.

Many companies actively engaged in clinical trial data transparency initiatives are using SAS to perform de-identification. Additionally, they have published their individual de-identification strategies in order for patients to understand how their confidentiality will be protected, and to inform researchers how they will need to prepare to analyze the data. This paper will review the company strategies and the various SAS approaches to de-identification in use today.

INTRODUCTION

Clinical trial data transparency initiatives have rapidly evolved from the briefest of conversations to widely adopted business solutions across the biopharmaceutical industry. Critical to the successful implementation of these transparency solutions is the ability to share data that has been properly de-identified in order to protect patient confidentiality and comply with national regulations, while still supporting investigation and analysis. In this emerging area, however, the rules regarding clinical trial data de-identification can be confusing, open to interpretation and difficult to understand.

Patient-level data de-identification by itself, however, is not new. Biopharmaceutical companies have been de-identifying data for years for either internal use by other departments, or to share within external researchers on a very limited basis. What has changed, however, is the energy and focus on clinical trial data transparency, and, by extension, on the patient-level data de-identification rules and accompanying processes that provide the data to support clinical trial data transparency.

There are currently three well-documented efforts regarding the development of de-identification industry rules:

- A number of leading biopharmaceutical companies are publishing their de-identification rules as part of a shared clinical trial data transparency website (www.clinicalstudyrequest.com)
- There is a dedicated de-identification project for SDTM-structured data within the PhUSE organization
- TransCelerate is developing a de-identification methodology.

While there is certainly overlap among those companies participating in these efforts, it's clear that there is no current consensus regarding the canonical set of de-identification rules that are applicable to clinical trials data.

Even with a clear set of rules, many companies will struggle to match their de-identification capabilities with the necessary effort to deliver de-identified data. Historically, delivering de-identified data has been accomplished by writing one-off SAS programs (or in some cases, developing a library of SAS macros), but this approach has been difficult due to the episodic nature of de-identification activities. Until recently, de-identification work was performed only intermittently. The resources assigned to the task typically were not facing significant delivery pressures, and were frequently unfamiliar with the details of the de-identification process. Because de-identification was not a high priority, no single community of users owned this responsibility, and each de-identification activity became a one-off project, even if a library of SAS tools was available. With the emergence of transparency initiatives, de-identification has become a critical and highly visible task that must be completed quickly, efficiently and accurately, and with the necessary documentation and controls that enable confidence in the de-identification process.

EVOLVING CLINICAL TRIAL DE-IDENTIFICATION RULES

The site *ClinicalStudyDataRequest.com* identifies eleven companies that have taken a collaborative approach to clinical trial data transparency. Of these, eight have published their de-identification strategies. While other biopharmaceutical organizations are certainly performing de-identification activities, these eight companies provide an excellent starting point for understanding the complexities associated with patient-level data de-identification.

CONSISTENT DE-IDENTIFICATION RULES

A variety of data fields are handled similarly across the eight companies. These include the 18 fields that must be de-identified as documented by HIPAA, but which typically have very little presence in clinical trials data. Included in these fields are *name*, *address*, *social security number*, etc. While any de-identification effort should certainly be diligent in ensuring such information is not present in the de-identified data, these fields are not typically the direct focus of clinical trial de-identification activities.

There are, however, some fields that are de-identified consistently among the eight companies. It should be noted that, in the case of data that conforms to CDISC structures, the fields may be identified with the same name. For other data, however, it will be more appropriate to align the field definition with the de-identification rule. The fields that are handled consistently across the eight companies are shown in Table 1.

Data Type	**De-Identification Action**
Personally identifiable information (PII, such as HIPAA-specified identifiable data)	Remove
Unique subject ID	Re-code
Subject ID	Re-code
Initials	Remove
Investigator names and locations	Remove
Age > 89	Remove age and categorize as Age >=90
Dictionary coded terms	Retain

Table 1: Consistently De-identified Fields among the Eight Biopharmaceutical Manufacturers

INCONSISTENT DE-IDENTIFICATION RULES

There are many other fields (or types of fields) that are handled differently among the eight companies. These fields are shown in Table 2. The different de-identification actions are not indicative of a right or wrong approach, but are representative of the emerging nature of the de-identification process. In previous years, each company handled patient-level data de-identification as an internal process, and any de-identification actions were typically shared only with the downstream data consumers. The emergence of clinical trial data transparency has changed an internal process to a much more public process, and it is only natural that there is divergence on the rules and methodology associated with patient level de-identification.

