JOURNAL OF CLINICAL RESEARCH BEST PRACTICES

Vol. 7, No. 11, November 2011 "Can You Handle the Truth?"

Degrees of De-identification of Clinical Research Data

By Jeanne M. Mattern

Two sets of U.S. government regulations govern the protection of personal data of clinical research study subjects:

- The Common Rule (45 CFR 46)[1]
- HIPAA Privacy Rule and Security Rule (45 CFR 160 and 164 Subparts A and E)[2]

Both the Common Rule and HIPAA allow certain transfers of personal data after they have been processed to protect privacy. Researchers and statisticians use deletion, coding, encryption and aggregation techniques to create "de-identified" or "anonymized" datasets to protect subject privacy to various degrees. These datasets are useful to researchers in answering important public health questions that do not require knowing study subject identities. Researchers must decide how much de-identification is required to satisfy the requirements for different situations.

As shown in Table 1, there are five main levels of de-identified data, two of which can be called anonymized. The information in this article can help researchers decide which method, if any, is best for a given set of data and research objective.[3]

Table 1. Types of De-Identified Data

Level of De-identification	Regula-tion	De-identification Method	Identifiability	HIPAA Authorization / DUA / Privacy Board OK Required
Original Data	NA	None	Direct	Yes / Yes / NA
Limited Data Set (LDS)	HIPAA	Most identifiers removed; public health data may be retained	Direct	Yes / Yes / Usually
Statistical De-Identification (SDD)	HIPAA	Statistician certifies "re-identification is highly unlikely"	Indirect	No / No / Sometimes
De-identified Data Set (DDS)	HIPAA	18 identifiers removed, substitute identifiers ("safe harbor")	Indirect	No / No / Sometimes
Anonymized	Common Rule	Dates and ZIP codes allowed; may be aggregated	None	No / No / No
Anonymized	HIPAA	18 identifiers removed; may be aggregated	None	No / No / No

Determining the Appropriate Level of De-identification

The appropriate level of de-identification is determined by two factors: the needs of the research project and the risk of de-identification. In turn, the risk of de-identification is determined by two factors: the sensitivity of the data (i.e., how private it is) and the chance of re-identification or "reversibility" (i.e., how secure it is). Some research projects are impossible to conduct because they need data at a lower level of de-identification than would be appropriate given the de-identification risk. Table 2 is an example table for assessing de-identification risk.

Table 2. De-identification Risk

		Sensitivity		
Identifiability		**Low**	**Medium**	**High**
	Low	Lowest	Low	Moderate
	Medium	Low	Moderate	High
	High	Moderate	High	Highest

In addition to the de-identification protections built into the data set, the person or organization receiving the data set and their protections (e.g., access, training, physical security, and further transfers) for the data set are also important considerations. Data use agreements (DUAs) describe such protections.

It may be possible to write the informed consent form and collect data in ways that facilitate the future creation of de-identified data sets. Thus, the best time to consider the potential for de-identified data sets is before the study design is complete, the informed consent form is finalized, and any data is collected.

The following are some questions to consider for protecting clinical research data:

- What type of data will be collected?
- How sensitive is the data to be collected?
- What is the source of the data to be collected?
- How will data be collected, captured and stored?
- What personal identifying variables (direct-identifying (e.g., name, address, health insurance number); indirect or quasi-identifying (e.g., diagnosis, visit date)) will the database contain?
- Will a link between personal identifiers and the other data be kept? Will researchers ever need to work in reverse to identify a subject?
- Will data be de-identified before or after data collection?
- How will systems be maintained and secured?
- Will the data be shared? If so, is a DUA in place?
- What results will be published?

One of the main questions to answer is whether de-identification of identifiers should occur before or after data collection. If the decision is to de-identify before data collection, some rules should be followed to ensure that the data is collected anonymously. Identification risk is strongly related to the type of variables collected. A good rule of thumb in determining whether a variable directly or indirectly identifies a subject is to ask whether the variable is important for data analysis. When de-identifying variables is required, different de-identification techniques work best for direct and indirect identifiers. Use randomization and

coding for direct identifying variables; apply analytics to quasi or indirect identifying variables. Once the de-identification risk has been assessed, the appropriate level of de-identification can be determined along the identifiability-anonymization continuum in Table 3:

