Health Informatics

Lecture 5

Samantha Kleinberg
samantha.kleinberg@stevens.edu
• Two weeks: Midterm!
• Participation: 15% of your grade
• Homework: 10% of your grade
HIPAA

Health Insurance Portability and Accountability Act

Includes definitions for protected health information (PHI), and what can be shared and with whom
• Who’s covered by HIPAA?
  – Ex: healthcare provider, researchers working with PHI from hospital

• What’s required?
  – Usually need consent, unless waiver from IRB or meet certain other criteria

• Key component: data de-identified and you don’t have reason to believe that they can be re-identified. If this satisfied, no longer covered by HIPAA
18 HIPAA identifiers

1. Names
2. Certain geographic information*
3. Dates other than year; all ages over 89 and all elements of dates (including year) indicative of such age, except when aggregated into a category of age 90 or older;
4. Phone numbers;
5. Fax numbers;
6. Electronic 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 and serial numbers, including license plate numbers;
13. Device identifiers and serial numbers;
14. URLs
15. IP addresses
16. Biometric identifiers, including finger and voice prints;
17. Full face photographic images and any comparable images
18. Any other unique identifying number, characteristic, or code
• No longer covered (safe harbor)
  – Remove 18 HIPAA identifiers, and no knowledge that people can be re-identified
  – Or guarantee “small” chance of re-identification

• Limited dataset
  – 16 of 18 removed (can keep dates)
  – No consent needed
  – Re-identification prohibited
• Not required to be removed
  – Genetic information
  – Text
• Use imperfect method or just give up?
• Should people be able to consent?
De-identifying data in practice

• Basic: replace all IDs with research ID, change dates, remove HIPAA identifiers
• Different date shift/function for each patient

• Structured data
  – Add noise
• Text
• Genome
  – Generalize sequence
• Images
  – Facial blur
Limits of deidentification

"87% of the U.S. Population are uniquely identified by {date of birth, gender, ZIP}.

http://latanyasweeney.org/work/identifiability.html
Matching Known Patients to Health Records in Washington State Data

Latanya Sweeney

(Submitted on 4 Jul 2013 (v1), last revised 5 Jul 2013 (this version, v2))

The State of Washington sells patient-level health data for $50. This publicly available dataset has virtually all hospitalizations occurring in the State in a given year, including patient demographics, diagnoses, procedures, attending physician, hospital, a summary of charges, and how the bill was paid. It does not contain patient names or addresses (only ZIPs). Newspaper stories printed in the State for the same year that contain the word "hospitalized" often include a patient's name and residential information and explain why the person was hospitalized, such as vehicle accident or assault. News information uniquely and exactly matched medical records in the State database for 35 of the 81 cases (or 43 percent) found in 2011, thereby putting names to patient records. A news reporter verified matches by contacting patients. Employers, financial organizations and others know the same kind of information as reported in news stories making it just as easy for them to identify the medical records of employees, debtors, and others.
<table>
<thead>
<tr>
<th>Name of Covered Entity</th>
<th>State</th>
<th>Individuals Affected</th>
<th>Breach Date</th>
<th>Type of Breach</th>
<th>Location of Breached Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supportive Concepts For Families, Inc.</td>
<td>PA</td>
<td>593</td>
<td>02/06/2013 - 12/16/2013</td>
<td>Unauthorized Access/Disclosure</td>
<td>Network Server</td>
</tr>
<tr>
<td>Health Care Solutions At Home Inc.</td>
<td>OH</td>
<td>1139</td>
<td>12/17/2013</td>
<td>Other</td>
<td>Other</td>
</tr>
<tr>
<td>St. Vincent Hospital And Healthcare, Inc</td>
<td>IN</td>
<td>1142</td>
<td>12/23/2013</td>
<td>Theft</td>
<td>Laptop</td>
</tr>
<tr>
<td>Missouri Consolidated Health Care Plan</td>
<td>MO</td>
<td>10024</td>
<td>03/23/2012 - 01/22/2014</td>
<td>Unauthorized Access/Disclosure</td>
<td>Network Server</td>
</tr>
<tr>
<td>The Clorox Company Group Insurance Plan</td>
<td>CA</td>
<td>520</td>
<td>04/16/2012 - 01/22/2014</td>
<td>Unauthorized Access/Disclosure</td>
<td>Network Server</td>
</tr>
<tr>
<td>Inspira Health Network Inc.</td>
<td>NJ</td>
<td>1411</td>
<td>12/23/2013 - 12/23/2013</td>
<td>Theft</td>
<td>Desktop Computer</td>
</tr>
<tr>
<td>Nissan North America, Inc.</td>
<td>TN</td>
<td>1511</td>
<td>05/08/2012 - 01/22/2014</td>
<td>Unauthorized Access/Disclosure</td>
<td>Network Server</td>
</tr>
<tr>
<td>Health Source Of Ohio Inc.</td>
<td>OH</td>
<td>8845</td>
<td>11/18/2013 - 12/24/2013</td>
<td>Unauthorized Access/Disclosure, Other</td>
<td>Other</td>
</tr>
<tr>
<td>Cornerstone Health Care, Pa</td>
<td>NC</td>
<td>548</td>
<td>12/31/2013</td>
<td>Theft, Loss</td>
<td>Laptop</td>
</tr>
</tbody>
</table>
Identifying Participants in the Personal Genome Project by Name

