Data Driven Clinical Research: If Only It Were So Simple...

John H. Holmes, PhD
University of Pennsylvania Perelman School of Medicine
Institute for Biomedical Informatics
Center for Clinical Epidemiology and Biostatistics
Our agenda

• A biomedical informatics-inspired framework
• The health information ecosystem
• Challenges for data-driven research
• Opportunities at the interface of intelligent analysis and biomedical research
• Conclusion
A Biomedical Informatics-Inspired Framework
What is “biomedical informatics”?

http://www.amia.org/biomedical-informatics-core-competencies

The interdisciplinary field that studies and pursues the effective uses of biomedical data, information, and knowledge for scientific inquiry, problem solving and decision making...

...motivated by efforts to improve human health.
A Biomedical Informatics-Inspired Framework For Data-Driven Clinical Research

Providers Caring for Patients

Information, Decision-Support, and Order-Entry Systems

Electronic Health Records

Regional and National Public Health and Disease Registries

Creation of Protocols, Guidelines, and Educational Materials

Standards for Prevention and Treatment

Biomedical and Clinical Research
An Integrated View of Biomedical Informatics

- Computational Biology
- Translational Bioinformatics
- Public Health Informatics
- Consumer Health Informatics
- Clinical Research Informatics
- Clinical Informatics

Biomedical Informatics
But they (should) exist in an information ecosystem
The ecosystem of Next Generation Biomedical Informatics encompasses:

- Computational biology
- Bioinformatics
- Medical informatics
- Health information technology
- Computer science
- Biological, physical, clinical and social sciences, and more…
Conceptualizing the informatics ecosystem: 
The Biomedical Informatics Pipeline

Adapted from
http://www.amia.org/about-amia/science-informatics
The Biomedical Informatics Pipeline

• Translational Bioinformatics
• Clinical Informatics

Electronic Health Records

• Computational Biology
• Translational Bioinformatics
• Clinical Research Informatics

Biomedical and Clinical Research

Public Health and Disease Registries

Public Health Informatics

Standards for Prevention and Treatment

Clinical Informatics

• Clinical Informatics
• Consumer Health Informatics

Protocols, Guidelines, and Educational Materials

Clinical Informatics

Information, Decision-Support, and Order-Entry Systems

Providers and Patients
Every step in this pipeline produces data!
The Health Information Ecosystem

The Electronic Medical Record
Patient Portals
Wearable Technology and Telemedicine
The Internet of Things
Non-Health related data
The Electronic Medical Record

• There is rarely a single EMR system
  – Clinical
  – Laboratory
  – Radiology
  – Monitoring
  – ...

• These systems are often brought together in a clinical data warehouse
Patient Portals

• (Usually) web-based interface to allow patients to
  – Communicate with providers
  – Request appointments, drug refills, etc.
  – Post diaries
  – Engage in research studies
  – Obtain educational materials

• Rich source of data!!
Wearable technology and telemedicine

• In-home 24-hour monitoring
  – Oxygen saturation
  – Blood pressure
  – Blood glucose
  – Cardiac rhythms
  – ...

• Remote diagnosis
  – Radiology
  – Dermatology
  – ...

Internet of Things

• Linkages between wearable devices and...
  – Appliances
  – Data systems
  – Chains of monitoring and therapeutic devices
  – Other people’s wearable devices
Non-health related data

• Built environment

• Weather

• Financial systems

• News
Back to the Electronic Medical Record...

An Example
The Penn Research Data Warehouse

**Penn Data Store**
- Phenotype and Limited Genotype Data from Health System Operations

**Research Data**
- Clinical Trials Registries
- Bio-banks
- ‘Omics

**Precision Medicine Warehouse**
- Studies
- Outcomes
- Consents
- ‘Omics
- Samples
- Protocols
- Tumors
- Subjects / Patients
- Treatment