Table 3. Identifiability-Anonymization Continuum

	Sensitivity	**Reversibility**
Completely Identifiable	Highly sensitive	Original data, so reversibility not needed
Code-linked	Very sensitive	Depends on the codes that mask identity created and how they are controlled
Quasi-identifiable	Somewhat sensitive	Encryption; low reversibility (depending on who holds the encryption key)
Completely Anonymized	Not sensitive	Completely unidentifiable and irreversible

HIPAA De-identification

Current HIPAA Privacy Rule regulation approves only two options for safeguarding data while preserving its usability: (a) de-identified data set (DDS) or (b) limited data set (LDS).[2] Data that has been de-identified involves removal of personally identifying information to protect privacy. Sometimes, the term "de-identified" is used synonymously with the term "anonymization." But sometimes the meanings are quite different, depending on the source. For example, the HIPAA Privacy Rule specifies 18 identifiers that need to be removed for de-identified data to exist, permits the creation of and protecting of a key that could be used to map your original subject identifiers to the new identified identifiers in the future, but prohibits new de-identified identifiers to be derived from the original subject identifiers. On the other hand, the Common Rule defines anonymized data as containing no personal identifiers, with the exception of dates and ZIP codes, permits new de-identified identifiers to be derivatives of the original subject identifiers, and asserts that the re-linking key must not be retained for the data to be regarded as anonymous.[3]

With the DDS method, data stripped of all common identifiers can remain exempt from privacy regulation. A DDS can be established either through the Safe Harbor method (removal of all common identifiers) or through Statistical De-Identification (SDD) (HIPAA 164.514 (b)(1)). With the SDD method, an expert statistician applies and documents statistical principles and methods to determine that the risk of identifying individuals is very small that the data could be used alone or in combination with other information. Therefore, the SDD method can be used to retain some of the Safe Harbor "prohibited identifiers," provided that the risk of re-identification is very small.[3,4]

The second type of not fully identifiable data, the limited data set or LDS, has many common identifier categories stripped away, but it retains some that are necessary for the research study at hand (e.g., treatment date, DOB, ZIP code, etc.).[4] Recipients of an LDS are required to sign a DUA (data use agreement) contract with the data owner prior to use. A DUA should be limited in scope to hold entities accountable for the data as well as disclosure of it. To do otherwise jeopardizes subjects' privacy, the reputations of everyone involved in a study (subject, researcher, institution), as well as the overall value of the research. The CE (covered entity) can use the dataset for research purposes, but it must

take precautions to safeguard the data since the remaining identifiers are still defined as Personal Health Information (PHI) under HIPAA.

De-identification or anonymization, therefore, can refer to the process of completely and permanently removing personal identifiers from data, thus converting it to aggregate data. As a result, it is no longer associated with individuals and can *never* be re-associated with the original data or with any subject. Sophisticated randomization techniques and irreversible coding methods, which replace identifying variables with unique or random pseudonyms, are useful, especially for complex relational databases. If performed correctly, these techniques should yield complete irreversibility.

At times, however, there may be a need to preserve the possibility of re-identifying data. The issue here is how re-identification risk can be adequately controlled, not whether re-linking should be prohibited. In such cases, the originator of the data would hold the key to the re-association (also referred to as "reversible pseudonymization"). Should there ever be a need to reidentify a subject, re-identification could occur through record linkages that would match or re-link records in separate data sets through a common "key" or data field. A related technique is to match records in a released dataset with records from a population registry. Since the dataset and the population registry have common identifiers, one could match the records in both datasets for re-identification. Analytic tools, such as algorithms and dataset subsamples, are other ways to accomplish de-identification reversibility; these methods employ sophisticated suppression and generalization techniques.

Achieving Anonymization

The owner of a database can completely remove all identifying information at the source side and, if correctly administered, that data will no longer be considered PHI. Privacy protections would not apply. However, the de-identification process may or may not be reversible. The data may be irreversibly changed with new identifiers assigned through the use of encryption, but this may not be fool proof. Besides assignment of pseudo identifiers that, when completed correctly, preserves data characteristics required for statistical analyses, there are other available techniques. One method spreads data over several databases. This limits access to all potentially identifiable data at the same time. Another technique groups data into different databases by certain criteria or variables. In this way, data is split among databases. There are also PETs (privacy enhancing software products) and diagnostic tools that may be used to gauge the risk of re-identification of data. These can check to assure that data privacy is not compromised. Stand-alone commercial applications that can automatically remove all identifiers, rendering datasets completely de-identified, are currently rare. One example of such an attempt is a product developed by De-ID Data Corporation, which claims its application, developed at the University of Pittsburgh, works with data management systems to strip away *all* categories of PHI identifiers from all types of data, structured and unstructured, contained in an uploaded dataset, rendering it de-identified and HIPAA compliant while safeguarding data access and preserving data integrity.[7] How claims of such products fare in light of a sensitivity analysis, as well as reversibility issues on the identifiable-anonymization continuum just discussed, remains to be seen.

