Is Big Data ready to improve outcomes or is it a new generation of garbage in/garbage out?

Yves A. Lussier, M.D., Fellow ACMI

Professor of Medicine, Assoc. Vice-President for Health Sciences & Chief Knowledge Officer Director for Precision Health and Cancer Informatics, (Cancer Ctr) Associate Director, BIO5 Institute University of Arizona

Fellow, IGSB, The University of Chicago & Argonne Laboratory

Lead Investigator, Beagle Supercomputing, CI, Un of Chicago & Argonne Lab

Chair, Grant Review Committee, NIH National Library of Medicine
No conflicts of interest to declare
Take Home Points

• Escalating the value chain
  – Data ➔ Information ➔ Evidence/knowledge

• Big Data
  – 5 Vs: Value>Veracity>Volume>Variety>Velocity
  – Information value is key
    • Messy clinical/history data requires transform to information
    • Big informative data: imaging / genomic

“Computational precision medicine: Data science for healing humanity - one person at a time”

- Lussier Group
Plan

- Definitions
- Challenges of BIG DATA
  - Flexible models of data representation and exchange
  - Reductionist vs. systems-level science
- Opportunities / Paradigm Shift
  - Drug Repurposing
  - Precision Medicine
  - Learning Health System
- Discussion
Big Data: classically defined by 3 “V’s”

- Volume
- Velocity
- Variability

Big Health Data: 2 additional “V’s”

- Value
- Veracity

Modified from Philip Payne Wash U
But Reasoning on Big Data Is Hard…

Unexpected problems
• Algorithms behave differently
• Applicability of convention metrics
  • P-values don’t mean a lot in peta-byte scale data sets
• Signal vs. noise
  • Detection
  • Understanding of patterns

Physical computing
• Data storage
• Computational performance

Modified from Philip Payne Wash U
The Role of Data Science: Generating Information and Knowledge

Data + Context

Information + Application

Knowledge

Modified from Philip Payne Wash U
Core Platforms Supporting Virtual Organizations

- Data Sharing Infrastructure
- Knowledge Management Tools
- Knowledge-Anchored Applications

Modified from Philip Payne Wash U
Readiness of technology infrastructures for Data Science

- Distributed Data & Knowledge
- Syntactic & Semantic Interoperability
- Security & Regulatory Frameworks
- Socio-technical Factors

Modified from Philip Payne Wash U
Exemplar Value Proposition for Data Science: Software-Oriented Architecture Approaches to Data Federation

• **Reduced need to replicate data**
  - Data “lives” where it is initially generated or stored
  - Lowers infrastructure costs

• **Increased ability for data stewards to oversee access**
  - Fine-grained and policy-based access control
  - User-centered locus of control

• **“Elasticity”**
  - Ability to expand or contract resources based on current needs (e.g., plug and play)

• **Adaptability**
  - Platform-independent design allows for rapid evolution
The "Omic" Funnel

Raw "Omics" Data

Information

Knowledge

Action

High-Throughput Sequence Data, Methylation, Tissue Array, Tertiary Structure, etc.

SNPs, Network Activation, Indels, CNVs, Rearrangements, etc.

Filter for Actionable Clinical Significance

Clinically Relevant "Omic" Findings

EHR Integration

Personalized Health Care

Patient Specific Clinical and Environmental Data

Scientific Literature

National DB of Clinically Significant Variants

National DB of Omic CDS Rules

ACMI Meeting. 2014 (EMERGE – J. Starren)
Sources & Dimensions of Health Data

- **Physiological Scale of Measure**
  - Physical scale
  - Time/Space
  - 2D-3D Space

- **Clinical Data**
  - Symptoms
  - Clinical Signs
  - Neurophysiologic Tests
  - Radiology Reports
  - Pathology Reports
  - Clinical Biochemistry
  - ‘Omics

- **Technology & Monitoring**
  - Cell Phone GPS
  - fMRI, EEG, ECG
  - Scintigraphy
  - Wearable, ICU & "Holter" monitors
  - Bacteriologic Profile, Antibiogram
  - O2 Monitors, Continuous Glucose Monitor

- **4D**
  - Geospatial pollutant & meteo monitoring, social media
Current Repositories & Warehouses

- EMR
- PACS
- LIMS

PHYSICAL SCALE OF MEASURE

- Scales of biology
- 4D (Geospatial pollutant & meteo monitoring, social media)

Symptoms
Clinical signs
Neurophysiologic tests
Wearable, ICU & "Holter" monitors
Bacterial profile, antibiotic
O2 monitors, continuous glucose monitor

Electron microscope
Brain "omics"

THE UNIVERSITY OF ARIZONA
Center for Biomedical Informatics & Biostatistics

SLUSSIER GROUP
Building an Argument for Translational Data Science: Current Trends