**Identified**
- Industry

**De-Identified**
- De-Identified Penn Data Store
- Search / Analytics "Cohort Explorer"

**Cohort Identification**
- ORACLE HEALTH SCIENCES
- TriNetx

**Penn Researchers and Clinicians**
- Discrete Data Only
## PennOomics Content

<table>
<thead>
<tr>
<th>Data</th>
<th>Start Date</th>
<th>Objects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patients</td>
<td>2005</td>
<td>3,071,933</td>
</tr>
<tr>
<td>Encounters</td>
<td>2008</td>
<td>93,469,891</td>
</tr>
<tr>
<td>Diagnosis</td>
<td>2008</td>
<td>52,073,170</td>
</tr>
<tr>
<td>Procedures</td>
<td>2008</td>
<td>14,831,535</td>
</tr>
<tr>
<td>Labs</td>
<td>2008</td>
<td>250,824,430</td>
</tr>
<tr>
<td>Microbiology</td>
<td>2008</td>
<td>9,864,271</td>
</tr>
<tr>
<td>Administered Medications</td>
<td>2008</td>
<td>61,884,739</td>
</tr>
<tr>
<td>Vital Signs</td>
<td>2011</td>
<td>83,840,084</td>
</tr>
<tr>
<td>Tumor Registry</td>
<td>2010</td>
<td>85,308 – Patients</td>
</tr>
<tr>
<td></td>
<td></td>
<td>375,032 – Stage/Grade</td>
</tr>
<tr>
<td>Studies</td>
<td>2001</td>
<td>1,003</td>
</tr>
<tr>
<td>CPD Patients</td>
<td>2013</td>
<td>3,921</td>
</tr>
<tr>
<td>Research Genomic Sequences</td>
<td>Varies</td>
<td>6,770</td>
</tr>
<tr>
<td>Bio-bank Samples</td>
<td>Varies</td>
<td>Eventually millions</td>
</tr>
</tbody>
</table>
But this is just one EMR system in one institution!

What if we need more data?
Distributed Research Networks

- Systems that allow secure remote analysis of separate data sets each derived from a different medical organization’s or health plan’s records
- Facilitate access to and use of clinical data for research
- Facilitate multi-site and interdisciplinary collaboration
A Generic DRN

Researcher → Portal

Researcher

Research question

Portal

Dataset

Query

Participating Sites

Clinical Site

Clinical Site

Clinical Site

...

Query results
Features of DRNs

- Distributed
- Scalable
- Client-server architecture
- User portal
- Confidentiality and privacy protection
- Centralized operational governance
- Often adhere to data warehouse paradigms
A “Virtual” Data Warehouse Approach

Researchers

DRN portal

Common data model and governance

“Virtual Data Warehouse”

Health Care Entity 1

Health Care Entity 2

Health Care Entity 3

... Health Care Entity n

Query

Site-specific Result

Aggregated Results
How are DRNs used?

- Activities preparatory to research
- Surveillance
- Comparative effectiveness research
- Administrative purposes
One example:
The Health Care Systems Research Network

http://www.hcsrn.org/en/

- **Nineteen healthcare organizations**
  - Integrated delivery systems that provide both health care and insurance coverage
  - >15 million people receiving health care
  - Comprehensive electronic medical records (EMR)

- **Formal research programs**
  - More than 400 researchers
  - ~1,500 active research projects
  - In-house survey research programs and research clinics
  - Standardized data sets
Challenges for Data-Driven Research

Data Types
Data Availability
Data Quality
Data Heterogeneity
Data Linkage
Challenge: Data Types

• Discrete
• Continuous
• Free-text
• Image
• Waveform
• Often all in the same record!
Challenge: Data Availability

• Regulatory issues
  – Confidentiality
  – Privacy

• Proprietariness issues
  – Data ownership

• Lack of awareness!
Challenge: Data Quality

• Usually, the data weren’t collected for you!

• Multiple users, multiple representations

• Missing data is the bane of our existence
Challenge: Data Heterogeneity

• Different sources will yield different data

• Syntactical heterogeneity

• Semantic heterogeneity

• One approach: Ontology
Challenge: Data Linkage

• Deterministic
  – Assumes a unique identifier

• Probabilistic
  – Assumes participating data elements are syntactically similar

• Regulatory issues
  – Privacy and confidentiality

• Linkage can be really, really hard…
A (perilous) example: Fatality Analysis Reporting System (FARS)

- Prospective surveillance database of all fatal vehicle accidents occurring in the US, territories, and possessions Rico since 1975


- Person-, vehicle-, and crash-level data
Objective of FARS

To provide an overall measure of highway safety, to help suggest solutions, and to help provide an objective basis to evaluate the effectiveness of motor vehicle safety standards and highway safety programs.
FARS inclusion criteria

• Crash must involve a motor vehicle traveling on a traffic-way customarily open to the public and result in the death of a person within 30 days of the crash

• Occupant of a vehicle or a non-occupant
Where do the data come from?