Latanya Sweeney, Akua Abu, Julia Winn

Harvard College
Cambridge, Massachusetts
latanya@fas.harvard.edu, aabu@college.harvard.edu, jwinn@post.harvard.edu

We linked names and contact information to publicly available profiles in the Personal Genome Project. These profiles contain medical and genomic information, including details about medications, procedures and diseases, and demographic information, such as date of birth, gender, and postal code. By linking demographics to public records such as voter lists, and mining for names hidden in attached documents, we correctly identified 84 to 97 percent of the profiles for which we provided names. Our ability to learn their names is based on their demographics, not their DNA, thereby revisiting an old vulnerability that could be easily thwarted with minimal loss of research value. So, we propose technical remedies for people to learn about their demographics to make better decisions.

and thousands of people get subsequently harmed doing so, policy makers may respond and take away the freedom to make personal data sharing decisions, thereby depriving society of individual choice. To make smarter decisions, people need an understanding of actual risks and ways technology can help.

BACKGROUND

Launched in 2006, the Personal Genome Project (PGP) aims to sequence the genotypic and phenotypic information of 100,000 informed volunteers and display it publicly online in an extensive public database [1]. Information provided in the PGP includes DNA information, behavioral
The False Security of Blind Dates

Chrononymization’s Lack of Impact on Data Privacy of Laboratory Data

J.J. Cimino
NIH Clinical Center, US National Institutes of Health, Bethesda, Maryland, USA

Keywords
Patient data privacy, data adjustments, clinical research, clinical informatics, health policy, anonymization, de-identification, dates

Summary
Background: The reuse of clinical data for research purposes requires methods for the protection of personal privacy. One general approach is the removal of personal identifiers from the data. A frequent part of this anonymization process is the removal of times and dates, which we refer to as “chrononymization.” While this step can make the association with identified data (such as public information or a small sample of patient information) more difficult, it comes at a cost to the usefulness of the data for research.

Objectives: We sought to determine whether removal of dates from common laboratory test panels offers any advantage in protecting such data from re-identification.

Methods: We obtained a set of results for 5.9 million laboratory panels from the National Institutes of Health’s (NIH) Biomedical Translational Research Information System (BTRIS), selected a random set of 20,000 panels from the larger source sets, and then identified all matches between the sets.

Results: We found that while removal of dates could hinder the re-identification of a single test result, such removal had almost no effect when entire panels were used.

Conclusions: Our results suggest that reliance on chrononymization provides a false sense of security for the protection of laboratory test results. As a result of this study, the NIH has chosen to rely on policy solutions, such as strong data use agreements, rather than removal of dates when reusing clinical data for research purposes.
Good practices

• Physical security
  – Locked cabinet/office, don’t print records and leave them in common area/on desk

• Encryption
  – Losing laptop/thumbdrive, sending harddrive for repair can be disastrous

• Restricted access to servers with PHI

• PHI in segregated, labeled, directories
Privacy protection approaches

• De-identification (i.e. HIPAA safe harbor)
  – Remove explicit and quasi-identifying data

• Trusted third party
  – Used for identifying subjects
    Researcher -> physician -> TTP -> researcher
  – Restricted version where 3rd party only has encrypted data
Uses for medical data

• Finding long-term risk factors for disease
• Drug-drug interactions, side effects (post-market)
• Population health
and more!
Uniqueness of Medical Data Mining

• Can’t just apply methods from data mining/ML
• Main points
  – Heterogeneity
  – Ethical, legal, social issues
  – Statistical philosophy
  – Special status of medicine
Heterogeneity of data

• Structured/unstructured, Imaging
  – Many data mining methods can handle only one of these

• How to combine qualitative/quantitative information?
Heterogeneity of data

• Importance of interpretation

• If we see CHF in record – what did clinician mean?
  – Could be suspected CHF
  – Hypothesis explaining symptoms
  – Past or family history...

