PhUSE 2015

Paper PP30

Data De-Identification: Balancing Patient Privacy, Data Utility and Good Programming Practices

Dave Handelsman, d-Wise, Research Triangle Park, USA

ABSTRACT

Clinical trial data transparency requires data de-identification to protect patient confidentiality and comply with regulators, while supporting data analysis. The rules regarding de-identification can be confusing, open to interpretation and difficult to understand.

At a basic level, de-identification means that someone reviewing 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 masking all dates, and applying a variety of additional rules. These data modifications must be done such that the data can still be successfully analyzed on its own, or when combined with additional trial data that may span multiple companies.

A variety of de-identification rule sets have been published recently, and companies are applying different technical strategies to perform de-identification. This paper will review the state of de-identification within the industry, and assess the strategies being used.

INTRODUCTION

For many years, biopharmaceutical companies have been de-identifying data for either internal use by other departments, or to share with external researchers on a very limited basis. In the past two years, however, the energy and focus on clinical trial data transparency have demonstrated the need to expedite the process of patient-level data de-identification, Simply put, the old ways of de-identifying data have had little consistency between companies and don't scale to meet the growing need for de-identified data.

With the emergence of transparency initiatives (and other data sharing strategies), de-identification has become a critical and highly visible task that must be completed quickly, efficiently and accurately. Additionally, the necessary documentation and controls that enable confidence in the de-identification process across the industry must be directly connected with the de-identification processing itself.

DE-IDENTIFICATION RULE SETS

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

- A number of leading biopharmaceutical companies have published their de-identification rules as part of a shared clinical trial data transparency website.[1]
- There is a dedicated de-identification project for SDTM-structured data within the PhUSE organization.[2]
- TransCelerate has developed a clinical trial de-identification methodology.[3]

These rule sets all build on existing strategies related to HIPAA, Safe Harbor and Expert Determination. While there is certainly overlap among those companies currently de-identifying clinical trial data, it's also clear that there is no current consensus regarding a canonical set of de-identification rules that are applicable across the pharmaceutical industry.

HIPAA, SAFE HARBOR AND EXPERT DETERMINATION

Clinical trial de-identification strategies generally fall into two broader approaches: Safe Harbor and Expert Determination, both of which are associated with HIPAA section 45 part 164.514 of the Code of Federal Regulations (45 CFR 164.514). Each of these strategies has been designed to address a de-identification approach that is broader than that which is relevant to clinical trial data, and neither is an ideal fit for either the data or risk associated with clinical trial data re-identification.

[1] www.clinicalstudyrequest.com
[2] http://www.phuse.eu/Data_Transparency.aspx
[3] http://www.transceleratebiopharmainc.com/wp-content/uploads/2015/04/CDT-Data-Anonymization-Paper-FINAL.pdf

Safe Harbor describes 18 categories of data that require de-identification. These categories include, as an example, name, address, social security number, email address and URLs, most of which are unlikely to appear in clinical trial data. Dates, however, are specifically identified and have relevance to clinical trials. These identifying categories may also be extended to investigator information included in the clinical trial data. Because Safe Harbor and clinical trial data are not fully aligned, simply applying the Safe Harbor rules is insufficient to properly de-identify clinical trial data.

Expert Determination applies a more statistical and hands-on approach to the task of de-identification. In this case, the data is reviewed to understand the risk that it might lead to re-identification, and this risk is calculated not only on the data values themselves, but on the context of the data, available secondary sources of similar data, as well the intended use of the data. Data, for example, that will only be accessible internally within a company would be assessed to be at a different (and presumably, lower) risk for re-identification than data to be published on a portal accessible to the general public.

While Expert Determination may provide additional confidence regarding de-identification for clinical trials, the effort associated with applying Expert Determination is significantly more than that associated with a Safe Harbor or enhanced Safe Harbor strategy. It is not clear that project-by-project Expert Determination provides sufficient differentiated value to justify the associated cost, resource and time commitments.

COMPANY-PUBLISHED DE-IDENTIFICATION STRATEGIES

The site ClinicalStudyDataRequest.com identifies twelve companies that have taken a collaborative approach to clinical trial data transparency, and each company has published its own de-identification strategy. While other biopharmaceutical organizations are certainly performing de-identification activities, these twelve companies provide an excellent starting point for understanding the complexities associated with patient-level data de-identification.

Within these companies' published rules, there are a number of similarities regarding how de-identification is handled. These similarities include the 18 fields that must be de-identified according to Safe Harbor but which typically have very little presence in clinical trials data. However, fields such as *unique subject ID*, *subject ID*, *initials*, *investigator names*, *age* and *dictionary terms* tend to be addressed consistently from a de-identification perspective.

The de-identification of other fields, however, varies considerably across these companies. These differently-handled fields include *date of birth*, other date fields, free-text values, and site-related operational data (patients per site, sites per study, sites per country, etc.). Notably there are fewer differences among those companies that were the earliest adopters of this initial collaborative approach. As more companies have begun participating, each has brought its own diverging de-identification strategy.

