Building Systems to Support Health Information Privacy

Brad Miller, Ph.D.
Assistant Professor
Department of Biomedical Informatics, School of Medicine
Department of Computer Science, School of Engineering
Vanderbilt University

Disclaimer

- Privacy is an overloaded word
- Today: privacy in the context of a specific domain

- Healthcare
  - Health Insurance Portability Accountability Act (HIPAA)
  - NIH Data Sharing Policy
  - NIH Genome-wide Association Study Data Sharing Policy

What’s Going On?

- We study “privacy” in various operational realms
  - Primary Care
    - Clinical Information Systems Design
    - “Intelligent” Auditing
  - Secondary Uses
    - De-identification / Re-identification / Anonymization
    - Secure Data Integration and Analysis

Privacy Everywhere

- We do not always control who gets, and has access to, our information
- Legally, however, data collectors may be required to maintain your privacy

Privacy Everywhere

- Let’s begin with data already in the system
Electronic Medical Records – Hooray!

• At a glance
• Access control
• Improve overview

Clinical Information Systems Design
(Duncavage 2007; Mate et al., 2008; Werner et al., 2007, 2008)

Surveillance (Pauket, Malin)

• Very little role-based access control in large academic medical centers (why?)
• Most auditing is done manually (why?)
> 1.5 million patient records
> 20,000 authorized users

Jan 1, 2006
• Users linked if accessed 1 common patient (~ 900 users, 2000 patients)

HORNET: Healthcare Organizational Research Toolkit
(http://code.google.com/p/homet/)

Learning Policies and Anomalies from EMR Access Logs
• Extracts patterns from medical record access logs to model policies & detect “privacy” violations.
What’s Going On?

- “Privacy” in various operational settings
  - Primary Care
    - Clinical Information Systems Design
    - "Intelligent" Auditing
  - Secondary Uses
    - De-identification / Re-identification / Anonymization
    - Secure Data Integration and Analysis

Information Integration

Research Support & Data Collection

Holy Hand Grenades! How Did You...

- Initially an institutionally funded project
- Office for Human Research Protections designation as Non-Human Subjects Research under 45 CFR 46 ("HIPAA Common Rule")*
  - Samples & data not linked to identity
  - Conducted with IRB & ethics oversight

HIPAA - Secondary Data Sharing

• Safe Harbor
• Limited Release
• Statistical or Scientific Standard

HIPAA Safe Harbor

• Data that can be given away without oversight
  - Names/initials
  - Street address, city, county, postal code and equivalent geocodes
  - All elements of dates, except year, and ages over 89
  - Social Security number, Medical Record, Health Plan ID, Account, License, Serial, Device
  - Web, e-mail, URL, IP addresses
  - Biometric identifiers: fingerprints, face prints
  - Full face photographs and comparable images
  - Any other unique identifying number, characteristic, or code
  • A code is an identifier if the person holding the coded data can re-identify the individual

HIPAA Limited Dataset

• Includes more specific information than Safe Harbor
• Can include
  - Dates of birth, death, service
  - Geographic Info: Town, Zip code, County
• Requires Contract: Research entity provides assurances that it will not use or disclose the information for purposes other than research and will not identify or contact the individuals who are the subjects

HIPAA Statistical / Scientific Standard

• Certified via “generally accepted statistical and scientific principles and methods” that the risk is very small that the information could be used, alone or in conjunction with other reasonably available information, by the anticipated recipient to identify the subject of the information.”

• “Must document the methods and results of the analysis that justify such a determination”

• “Must not disclose the key or other mechanism that would have enabled the information to be re-identified”
  - Includes past pseudo-random number algorithms and seed attacks

"Scrubbing" Medical Records

Rules
Regular Expressions
Dictionaries
Exclusions

Machine Learning – Conditional Random Fields

“Scrubbed” Medical Record

Unknown residual re-identification potential (e.g. “the mayor’s wife")
Technology + Policy

- Data bank access restricted to Vanderbilt employees
- Must sign use agreement that prohibits "re-identification"
- Operations Advisory Board and Institutional Review Board approval needed for each project
- All data access logged and audited per project

