De-identification Koans

ICTR Data Managers

Darren Lacey

January 15, 2013

Disclaimer

- There are several efforts addressing this issue in whole or part
- Over the next year or so, I believe that the conversation will result in operational policies and procedures
- Nothing I say here today should be taken as the official position of anybody official
- This may not be a settled issue for several years

Where to start on this

- HHS has issued guidance – 11/26/12 --*Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule*
- NIST SP 800 -122 – *Guide to Protecting the Confidentiality of PII*
- www.insidehopkinsmedicine.org/hipaa
- http://www.hopkinsmedicine.org/institutional_review_board/

First the easy stuff (18 de-identifiers)

1. Names
2. Geography
3. Dates
4. Phone numbers
5. Fax numbers
6. E-mail addresses
7. SSN
8. Medical record #
9. Health plan #
10. Account #
11. Cert/license #
12. Vehicle identifier
13. Device identifier
14. URL
15. IP Address
16. Biometric (voice, finger)
17. Image/photos

And anything else…

18. Any other unique identifying number, characteristic, or code (this will be important in a second)

Some wrinkles -- Geography

2. All geographical 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 (2) The initial three digits of a zip code for all such geographic units containing 20,000 or fewer people is changed to 000.

Another wrinkle -- Dates

3. 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 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.

A few other things – and I am out of my depth here

- HIPAA/PHI must involve data associated or derived from a healthcare event entered into the medical record and where subject is not informed of results
- Much of this information (e.g. genetics) would still fall under Common Rule
- Non-PHI health information becomes PHI when associated with one of the identifiers

HIPAA Privacy Rule – 164.514

- De-identification Standard -- 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.

Method 1 – Expert Determination

- Statistical expert
- Apply statistical principles
- Risk measure – “risk is very small that the information could be used, alone or in combination with other reasonably available information, by an anticipated recipient to identify an individual who is a subject of the information.”
- Document process and determination

Method 2 – Safe Harbor

- Pull out the 18 identifiers
- Can assign a re-identification code that will allow for re-identification later
- Applies to structured and unstructured data
- This is more closely involved with “minimum necessary”
- And we are primarily going to discuss the safe harbor (or heuristic method)

Limited Data Set

- Removes 16 identifiers (mostly direct identifiers)
- May include:
  - Geography – (town/city, five digit zip code)
  - Dates – (admission/discharge, etc.)
- Requires data use agreement for disclosure

Reidentification “Attacks” -- AOL

- AOL releases search query terms with pseudonyms
- Search terms include: ‘tea for good health’, ‘numb fingers’, ‘hand tremors’, ‘60 single men’, ‘dog that urinates on everything’, ‘landscapers in Lilburn, Georgia’, ‘homes sold in shadow lake subdivision gwinnett county georgia’
- NY Times found her

Attribute vs Identity Disclosure

- Attribute disclosure -- find out something new about individual in database without certainty who
  - Probabilistic (can be statistically significant)
  - Thus can have stigmatization issue (99% of individuals have this characteristic)
- Identity disclosure – determine which record in the database belongs to a particular individual
- HIPAA covers identity disclosure only

De-identification Process

- Remove identifiers from individual records
- Check for records with unique data that might identify
- Provide a unique code to each data record that will match to an individual (hard)
- Data scrubbing text

One way hashing

- Jim Subject --“Glucose 75” “12/25/12”
- 767384672393782 --“Glucose 75” “12/25/12”
- You can use the hash each time to preserve patient context across records
- Or hash the whole record or parts thereof
- Vulnerable to “dictionary attack”
- HIPAA Privacy Rule prohibits codes that “derive” from the PHI itself

Uniqueness

- Ages over 89 make it statistically likely that the number of individuals in this category is small enough to identify
- Determining safe levels of uniqueness is part of that expertise component in the first method
- I don’t trust myself to figure these out in many cases

Implications – things that don't really work

- Character scrambling – code, uniqueness
- Truncation or masking -- uniqueness
- Date shifting – code
- Sloppy pseudonyms – code

What does work (most of the time)-- Randomization

- It is not random if it derives from the master data set
- Randomized data elements can be difficult to track
- Requires master re-identification file
- Sequencing longitudinal data

Zero Knowledge Protocol

- Resolves a question without identifiable context, just the question being answered
- New “identifier” through the sum of a random number and an identifier, and you will need a matching list
- So the new code becomes the unique random code

Data Scrubbing

- Black list – remove unacceptable words or phrases
  - Most common
  - Slow
  - Does not fully de-identify
- White list – only keep acceptable stuff and you need
  - Might fully de-identify
  - Often produces poor quality results

Statistical De-identification Method

- Uses certification method
- Requires uniqueness analysis and several others
- Privacy Analytics – PARAT tool provides quantifiable analysis of this
  - Can look at types of data and distribution of re-identification probabilities
  - Can look at potential use value

Manage Re-identication Risk

- Amount of De-identification
- Invasion of Privacy
- Motives and Capacity
- Mitigating Controls

Mitigating Controls – Privacy and Security Practices

- Requires a privacy and security plan
  - Need not be formal or long
  - It should be communicated and signed off on by the team including any third parties
  - It should include the same types of components as the statistical model – including privacy risk, data disposal, etc.
  - It may not be enough to cut and paste from IRB documentation
- Security plan may be subject to regulation or standard (e.g. NIST SP 800-18, 21 CFR Part 11)

PHI/Including Re-identification File

- On access-logged server or host in secure location
- Highest level of security and day-to-day access should be minimized
- Full disc encryption is appropriate
- But you should also have file encryption (with different authentication)

General Security Controls

- Device Security – including mobile devices
- Data Access Management
- Physical Security
- Server Protection
- Transmission Security
- Training and Awareness
- Annual Review of Plan
- Data Disposal

Common Issues

- Protection of Re-identification file
- Role of research clusters (here and elsewhere)
- Third party data (including government data)
- FIPS 199 data management plans
- File transfer mechanisms
- Rapidly changing access control lists
- Expansion of scope beyond first data set

Last Slide

Darren Lacey

dll@jhu.edu