Department of Health & Human Services
Office of the National Coordinator for
Health Information Technology

The Safe Harbor Method of De-Identification

An Empirical Test

October 08, 2009

Deborah Lafky, MSIS, Ph.D., CISSP
Program Officer, Security & Cybersecurity

The Project Staff and Funding

NORC at the UNIVERSITY OF CHICAGO

- NORC staff working on the Project
  - Avi Singh, Principal Investigator
  - Michael Davern, Project Director
  - Elizabeth Hair, Project Manager
  - Peter Kwok, Lead Statistician
  - Joshua Borton, Statistician
  - Amanda Yu, Research Scientist
  - Craig Holden, Research Analyst

- ARRA funding provided by the Office of the National Coordinator for Health Information Technology.

HHS.GOV/Recovery

Department of Health & Human Services
Office of the National Coordinator for
Health Information Technology

Agenda

- What is the “Safe Harbor” method of de-identification?
- Why are we testing it now?
- What are we testing?
- How did we do the tests?
- What did we find?
- What does it all mean?

Department of Health & Human Services
Office of the National Coordinator for
Health Information Technology

What is the “Safe Harbor” method of de-identification?

- Alternative to “expert determination” method
- HIPAA Privacy Rule §164.514(b)(2)(i)
- 18 direct and indirect identifiers must be removed and there must be no actual knowledge that information can be identified*

1. Names
2. Geographic subdivisions smaller than state
3. All elements of dates except year
4. Telephone numbers
5. Fax numbers
6. E-mail addresses
7. Social Security numbers
8. Medical record numbers
9. Health plan beneficiary numbers
10. Account numbers
11. Certificate/license numbers
12. Vehicle identifiers/serial numbers
13. Device identifiers/serial numbers
14. URLs
15. IP addresses
16. Biometric identifiers
17. Full face photographic images
18. Any other unique identifying number, characteristic, or code

* This list does not present the full detail of each of these. Refer to the regulation text for additional specifications and requirements.

Department of Health & Human Services
Office of the National Coordinator for Health Information Technology

Why is HHS testing the Safe Harbor Method?

- De-identified data sets are not protected health information under HIPAA Privacy Rule.
- Recent authors have questioned whether the Safe Harbor method is still strong enough to prevent re-identification; availability of 3rd party data has increased since the method was developed.
- ONC is providing technical input to OCR with respect to de-identification policy.
- Results will inform departmental policy.
  - HITECH requires guidance on de-identification.

Department of Health & Human Services
Office of the National Coordinator for Health Information Technology

What is HHS testing?

Can a Safe Harbor de-identified data set be combined with readily available outside data to re-identify data set subjects?

- *Some researchers and others have stated that increased personal data availability, e.g. on the Internet, makes re-identification easy, but there has been little empirical evidence to support that claim.*

Department of Health & Human Services
Office of the National Coordinator for Health Information Technology

Why are people concerned about re-identification?

- Loss of privacy
- Material impacts
  - Health/life insurance
  - Employment
- Is secondary use safe?
  - Does public acceptance of secondary use depend on the context of that use?
    - Public good vs. other types of use

What is HHS testing?

- Two basic scenarios:
  1. Safe Harbor method de-identified data are obtained by someone with no knowledge except that which is available to the general public (low knowledge scenario).
     - e.g. a thief who steals a laptop just because the opportunity presents itself
  2. Safe Harbor method de-identified data are obtained by someone who has some knowledge about information it may contain (high knowledge scenario).
     - e.g. a research assistant seeking information on a celebrity known to be in the data set

Department of Health & Human Services
Office of the National Coordinator for Health Information Technology

What is HHS testing?

- Two basic contexts:
  1. Re-identify all (or as many as possible) individuals in the data set (high yield scenario).
     – e.g. To obtain material for identity theft

  2. Re-identify particular individual(s) suspected to be in the data set (targeted yield scenario).
     – e.g. To obtain damaging information on a public figure.

4 Classes of Risk

With higher knowledge, try to identify a targeted individual	With higher knowledge, try to identify as many people as possible.
With little knowledge, try to identify a targeted individual	With little knowledge, try to identify as many people as possible.

HIGH

LOW

Knowledge

Yield

Department of Health & Human Services
Office of the National Coordinator for Health Information Technology

4 Classes of Risk

Effort to re-identify is allocated to the desired payoff.

