Genomics and Informatics:
Technology Solutions to Compute with Sensitive Data

April 21, 2015
Innovations in Healthcare Informatics

NIH U54 HL108460

Lucila Ohno-Machado, MD, MBA, PhD
Biomedical Informatics, University of California San Diego
• Biomedical science is moving towards data-driven approaches
  » 60+ genetic variants are adopted in clinical practice
  » Sequencing is getting cheaper
  » Family history data is considered as "the best genetic test available"

slide from Dr. Xiaoqian Jiang
USE OF MEDICAL INFORMATION AND SPECIMENS:

I understand that my medical information, photographs, and/or video in any form may be used for other [INSTITUTION] purposes, such as quality improvement, patient safety and education. I also understand that my medical information and tissue, fluids, cells and other specimens (collectively, "Specimens") that [INSTITUTION] may collect during the course of my treatment and care may be used and shared with researchers.
Health Insurance Portability & Accountability Act (HIPAA)

- **Security**
  - Risk of disclosure
  - Liability
- **Privacy**
  - Risk of re-identification
  - Informed consent

**HIPAA ‘De-identified’ data**
- removal of 18 identifiers, such as dates, biometrics, names, etc.
- expert certification of low risk of re-identification
- ‘Limited’ data sets have dates

Biometrics are Protected Health Information (PHI)

PHI requires HIPAA

• Biometrics require HIPAA
Genomes are Biometrics

Biometrics are Protected Health Information (PHI)

PHI requires HIPAA

• Biometrics require HIPAA
What is the problem?

• Imagine a public data set of genomes about a certain disease (e.g., Alzheimer’s)

• Can you find out if your neighbor has Alzheimer’s?

• Same problem happens with clinical data, which can be re-identified to a target patient
  » Dates of visits
  » Combination of patient characteristics
New DNA tests on a secret sample collected from a relative of suspect Albert deSalvo triggered the exhumation after authorities said there was a “familial match” with genetic material preserved in the killing of Sullivan...

Authorities made the match through DNA taken from a water bottle thrown away by DeSalvo’s nephew...

But a lawyer for the DeSalvos told CNN the family was “outraged, disgusted and offended” by the decision to secretly take a DNA sample of one of its members...
Bigger data, bigger challenge

- High dimensional data (e.g., whole genomic sequencing)
- Privacy rules (PHI including genomic data need to be protected under HIPAA)
- Institutional processes (internal review boards)
- Patient preferences: Tiered information consent
- Federated clinical/genome data research networks
  » Secure multiparty computation
  » Homomorphic encryption


adapted from Dr. Xiaqian Jiang
Our Goals

- Share access to data and computation
- Train the new generation of data scientists
- Provide innovative software, platform, and infrastructure
- Protect privacy

Develop
  » Algorithms
  » Tools
  » Infrastructure
  » Policies

Supported by the NIH Grant U54 HL108460 to the University of California, San Diego
Models for ‘Sharing’ Data

Clinical Data Network – UC-ReX

- Clinical Data Warehouses from 5 Medical Centers and affiliated institutions exchange (>12 million patients)
- Improve patient safety surveillance, quality improvement, translational research

Funded by the UC Office of the President
Data Matching Function: Map D onto data dictionaries, use Data Warehouses derived from EHR

- Currently
  - Count queries
- Future steps
  - Individual-level patient data delivery
  - Cross-institutional analytics
Quality improvement/health services research?

User requests data for Quality Improvement or Research

Are the data accessible?

Trusted Broker(s)
- Identity & Trust Management
- Policy enforcement

Security Entity

Diverse Healthcare Entities in 3 different states (federal, state, private)

How many patients over 65 are on Warfarin or Dabigatran?

What are the major and minor bleeding rates for patients on these drugs?

Is CYP2C9 mutation an important factor?

Count queries and statistics across data warehouses

AHRQ R01HS19913 / EDM forum

SCANNER
• Policy embedded in software

User requests data for Quality Improvement or Research

Trusted Broker(s)
• Identity & Trust Management
• Policy enforcement

Security Entity

Kim K et al. Data Governance Requirements for Distributed Clinical Research Networks: Triangulating Perspectives of Diverse Stakeholders. JAMIA 2014
Meeker et al. A System to Build Distributed Multivariate Models and Manage Disparate Data Sharing Policies: Implementation in the Scalable National Network for Effectiveness Research. JAMIA (accepted)
Kim K et al. Comparison of Consumers’ Views on Electronic Data Sharing for Healthcare and Research. JAMIA (accepted)
- Predictive modeling and adjustment for cofounders require lots of data
- Some institutions cannot move data outside their firewalls, we can bring computation to the data