Data Type	De-Identification Action
Date of Birth	There is little consensus on how to de-identify *date of birth* except for the fact that it must be de-identified. Approaches include deleting month and day while keeping the year, removing the entire date value, and replacing the value with a calculated age based on a company defined reference date field.
Date fields, except for Date of Birth	There are two main approaches regarding the de-identification of dates. For several companies, each date value is replaced with a calculated study day value based upon a series of prioritized anchor date fields. Each patient would have an anchor date value selected from the prioritized date fields, and all study day calculations for that patient would be based upon the selected anchor date. Alternatively, an offset date approach is used by several companies. In this case, a patient is assigned a random offset value in some pre-specified range, and that offset value is used to calculate a de-identified date value from each original data value. In this way, the timing between dates is preserved while the actual dates are de-identified.
Free-text values	Several companies indicate that free-text (verbatim) values may be reviewed to determine if the identifiable segment of a free-text value should be selectively redacted in order to preserve the scientific value of the free text value. Other companies simply remove every free-text value.
Patients per Site, Sites per Study, Sites per Country	The least definitive approach to de-identification is related to addressing sites with small numbers of patients and, similarly, countries with small numbers of sites, and similar low-frequency situations. In these cases, the relatively low frequency of information increases the risk of re-identification. Some companies indicate that sites with fewer than 10 patients will be "addressed" (aggregated or removed), but in general there is little consistency between companies for this situation.

Table 2: Inconsistently De-identified Fields among the Eight Biopharmaceutical Manufacturers

Additionally, there are some de-identification rules that are referenced in only a subset of the eight published documents. Examples include:

- Variable names (not values) should be reviewed for location information. For example, the variable name may indicate a particular laboratory location, which could in turn indicate a site or country.
- Extension and other follow-on studies should consistently re-code patient identifiers across studies in order to maintain the continuity of data for an individual patient.

ALTERNATIVE DE-IDENTIFICATION RULE SETS AND STRATEGIES

Formal industry collaborations regarding de-identification are in process. The PhUSE De-Identification Working Group has recently closed comments on its draft de-identification rules, and the final document is expected to be published in Spring 2015. This document is focused exclusively on the de-identification of CDISC SDTM 3.2, however, and will only serve as a rough guide for data that conforms to other CDISC data models, and an even rougher guide for older data that may vary considerably from trial to trial. TransCelerate is expected to publish its model approach to de-identification and anonymization later in 2015.

In other industries, data is de-identified using much more complex algorithms and processes. In these cases, not only is the more readily identifying information addressed, but the core data itself is statistically modified to be representative of (but different than) the original data. While this may decrease the risk of re-identification, it also increases the complexity of the de-identification process, frequently reduces the utility of the data and may interfere with the downstream goals associated with the entire clinical data transparency process. To date, this approach has not been the preferred methodology to support patient-level data de-identification as it relates to clinical trials data.

Similarly, the use of *expert determination*, as defined by HIPAA, is not widely adopted for clinical trials data.

SAS AND THE DE-IDENTIFICATION PROCESS

Because SAS data sets and transport files are the common data transfer mechanism within the industry, SAS is the ideal transformation engine for de-identification purposes. There are many expert users with SAS experience, and clinical trial data is easily read by SAS.

Some organizations are building SAS macro libraries to perform de-identification tasks (for example, macros to

calculate study day or a date offset), especially as these organizations see the volume of de-identification work increasing. Other organizations continue to look at SAS de-identification programs as one-off exercises, and treat each project semi-independently. That is, there is no library of macros to be called, but previously developed programs may be copied to a new de-identification project.

In either case – whether macros are being used or programs are being treated as one-offs – validation processes must be followed. Macros must be validated and locked down before they can be readily used in a new project, or extensive validation must be applied during the use of one-off SAS programs, in order to be satisfied with the reliability of the de-identification process.

It has been argued that "traditional" validation and quality control processes associated with clinical trials are not necessary for de-identification because the data is being used outside of the normal submission process. Since the externalized analysis may not follow a rigorous validation process itself, there would be no reason to enforce such a process as de-identification is applied. However, it's important to remember the critical reason driving the growth of patient-level data de-identification. The emergence of clinical trial data transparency is accelerating the adoption of transparency-related processes, and there is significant scrutiny regarding how the pharmaceutical industry will support transparency in practice. Even though the de-identified data is not being used for submission purposes, it would be foolish to deliver data that could have lower quality (in terms of cleanliness and accuracy) than what is used for submission purposes. The outcry from industry skeptics would be loud and damaging. The effort to validate the data de-identification process is certainly worth the resources and time required.