- Cooperative agreement with an agency in each state government to provide information in a standard format on fatal crashes in the state
  - Police and/or fire departments
  - Other first responder agencies
  - State public health agencies
The FARS data model

Crash → Vehicle → Person
Some characteristics of FARS

• Denormalized
  – The Person File contains pertinent data from the Vehicle and Crash Files
• Large
  – >100,000 person records
  – >100 variables
• Unbalanced
  – 38,000 to 42,000 deaths, depending on year
• Many missing values
  – Bicycles don’t have airbags!
• Some variables are continuous
  – Require discretization for some methods
• Interactions
  – Passenger airbag deployment vs. year of vehicle
• Prospective
  – Even within a given year, new patterns emerge over time
Some possible research questions emanating from FARS

• What crash characteristics are associated with fatality in a fatal crash?

• What environmental characteristics are associated with fatality in a fatal crash?

• What demographic characteristics are associated with fatality in a fatal crash?

• Do certain vehicle types or makes/models contribute more fatality in a fatal crash?

• Many, many others…
An example

What types of fatal head injury are found in fatal crashes?

• Problem 1: FARS doesn’t contain ICD codes!

• Problem 2: Death-related datasets don’t contain the depth and breadth of FARS!

• Solution?
  – Record linkage
Two data sources

- **FARS (for the crash data)**
- **Death-related data**
  - National Death Index
    - Advantage:
      - Census of all deaths occurring in US
    - Disadvantage:
      - Linkage requires any two of DOB, SSN, or Name)
  - Multiple Causes of Death File
    - Advantage:
      - Census of all deaths occurring in US
      - Linkage isn’t as restrictive as NDI
    - Disadvantage:
      - Linkage must be accomplished using probabilistic methods
In order to do the linkage...

<table>
<thead>
<tr>
<th>FARS</th>
<th>MCOD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>Age</td>
</tr>
<tr>
<td>Sex</td>
<td>Sex</td>
</tr>
<tr>
<td>Race</td>
<td>Race</td>
</tr>
<tr>
<td>Hispanic ethnicity</td>
<td>Hispanic ethnicity</td>
</tr>
<tr>
<td>County of crash</td>
<td>County of death</td>
</tr>
<tr>
<td>State of crash</td>
<td>State of death</td>
</tr>
<tr>
<td>Date of death</td>
<td>Date of death</td>
</tr>
</tbody>
</table>
But, there are problems!

- Date of death is coded differently
  - FARS is the usual MM/DD/YYYY
  - MCOD uses only the day of the week within month and year, not the actual date

- Geographic location coding systems are different
  - Requires construction of a “crosswalk” file
FARS-MCOD Linkage
Constructing the Crosswalk Table

Join based on exact state name match and Soundex match on county name

FIPS (abbreviated list)
- FIPS_State
- FIPS_Place
- State_Alpha
- Place_Name
- FIPS_County
- Name_Of_County
- Zip_Code
- GSA_Code

GSA
- State_Code
- County_Code
- City_Code
- State_Name
- County_Name
- City_Name
- Identifier

Crosswalk (GSA+FIPS)
- GSASateCode
- GSACountyCode
- FIPSStateCode
- FIPSCountyCode
- StateName
- CountyName
But alas, it’s illegal to link these!

Here’s how we handled it...
There was an existing linked DB

- Created by NHTSA
- Managed by NCHS
- Requirements
  - Permission from the jurisdictions to use DCs
    - Letter of support from NAPHSIS
    - Jurisdictional IRB approval
  - Data available for use only at NCHS Research Data Center
    - No local datasets!
    - And you think the TSA is demanding…
Opportunities at the interface of intelligent analysis and biomedical research

New Types and Sources of Data
Opportunities for Intelligent Data Analysis
Opportunities for Data Science
Electronic Phenotyping
Opportunity: New types and sources of data

- Environmental data
  - Maps
  - Terrains
  - Photographs
- Transportation data
- Retail purchasing data
- Social media
Some thorny regulatory issues confronting social media researchers

- Consent!
- Protecting user privacy and confidentiality
  - Need for anonymization poses computational difficulties
- Complying with provider requirements
  - Terms of Service and Data Use Agreements
- Serendipity
  - What to do when a serious event is uncovered, outside the scope of an approved protocol?
Opportunity:
Need for intelligent data analytics

• Machine learning
• Statistical classification
• Distributed analytics
• Intelligent visualization
• Naturally-inspired methods
Opportunity: Expansion of the Data Sciences

• Big Data initiatives
• Incorporation of intelligent analytics
• Hybridization of
  – Statistical
  – AI/ML
  – Visualization
  – Human expertise
Bringing all of this together...
Electronic Phenotyping for Cohort Identification

Leveraging the EHR for clinical research through advanced intelligent analytics
What is phenotyping?