• If we don’t see indicator for CHF does that mean patient doesn’t have it?
  – May not be billed for, may not be treated (if more pressing problems)
Heterogeneity of data

• Standardization
  – Of data (recall ICD9 for example)
  – Of outcomes

• Example:
  – One group evaluates glucose control system by calculating how often glucose is within 70-150, another group uses 80-140. How to compare?
Heterogeneity of data

• Difficulty applying precise labels

• Uncertainty
  – Does a particular billing code mean patient definitely has illness?
  – When did the illness start?

• **Test vs diagnosis**
  – We see imperfect indicators for a disease, not the disease itself
  – For prediction, target event may be diagnosis NOT onset of disease
Ethics, legal and social factors

• Access to data (researchers, patients)
• Concern about liability
  – Affects what data can be collected, which tests can be done
• Privacy
  – Can we use all the data we want?
  – Can we analyze in Amazon cloud? Give text to mechanical turkers? Scrape data from password protected websites?
Statistical methods

• Do the assumptions of computational methods hold? Do we need new domain-specific ones?
  – E.g. Clinical NLP as distinct subset of NLP

• Importance of domain knowledge

• Error
Statistical methods

Missing data
- Very common and the reason for it changes interpretation

- Chronic vs acute conditions
- Expensive test (e.g. could be missing due to insurance coverage)
Statistical methods

Redundancy, inconsistent data

- conflicting lab results
- note says patient is Male, later says Female
Special status of medicine

- Impact/risk of false findings: good enough for a paper vs good enough to treat a patient
- Cannot do any conceivable test
Believe it or not: how much can we rely on published data on potential drug targets?

Florian Prinz, Thomas Schlange and

A recent report by Arrowsmith noted that the success rates for new development projects in Phase II trials have fallen from 28% to 18% in recent years, with insufficient efficacy being the most frequent reason for failure (Phase II failures: 2008–2010. *Nature Rev. Drug Discov.* **10**, 328–329 (2011))\(^1\). This indicates the limitations of the predictivity of disease models and also that the validity of the targets being investigated is frequently questionable, which is a crucial issue to address if success rates in clinical trials are to be improved.

Candidate drug targets in industry are derived from various sources, including in-house target identification campaigns. into ‘feasible/marketable’, and the financial costs of pursuing a full-blown drug discovery and development programme for a particular target could ultimately be hundreds of millions of Euros. Even in the earlier stages, investments in activities such as high-throughput screening programmes are substantial, and thus the validity of published data on potential targets is crucial for companies when deciding to start novel projects.

To mitigate some of the risks of such investments ultimately being wasted, most pharmaceutical companies run in-house target validation programmes. However, validation projects that were started in our company

This analysis revealed that only in ~20-25% of the projects were the relevant published data completely in line with our in-house findings.
Fifty-three papers were deemed ‘landmark’ studies ... scientific findings were confirmed in only 6 (11%) cases.
Much of the epidemiological data underpinning the government’s dietary advice comes from studies run by Harvard’s school of public health. In 2011, directors of the National Institute of Statistical Sciences analyzed many of Harvard’s most important findings and found that they could not be reproduced in clinical trials.
Deming, data and observational studies

Table 1. We have found 12 papers in which claims coming from observational studies were tested in randomised clinical trials. Many of the trials are quite large. In most of the observational studies multiple claims were tested, often in factorial designs, e.g. vitamin D and calcium individually and together along with a placebo group. Note that none of the claims replicated in the direction claimed in the observational studies and that there was statistical significance in the opposite direction five times.

