Enabling Data Driven Research at UW Medicine

Sean D. Mooney, PhD
Professor, Department of Biomedical Informatics and Medical Education
Chief Research Information Officer, UW Medicine
Stanford recently created a Department of Data Science in their School of Medicine
Harvard recently created a Department of Biomedical Informatics in their Medical School
What is Digital Health?
Digital Health is the convergence of the digital and genomic revolutions with health, healthcare, living and society - wikipedia
Data is transforming both research about healthcare and its delivery

As the scope and breadth of project research data increases, secure management becomes increasingly important.

Where does disclosed clinical data that is used for research go?

How do we keep up with federal data management and security standards?

Tony Black
Director of Research IT

Reddit is a anonymous social network

Example
Reddit is a anonymous social network

- 6% of all Americans are on Reddit
- 32 Million Users/Month
- 1 Billion page views/Month
- Anonymous – users are behind pseudonyms
Reddit is a anonymous social network

We collected 1,700,000,000+ comments from 14,000,000+ Reddit pseudonyms
Here is the weight and gender distribution of a subset of 15,000+ Reddit users.
The weights are derived from parsing user comments in /r/loseit/ the weight loss subreddit
The genders are predicted with 96% accuracy
An inconvenient truth...

/r/gaming is the most associated subreddit to user weight
Biomedical Informatics faculty member Annie Chen employs social media analytics to derive insights into patient experience.
Using social media and combining text mining and visualization techniques, she can lead to insights about the user experience of cigarette smoking vs. e-cigarettes, which can in turn inform policy decisions.

Medical record data is increasing being stored in Electronic Health Records
Electronic Health Record use has increased dramatically over the past 15 years

http://www.cdc.gov/nchs/data/databriefs/db143.htm
Different clinics, hospitals and even departments often use their own systems.
Different EMR datasets are integrated into data warehouses (ideally) or secondary repositories.

Clinical Data Repository or Electronic Data Warehouse

ETL – Extract, Transform and Load
UW Medicine Clinical Data Repository

Adam Wilcox, PhD
Chief Analytics Officer

Peter Tarczy-Hornoch, MD
Chair, BIME
UW Medicine

UNIVERSITY OF WASHINGTON MEDICAL CENTER

HARBORVIEW MEDICAL CENTER

NORTHWEST HOSPITAL & MEDICAL CENTER

Hospitals

UW Medicine

Clinical Data Repository
4.3 million patients

>50 TB of data

>150 data streams
In order to compute with EHR data, it is coded.
The International Classification of Diseases (ICD) is likely the most used ontology in the world.
- Coded data
  - Patient
    - Demographics
    - Ongoing problems
  - Visit
    - Vitals
    - Diagnoses
    - Medications
    - Lab Tests
20% of the data is coded
80% of the data is unstructured
20% of the data is coded
80% of the data is unstructured

Unstructured data is contained in clinical notes (text) and images
Biomedical Informatics faculty member Meliha Yetisgen builds models to predict things about patients (called phenotypes) From unstructured data in the EHR
Did a patient have a phenotype?

**Task:** decide whether the patient *had* a specific phenotype (after the patient was discharged)

**Dataset:** all patient reports

**Application:** clinical research
Does the patient have a phenotype?

**Task:** decide whether the patient has a specific phenotype at timestamp \( t \)

**Dataset:** all reports from timestamp 0 to \( t \)

**Application:** phenotype surveillance
Will the patient have a phenotype?

**Task:** At timestamp $t$, decide whether the patient will have a specific phenotype in the future (e.g., at timestamp $t+1$, $t+2$).

**Dataset:** all reports from timestamp 0 to t

**Application:** clinical decision support
Mobile health devices are being increasingly used
Example: Surgical site infections are a burden to healthcare

- Bill Lober, MD Health Services
- Heather Evans, MD Surgery

- >300,000 seen annually
- 2-11 times risk of death
- 7-10 extra days in the hospital
- $20K per infection
- Est. annual cost $3-10 billion

[Sources: Anderson et al Infect Control Hosp Epi 2008; Zimlichman et al JAMA Intern Med 2013]

[Images: Scarred surgical site, links to data sources]
mPower is an app that can reduce complications by allowing surgeons to review surgical sites from pictures taken by patient smartphones.
Your Genome

A DNA sequence of 4 letters, A, T, C and G. It’s more than 3,000,000,000 letters long!
There are now more than 5,000 known rare diseases!

To be rare, a disease must be found in less than 1 in 2,000 people.

It is estimated that 6-8% of all Europeans have a rare disease, so even though each disease is unlikely, in aggregate, they are common!

There are so many rare diseases, that no individual doctor can have knowledge of them all.

80% of rare diseases have genetic origins.
These diseases are significant to human health

It is estimated, that 1 in 4 patients with rare diseases wait more than 30 years for the correct diagnosis.

Further, it is estimated that 1 in 6 underwent surgical treatment based on a wrong diagnosis.

And, it is estimated that 1 in 10 underwent psychological treatment based on a wrong diagnosis.

“Rare Diseases: Understanding this Public Health Priority”, Eurordis, November 2005 – www.eurordis.org
Hemophilia

Tay Sachs

Phenylketonuria

Androgen Insensitivity Syndrome

Osteogenesis Imperfecta

Cystic Fibrosis

Lysosomal Storage Disorders

Inherited cancers

Inherited Parkinsons

Sickle Cell Anemia
People have differences in their DNA sequence

This sequence – 3 billion A’s, T’s, G’s and C’s – has differences between individuals.

What are those differences?
People have differences in their DNA sequence

This sequence – 3 billion A’s, T’s, G’s and C’s – has differences between individuals.

What are those differences?

