De-identification and the Sharing of Big Data

Save to myBoK

By Susan E. White, PhD, CHDA

One of the newest buzz words in data analytics is "Big Data," and the data created through the healthcare industry is some of the "biggest" around. The widespread implementation of electronic health records (EHRs) and the need to share data to measure quality and manage accountable care organizations (ACOs) brings to light all of the privacy issues surrounding sharing patient data. In order to fully leverage Big Data, that data must be shared and combined in a way that preserves its utility for research and performance analytics.

The HIPAA privacy rule includes standards for the release and use of protected health information (PHI). The rule allows for the sharing of data if it is de-identified so that the individual patient's identity remains protected. Specifically, the statute includes the following standard for the de-identification of data:[1]

> (a) Standard: de-identification of protected health information. Health information that does not identify an individual and with respect to which there is no reasonable basis to believe that the information can be used to identify an individual is not individually identifiable health information.

The rule goes on to specify two methods for de-identification: Safe Harbor (CFR164.514(b)(2)) and Expert Determination (CFR164.514(b)(1)). Once data is de-identified, it is no longer covered under the HIPAA rule since it no longer fits the definition of PHI. The HIPAA rule also includes implementation specifications for re-identification of the data.

The difficulty in interpreting this portion of the HIPAA rule caused many covered entities to err on the side of caution and prohibit the release of patient data for any non-reimbursement purpose, or to take the "Safe Harbor" approach. This approach requires the removal of the 18 data elements that are considered identifiers. These data elements are listed in Table 1. The "Safe Harbor" method ensures compliance with HIPAA but severely limits the utility of patient data for comparative effectiveness projects, health disparity studies, and a number of other valuable research applications.

Most providers did not attempt to apply the "Expert Determination" method because it requires the application of statistical principles to ensure that the risk of identifying an individual is very small.

Sixteen years after the HIPAA bill was signed in 1996, the Office for Civil Rights (OCR) released guidance to help practitioners determine what data elements could be released and under what conditions. This guidance is intended to assist covered entities to understand what de-identification is, the general process by which de-identified information is created, and the options available for performing de-identification.

In developing this guidance, OCR solicited input from stakeholders with practical, technical, and policy experience in de-identification at a workshop consisting of multiple panel sessions held March 8-9, 2010, in Washington, DC. Each panel addressed a specific topic related to the HIPAA Privacy Rule's de-identification methodologies and policies.

The guidance was posted to the Department of Health and Human Services (HHS) website at http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/guidance.html.

Table 1: Safe Harbor Data Elements

The Safe Harbor method requires the removal of "any other unique identifying number, characteristic or code" from personal health information. The HHS guidance on this data element suggests that clinical trial record numbers, bar codes for medications, or even occupation might be considered an identifier. Prior to the release of patient data under Safe Harbor, the entity must also consider if data elements may be combined with other data sources to determine a patient's identity.

Restricted Data Elements:

<table>
<tr><td colspan="2">(A) Names</td></tr>
<tr><td colspan="2">(B) All geographic subdivisions smaller than a state, including street address, city, county, precinct, ZIP code, and their equivalent geocodes, except for the initial three digits of the ZIP code if, according to the current publicly available data from the Bureau of the Census:

(1) The geographic unit formed by combining all ZIP codes with the same three initial digits contains more than 20,000 people; and
(2) The initial three digits of a ZIP code for all such geographic units containing 20,000 or fewer people is changed to 000</td></tr>
<tr><td colspan="2">(C) All elements of dates (except year) for dates that are directly related to an individual, including birth date, admission date, discharge date, death date, and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older</td></tr>
<tr><td>(D) Telephone numbers</td><td>(L) Vehicle identifiers and serial numbers, including license plate numbers</td></tr>
<tr><td>(E) Fax numbers</td><td>(M) Device identifiers and serial numbers</td></tr>
<tr><td>(F) E-mail addresses</td><td>(N) Web Universal Resource Locators (URLs)</td></tr>
<tr><td>(G) Social security numbers</td><td>(O) Internet Protocol (IP) addresses</td></tr>
<tr><td>(H) Medical record numbers</td><td>(P) Biometric identifiers, including finger and voice prints</td></tr>
<tr><td>(J) Account numbers</td><td>(Q) Full-face photographs and any comparable images</td></tr>
<tr><td>(K) Certificate/license numbers</td><td>(R) Any other unique identifying number, characteristic, or code, except as permitted by paragraph (c) of the guidance document</td></tr>
</table>

Source: Department of Health and Human Services. “Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule.” http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/guidance.html#guidancedetermination.

Applying the Safe Harbor Method

The data elements that create the most confusion in applying the Safe Harbor method are dates and geographic location. ZIP codes are often used in health disparity studies to link to socioeconomic data from the Census Bureau or other sources. Dates

are used to determine the timing within a treatment episode and readmission rates.