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. Historically, each company handled patient-level data de-identification as an internal and individual process, and any de-identification actions were typically shared only with the downstream data consumers. The emergence of clinical trial data transparency has transformed an internal process to a much more public process, and it is only natural that there is currently divergence in the rules and methodology associated with patient level de-identification.

PhUSE WORKING GROUP DE-IDENTIFICATION RECOMMENDATIONS

In May 2015, the PhUSE De-Identification Working group published a collection of de-identification recommendations that were developed by more than 20 active participants across a broad spectrum of companies and organizations associated with clinical trials activities. These recommendations are specifically associated with data that conforms to the CDISC SDTM 3.2 data model, and may additionally serve as a rough guide associated with other de-identification rule set strategies.

The PhUSE working group identified a total of 11 broad rules which were then aligned to the hundreds of variables defined in the CDISC SDTM 3.2 Implementation Guide. Many of these 11 rules are similar to the published rules associated with the companies participating in clinical trial transparency initiatives. This should not be surprising as many of these same companies also participate in the PhUSE working group. What PhUSE has additionally provided are the explicit instructions for how to address each expected data variable.

The PhUSE guidelines present an approach to reconcile the Safe Harbor and Expert Determination de-identification strategies, where an enhanced Safe Harbor approach is recommended for the majority of the data and an Expert

Determination approach is described for low-frequency data that may increase the risk of re-identification. It is not clear, however, how much industry fully supports the added level of complexity and cost associated with Expert Determination. Based on conversations with individual working group members after the PhUSE guidelines were published, there still appears to be considerable disagreement among participants, and industry as a whole, as to the actual value and likely adoption of this additional level of de-identification for primary de-identification business use cases.

TRANSCELERATE DE-IDENTIFICATION MODEL

In April 2015, TransCelerate published their recommendations regarding the de-identification of clinical trial data. These recommendations do not go into the same level of detail as the PhUSE recommendations, nor are they focused exclusively on CDISC SDTM 3.2 data. Instead, the TransCelerate model provides a number of high-level recommendations that are consistent with both the published de-identification rules associated with the companies participating in the Transparency initiatives as well as the PhUSE recommendations.

The TransCelerate model additionally describes the importance of consistency between patient identifiers across an entire study, or associated with extension studies. Within a single study, all data deliveries (e.g., submission data and analysis data) that are de-identified should use consistent patient identifiers across these data sets. Similarly, de-identification activities associated with originating and extension studies should maintain consistency for the re-coded patient identifiers so that a patient is correctly tracked between the associated studies.

Both Safe Harbor and Expert Determination are discussed within the TransCelerate de-identification model, and, in general, an enhanced Safe Harbor strategy is recommended due to the nature of the data itself, and the controlled environment for which it is intended. Expert Determination is cited as not necessary in most cases.

DE-IDENTIFICATION VS. ANONYMIZATION

Although these two terms are frequently used interchangeably, there is growing recognition that they refer to two related, but different, concepts. De-identification is the process of recoding identifying terms, redacting identifying data, offsetting dates and other similar techniques described under the previous sections of this document. Anonymization is a subsequent step in the process whereby the key that links the original identifying data (specifically, the patient identifier) to the de-identified identifying data is destroyed. The destruction of this key is designed to further ensure that there is no way to map the de-identified data back to the originating data. Again, there are ongoing industry discussions regarding which approach is most aligned with the goals of patient-level data de-identification.

IMPLEMENTATION STRATEGIES

SAS PROGRAMMING

As companies look to de-identify clinical trial data, it is natural to gravitate toward SAS programming scripts as the most expedient way to deliver de-identified data sets. However, while SAS may be the preferred tool among the hands-on user community, skilled SAS resources are, as ever, in very high demand. By developing de-identification tools written in SAS, the task of de-identification gets immediately subjected to the high cost of delivering content driven by scarce and limited resources

. For most organizations, SAS programmers provide the highest value when they are delivering data and statistical analyses that have a direct impact on moving therapies through the research and development process on their way to submission. Clinical trial data de-identification is not typically on this highest-value path.

The development of SAS-based de-identification tools is a costly and time-consuming distraction from the core biopharmaceutical goals – demonstrating safety and efficacy of new therapies. Once developed, these SAS-based tools require ongoing maintenance and support, and frequently become stale as maintenance activities are inadequately resourced over time. For some organizations, these SAS-based tools may already exist, but they are not designed to scale with the growing need to efficiently de-identify data. .

Clinical trial data-related activities carry a high burden when it comes to process control, workflow, documentation and auditing, and the perpetuation of SAS coding scripts for new tasks like de-identification adds directly to this burden. For patient-level data de-identification, where exposing a patient's true identity can be at risk, biopharmaceutical companies cannot (and should not) rely on manual processes to ensure that all data transformation activities are performed correctly, and are documented thoroughly.

AN APPLICATION-BASED APPROACH TO DE-IDENTIFICATION

De-identification, like many other modern business processes, lends itself directly to the use of a point-and-click application that does not require programmer intervention or custom script development. A robust de-identification solution enables 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. This liberates companies from being dependent on scarce SAS programming resources. The basic business process associated with de-identification is depicted in Figure 1. These processes include not only the application of the de-identification algorithms but also the surrounding processes associated with loading, reviewing, approving and ultimately saving the de-identified data sets.