<table>
<thead>
<tr>
<th>ID no.</th>
<th>Pos.</th>
<th>Neg.</th>
<th>No. of claims</th>
<th>Treatment(s)</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>Vit E, beta-carotene</td>
<td>NEJM 1994; 330: 1029–1035</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>3</td>
<td>4</td>
<td>Hormone Replacement Ther.</td>
<td>JAMA 2003; 289: 2651–2662, 2663–2672, 2673–2684</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>Vit E, beta-carotene</td>
<td>JNCI 2005; 97: 481–488</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>Vit E</td>
<td>JAMA 2005; 293: 1338–1347</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>Low Fat</td>
<td>JAMA. 2006; 295: 655–666</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>Vit D, Calcium</td>
<td>NEJM 2006; 354: 669–683</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>Folic acid, Vit B6, B12</td>
<td>NEJM 2006; 354: 2764–2772</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>Low Fat</td>
<td>JAMA 2007; 298: 289–298</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>0</td>
<td>12</td>
<td>Vit C, Vit E, beta-carotene</td>
<td>Arch Intern Med 2007; 167: 1610–1618</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>0</td>
<td>12</td>
<td>Vit C, Vit E</td>
<td>JAMA 2008; 300: 2123–2133</td>
</tr>
<tr>
<td>11</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>Vit E, Selenium</td>
<td>JAMA 2009; 301: 39–51</td>
</tr>
<tr>
<td>12</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>HRT + Vitamins</td>
<td>JAMA 2002; 288: 2431–2440</td>
</tr>
</tbody>
</table>

Totals 0 5 52
Replications still somewhat rare in health informatics

Why?
Replication

Repeating same method under same conditions
Tells us method sufficiently documented and finding stable
Reproduction

Demonstrating *same finding* in another setting, under different conditions

Tells us about generalizability of results
Goals

• Uncover bias
• Find confounders
• Detect limits, generalizability
• Validate methods/findings
Why is replication challenging?

• Patients heterogeneous
• Populations have different characteristics
• Sample size
• Not all data available everywhere (+ retrospective)
• Data quality (noisy, sparse, missing)

Replication doesn’t validate methods
  – Need ground truth
  – Need controlled data
Today’s paper

S. Kleinberg and N. Elhadad (2013) Lessons Learned in Replicating Data-Driven Experiments in Multiple Medical Systems and Patient Populations. AMIA Annual Symposium.
More on challenges

• Observational Data in general
  – Nonstationarity
  – Selection bias
  – Choosing variables
  – Seeing vs doing

• Biomedical Data
  – Biased approximation of truth
  – Sample size
  – Censoring (left, right)
  – Institutional differences
  – Fragmentation of data
Nonstationarity

The data are not static:

• EHR systems change
• Terminology changes
• New tests developed
Nonstationarity

Risk Calculator for Cholesterol Appears Flawed

By GINA KOLATA
Published: November 17, 2013 | 794 Comments

Last week, the nation’s leading heart organizations released a sweeping new set of guidelines for lowering cholesterol, along with an online calculator meant to help doctors assess risks and treatment options. But, in a major embarrassment to the health groups, the calculator appears to greatly overestimate risk, so much so that it could mistakenly suggest that millions more people are candidates for statin drugs.

The apparent problem prompted one leading cardiologist, a past president of the American College of Cardiology, to call on Sunday for a halt to the implementation of the new guidelines.

“It’s stunning,” said the cardiologist, Dr. Steven Nissen, chief of cardiovascular medicine at the Cleveland Clinic. “We need a pause to further evaluate this approach before it is implemented on a widespread basis.”
Other ways data may be nonstationary

• Changes in record keeping
• Changes in patient population over time
• Changes in terminology (gallop vs third heart sound; crackles vs rales)
• More accurate tests
• New diagnoses/risk factors

• Note diff between physiology and our observation of it!
<table>
<thead>
<tr>
<th>Treatment</th>
<th>Dead</th>
<th>Alive</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>85</td>
<td>215 (72%)</td>
</tr>
<tr>
<td>B</td>
<td>59</td>
<td>241 (80%)</td>
</tr>
<tr>
<td>Total</td>
<td>144</td>
<td>456</td>
</tr>
</tbody>
</table>
## Simpson’s paradox

Bias in graduate admissions?

<table>
<thead>
<tr>
<th></th>
<th>Admit</th>
<th>Deny</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>3738</td>
<td>4704</td>
</tr>
<tr>
<td>Female</td>
<td>1494</td>
<td>2827</td>
</tr>
</tbody>
</table>

\[
P(\text{admitted}) = \frac{5232}{12763} \approx 0.41
\]

\[
P(\text{admitted} | \text{female}) = \frac{1494}{4321} \approx 0.35
\]

\[
P(\text{admitted} | \text{male}) = \frac{3738}{8442} \approx 0.44
\]

Alice’s patient outcomes

Bob’s patient outcomes
Selection bias

Alice’s patient population

Bob’s patient population
A+H → Hospital

Find association between A, H because not seeing A and H separately in hospital population
Selection bias

• Drug given to sicker patients -> worse outcomes?

• Hospital outcomes depend on population: socioeconomic factors, medical insurance, etc.