The characterization of an individual’s condition based on data.
And electronic phenotyping?

Using the EHR and any other machine-readable data to characterize a patient’s condition
Types of data used in electronic phenotyping

- Molecular and genomic data
- Images
- Structured clinical data
- Unstructured clinical data
- Self-reported data
- Environmental data
- ...

Thus, highly heterogeneous, potentially difficult to obtain, and demanding of novel data management and analytic strategies.
A Case Study: Hyperlipidemia

• High levels of cholesterol and/or triglycerides in the blood

• Numerous types
  – Hypercholesterolemia
  – Hypertriglycerideridemia
  – ....

• Potentially serious condition

• Mediated by numerous factors from genes to environment
Why would we do electronic phenotyping for this?

• We want to identify and characterize accurately a cohort of patients with hyperlipidemia

• We don’t have genotypes of every patient
  – We will need to genotype those who aren’t typed

• We want to use the phenotype data in conjunction with genotype data to completely characterize hypercholesterolemic patients
  – Early biomarkers of disease
  – Possible targets for treatment
A life-cycle approach to electronic phenotyping

1. Data Preparation

- Information extraction
  - Query development and application
- Ontology
  - Table and variable definition
  - Variable standardization
- Data integrity
  - Range and logic checks
  - Missing value characterization
- Imputation (yes or no?)
- Variable transformation
- Identification of spurious relationships
  - Association rule mining
  - Visual data checks
A life-cycle approach to electronic phenotyping

2. Dimensionality Reduction

- Feature selection
  - Decision trees
  - Correlational feature selection
  - Stepwise methods
  - Metaheuristic methods
- Structure learning
  - Directed Acyclic Graphs
  - Bayesian networks
- Expert input and validation!!!!
A life-cycle approach to electronic phenotyping

3. Phenotype discovery: Structured data

• Goal: Cohort identification for further analysis
  – Need to identify the candidate phenotype patterns
  – We know the class: Hyperlipidemia

• Supervised methods
  – Association rule mining
  – Statistical classifiers
  – SVMs, HMMs, Naïve Bayes

• Unsupervised methods
  – Clustering (k-means, COBWEB, and E-M)
A life-cycle approach to electronic phenotyping

3. Phenotype discovery: Text

- Clinical notes and reports
- Patient self-reports
- Social media
- Methods
  - Information extraction
  - Keyword retrieval
  - Latent semantic indexing
  - Sentiment Analysis
A life-cycle approach to electronic phenotyping
3. Phenotype discovery: Images

- Pathologic images
- Radiologic images
- Photographs
- Methods
  - Feature definition and assignment (tagging)
  - Feature extraction
  - Clustering methods
  - Expert validation
A life-cycle approach to electronic phenotyping

3. Phenotype discovery: Environmental data

- Wearable devices
- Neighborhood data
  - Availability of food markets, sidewalks, and recreational activities
- Housing characteristics
- Weather data
- Many other sources…
A life-cycle approach to electronic phenotyping

4. Evaluation

• **Metrics**
  – Sensitivity, specificity, predictive values, F-value
  – Receiver operating characteristic curves
    • Visual
    • c-statistic (area under the curve)

• **Expert evaluation**
  – Critical to evaluation!
  – Panel of 3-10 experts in the clinical domain
Challenges encountered in electronic phenotyping

- Availability of data
- Data used for phenotyping are often temporal
- Syntactic and semantic interoperability
- Privacy and confidentiality
- Phenotype interpretation
The future of electronic phenotyping

- Integration of more types of data sources
- Real-time surveillance and cohort definition
- Better integration of NLP
- Decision support for clinicians
  - “Patients like you” and “Patients like mine”
- Decision support for researchers
  - Intelligent cohort identification processes and workflows
Conclusion

• The health information ecosystem is complex
  – Numerous sources of data
  – Much noisy data
  – Highly temporal

• Many challenges confront the AI researcher

• Many opportunities too!
Questions?
Online Discussion Forums as Potential Sources of Adverse Drug Event Data
Some preliminary considerations

• What is a discussion forum?
  – Online resource where users can participate in “conversations” on a particular topic

