

DE-Identifying Data when Sharing Clinical Data

Pushpa Saranadasa
Merck & Co., Inc.

ABSTRACT

In any particular clinical study a great deal of information is collected that may be of use in other research endeavors. If analysts other than the originating clinical team review the same data from different perspectives then it is more likely that additional assessments and applications will emerge: New investigational approaches to a compound or treatment, perhaps from safety testing to efficacy testing, may come upon new applications accruing to the benefit of the medical community. The caveat to data sharing is that the data transparency necessary within the originating research team may limit the information to be provided to other public and/or private partners interested in using the data in their research.

The sharing of knowledge may involve the patient-level detailed data used in clinical studies. For many reasons, not the least of which is patient privacy, any shared data must first be de-identified. The process of de-identification and creating substitute identifiers involves masking the original identifiers in a way that preserves the utility of the data while at the same time minimizing the risk of re-identification of the patients and any of their unique data intrinsic to the originating study. This paper will explicitly discuss: 1) what is meant by data privacy and de-identification, 2) how de-identification is accomplished and the robustness of the process and 3) the challenge of ensuring study subject anonymity while preserving data base utility.

INTRODUCTION

Data and knowledge sharing involves the trust and privacy of study subjects. When outside researchers are seeking data to pursue additional avenues of investigation it is extremely important to protect the privacy of the participants in the clinical trial while maintaining the utility of the data. Throughout the “data sharing” process the study programmer plays a significant role, communicating with the assigned statisticians, finding related information, and preparing documents to be included in the final delivery. Integral to this is the de-identification of the data in order to comply with the protections articulated in the Health Insurance Portability and Accountability Act. An essential part of the de-identification process is anonymization, which, to the greatest extent possible, prevents unraveling the de-identification to reveal the identities of study subjects. In this essay the process of de-identification is described and the challenges therein are discussed.

1) DATA PRIVACY AND DE-IDENTIFICATION

1.1 What is Data Privacy?

The **Health Insurance Portability and Accountability Act (HIPAA) of 1996** and the Privacy Rule of 2003 are the defining documents for the sharing of protected health information (PHI). The Act in conjunction with the Privacy Rule and the Administrative Simplification documents define the circumstances under which PHI can be shared between health institutions and researchers without prior express written permission from the concerned individual. Essentially the Act, in its various parts, outlines the permissible uses and disclosures that may occur upon the de-identification of the individuals in the shared data. The Act also provides an implicit enumeration of the identifiers that must be removed or masked before data may be shared; the process of removing identifiers and masking study participants so that data may be shared is known as de-identification.

1.2 De-identification and Anonymization

De-identification of data refers to the process of removing or obscuring any personally identifiable information from individual records in a way that *minimizes* the risk of unintended disclosure of the identity of individuals and information about them. There are 18 types of identifiers specified by HIPAA, that should be considered for removal or recoding.

Anonymization of the data refers to the process of data de-identification that produces data in which individual records cannot be linked back to an original owner's data and thereby revealing protected information about study participants. The de-identified data set does not include sufficient information as to identify study participants as it does not include the required translation variables to do so. For example, part of de-identification is creating a subject id random "seed" that is used to populate subject id. Only the sponsor can keep that seed, key, information for future requirements. This is the code that links the de-identified datasets to the original datasets.

2) THE COMPONENTS OF THE PROCESS AND ITS ROBUSTNESS

If the requisite guidelines and protocols are in place at a sponsor company then they serve as a legally binding data sharing agreement between the data sponsor and outside user organizations. The sponsor should ensure that there are processes in place in the company to review research requests, assess how feasible the request is, whether the request has scientific value to the medical community, and how qualified statistical and programming resources are to be allocated by the sponsor's research team. Thereafter the sponsor may choose to approve valid requests.

The following steps will commence following the approval of a signed data sharing agreement.

2.1 Identifying what is requested:

The request normally consists of two components:

Datasets: Raw datasets: SDTM (Study Data Tabulation Model). These are the data collected in a clinical trial. ADaM data sets: Analysis Ready datasets.

Supporting documents: Dataset specifications, Annotated Case Report form, Clinical Study Report (CSR), Statistical Analysis Plan (SAP), and any other additional documents that may be necessary to use the data.

