biomedical and healthCAre
Data Discovery Index Ecosystem

Lucila Ohno-Machado
University of California San Diego

Research Data Alliance, San Diego
BoF Biomedical Data
3/10/15
Core Team

- U Michigan
  - George Alter
- U Oxford
  - Susanna Assunta-Sansone
- U California San Diego
  - Maryann Martone
- U Texas Houston
  - Hua Xu

- NIH
  - Ian Fore
  - Jennie Larkin
  - Dawei Lin
  - Ron Margolis
  - Julia Puzak
  - Heidi Sofia
  - Alison Yao
Aims

1. Help users find accessible data

2. Assist data producers on how to publish data for maximal discoverability

3. Build a prototype to dock pilot products

PubMed of Data = DataMed
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”
Types of (Big) Data

- Genomes
- Transcriptomes
- Proteomes
- Metabolomes
- Microbiomes
- Images
- Physiologic signals
- Electronic Health Record
- Claims data
- Patient-reported data
  - Mobile health sensors
- Environment
  - GPS
  - Air quality
Short-Term Goals

- Develop a functioning prototype of a DDI

- Serve as a community hub, incubator and clearinghouse for examining issues, developing and testing innovative strategies, and establishing appropriate partnerships for creating a DDI
What about privacy?

- 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
Health Insurance Portability & Accountability Act

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

Presidential Commission for the Study of Bioethical Issues
- Privacy and progress in whole genome sequencing. [http://www.bioethics.gov](http://www.bioethics.gov) covers genome sequencing, exome sequencing, genome-wide SNV analysis, and data from large scale genomic studies

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 + ‘de-identified’ data
Breaches Affecting 500 or More Individuals

As required by section 13402(e)(4) of the HITECH Act, the Secretary must post a list of breaches of unsecured protected health information affecting 500 or more individuals. These breaches are now posted in a new, more accessible format that allows users to search and sort the posted breaches. Additionally, this new format includes brief summaries of the breach cases that OCR has investigated and closed, as well as the names of private practice providers who have reported breaches of unsecured protected health information to the Secretary. The following breaches have been reported to the Secretary:

Full Dataset CSV format (18 KB) XML format (57 KB)

Select a column head to sort by that column. Select again to reverse the sort order. Select an individual record to display it in full below the table.

Filter: 556 records showing

<table>
<thead>
<tr>
<th>Name of Covered Entity</th>
<th>State</th>
<th>Individuals Affected</th>
<th>Date of Breach</th>
<th>Type of Breach</th>
<th>Location of Breached Info</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st response Medical Transtop Corp.</td>
<td>MD</td>
<td>552</td>
<td>06/15/2012-10/01/2012</td>
<td>Unauthorized Access/Disclosure</td>
<td>Desktop</td>
</tr>
<tr>
<td>ABQ HealthPartners</td>
<td>NM</td>
<td>778</td>
<td>2012-12-20</td>
<td>Theft</td>
<td>Computer</td>
</tr>
<tr>
<td>Accendo</td>
<td>AZ</td>
<td>175,350</td>
<td>2011-01-01</td>
<td>Unauthorized Access/Disclosure</td>
<td>Paper</td>
</tr>
<tr>
<td>Access Medical Group</td>
<td>PR</td>
<td>7,606</td>
<td>2012-01-11</td>
<td>Theft</td>
<td>Laptop</td>
</tr>
</tbody>
</table>
Biometrics and PHI

- Biometrics are Protected Health Information (PHI)

- PHI requires HIPAA

- Biometrics require HIPAA
Genomes are Biometrics

Biometrics are Protected Health Information (PHI)

• Biometrics require HIPAA

PHI requires HIPAA
In 2.5 years we will have

- A free, user-friendly means for users to locate data sets of interest
  - NIH-funded projects (BD2k Centers of Excellence, Administrative Supplements)
  - Essential Repositories
- Standardized, searchable information (meta-data) about the contents of a data set
  - Built from existing meta-data in repositories and indices
  - Include meta-data about accessibility of the data
Data Discovery Index

- Engage end users
- Leverage existing efforts
- Operate in the ‘NIH commons’
- Will deliver a prototype
- We expect NIH to incorporate the DDI

- Funded by NIH U24AI117966
Thank you!

Questions?