HIGH	With higher knowledge, try to identify a targeted individual	With higher knowledge, try to identify as many people as possible.
LOW	With little knowledge, try to identify a targeted individual	With little knowledge, try to identify as many people as possible.

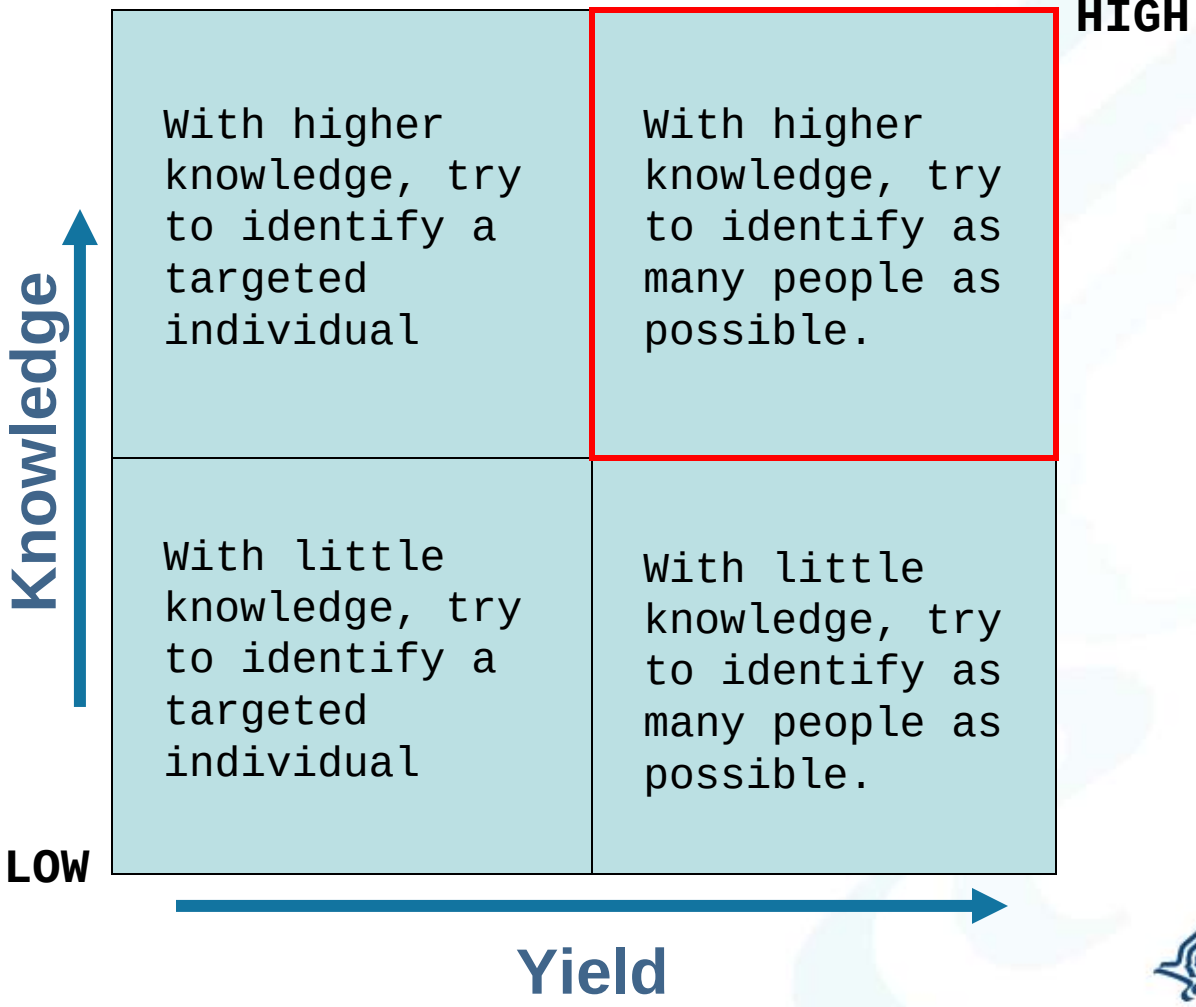
Knowledge

Yield

Department of Health & Human Services
Office of the National Coordinator for Health Information Technology

4 Classes of Risk

How much effort is required to produce a high yield from Safe Harbor de-identified data?



Department of Health & Human Services
Office of the National Coordinator for Health Information Technology

Two different challenges

1. Given a Safe Harbor Method de-identified data set, how many of the records can be accurately linked back to specific patients?

