De-Identification and the Health Insurance Portability and Accountability Act (HIPAA)

Overview and framing of current issues

Simson L. Garfinkel, Ph.D.
Information Technology Laboratory
National Institute of Standards and Technology

Subcommittee on Privacy, Confidentiality & Security
National Committee on Vital and Health Statistics
May 24, 2016

DISCLAIMER: Specific products and organizations identified in this report were used in order to perform the evaluations described. In no case does such identification imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that identified are necessarily the best available for the purpose.
National Institute of Standards and Technology

Founded in 1901

Non-regulatory federal laboratory.

Mission:
“To promote US innovation and industrial competitiveness by advancing measurement science, standards, and technology in ways that enhance economic security and improve our quality of life.”
NISTIR 8053: De-Identification of Personal Information

Covers:

• Why de-identify?
• De-identification terminology
• Famous re-identification cases
• De-identifying and re-identifying *structured data* (e.g. survey data, Census data, etc.)
• Challenges with de-identifying *unstructured data* (e.g. medical text, photographs, medical imagery, genetic information)

October 2015
vi+46 pages
Today there is a significant and growing interest in de-identification.

Controlled Sharing

Open Science

Data Publishing
Big-data is not a new science—it’s the future of all science.

“... Qualified researchers from many organizations will, with appropriate protection of participant confidentiality, have access to the cohort’s de-identified data for research and analysis.”

Request for Information: NIH Precision Medicine Cohort
NOT-OD-15-096
Under the current HIPAA Privacy Rule, de-identified Protected Health Information can be distributed without restriction.

De-identification:

- name
- address
- birthday
- medical record number
- etc.

Medical Records → De-identification → Public Internet
Interest in de-identification extends far beyond healthcare.

Social Science Data
https://www.flickr.com/photos/usdagov/4423599680

Consumer Financial Data

Website
“We will never share your personal information…”
De-identified data can be re-identified.

Sometimes data are not properly de-identified.
De-identified data can be re-identified

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>24</td>
<td>Grover Cleveland</td>
</tr>
<tr>
<td>25</td>
<td>William McKinley</td>
</tr>
<tr>
<td>26</td>
<td>Theodore Roosevelt</td>
</tr>
<tr>
<td>27</td>
<td>William Howard Taft</td>
</tr>
<tr>
<td>28</td>
<td>Woodrow Wilson</td>
</tr>
</tbody>
</table>

Sometimes de-identified data can be *linked* to another dataset.
Simple statistics can be identifying.

<table>
<thead>
<tr>
<th>Title</th>
<th>Age</th>
<th>Sex</th>
<th>Address</th>
<th>ICD-10</th>
<th>Diagnosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lab Tech</td>
<td>35</td>
<td>M</td>
<td>100 Utah St.</td>
<td>K25.0</td>
<td>Gastric Ulcer with hemorrhage</td>
</tr>
<tr>
<td>Lab Tech</td>
<td>56</td>
<td>F</td>
<td>653 Pleasant St.</td>
<td>J00</td>
<td>Acute nasopharyngitis [Common Cold]</td>
</tr>
<tr>
<td>Professor</td>
<td>35</td>
<td>M</td>
<td>564 Main St.</td>
<td>C64.1</td>
<td>Malignant neoplasm of right kidney</td>
</tr>
<tr>
<td>Professor</td>
<td>69</td>
<td>F</td>
<td>202 Sky Lane</td>
<td>C64.1</td>
<td>Malignant neoplasm of right kidney</td>
</tr>
<tr>
<td>Contracts Specialist</td>
<td>52</td>
<td>F</td>
<td>956 Diablo Rd.</td>
<td>L30.9</td>
<td>Dermatitis, unspecified [Eczema]</td>
</tr>
<tr>
<td>University President</td>
<td>56</td>
<td>F</td>
<td>451 Termo Dr.</td>
<td>C64.1</td>
<td>Malignant neoplasm of right kidney</td>
</tr>
</tbody>
</table>

Hypothetical dataset from university healthcare system
Re-identified information can link with other data.