THE CHALLENGE WITH LIBRARY SAS MACROS

Libraries of SAS macros can be valuable tools in terms of being able to provide repeatable processing when the incoming data matches a known and expected structure. In many cases, however, clinical trial data varies from that known structure. Because of the validation effort that is typically required to get macros promoted into a production library, there is reluctance to update the macros except under certain circumstances (e.g., a critical bug is identified) or on a more modest schedule that can accommodate emerging de-identification rules. Instead, what frequently occurs is that macros are copied down to a project and modified, which creates the need for additional validation on a per-project basis – the exact problem the development of the production macros had been expected to address.

Alternatively, considerable data manipulation may be required to transform the data to match an existing library macro's expectation. This is especially true for older trial data that may have little conformance to CDISC data, but is likely to be true for even modern data as the CDISC models still leave considerable room for interpretation.

In either case, library SAS macros are limited in their ability to provide the efficiency that is desired when it comes to developing and executing de-identification processes.

THE CHALLENGE WITH PROJECT-SPECIFIC SAS PROGRAMMING

It is certainly possible to perform de-identification processes with "one-off" SAS programs. As with many clinical trials, the SAS programming effort tends to be similar to other projects, and it is reasonable that programs from one project could be copied to another project, used as a starting point and modified as necessary, as is frequently done for new clinical trials. The difference here is that the rules and expectations associated with the de-identification process are fairly narrow, and a specific understanding of how to implement the de-identification process must be provided. Because the nature of the work tends to be on as-needed basis, it is difficult to develop momentum and efficiency around a particular de-identification project or process, and frequently the programmer assigned to the work must start at "square one" in terms of understanding the nature of the project and the previous programs that may have been used.

This challenge exists with the library of SAS de-identification macros as well, but is much more pronounced in this situation. Here, the problem is not as simple as re-structuring data to fit a macro call, but requires adapting an existing series of SAS programs to match an emerging business problem on a project-by-project basis.

OUTSOURCING

Several biopharmaceutical companies are looking at the business problem of de-identification as yet another process to be outsourced. The outsourcing company performing this work, however, will not only have the same challenges regarding SAS as described above, but will have to adapt any existing SAS programs to match the specific needs of their different customers. This will influence the effort associated with the work and the cost associated with outsourced project.

AN APPLICATION APPROACH BASED ON SAS

De-identification, like many other processes, lends itself directly to the development of an application that does not require programmer intervention or custom-coding. A robust de-identification solution will enable the process of de-identification to be performed by trained users that are familiar with clinical trial data and the basic tenets of de-identification methodologies. The basic business process associated with de-identification is depicted in Figure 1.