To Vanderbilt ... and Beyond

The eMERGE Network

electronic Medical Records & Genomics

A consortium of investigators linked to electronic medical record data for the discovery of genetic predispositions to common diseases

Consortium members [http://www.gwas.net]
- Group Health of Puget Sound (GHP)
- Mayo Clinic
- Northwestern University
- Vanderbilt University

Funding conditions: contribute de-identified genomic and EMR-derived phenotype data to database of genotype and phenotype (dbGAP) at NCBI, NIH

Data Sharing Policies

- Feb ‘03: National Institutes of Health Data Sharing Policy
  - "data should be made as widely & freely available as possible"
  - researchers who receive $500,000 must develop a data sharing plan or describe why data sharing is not possible
  - data must be shared in a manner that is devoid of "identifiable information"

- Aug ‘06: NIH Supported Genome-Wide Association Studies Policy
  - Researchers who received >= $0 for GWAS

Healthcare Reform At Work

- In 1997, approx. 44 of 50 states collected and disseminated hospital discharge data
- In 2005, approx. 47 of 50 states " "
- Attributes recommended by National Association of Health Data Organizations for disclosure
  - Patient Zip Code
  - Patient Birth Date
  - Patient Gender
  - Patient Race/Background
  - Patient Number
  - Visit Date


The Face that Launched a Thousand Ships
Case Study – “Quasi-identifier”

Back in the ‘90s

Hospital Discharge Data

City of Cambridge, MA Voter Registration Records

Case Study – “Quasi-identifier”

Back in the ‘90s

Voter List

5-Digit Zip Code
+ Birthdate
+ Gender

63-87% of US estimated to be unique


Case Study – “Quasi-identifier”

Re-identification of William Weld

Hospital Discharge Data

Voter List

And Now, It’s A Phenomenon!!!

The AOL® Search Log Case (2006)

• Goal: Support web search research
• 650k customers, 20 million queries, 3 month period
• Names replaced with persistent pseudonyms
Barbaro & Zeller. A face exposed for AOL searcher no. 4417749.

User 4417749 issued hundreds of searches with a particular syntax.

Thelma Arnold & Dudley

The Netflix Challenge (2008-2009)

- Netflix published movie selections of ~450,000 pseudonymized subscribers
- Re-identification via uniqueness of movie combinations

- Class action filed December 2009

Now Back to Your Regularly Scheduled Programming

Attacks on Demographics
- We use population estimates according the US Census
- It’s not perfect, but it’s a start.

Beyond “unique”


http://factfinder.census.gov/
Case Study: Tennessee

![Graph showing population distribution under limited and safe harbor policies.]

Policy Analysis via “Trust Differential”...

Risk (Limited Dataset)
Risk (Safe Harbor)

- Unique states:
  - Delaware’s risk increases by a factor ~1,000
  - Tennessee’s risk increases by a factor ~2,300
  - Illinois’s risk increases by a factor ~65,000

- ≤20,000:
  - Delaware’s risk does not increase
  - Tennessee’s risk increases by a factor ~8
  - Illinois’s risk increases by a factor ~37

Cost?

<table>
<thead>
<tr>
<th>State</th>
<th>Limited Dataset</th>
<th>Safe Harbor</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Marker Risk</td>
<td>Cost per Re-id</td>
</tr>
<tr>
<td>VA</td>
<td>3159764</td>
<td>$0</td>
</tr>
<tr>
<td>NY</td>
<td>2905697</td>
<td>$0</td>
</tr>
<tr>
<td>SC</td>
<td>2231973</td>
<td>$0</td>
</tr>
<tr>
<td>WI</td>
<td>72</td>
<td>$174</td>
</tr>
<tr>
<td>WV</td>
<td>55</td>
<td>$309</td>
</tr>
<tr>
<td>NH</td>
<td>10</td>
<td>$827</td>
</tr>
</tbody>
</table>

A Couple of Parting Thoughts

- The application of technology must be considered within the systems and operational processes they will be applied.

- One person’s vulnerability is another person’s armor (variation in risks).

- It is possible to inject privacy into health information systems, but it must be done early (see “privacy by design”).

- Sometimes theory needs to be balanced with practicality.