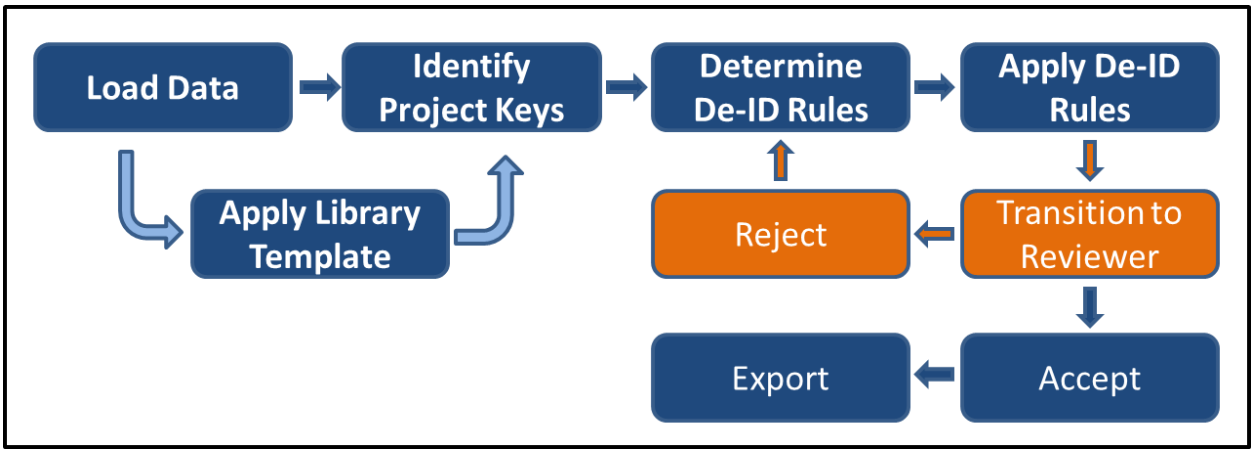

Figure 1. De-Identification Workflow

In order for an application-based approach to be more successful than a scripting-based approach, the application must enable achieving the following key goals:

1. **INCREASED EFFICIENCY AS IT RELATES TO COST, EFFORT AND STAFFING**. In short, the application must present a measurable goal (e.g, reduced cost) that aligns with a strategic initiative (de-identifying data). For many companies, this may be a simple case of understanding the expected savings in terms of time, cost or resources. However, value can also be measured in terms of opportunity cost (for example, what other work could be accomplished if the highly skilled analytical programmers were assigned to other tasks?), as well as increases in de-identification quality and confidence.

2. **FLEXIBILITY**. As clinical trial de-identification matures, an application must provide the means to evolve as well. This means working with new sources of data, new types of data, and new de-identification rules.

3. **AUTOMATION AND INTEGRATED WORKFLOW**. To date, de-identification has been almost exclusively a manual process in terms of preparing data, editing programs, testing results, reviewing the de-identification, signing off and documenting the process. An application provides the ability to connect all of the relevant de-identification processes into a common and integrated workflow that enables automation, transition and documentation of processes within the system.

4. **PROPERLY MANAGING DE-IDENTIFICATION KEYS**. The verdict is still out regarding whether or not data should be fully anonymized and the key then destroyed. An application-based approach can provide the best of all worlds. The decision to destroy they key, and the action to actually destroy it can be actively managed and documented within an application. If the goal is to preserve the key (either temporarily or indefinitely), this can be managed within the application without forcing the user to transfer this very secure information to an external environment. Importantly, with the key preserved within the application, being able to apply that same key to another study at a later time becomes very straightforward.

CONCLUSION

Clinical trial data de-identification has come to the forefront as an emerging and important business process. Old strategies that dealt with de-identification on a one-off basis are being replaced by a collection of industry-developed de-identification strategies. While these strategies have much in common, they still differ in a variety of areas.

As companies move to deploy de-identification solutions, their approach should be flexible in nature, and easily adaptable as the published industry de-identification rules converge. In many cases, the quick approach to implementing these rules – developing, testing and maintaining complex SAS programming scripts – is not the best

approach. SAS scripts require a commitment of not only development, testing and maintenance resources to produce a validated collection of programs, but also ongoing execution support from scarce SAS experts to apply these scripts on an ongoing basis. The execution of these scripts and the accompanying validation and compliance activities will inevitably be manual in nature, and time-consuming to complete.

d-Wise has developed Blur, an application-based de-identification solution, to not only apply the emerging rules for de-identification, but to also integrate the execution of these rules into a comprehensive workflow-driven process that provides automated documentation, traceability and audit trails. As an industry solution, Blur will also stay up-to-date as de-identification rules evolve and converge, and will provide a modern and superior alternative to the development and use of SAS-based de-identification scripts.

CONTACT INFORMATION

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

Author Name:	Dave Handelsman
Company:	d-Wise
Location:	Research Triangle Park, United States of America
Work Phone:	919 825 4775
Email:	dave.handelsman@d-wise.com
Web:	www.d-wise.com/blur