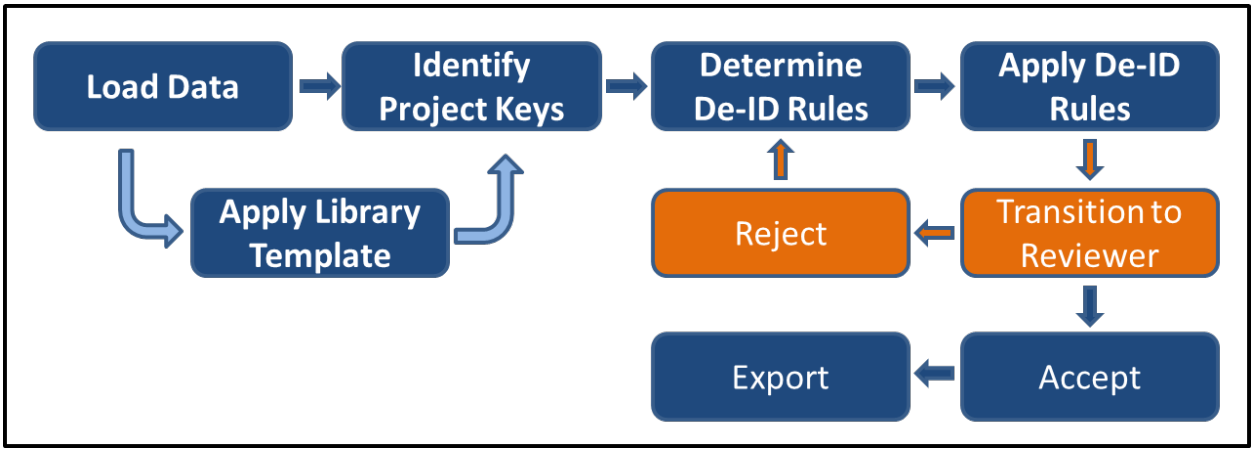

Figure 1 De-Identification Process

Because the application does not provide an interface to develop and execute custom SAS programs, it is important that all process functionality be readily accessible through the application. As an example, the user must explicitly identify the key variable(s) that logically link all data tables in order to ensure that the data is handled properly. The explicit patient identifier numbers (for example, *usubjid* for SDTM-compliant data, or perhaps *subject+patient* for older trials) will be de-identified as part of the process, and it is critical that the patient identifiers are consistently handled across tables. The application must additionally be sufficiently intelligent to enable the user to identify key variables that may be named differently between tables. This is especially necessary for older trials, where it's possible that lab data, for example, may not have identifying information that is consistent with the other project data tables.

FLEXIBILITY

Because the industry expectations regarding de-identification rules continue to evolve, the application must provide a collection of selectable rules that can accommodate different scenarios or customers. For example, some organizations may prefer to de-identify dates via a study day calculation, while others may want to offset the date. At the present time, both approaches are acceptable, and a solution that accommodates multiple approaches will provide the best means to address these evolving rules. Similarly, the downstream users of de-identified data may influence the project-specific de-identification strategy – especially if the users represent either internal or external audiences. For internal data exploration exercises, the data may require "less" de-identification than for external researcher access where there is greater risk of unexpected disclosure and re-identification.

EFFICIENCY

A critical aspect of any de-identification process is efficiency. While the details of de-identification will vary from trial to trial, there is likely to be some level of similarity between trials. This is especially true of more modern trials that conform to CDISC standards. A successful de-identification application will enable organizations to build a library of de-identification templates (keys, relevant rules and the associated fields, etc.) that can be applied to new de-identification projects as necessary. When a new project is being started, the uploaded data can be scored against the templates and the best match selected to jump-start the de-identification work. This is critical due to the episodic nature of de-identification work. Rather than being intrinsic to the clinical trial process, de-identification is performed as necessary based upon researcher demands. As can be seen in the recent surge in clinical trial data transparency projects, the demand for de-identified data has increased dramatically but the resources available to perform the work have not. A well-designed application that enables a user to quickly find a similar project for this episodic work will provide tremendous efficiency over the current business processes being used. As shown in Figure 2, this can be represented as a score that shows which template is the best fit for a particular de-identification project. In this case, the data structure of this new project is a 93% match for the *CDISC STDM* template, but only a 21% match for the *Hypertension* template.