• Usually topical

• Often used for seeking advice, reporting phenomena, and offering support
<table>
<thead>
<tr>
<th></th>
<th>Discussion Forum</th>
<th>Facebook</th>
<th>Twitter</th>
<th>Chat</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topical?</td>
<td>Yes</td>
<td>Not always</td>
<td>Not always</td>
<td>Often</td>
</tr>
<tr>
<td>Can elucidate conversation?</td>
<td>Yes</td>
<td>Often</td>
<td>Not in sampled data</td>
<td>Yes</td>
</tr>
<tr>
<td>Can elucidate social network?</td>
<td>Difficult</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Mode</td>
<td>Asynchronous</td>
<td>Asynchronous</td>
<td>A/synchronous</td>
<td>Synchronous</td>
</tr>
<tr>
<td>Privacy concerns?</td>
<td>++++</td>
<td>++++</td>
<td>++++</td>
<td>++++</td>
</tr>
<tr>
<td>Data available?</td>
<td>+++</td>
<td>+/-</td>
<td>++++</td>
<td>+/-</td>
</tr>
<tr>
<td>Syntax issues</td>
<td>+++</td>
<td>++</td>
<td>++++</td>
<td>+++</td>
</tr>
<tr>
<td>Focus of social media research</td>
<td>+/-</td>
<td>+++</td>
<td>++++</td>
<td>+++</td>
</tr>
</tbody>
</table>
Why study health-related discussion fora for adverse drug events?

- Which drugs and supplements people use
- Therapeutic effects experienced by posters or others
- Side effects posters or others have experienced
- Attitudes, beliefs, and sentiments
- Identification of possible sentinel events
- Conversational platform
Threats to discussion fora research

- Restrictive terms of service policies
- User aliasing (“sock puppets”)
- Thread drift
- Discussion continuing across threads
- Syntax
  - 😊 😞 😁
  - Srly, bbl, IMHO
  - “tammy” (synonym for Tamiflu… or someone’s name!)
How do we do research on discussion fora?
Goal: Understand the frequency and content of side effects and associated adherence behaviors discussed by breast cancer patients related to using aromatase inhibitors (AIs), with particular emphasis on AI-related arthralgia.

Methods

• Mixed methods study to examine content related to AI associated side effects posted by individuals on 12 message boards between 2002 and 2010

• Determined frequency of and association between side effects and AIs and identified themes using content analysis

• 1,000 randomly selected messages related to arthralgia were coded by two independent raters
The only SEs i had on Tamoxifen were weight gain and hot flashes/night sweats. Has anybody been suffering from chapped, cracked lips, especially at the corners of your mouth, since being on Arimidex? I had this problem many years ago but always could fix it by taking vitamin B complex. The vitamin B complex isn’t helping any more and I am wondering if Arimidex is causing this. My eye doctor says my dry eyes are probably from age --- but he thinks Arimidex made them worse and sent a letter to my DR. saying so.

2/07 Moved to Aromasin & Zometa because of Arimidex triggering RA & Lupus. I've also had conjunctivitis 4 times. My onc said this wasn't a SE of arimidex but I know it is!

I'm now on arimidex. I am doing fine with BC but have developed fibromyalgia. After about 1.5 years on Aromasin, my cholesterol which is normally 186 was at 254. Something I have not noticed anyone talking about is high cholesterol. oh, should add.....rather significant impact....of aromasin - was vaginal dryness.

I am on aromasin and have developed dry eyes. does anyone else have this problem?

slammed back with flushings, terrible night sweats, bad mood swings, now have bad joint pain(mostly in my ankle and right hand/thumb). oh ugh with the femara! Is there a side effect, that is B9, from Femara that causes vaginal discharge? I just got home from a cruise to AK and upon wake up this a.m. [sic] I had a discharge. Not real bloody but kind of like the very last day of a light period.
Results

• Among 25,256 posts related to AIs, 4,589 (18.2%) mentioned at least one side effect.
  – Joint/musculoskeletal pain (N = 5093)
  – Hot flashes (1498)
  – Osteoporosis (719)
  – Weight gain (429)
• 12.8% mentioned discontinuing AIs, while another 28.1% mentioned switching AIs
• Many also offered support and advice for coping with AI-associated arthralgia
What can we glean from this?

• Patients taking AIs seem to experience side effects more frequently than noted on the labels.
• There is no denominator: interpret cautiously!
• However, these results are suggestive of important phenomena.
• Validation is needed!