2. Is Person X in this de-identified data set?

Two different challenges

1. Given a de-identified data set, how many of the records can be accurately linked back to specific patients?

2. Is Person X in this de-identified data set?

Department of Health & Human Services
Office of the National Coordinator for Health Information Technology

Research Question

How likely is it that any particular record in a HIPAA Safe Harbor de-identified data set can be correctly re-linked to a person?

Is it easy or hard?

Department of Health & Human Services
Office of the National Coordinator for Health Information Technology

How is this testing being done?

- A set of ~15,000 Safe Harbor method de-identified patient records were pulled from a large academic health center serving a multi-county region of about 1.6 million.
  - To increase the likelihood of an “easy” match, all subjects were drawn from a pool who self-identified as part of a large minority ethnic group
  - The NORC research team did not have access to the real identities of the subjects
- A matched list of individuals in the same geographic area and of the same ethnic group was obtained from a commercial data repository (considered reliable enough by the US Census to be used to verify and cross-check its household data).

How is this testing being done?

- NORC researchers tried to match de-identified records with identifiable records in the purchased database.
  - 2-step process
    1) To get an accurate linkage, there must be uniquely correlating information
       - People who have many traits in common are very difficult to correlate with any certainty.
       - People who have unique or near-unique “profiles” are easier to match.
    - Therefore, Step 1 is to search for unique profiles
    - Out of ~15,000 de-identified records, this data set produced 216 “uniques”.

How is this testing being done?

- Step 2
  1) Manually search through the external source data (e.g. InfoUSA) to see if any of the records align with any of the “uniques” in the de-identified data set.
  2) Send the possible matches back to the health center data team for verification that a true match was made.

All done with IRB approval.

What are the findings?

- 216 unique profiles found in the de-identified data (1.5%)
  - As data sets grow larger, unique profiles are fewer.
  - Only 84 unique profiles out of 32,549 (0.25%) InfoUSA records in the same geographic area and same ethnic group
- 28 potential pairs were found after combing through the data manually
  - There are no matching algorithms the team knows of that are more accurate than using human judgment because

  (a) **contextual knowledge** is essential and

  (b) data sources are “dirty”
- Only 2 were verified to be correct matches…

  …for a match rate of less than 0.01%

Department of Health & Human Services
Office of the National Coordinator for Health Information Technology

What does this all mean?

- Matching up Safe Harbor de-identified records to publicly available data is:
  - Labor-intensive
  - Costly
  - Has a low yield

  These facts are a deterrent to identity thieves

  Some provisos apply:
  - The larger the data set, the safer it is (safety in numbers)
  - The more extra knowledge an intruder has, the better they will be able to match the data

Department of Health & Human Services
Office of the National Coordinator for Health Information Technology

Notes

- Data sets should be handled such that if they were to fall into the wrong hands, correlating information that would assist in re-identification is not present
  - e.g. do not ship a de-identified data set together with a copy of a corresponding third-party data source
- Smaller data sets should be treated carefully if they contain a higher proportion of unique profiles.
  - Phase 2 of this research looks at ways to apply additional treatment to data sets to reduce the likelihood of re-identification

Notes

- Two types of highly targeted attacks are extremely difficult to foil.

1. Focused attack on a specific individual, e.g. a celebrity.

   It is probably a bad idea to include Britney Spears in a de-identified data set, for example.

2. An attack that merely attempts to prove that de-identification is not perfect.

   No method is perfect and a determined attacker, given enough time and money, is likely to be able to demonstrate this acknowledged fact.

Notes

- Under most circumstances HIPAA Safe Harbor method of de-identification protects against re-identification.
  - Best practice may include additional steps, beyond removal of Safe Harbor Method identifiers to further reduce risk in certain circumstances
    - e.g. selective perturbation of some of the variables
- This study was predicated on de-identified data used in medical research.
  - Uses for commercial purposes have different dynamics
    - Patient sensitivity to re-identification risk
    - Motivation and opportunity to try re-identification

Resources

Office for Civil Rights De-Identification Workshop

Researchers have been developing methods to treat data sets so that re-identification risk is even further reduced while maintaining as much utility as possible.

Webcast and more available at:

http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/deidentificationworkshop2010.html

---

ONC Office of the Chief Privacy Officer
Main Number: (202) 690-7151

NORC
Main Number: (301) 634-9300

Department of Health & Human Services
Office of the National Coordinator for Health Information Technology