• Single position changes
• Small insertions or deletions
• Large insertions or deletions
• Simple repeats
• Rearrangements
Each of us differs by several million nucleotides

(3-4.5M)
We now know a lot of differences!

The national human genetic variation database (dbSNP) currently contains more than 100,000,000 sites on the genome that vary between individuals.

The HGMD contains 156,932 variants associated with human disease.

PharmGKB contains hundreds of variants associated with pharmacogenetic effects.

COSMIC now contains more than 2,000,000 somatic (spontaneous) variants in cancer.
Sequencing Cost Is Exceeding Moore’s Law

(Figure from Eric Lander and http://pew.org)
Receiver Operator Characteristics plot (ROC) measures prediction accuracy
We can predict disease causing variants on the genome with >80% accuracy!

How do we do it?

Unpublished data from Vikas Pejaner, Mooney and Radivojac lab
Does a mutation look more like neutral or disease?
We can also predict their mechanisms
With EHRs, research discoveries can now be integrated into the point of care.
Example: pharmacogenetics alerts in the EHR notify providers of risks of prescription response when genetic data is present.
Electronic Medical Records and Genomics (eMERGE)

N > 100,000

Genetic testing results returned to patients

EHR Integration

Clinical decision support

Process outcomes

Ethical, Legal and Social Implications

Prof. David Crosslin
BIME

© 2016 – Sean D. Mooney
How can you participate?

Community challenges solve problems faster and in more innovative ways than traditional project models.
The Personal Genome Project

George Church, Harvard Medical School.
The PGP Challenge

CAGI: Critical Assessment of Genome Interpretation

2015 is the 4th iteration of the CAGI: Personal Genome Project challenge

Challenge: What can we predict from real genomes and phenotypes blinded?

Dataset provided by Personal Genome Project and George Church, Harvard Medical School
Phenotypes provided by CAGI organizers in consultation with George Church
1. Genome
2. Binary Traits: Leukemia, Cavities etc. Y/N
3. Numerical Traits: age, triglyceride etc.
4. Categorical Traits: blood type, eye color, ancestry etc.
CAGI 2013
16 Submissions
We built a matching game in 2013

• Each genome contains variants covering the whole genome.

• Predicted profiles are assigned probability values.

• Each profile contains 239 binary traits.

• There are 214 decoy profiles that match no given genome data.
### PGP Challenge 2013

<table>
<thead>
<tr>
<th>Submission</th>
<th>Without Decoys and 23andMe</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Correct</td>
</tr>
<tr>
<td>S1</td>
<td>0</td>
</tr>
<tr>
<td>S2</td>
<td>0</td>
</tr>
<tr>
<td><strong>S3</strong></td>
<td><strong>6</strong></td>
</tr>
<tr>
<td>S4</td>
<td>4</td>
</tr>
<tr>
<td>S5</td>
<td>1</td>
</tr>
<tr>
<td>S6</td>
<td>0</td>
</tr>
<tr>
<td>S7</td>
<td>0</td>
</tr>
<tr>
<td><strong>S8</strong></td>
<td><strong>5</strong></td>
</tr>
<tr>
<td>S9</td>
<td>2</td>
</tr>
<tr>
<td>S10</td>
<td>0</td>
</tr>
<tr>
<td>S11</td>
<td>0</td>
</tr>
<tr>
<td>S12</td>
<td>5</td>
</tr>
<tr>
<td>S13</td>
<td>2</td>
</tr>
<tr>
<td>S14</td>
<td>1</td>
</tr>
<tr>
<td>S15</td>
<td>5</td>
</tr>
<tr>
<td>S16</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0</td>
</tr>
</tbody>
</table>
The federal government is creating a Precision Medicine Initiative.
A cohort of 1,000,000 volunteers who will be tracked over time to assess disease risks and treatment.
How do we invest in Digital Health at UW?
Columbia University

University of Rochester

Regenstrief Institute
Indiana University
We need to invest in Information Platforms

• For example, REDCap (electronic clinical data collection web portal) maintained by ITHS Research IT team (Bas de Veer)
  • Has grown from 500-3300 projects created in just over five years (~1000 in production)
We need to invest in **THIS** to enable innovative research translation.

Patient+Clinician Interaction

Data entered into an Electronic Health Record
Protected health information (PHI) is sensitive

FISMA and HIPAA have high standards for secure data management

Cloud is available but currently guidance is difficult to find
This is the most exciting time I have ever been in research because of advances in computing
Thanks!
sdmooney@uw.edu

I want to acknowledge everyone in Research IT, ITHS, ITS, my collaborators and the Mooney Research Group who made this talk possible

Research IT CIO James Fine, Tony Black, Jessica Yeung, Christie Fong, Nicholas Dobbins, Bas de Veer, Steve Senter, van Anandasakaran, Cole Bessee, Ron Shaker, Justin Prosser, Peter Tarczy-Hornoch, Meliha Yetisgen, Kari Stephens

Mooney Group Somying Promso (TCELS), Matt Mort (HGMD), Vidhya Krishnan, Kishore Kamati, Tal Oron, Yiqiang Zhao, Janita Thusberg, Judy Shi, Mat Fleisch, Tobias Wittkop, Biao Li, Darcy Davis, Artem Zykovich, Alex Alleovitch, Jackson Miller, Chet Seligman, Adrian Bivol, Gayathri Rajan, Julie Lin, Nikhil Gopal, Dae Lee, Binghuang Cai, Nikki Kiga, Abhi Pratap, Jonathan Moon, Don Smith, Tom Peterson

Collaborators Predrag Radivojac, Tom Peterson, Maricel Kann, Andrew Su, Chunlei Wu, Vikas Pejaver, David Cooper, Russ Altman, Teri Klein, Ann Oberg, many others