2.2 Internal Team discussion: This is the core part of the process.

Upon receiving the request it is important to discuss what has been requested and who will be responsible for providing the information. A detailed review is needed at this time. The response to the request may contain additional/supporting documentation in addition to what was described in 1) above. For example, how some measurements were collected. The statisticians and programmer involved in the study may discuss the relations between and within the SDTM and ADaM datasets and make any decisions as to how to avoid any potential risks in violating HIPAA. For example, how the 'relrec' domain can be used to find any related information. In some datasets going over the values of the variable may also be needed. For example, the "cforres" variable in the "CF" domain may include dates that reveal the date of a rare disease of subject patients. Another issue might be how to pool age groups so that the cohort of, e.g. "xx or older," will not give away sufficient information to identify any individual patient. Another instance of concern may be the existence of attributes which may be product, study phase and/or disease specific and should be addressed accordingly.

The gathered information should then be discussed with study clinicians for their opinion. For example, if requesters are asking for the dataset that contains adjudicated events then the team should discuss whether any filtering criteria should be implemented for those events that should not be exposed to the public. For example, how the AE should be handled corresponding to those events.

In this internal discussion companies may choose to exclude some information from the datasets. The important conclusion should be how to balance the protection of patients' privacy while helping the scientific community to utilize the shared data.

2.3 Understanding direct and indirect de-identifiers

There are two types of identifiers: direct and indirect. The identifiers listed below can all be categorized as direct identifiers because these can uniquely identify individuals. Indirect identifiers are fields that can identify individuals through inference but may be useful for researchers at the time of data analysis. Examples of these include dates, demographic information (such as race and ethnicity), and socioeconomic variables. This distinction is important because the techniques used to protect the variables will depend on how they are classified.

The direct identifiers (the list is not limited to the following):

1. Names and initials,
2. All elements of dates (except year) which can be directly associated with a specific individual: birthdate, date of death, adverse event date, admission date, discharge date, etc.
3. Kit numbers (diagnostic kits) and device numbers (devices used in the trials),
4. Geographic information such as place of work, trial site location, addresses, zip codes, etc.
5. Telephone numbers,
6. Email addresses,
7. Fax numbers,
8. Account numbers,
9. Social security numbers,
10. Health plan beneficiary numbers,
11. Medical record numbers,
12. Vehicle identifier numbers and serial numbers including license plate numbers,
13. Certificate and/or license numbers (marriage licenses, etc.)
14. Biometric identifiers including such as MRI, hand voice prints, etc.
15. Full face photographic images or comparable images
16. WWW Universal Resource Locators (URLs),
17. Internet Protocol (IP) addresses, and
18. Any other unique identifying number, code or characteristic.

Recoding identifiers:

- Allocation/Randomization identifiers should be assigned new random values.
- Site identifier should be recoded using new randomly generated identifiers.
- Specific site location can be recoded to the corresponding country.
- If the risk of re-identification for a country with a small population size is found to be too great, it can be combined with a country in the same geographical region.
- Investigator identifiers and corresponding contact information should be set to blank.
- Assess any therapeutic area specific fields that may contain personal health information and remove.

Given the type of privacy of the study, the sponsor may need to take action as to how indirect types of identifiers should be handled i.e., re-coded using new randomly generated identifiers, removed, or set to blank. Macro parameters can be set achieve those scenarios.

Handling Dates

When any study is conducted there are many different dates recorded in the SDTM and ADaM datasets. For example, date of birth, dates of admission, dates related to each patient's visits, death dates, adverse event dates, etc. Removal of the patient specific date information is required to protect study participant privacy. In order to preserve chronology of events and the passage of time between events there are two popular methods that are used to handle dates:

- a. Relative Study Day Method is calculated for each observation using admission date or a randomized start date. In order to maintain relationship between visits and events the same relative day algorithm should be applied across SDTM and ADaM datasets. After the relative days are populated all actual date variables are then removed from the datasets or set to blank programmatically.
- b. Off Date Method, that is replacing all original dates related to the subjects with new dates offset by a random number can be easily implemented in the program.

Handling age-related information

Although the date of birth field is removed, the de-identified data could contain exact age for any participants aged, say, 89 years or younger and set to blank for any participants of greater age than 89 years. Program should not report any specific age outliers, instead create a new variable that represents those ages by a range, for instance as "x to xx" or "xx or older."