Distributed Model

First derivative:

\[
l'_r(\beta) = \sum_{i=1}^{n} \left\{ s_{i,r} - d_i \frac{\sum_{j \in R(t_i)} z_{j,r} \exp^{\beta^T z_j}}{\sum_{j \in R(t_i)} \exp^{\beta^T z_j}} \right\}
\]

\[
l'_r(\beta) = \sum_{k=1}^{m} \sum_{i=1}^{n} \left\{ \sum_{z_j \in \text{site}_k} z_j - \frac{\sum_{i=1}^{n} d_i \sum_{k=1}^{m} \sum_{j \in R(t_i)} I(z_j \in \text{site}_k) z_{j,r} \exp^{\beta^T z_j}}{\sum_{k=1}^{m} \sum_{j \in R(t_i)} I(z_j \in \text{site}_k) \exp^{\beta^T z_j}} \right\}
\]

Hessian matrix:

\[
l''_{r,q}(\beta) = -\sum_{i=1}^{n} d_i \left\{ \frac{\sum_{j \in R(t_i)} z_{j,r} z_{j,q} \exp^{\beta^T z_j}}{\sum_{j \in R(t_i)} \exp^{\beta^T z_j}} \right\} - \frac{\sum_{j \in R(t_i)} z_{j,r} \exp^{\beta^T z_j}}{\sum_{j \in R(t_i)} \exp^{\beta^T z_j}} \frac{\sum_{j \in R(t_i)} z_{j,q} \exp^{\beta^T z_j}}{\sum_{j \in R(t_i)} \exp^{\beta^T z_j}}
\]

Conclusion: no patient data needs to be sent from the sites, only aggregates
Horizontal and Vertical Partitions

<table>
<thead>
<tr>
<th>Patient</th>
<th>Age</th>
<th>Insurance</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>45</td>
<td>X</td>
</tr>
<tr>
<td>A2</td>
<td>32</td>
<td>Y</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Patient</th>
<th>Age</th>
<th>Insurance</th>
</tr>
</thead>
<tbody>
<tr>
<td>B1</td>
<td>45</td>
<td>Y</td>
</tr>
<tr>
<td>B2</td>
<td>32</td>
<td>Y</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Patient</th>
<th>Age</th>
<th>Insurance</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>45</td>
<td>X</td>
</tr>
<tr>
<td>A2</td>
<td>32</td>
<td>Y</td>
</tr>
</tbody>
</table>
Privacy-preserving SVM for vertically partitioned data

Members: 5 UCs, National VA (VINCI resource), 3 Federally Qualified Health Systems in LA (USC), RAND
Cohorts: Kawasaki Disease, Congestive Heart Failure, Obesity

Funded by PCORI as part of PCORnet
Privacy-preserving GWAS study on horizontally partitioned data of Kawasaki disease
Models for ‘Sharing’ Data

iDASH On-Demand Resources

On-demand
Virtualized
Elastic
Resilient
Compute
And
Storage
Technology

**On-demand Resources**

- **Data**
- **Tools**
- **Recipes**

**Automated**
- Compute nodes
- Memory
- Disk storage
- Networking

**Powered by VMware**

- Compute request, direct upload & download of proprietary data, tool, recipe
- Upload & download data

**Middleware and HIPAA security developed by iDASH**

- Safe
- HIPAA-compliant
- Annotated
- Data deposit box
- Environment

**HIPAA and non-public data**

**Public data, tools, recipes**

Supported by the NIH Grant U54 HL108460 to the University of California, San Diego
iDASH On-Demand Resources

- 3 computation tiers
- 3 storage tiers
- 10GbE throughout
- Full redundancy
- RSA Two Factor Auth.
- Remote data replication

More planned:
- 600+ cores
- 4TB+ RAM
- 200TB+ storage

800+ cores
7TB+ RAM
600TB+ storage
Repeateable Results

Workflow
- Short reads
- Index reference
- Align to reference
- Call variants
- Annotate variants
- Pick high impact
- Deleterious SNPs

Blueprint
- Reference DB
- Test data
- Configuration
- Helper tools
- OS

Context

Workflows
- Short reads
- Index reference
- Align to reference
- Call variants
- Annotate variants
- Pick high impact
- Deleterious SNPs

Bookshelf

iDASH On-Demand Resources

4/20/2015
Supported by the NIH Grant U54 HL108460 to the University of California, San Diego
Repeatable Results

