De-Identification of Health Data under HIPAA:

Regulations and Recent Guidance

Deven McGraw
Director, Health Privacy Project
January 15, 201311

HIPAA Scope

- Does not cover all health data
- Applies to "covered entities"
  - Most health care providers, health insurers/health plans, and health care clearinghouses.
- Applies to business associates of covered entities (contractors who receive identifiable health information to perform certain services on behalf of a covered entity).
- So does not cover all health data – but does cover predominant sources of identifiable health information

Federal (HIPAA) Policy on "De-identification"

- Found in regulation – the "Privacy Rule", plus HHS Guidance (required by ARRA/HITECH)
- "De-identified data" = data that meets the Privacy Rule standard for de-identification
- Data that meets the HIPAA de-identification standard is <u>not</u> PHI and not regulated by HIPAA
- De-identification standard = no reasonable basis to believe the data can be used to identify an individual (45 CFR 164.514(a))
  - This is a legal standard – there is no specific % risk to target; the HHS Guidance addresses why there is no explicit numeral level (because whether the risk is "very small" is context dependent – pgs 10-11)

De-identified Data under HIPAA

- Two methods may be used to de-identify:
  - Statistician or expert method requires someone with statistical expertise to determine (and document) that the risk is "very small" that the information, on its own or in combination with other reasonably available information, could be used by an anticipated recipient to identify an individual (164.514(b)(1)).
  - Safe harbor requires the removal of 18 categories of data; in addition, data holder must not have actual knowledge that the data, either alone or in combination with other data, could identify an individual.(164.514(b)(2))

Assignment of Codes

- Covered entities may assign a code (or other means of record identification) to allow de-identified data to be re-identified by the covered entity, as long as
  - The code is not derived from, or related to, information about the individual and is not otherwise capable of being translated in a way that facilitates identification of the individual, and
  - The covered entity does not use or disclose the code or other means of identification for any other purpose, and does not disclose the mechanism for any other purpose. (164.514(c))

Safe Harbor Data Categories

- Names;
- 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 a 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
  - The initial three digits of a zip code for all such geographic units containing 20,000 or fewer people is changed to 000.

Safe Harbor Data Categories (2)

- All elements of dates (except year) for dates directly related to an individual, including birth date, admission date, discharge date, date of death; and all ages over 89 and all elements of date(including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older;
- Telephone and fax numbers;
- Electronic mail addresses;
- Social Security numbers;
- Medical record and health plan beneficiary numbers;

Safe Harbor Data Categories (3)

- Account numbers and certificate/license numbers;
- Vehicle identifiers and serial numbers, including license plate numbers;
- Device identifiers and serial numbers;
- Web Universal Resource Locators (URLs);
- Internet Protocol (IP) address numbers;
- Biometric identifiers, including finger and voice prints, and full face photographic images and any comparable images; and

Safe Harbor Data Categories (4)

- Any other unique identifying number, characteristic, or code, except as permitted above (see slide 5 - Covered entities may assign a re-identification code but can't disclose it to the de-identified data recipient).

Note: All of the above data categories must be removed in order for a data set to qualify as de-identified under the safe harbor method. If any of the above categories of data is needed to preserve utility, entities can use statistical methodologies to achieve the "very low risk" standard.

Legal Consequences

- Data that meets the "very low risk" standard is not regulated by HIPAA.
- Can use either method to meet the standard
- With use of safe harbor, data is "deemed" to meet standard.
- Under statistician/expert method, expert determines that data meets the standard.
- Inappropriate release of data that is not "de-identified" per legal standard would trigger potential liability for a covered entity or potentially a business associate (CDT is not aware of any enforcement actions to date)

What does the HHS Guidance Add?

- Reiterates the legal standard and the two methodologies for de-identification.
- Acknowledges that de-identified data retains a very small risk of re-identification – not required to get to zero risk.
- Provides guidance on use of both methodologies, and on use of a "code" that can facilitate re-identification.
- Clarifies circumstances under which a business associate can de-identify data: when authorized by the business associate agreement.
- Clarifies that a data use agreement restricting use or prohibiting re-identification is not required but can be used.

Guidance on Expert/Statistician Method (1)

- Covers "who is an expert" – no specific degree or training required, but with respect to enforcement, OCR would review the relevant professional experience and training of the expert
- Standard is "very low risk" - no specific numeric target
- Can derive multiple solutions from the same data set, as along as the expert has taken care that the data sets cannot be combined in ways that would increase the re-identification risk beyond the acceptable level.
- Provides illustrative general workflow for how an expert goes about de-identifying a data set.

Guidance on Expert/Statistician Method (2)

- Lists principles used by experts in determining identifiability and typical approaches used to de-identify (suppression, generalization/abbreviation, perturbation; k-anonymity offered as specific example).
- Can use data use agreement to limit distribution of de-identified data set, but the agreement is not a substitute for application of statistical methods (i.e., it's not a de-identification tool).
- Covered entities may assign re-identification codes – but cannot disclose these with the de-identified data. (In reg) Guidance clarifies that a covered entity "may disclose codes derived from PHI as part of a de-identified data if an expert determines that the data meets the de-identification requirements.

Guidance on Expert/Statistician Method (3)

- Question 2.9 – Use of codes
  - Bottom line: the data cannot be re-identifiable by the de-identified data recipient.
  - Thus, a covered entity can disclose codes that are derived from PHI as long as the expert determines that the data (including the code) meets the "very low risk" standard.
  - PHI can also be transformed into values derived by cryptographic hash functions using the expert method, as long as the keys are not disclosed, "including to the recipients of the de-identified information."

Guidance on Safe Harbor (1)

- Clarity on when zip codes can be used; must use most current publicly available data from the Census Bureau.
- Parts or derivatives of the list of identifiers cannot be disclosed (e.g., last 4 numbers of the SSN).
- Clarity on the prohibition on dates.
- Clarity in what constitutes "any other unique identifying information" – e.g., other identifying numbers like clinical trial record numbers; non-secure codes (like a hash function without a secret key), barcodes, unique identifying characteristics ("current President of the United States").

Guidance on Safe Harbor (2)

- Clarifying the standard for "actual knowledge" – clear and direct knowledge that the remaining information could be used, either alone or in combination with other information, to identify an individual who is the subject of the information.
  - Doesn't include knowledge of studies of methods to identify de-identified information. "OCR does not expect a covered entity to presume such capacities of all potential recipients of de-identified data." (p.28)
- Categories of data must be removed from free text as well as standardized data fields (no distinction in regulation).
- De-identification is aimed only at protecting the patient subjects of the data, not the providers or their staff

CDT on "Deidentification"

- White Paper (June 2009): "Encouraging the Use of, and Rethinking Protections for, De-Identified (and "Anonymized") Health Data": http://www.cdt.org/files/pdfs/20090625_deidentify.pdf
- Policy Post (shorter version of above) (6/26/09): http://www.cdt.org/policy/stronger-protections-and-encouraging-use-de-identified-and-anonymized-health-data
- iHealthBeat Perspectives (even shorter) (7/30/09): http://www.ihealthbeat.org/perspectives/2009/anonymized-medical-data-protects-privacy-improves-care.aspx
- Building Public Trust in Uses of HIPAA De-Identified Data,http://jamia.bmj.com/content/early/recent