The Continued Importance of De-Identified Data

The HIPAA Privacy Rule does not clearly define criteria for, or use of, de-identified or anonymized data. It is in the process of being revised, so it may become more closely aligned with the Common Rule.[1] On July 12, 2011, The Office of Management and Budget announced publication of a pre-rule by which "human subjects research apply the Common

Rule to all such research, regardless of funding source, revise categories of research that are exempt from review, and apply HIPAA protections to research data."[8]

Since study data do not always require protection by the HIPAA Privacy Rule, more directives are, indeed, necessary. Technology is chipping away at our expectations of privacy and protections. A balance of value and risk is not futile to think about. Researchers work to prevent privacy invasions and place limits on how and what data is shared, but a balance of value and risk needs to be achieved.[9] Besides working to stay atop current regulations, much can be done to re-examine the restrictions contained in the rules, especially redefining them as they apply to researchers entrusted with data.

The Center for Democracy and Technology (2009) advocates anonymization options that would encourage the use of "less than fully identifiable" data and "assure that data are accessible and disclosed in the least identifiable form possible for any given purpose."[6]

There have been recommendations for revising the current regulation with regard to de-identification of data that could encourage transparency of use, thereby making data more available, while actually better safeguarding privacy and instituting stronger risk protections. However, current best practice is to invoke increased security even for databases containing anonymous information, and to test anonymous systems to ensure that anonymous information cannot become re-identified. Additional security is becoming ever more important in light of advances in technology that enable re-identification of anonymous information.[9] Data that are unstructured (e.g., text, clinical notes, and comments) can also contain sensitive and directly identifiable information. Such data is much more difficult to auto-de-identify, although tools are available for de-identifying specific types of text (e.g., diagnosis, procedure or discharge notes).

Therefore, in dealing with sensitive data, the entire context of the data needs to be considered. De-identified data can provide strong protections of privacy, but further regulations are needed that provide specific ways to limit risk and safeguard privacy, including ways to help ensure that re-identification attempts will be detectable and unsuccessful. While poor data de-identification or anonymization can lead to bad decisions, bad outcomes, and bad science, properly de-identified data are invaluable to society. With the HITECH Act and current trends in healthcare reform, de-identified PHI data will become ever more important. This law provides requirements to avoid breach of privacy and lost research opportunities.[4] De-identified data supports discovery, innovation and healthcare delivery, as well as improving its effectiveness, efficiency and quality.

References

1. The Common Rule http://ori.dhhs.gov/education/products/ucla/chapter2/page04b.htm
2. HIPAA Privacy Rule 45 CFR 164.514
3. Shostak, J. (2006). De-identification of clinical trials data demystified. http://www.lexjansen.com/pharmasug/2006/publichealthresearch/pr02.pdf
4. 18 safe harbor exclusion elements 45 CFR 164.514(b)(2)(i)
5. HITECH Act 2009 http://www.hipaasurvivalguide.com/hitech-act-text.php
6. Center for Democracy & Technology, Washington, D.C. (June 2009). Encouraging the use of and rethinking protections for de-identified (and "anonymized") health data, page 2. http://www.cdt.org
7. De-IData: Health Data Safety Software http://www.de-idata.com/about
8. US GSA. "Human Subjects Research Protections: Enhancing Protections for Research Subjects and Reducing Burden, Delay, and Ambiguity for Investigators." http://www.reginfo.gov/public/do/eoDetails;jsessionid=9f8e89cb30d60d7f090df08c43f1

94b684939bb4da32.e34ObxiKbN0Sci0SaxaPch4Kchn0n6jAmljGr5XDqQLvpAe?rrid=120486

9. Porter, C. C. (2008). De-identified data and third party data mining: The risk of re-identification of personal information. *Washington Journal of Law, Technology & Arts, 5* (3) http://digital.law.washington.edu/dspace-law/bitstream/handle/1773.1/417/vol5_no1_art3.pdf?sequence=1

Author

Jeanne Mattern, Ph.D., LSW. CCRP is Project Staff/Regulatory Compliance Officer at Quantitative Health Sciences Department, Lerner Research Institute, Cleveland Clinic. Contact her at 1.216.445.5775 or matter@ccf.org.