The HHS guidance gives more details on exactly when ZIP codes may be released, as well as more details regarding the release of dates. The release of dates must be limited to only the year. The data must be reviewed to determine if any date of service might allow the end user of the data to infer that a patient is age 90 or older in combination with year of birth. This restriction on the release of dates severely limits the utility of patient-level data for research and quality measurement.

The Safe Harbor method requires the removal of "any other unique identifying number, characteristic or code," according to the guidance released by HHS.[2] The HHS guidance on this data element suggests that clinical trial record numbers, bar codes for medications, or even occupation might be considered an identifier.

Prior to a release of patient data under Safe Harbor, the entity must also consider if data elements could be combined with other data sources to determine a patient's identity. For instance, if birth year is released for a newborn and it could be combined with birth notices in a local paper, then the birth year becomes what is called a pseudo-identifier.

A pseudo-identifier is a data element that taken alone does not identify an individual but may be combined with publicly available records to infer the identification of a patient. Pseudo-identifiers are an issue for patients with rare diseases or who might be receiving a relatively rare treatment publicized in local or national media.

Table 2: Principles for Determining the Indentifiability of Health Information

Principle	Description	Examples
Replicability	Prioritize health information features into levels of risk according to the chance it will consistently occur in relation to the individual.	*Low:* Results of a patient's blood glucose level test will vary. *High:* Demographics of a patient (i.e., birth date) are relatively stable.
Data source, Availability	Determine which external data sources contain the patients' identifiers and the replicable features in the health information, as well as who is permitted access to the data source.	*Low:* The results of laboratory reports are not often disclosed with identity beyond healthcare environments. *High:* Patient name and demographics are often in public data sources, such as vital records-birth, death, and marriage registries.
Distinguishability	Determine the extent to which the subject's data can be distinguished in the health information.	*Low:* It has been estimated that the combination of *Year of Birth, Gender, and 3-Digit ZIP Code* is unique for approximately 0.04% of residents in the United States. This means that very few residents could be identified through this combination of data alone. *High:* It has been estimated that the combination of a patient's *Date of Birth, Gender, and 5-Digit ZIP Code* is unique for over 50 percent of residents in the United States. This means that over half of U.S.

		residents could be uniquely described just with these three data elements.
Risk assessment	The greater the replicability, availability, and distinguishability of the health information, the greater the risk for identification.	*Low:* Laboratory values may be very distinguishing, but they are rarely independently replicable and are rarely disclosed in multiple data sources to which many people have access. *High:* Demographics are highly distinguishing, highly replicable, and are available in public data sources.

Source: Department of Health and Human Services. "Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule." http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/guidance.html#guidancedetermination.

Applying the Expert Determination Method

The HHS authors define an "expert" in their guidance. They point to experts that may be found in statistics, mathematics, or other scientific areas.

They do not specify the value of "very small" in terms of the risk of identification. HHS suggests that the risk should be assessed based on the particular circumstances of the release and the particular data set released. The suggested process for determining the risk is:[3]

1. Expert works with covered entity to determine appropriate statistical or scientific methods to mitigate risk of identification
2. Expert applies method to mitigate risk
3. Expert assesses risk
4. Expert documents methods and results to justify determination

The principles used by experts to determine the likelihood that a patient may be identified include replicability, data source availability, distinguishability, and assessment of risk. These principles are further defined in Table 2.

HIM professionals are critical in this process. As the primary stewards of patient data, they understand which data elements are most likely to be either identifiers or pseudo-identifiers. Many of the experts who will participate in this process will have limited experience in the operational side of collecting and validating the data included in the patient-level data to be released. The principles listed in Table 2 require the expertise and context of HIM professionals to be applied in a consistent and effective manner. The goal of de-identification procedures should be to release a data set as detailed and robust as possible while still protecting the identity of the patient.

Two Methods to Achieve De-identification in Accordance with the HIPAA Privacy Rule

HIPAA Privacy Rule De-identification Methods	
Expert Determination 164.514(b)(1)	Safe Harbor 164.514(b)(2)

Apply statistical or scientific principles	Removal of 18 types of identifiers
Very small risk that anticipated recipient could identify individual.	No actual knowledge residual information can identify individual.

Source: Department of Health and Human Services. "Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule." http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/guidance.html#guidancedetermination.

Note

1. Department of Health and Human Services. "Other Requirements Relating to Uses and Disclosures of Protected Health Information." *Federal Register*. 45 CFR part 164.514. http://www.gpo.gov/fdsys/pkg/CFR-2007-title45-vol1/pdf/CFR-2007-title45-vol1-sec164-514.pdf.
2. Department of Health and Human Services. "Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule." http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/guidance.html#guidancedetermination.
3. Ibid.

Susan E. White (white.2@osu.edu) is associate professor, clinical HRS, at Ohio State University.

Article citation:
White, Susan E. "De-identification and the Sharing of Big Data" *Journal of AHIMA* 84, no.4 (April 2013): 44-47.