Handling Free Text Verbatim Terms and Subject Narratives

The program should remove all free text verbatim terms because these fields may contain subject specific information that could reveal the identity of the subjects.

Handling rare events

Rare events should be removed from the data and set to blank for those events. For this particular case the values of the macro variables, which are rare events need to be reviewed thoroughly.

2.4 Robustness in the de-identification process

De-identification to protect privacy can employ various methods ranging from doing the job 'by hand' to automating the process with software. Efficiency suggests that a SAS macro program can be written for application to collections of datasets intended to be shared. Although the intent of CDISC SDTM and ADaM has been to standardize data sets, there may remain idiosyncrasies resulting from study design, follow-up studies and ancillary analyses. As a consequence, any SAS macro will not be robust to all applications. Hence the sharing team will have to be sure that the data to be shared is fully compliant with principles of study participant privacy.

2.5 Validation

It is important that data sponsors perform a validation and review of their anonymization process to ensure that all necessary data have been de-identified appropriately and consistently between the SDTM and ADaM datasets and within the datasets. Given that the destruction of a key code is a uni-directional step (ie, cannot be reversed), this would need to be performed prior to destroying the key code that links the de-identified datasets to the original datasets.

2.6 Documents and Data Delivery

In order to perform data analysis researchers may require additional documentation prepared by the sponsor explaining the de-identification process. The sponsor needs to document the de-identification algorithm and the anonymization steps applied in the process. The sponsor should thoroughly review the document to ensure it does not reveal any privacy information about subjects. That is, anonymization must be preserved. Along with the process document it would be helpful to provide a spread sheet that contains content of the SDTM and ADaM (using "proc contents") datasets and flag the variables that have been de-identified by the sponsor.

2.7 Access to Final package

The other important final step is to ensure that data access is controlled by the sponsor to prevent unauthorized access. Only pre-approved personnel should be given access to download the data, and the time frame for doing so also should be restricted.

3) CHALLENGES

Most companies have standard macros for the process of de-identification, but careful attention should be given to each study as all studies won't fit into one model. It should be noted that a case-by-case assessment may be required by data owners to determine the appropriateness of disclosing study information. The resources that need to be allocated for the de-identification process may vary due to the complexity of the data. Sponsors should review their process on a regular basis in order to be sure the process is aligned with advanced available tools that may be capable of revealing private information.

The relationship between de-identification and anonymization deserves careful consideration when the process is applied to a project that may include a base (main) study, extension studies and follow-up studies. Once the key-code used to anonymize the de-identified data is destroyed it will no longer be possible to link the other studies back to the de-identified main

CONCLUSION

The process of balancing the utility of the shared data and preserving the privacy of the subjects in clinical trials needs careful consideration. The data sharing process takes resources and time. The utility of shared data is important if researchers are to be able to perform some statistical analysis using the data that are delivered, otherwise neither party will gain any benefits from the sharing. If sponsors greatly modify the original data using some one-size-fits-all standardized transformation algorithm it will significantly reduce the utility of data. This would defeat the over-arching principle of data sharing which is to make available information we already have in order that other researchers may explore avenues that may result in greater health benefits for the public.

REFERENCES

Nelson ,G. (2015). Practical Implications of Sharing Data: A Primer on Data .Privacy, Anonymization, and De-Identification.

<https://www.pharmasug.org/proceedings/2016/IB/PharmaSUG-2016-IB06.pdf>

Shostak, J. (2006). De-Identification of Clinical Trials Data Demystified. Paper presented at the PharmaSUG, Denver, CO.

<http://www.lexjansen.com/pharmasug/2006/publichealthresearch/pr02.pdf>

Rights, O. f. C. (2013). Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule.

<https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html>

TransCelerate BioPharma Inc (2013) Model Approach: De-identification and Anonymization of Privacy Information in Data.

<http://www.transceleratebiopharmainc.com/wp-content/uploads/2015/04/CDT-Data-Anonymization-Paper-FINAL.pdf>

ACKNOWLEDGMENTS

I would like to thank the managers in my programming group for reviewing and providing valuable input. Special thanks to Changhong Shi for encouraging me to present this paper, for reviewing the text, and her guidance regarding the contents.

CONTACT INFORMATION

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

Pushpa Saranadasa
Merck & Co., Inc.
351 North Sumneytown Pike
North Wales, PA 19454 -2505
Pushpa_saranadasa@merck.com