Workflow
- Short reads
- Index reference
- Align to reference
- Call variants
- Annotate variants
- Pick high impact
- Deleterious SNPs

Blueprint
- Reference DB
- Test data
- Configuration
- Helper tools
- OS

Context

Bookshelf

MyDATA

iDASH On-Demand Resources

Repeatable Results
iDASH’15 Privacy Protection Challenge

- Focus on secure outsourcing and secure data analysis in a distributed setting ([humangenomeprivacy.org](http://humangenomeprivacy.org))

**IDASH Privacy & Security Workshop 2015**

Secure Genome Analysis Competition

- Task 1: Homomorphic encryption (HME) based secure genomic data analysis
- Task 2: Secure comparison between genomic data in a distributed setting

Enter the Competition

- See Media and news coverage in GenomeWeb
- CADPP15 is supported by iDASH U54HL108460, iDASH linked R01HG007078 and NHGRI K99HG008175
Models for ‘Sharing’ Data

‘De-Identification’ (microdata release)

<table>
<thead>
<tr>
<th>Name</th>
<th>Age</th>
<th>Education</th>
<th>Hours/week</th>
<th>...</th>
<th>HTN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frank</td>
<td>42</td>
<td>6</td>
<td>40</td>
<td>...</td>
<td>Y</td>
</tr>
<tr>
<td>Bob</td>
<td>31</td>
<td>10</td>
<td>60</td>
<td>...</td>
<td>Y</td>
</tr>
<tr>
<td>Dave</td>
<td>43</td>
<td>9</td>
<td>40</td>
<td>...</td>
<td>N</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>

- HIPAA compliant methods
  - Safe harbor dataset
    - Removal of 18 safe harbor identifiers
  - Limited dataset
    - Removal of direct identifiers
  - Statistical methods
    - Removal/grouping of attributes
    - “Risk reasonably low”
- Re-identification and disclosure risks

Courtesy of Li Xiong
Statistical Data Release (macrodata release)

<table>
<thead>
<tr>
<th>Name</th>
<th>Age</th>
<th>HIV+</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frank</td>
<td>42</td>
<td>Y</td>
</tr>
<tr>
<td>Bob</td>
<td>31</td>
<td>Y</td>
</tr>
<tr>
<td>Mary</td>
<td>28</td>
<td>Y</td>
</tr>
<tr>
<td>Dave</td>
<td>43</td>
<td>N</td>
</tr>
</tbody>
</table>

# of HIV+ patients

Mohammed N. Privacy Preserving Heterogeneous Health Data Sharing. J Am Med Inform Assoc 2013

Courtesy of Li Xiong
Statistical Data Release: Disclosure Risk

Original records

Original histogram

[Table]

<table>
<thead>
<tr>
<th>Name</th>
<th>Age</th>
<th>HIV+</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frank</td>
<td>42</td>
<td>Y</td>
</tr>
<tr>
<td>Bob</td>
<td>31</td>
<td>Y</td>
</tr>
<tr>
<td>Mary</td>
<td>28</td>
<td>Y</td>
</tr>
<tr>
<td>Dave</td>
<td>43</td>
<td>N</td>
</tr>
</tbody>
</table>

[Diagram]

# of HIV+ patients

[Histogram]
Statistical Data Release: Differential Privacy

<table>
<thead>
<tr>
<th>Name</th>
<th>Age</th>
<th>HIV+</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frank</td>
<td>42</td>
<td>Y</td>
</tr>
<tr>
<td>Bob</td>
<td>31</td>
<td>Y</td>
</tr>
<tr>
<td>Mary</td>
<td>28</td>
<td>Y</td>
</tr>
<tr>
<td>Dave</td>
<td>43</td>
<td>N</td>
</tr>
</tbody>
</table>

Original records

<table>
<thead>
<tr>
<th>Name</th>
<th>Age</th>
<th>HIV+</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alice</td>
<td>43</td>
<td>Y</td>
</tr>
<tr>
<td>Frank</td>
<td>42</td>
<td>Y</td>
</tr>
<tr>
<td>Bob</td>
<td>31</td>
<td>Y</td>
</tr>
<tr>
<td>Mary</td>
<td>28</td>
<td>Y</td>
</tr>
<tr>
<td>Dave</td>
<td>43</td>
<td>N</td>
</tr>
</tbody>
</table>

Original histogram

Perturbed histogram with differential privacy

3 if we add Alice

Courtesy of Li Xiong
Differential Privacy (Dwork et al)