Research Database:

Patient 234-334-11
Diagnoses: A98.4, J00, L30.9

Patient 234-334-11
Age: 35
Genetic History:

Patient 234-334-11
Psychological Records

Patient 234-334-11
Social Services History

<table>
<thead>
<tr>
<th>Ebola Patients</th>
<th>ICD-10</th>
<th>Diagnosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alice</td>
<td>30</td>
<td>F</td>
</tr>
<tr>
<td>Bob</td>
<td>35</td>
<td>M</td>
</tr>
<tr>
<td>Carol</td>
<td>40</td>
<td>F</td>
</tr>
</tbody>
</table>
Techniques for limiting identity disclosure:

<table>
<thead>
<tr>
<th>Title</th>
<th>Age</th>
<th>Sex</th>
<th>Address</th>
<th>ICD-10</th>
<th>Diagnosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>University President</td>
<td>56</td>
<td>F</td>
<td>XXX</td>
<td>C64.1</td>
<td>Malignant neoplasm of right kidney</td>
</tr>
</tbody>
</table>

**Generalization:**
- University President $\Rightarrow$ Senior Administrator
- Age: 56 $\Rightarrow$ Age: 50-59

**Field Swapping:**
- Age: 52 $\Rightarrow$ Age: 56
- Age: 56 $\Rightarrow$ Age: 52

**Noise Addition:**
- University President $\Rightarrow$ VP Finance
- Age: 56 $\Rightarrow$ Age: 58 ±5

**Suppression:**
- University President $\Rightarrow$ XXXXXXXXXXXXXXXX
- Age: 56 $\Rightarrow$ Age: XXX
Lowering identifiability lowers data quality.

- Identified & high quality.
- Poor privacy protection.
- Naïve de-identification
- All quality gone.
- Bad Science,
  Bad Decision
HIPAA Privacy Rule “Safe Harbor” Provision:
Medical records are de-identified if 18 data elements are removed

Direct Identifiers:
— Names
— Individual numbers: phone, fax, SSN, medical record, account #s, etc.
— Email addresses, IP address, URLs
— Biometrics: fingerprints, voiceprints, photographs, etc.
— Any other uniquely identifying number, characteristic or code.

Indirect Identifiers:
— Geographic subdivisions smaller than a state, except first 3 digits of ZIP, provided the combined ZIP codes contain more than 20,000 people.
— Dates directly related to an individual (except for “age 90 or older”)
# Geographic information requires special attention

## Hypothetical dataset from university healthcare system

<table>
<thead>
<tr>
<th>Title</th>
<th>Age</th>
<th>Sex</th>
<th>Address</th>
<th>ICD-10</th>
<th>Diagnosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lab Tech</td>
<td>35</td>
<td>M</td>
<td>100 Utah St. Anytown, 20124</td>
<td>K25.0</td>
<td>Gastric Ulcer with hemorrhage</td>
</tr>
<tr>
<td>Lab Tech</td>
<td>56</td>
<td>F</td>
<td>653 Pleasant St. Uptown, 20321</td>
<td>J00</td>
<td>Acute nasopharyngitis [Common Cold]</td>
</tr>
<tr>
<td>Professor</td>
<td>35</td>
<td>M</td>
<td>564 Main St. Nassis, 25312</td>
<td>T25.332S</td>
<td>Burn of third degree of left toe</td>
</tr>
<tr>
<td>Professor</td>
<td>69</td>
<td>F</td>
<td>202 Sky Lane Katap, 20134</td>
<td>C64.1</td>
<td>Malignant neoplasm of right kidney</td>
</tr>
<tr>
<td>Contracts Specialist</td>
<td>52</td>
<td>F</td>
<td>956 Diablo Rd. Quirky, 23990</td>
<td>L30.9</td>
<td>Dermatitis, unspecified [Eczema]</td>
</tr>
<tr>
<td>University President</td>
<td>56</td>
<td>F</td>
<td>451 Termo Dr. Boltz, 25333</td>
<td>C64.1</td>
<td>Malignant neoplasm of right kidney</td>
</tr>
</tbody>
</table>
Safe Harbor allows ZIP3 (assuming there are 20,000 people living in the area)

<table>
<thead>
<tr>
<th>Indirect identifiers</th>
<th>Direct identifier</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Title</strong></td>
<td><strong>Age</strong></td>
</tr>
<tr>
<td>Lab Tech</td>
<td>35</td>
</tr>
<tr>
<td>Lab Tech</td>
<td>56</td>
</tr>
<tr>
<td>Professor</td>
<td>35</td>
</tr>
<tr>
<td>Professor</td>
<td>69</td>
</tr>
<tr>
<td>Contracts Specialist</td>
<td>52</td>
</tr>
<tr>
<td>University President</td>
<td>56</td>
</tr>
</tbody>
</table>

Hypothetical dataset from university healthcare system
Results of the 2010 Office of the National Coordinator for Health Information Technology Safe Harbor Re-Identification Test:

15,000 Hispanic Patients
- 216 distinct by Sex, ZIP3 & age

30,000 Records from InfoUSA
- 84 distinct by sex, ZIP3 & age
- 20 match on sex, ZIP3 & age
- 2 actual matches on last name, street address, and phone

Data from 2004-2009
K-anonymity: assure at least “k” records have the same set of indirect identifiers.