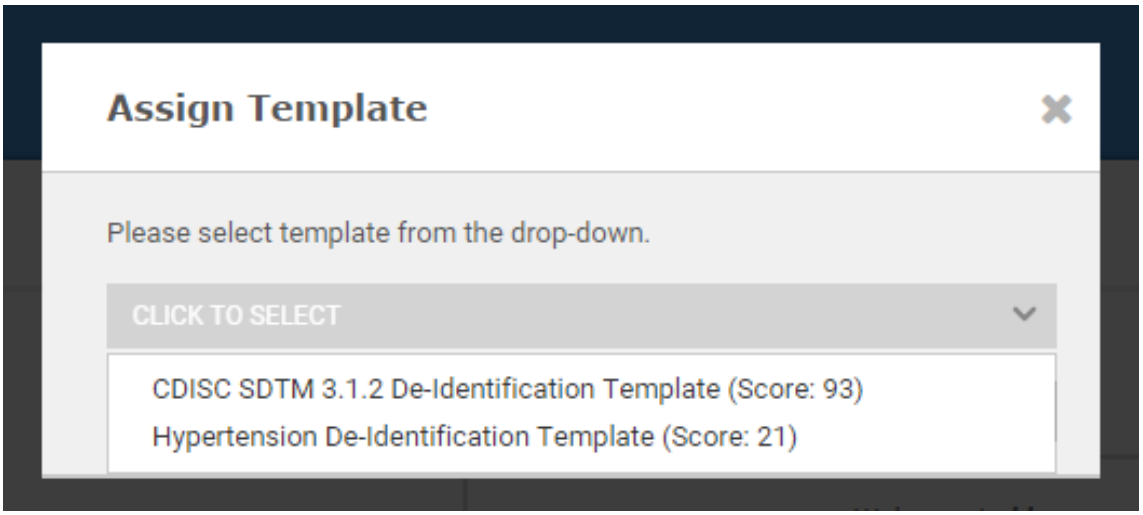

Figure 2 Assigning templates to enable efficiency

INTEGRATED WORKFLOW AND REPORTING

As with all clinical trial activities that involve transforming or deriving data, there is an industry expectation that there will be a primary author responsible for the work and a reviewer responsible for verifying the accuracy of the primary author's work product. Patient-level data de-identification is no different, and ideally the transition and sign-off processes between development and finalization will be supported via an integrated workflow. This could, for example, include an interface for the primary author to sign off on his/her work, and then request a review from another member of the team. Similarly, the reviewer would complete their assessment and sign-off directly within the application, which would indicate the project is complete, or would indicate the work was not completed satisfactorily and it would be returned to the primary author.

The final approval, as well as a full record of the applied de-identification rules, and other critical information about the de-identification process would be fully documented in a project-specific report that accompanies the finalized de-identification data as shown in Figure 3. This report should not be overlooked, as it will provide important answers to the inevitable questions that will be raised by the downstream data consumer.

Project Keys

Project ***Study_100*** used the following tables/columns for anchor dates in order of priority:

Name	Description	Columns	Method
Primary Key		USUBJID	Key

De-identification Methods

The following methods were used for de-identification for project ***Study_100*** :

Table	Column	Method	Parameters
ae	USUBJID	Key	
cm	USUBJID	Key	
dm	USUBJID	Key	
	AGE	Age Band	
	DMDTC	Study Day Offset	
ds	USUBJID	Key	
ex	USUBJID	Key	

No other columns for this project had de-identification rules assigned.

Sign-off History

The project was requested for review by user dave_deid on 14Apr2015 02:36 EDT, and was Accepted by user reviewer on 14Apr2015 02:42 EDT.

Figure 3 Sample De-Identification Report

MANAGING DE-IDENTIFICATION KEYS

The *key* fields in a de-identification project are those patient identifying fields that uniquely link all patient data across tables. When the original *key* for a patient is de-identified, the original values are replaced with a collection of randomized values that represent the new patient *key*. There is considerable industry discussion regarding what happens to the link (which may be thought of as a look-up table or index) between the original key and the de-identified key at the conclusion of a de-identification project.

At some companies, this look-up table is destroyed at the conclusion of the de-identification project in order to reduce the risk that the de-identified patient data will be re-identified in the future. By destroying the table, the primary means of re-identifying the data is eliminated. Alternatively, some companies are preserving the look-up table in the eventuality that there is a need to re-identify a patient if a safety issue is identified during downstream analysis. In this case, care must be taken to restrict access to this look-up table in order to minimize risk.

Until there is a clear industry expectation regarding the disposition of the *key* look-up table, any application will need to accommodate both scenarios.

CONCLUSION

The old ways of using custom SAS macros or one-off SAS programs is not sustainable or efficient as the business need for patient-level data de-identification accelerates. Existing business processes are too manual, and the efficiency between projects is not sufficient to perform the work in a timely and cost-effective manner. A better approach to addressing the business challenge of de-identification is the development of a clinical trials-focused application. This application should integrate the transformation functions of de-identification with fit-for-purpose workflow capabilities that provide the efficiency that is needed and expected in this critical business process.

REFERENCES

- Handelsman, Dave, An Analysis of Data De-Identification Practices in Use at Leading Biopharmaceutical Companies, Silvery Spring, MD, PhUSE Computational Science Symposium, 9-March-2015. http://www.phusewiki.org/docs/CSS2015Presentations/PP02FINAL.pdf
- http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/hhs_deid_guidance.pdf

CONTACT INFORMATION

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

Name:	Dave Handelsman
Enterprise:	d-Wise
Address:	1500 Perimeter Park Drive
City, State ZIP:	Morrisville, NC 27560
Work Phone:	919 825 4775
E-mail:	dave.handelsman@d-wise.com
Web:	www.d-Wise.com
LinkedIn:	https://www.linkedin.com/in/davehandelsman