A privacy mechanism $A$ gives $\epsilon$-differential privacy if for all neighbouring databases $D, D'$, and for any possible output $S \in \text{Range}(A)$,

$$\Pr[A(D) = S] \leq \exp(\epsilon) \times \Pr[A(D') = S]$$

• $D$ and $D'$ are neighboring databases if they differ on at most one record

Courtesy of Li Xiong
Laplace Mechanism

For example, for a single count query $Q$ over a dataset $D$, returning $Q(D) + \text{Laplace}(1/\epsilon)$ gives $\epsilon$-differential privacy.
Evaluate solutions of **guaranteed privacy protection** for protecting the **output** of genomic data analysis

- Task 1: Privacy-preserving SNP Data Sharing
- Task 2: Privacy-preserving release of top K most significant SNPs
Which statistics to compute?

How to apportion privacy budget?

Statistical Data Release under Differential Privacy

Released Statistics/Synthetic Records

Original Data

Data Privacy Interface

Courtesy of Li Xiong
SHARE: Statistical & Synthetic Health Information Release

Data challenges:
- High dimensionality
- Self-correlation

Target applications:
- Cross-sectional studies
- Longitudinal studies

Original Health Records → DP Interface → Low Dimensional Histograms (DPCube, ...)
→ High Dimensional Distributions (DPCopula)
→ Sequential Data Release (DPTrie)

Released Statistics/Synthetic Records

Low Dimensional Histograms (DPCube, ...)
High Dimensional Distributions (DPCopula)
Sequential Data Release (DPTrie)

Courtesy of Li Xiong
Models for ‘Sharing’ Data

Sharing

Consent Management System
Do I wish to disclose data D to U?

Sharing Look-up
I can check that U looked at my data D

Yes

Trusted broker
- Data use agreements
- Study registry

User U requests Data D on individual I

Healthcare Institutions

Patient Interface

Patient I
Some preliminary findings from a limited set of interviews

- Healthy volunteers do not want to share with commercially sponsored researchers
- Some want their medical information shared with only UCSD researchers, no others
- Many do not want to share at least 1 category of sensitive information
- Most common decline was genetic, followed by sexual & reproductive health
<table>
<thead>
<tr>
<th>What clinical data am I sharing?</th>
<th>Who can access the clinical data I share?</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>UCSD and VA Hospital</td>
</tr>
<tr>
<td>Demographics</td>
<td>Non-profit Organizations</td>
</tr>
<tr>
<td>Socioeconomic Information</td>
<td>For-profit Organizations</td>
</tr>
<tr>
<td>Sexuality</td>
<td>✓</td>
</tr>
<tr>
<td>Past Pregnancy</td>
<td>✓</td>
</tr>
<tr>
<td>Anthropometrics</td>
<td>✓</td>
</tr>
<tr>
<td>Vital Signs</td>
<td>✓</td>
</tr>
<tr>
<td>Current or Previous Disease or Condition</td>
<td>✓</td>
</tr>
<tr>
<td>Family's Current or Previous Disease or Condition</td>
<td>✓</td>
</tr>
<tr>
<td>Laboratory and Test Results</td>
<td>✓</td>
</tr>
<tr>
<td>Therapy or Treatment Procedures</td>
<td>✓</td>
</tr>
<tr>
<td>Medications</td>
<td>✓</td>
</tr>
<tr>
<td>Social History</td>
<td>✓</td>
</tr>
<tr>
<td>Health Care Encounter</td>
<td>✓</td>
</tr>
<tr>
<td>Tissue and Blood Sample Usage</td>
<td>✓</td>
</tr>
</tbody>
</table>
informed CONsent for Clinical data Use for Research
Which DNA variants are implicated in KD susceptibility in this population?

Emory, Genome Institute of Singapore, Imperial College

Do patients understand what they consented for?

San Diego State University

Privacy Preserving Analytics for KD in African-Americans

Consent for Data and Biosample Sharing in an Underserved Population

Partnership for Epidemiological Research on Latinos

Data and Biospecimen Sharing Privacy Preserving Computation

Does consent rate depend on who is obtaining the consent?

Maricopa Health System
Looking into the Near Future

Healthcare data and Biospecimen collections

Several terabytes of data are being generated every day and samples are being collected. Participants could decide what/when/with whom to share

- Genomes
- Images
- Body sensors
- Environmental sensors
- Electronic Health Records
- Personal Health Records
- Pharmacy
- Social media
Department of Biomedical Informatics at UCSD, funded by NIH (U54, UL1, U24, UH3, R21, U01, T15, R00, K22, K99, D43), AHRq UCBRAID/OP, PCORI, NVIDIA