<table>
<thead>
<tr>
<th>Indirect identifiers</th>
<th>Direct identifier</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Title</strong></td>
<td><strong>Age</strong></td>
</tr>
<tr>
<td>Lab Tech</td>
<td>35</td>
</tr>
<tr>
<td>Lab Tech</td>
<td>56</td>
</tr>
<tr>
<td>Professor</td>
<td>35</td>
</tr>
<tr>
<td>Professor</td>
<td>69</td>
</tr>
<tr>
<td>Contracts Specialist</td>
<td>52</td>
</tr>
<tr>
<td>University President</td>
<td>56</td>
</tr>
</tbody>
</table>

*Color background indicates values modified for k=2 k-anonymity*
“Tiger Teams” are another way to test re-identification.

Data from 2004-2009

15,000
Hispanic
Patients
216 distinct by
Sex, ZIP3 & age

30,000
Records from
InfoUSA
84 distinct by
sex, ZIP3 & age

20 match
on sex,
ZIP3 & age

2 actual matches on last name,
street address, and phone

Estimated Re-identification rate:
No verification: 20 in 15,000
Verification: 2 in 15,000
Re-identification tests assume data available to match. As more data become available, re-identification gets easier.

**Hypothetical Example**

- **300M records from provider “C” (Social Media?)**
- **15,000 Hispanic Patients**
- **30,000 Records from InfoUSA**
- **1M records From provider “B”**
A constellation of diseases can be an identifier

De-identified medical records from provider “N”

Linked records

- Smallpox
- Concussion at age 9
- Malaria at age 21
- Depression
- Fractured jaw

Hypothetical Example
A single identified blood test can be the link to dozens of de-identified records.
Blood tests can be de-identified by adding noise

<table>
<thead>
<tr>
<th>Test Name</th>
<th>Patient's Results</th>
<th>Ref. Range</th>
<th>Units</th>
</tr>
</thead>
<tbody>
<tr>
<td>Na</td>
<td>124</td>
<td>136-145</td>
<td>mEq/L</td>
</tr>
<tr>
<td>K</td>
<td>5.8</td>
<td>3.5-5.1</td>
<td>mEq/L</td>
</tr>
<tr>
<td>CO2</td>
<td>25</td>
<td>23-29</td>
<td>mEq/L</td>
</tr>
<tr>
<td>Cl</td>
<td>101</td>
<td>98-107</td>
<td>mEq/L</td>
</tr>
<tr>
<td>Glucose</td>
<td>107</td>
<td>74-100</td>
<td>mg/dL</td>
</tr>
<tr>
<td>Ca</td>
<td>10.1</td>
<td>8.6-10.2</td>
<td>mg/dL</td>
</tr>
<tr>
<td>BUN</td>
<td>17</td>
<td>8-23</td>
<td>mg/dL</td>
</tr>
<tr>
<td>Creatinine</td>
<td>0.9</td>
<td>0.8-1.3</td>
<td>mg/dL</td>
</tr>
</tbody>
</table>

(values for demonstration only)

Research database
“Differential Privacy” adds systematic noise to query results

Key concepts: Privacy Budget & Noise
The Census Bureau distributes synthetic data to protect privacy while preserving some data quality.
Can synthetic datasets designed to enable research also be used to promote accountability and transparency?

- Animated encounter data
- Synthetic tabular data
- Body-worn camera video with replaced faces
De-identification strategies should be formally evaluated.

Do they meet the stated policy goals?

Does the software faithfully implement the stated algorithm?

Are the statistical privacy guarantees actually met?

Is the necessary training in place?

Will there be monitoring and auditing?
De-identification of non-tabular data poses special problems.

Google claims 90% of faces and 95% of license plates removed through automated processing.

Medical imagery can be highly identifying.
More research is needed to determine if systems can protect privacy and allow for unlimited use of data.
Can raw data be transformed so completely that individuals cannot recognize their own data once they are in a crowd?


In summary:
We have learned a lot about de-identification in recent years.

The de-identification “toolkit” has several options:

- suppression, generalization  *commonly used in healthcare*
- field swapping, noise addition  *commonly used in vital statistics*

K-anonymity and Differential Privacy are formal models for evaluating the quality of de-identification.

We increasingly have the ability to:

- Modify data so that the data subjects’ identity is removed, leaving information that is somewhat useful.
- But the more useful it is, the more likely it can be re-identified.

We need procedures for:

- Evaluating the effectiveness of de-identification.
- Evaluating the usefulness of the data that remain.

We need these techniques for a wide range of data:

- Structured data, text, medical, video.

Lowering identifiability lowers data quality.