Learning Healthcare Systems
- Instrumenting the clinical environment
- Generating hypotheses
- Creating a culture of science and innovation

Precision Medicine
- Rapid evidence generation cycle(s)
- ‘omics’
- Analytics/decision support

Big Data
- System-level thinking
- Data science

Integrated and High Performing Healthcare Research and Delivery Systems

Learning from every patient encounter
- Leveraging the best science to improve care
- Identifying and solving complex problems

Modified from Philip Payne Wash U
Paradigm Shift: beyond opposing views of data analysis 1/2

Reductionism
- Components
- Time
- Space
- Context

Systems Science
- Interrelationships, Dynamics

Medical Treatments
- Disease-driven
- Aimed for normalcy (normal range)
- Additive

Medical Treatments
- Individualized
- Multidimensional use of drugs
- Time-sensitive
- Space-sensitive
- Synergistic

http://journals.plos.org/plosmedicine/article?id=info:doi/10.1371/journal.pmed.0030209
Paradigm Shift: beyond opposing views of data analysis 2/2

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Reductionism</th>
<th>Systems-Oriented Perspective</th>
</tr>
</thead>
<tbody>
<tr>
<td>Optimal</td>
<td>Conditions where one or few components are responsible for the overall behavior of the system</td>
<td>Conditions where interactions between components are responsible for the overall behavior of the system</td>
</tr>
<tr>
<td>Disease types</td>
<td>Acute, simple diseases</td>
<td>Chronic, complex diseases</td>
</tr>
<tr>
<td>Examples</td>
<td>Urinary tract infection</td>
<td>Diabetes</td>
</tr>
<tr>
<td></td>
<td>Appendicitis</td>
<td>Coronary artery disease</td>
</tr>
<tr>
<td></td>
<td>Aortic aneurysm</td>
<td>Asthma</td>
</tr>
<tr>
<td>Theoretical limitations</td>
<td>Disregards component–component interactions and dynamics</td>
<td>Costly in resources and time</td>
</tr>
</tbody>
</table>

DOI: 10.1371/journal.pmed.0030209.t001

http://journals.plos.org/plosmedicine/article?id=info:doi/10.1371/journal.pmed.0030209
**WHAT IS IT?**

**Precision medicine** is an emerging approach for disease prevention and treatment that takes into account people’s individual variations in genes, environment, and lifestyle.

The Precision Medicine Initiative® will generate the scientific evidence needed to **move the concept of precision medicine into clinical practice.**

**WHY NOW?**
WHY NOW?

The **time is right** because of:

- Sequencing of the human genome
- Improved technologies for biomedical analysis
- New tools for using large datasets
NEAR-TERM GOALS

Intensify efforts to apply precision medicine to **cancer**.

- Innovative **clinical trials** of targeted drugs for adult, pediatric cancers
- Use of **combination therapies**
- Knowledge to overcome **drug resistance**
LONGER-TERM GOALS

Create a research cohort of > 1 million American volunteers who will share genetic data, biological samples, and diet/lifestyle information, all linked to their electronic health records if they choose.

Pioneer a new model for doing science that emphasizes engaged participants, responsible data sharing, and privacy protection.
Precision Medicine Initiative Summit
White House, President Obama, 2/25/2016

• White House Announces UA's Involvement in National Precision Medicine Initiative
  https://uanews.arizona.edu/story/white-house-announces-ua-s-involvement-in-national-precision-medicine-initiative

• As part of its statewide programs, UAHS is launching new precision medicine initiatives:
  – Expand the clinical utility of its open-source, patient-centric analytic methods to aid physicians in interpreting the dynamic disease-associated gene expression changes arising from patients’ own DNA blueprint.
  – System-wide dissemination of an on-demand "case-based reasoning" system that intelligently searches and analyzes entire databases of electronic medical records. This will give clinicians the power to develop an individualized and effective treatment plan for unusual or complex clinical conditions, grounded on practice-based evidence.
  – Development of genetic assays to predict an individual's response to therapy and prevention of adverse reactions, termed "pharmacogenomics".
  – Partnership with five other institutions to advance the Sanford Pediatric Genomics Consortium to help families and their providers improve health-care decision-making through better understanding and integration of genomic evidence.
Case-based reasoning as a case of learning health system

<table>
<thead>
<tr>
<th>Outcome or Risk Factor</th>
<th>Keywords Used to Conduct Expedited Electronic Search</th>
<th>Prevalence of Thrombosis no./total no. (%)</th>
<th>Relative Risk (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Outcome — thrombosis</td>
<td>“Thrombus,” “Thrombosis,” “Blood clot”</td>
<td>10/98 (10)</td>
<td>Not applicable</td>
</tr>
<tr>
<td>Thrombosis risk factor</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Heavy proteinuria (&gt;2.5 g per deciliter)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Present at any time</td>
<td>“Nephrosis,” “Nephrotic,” “Proteinuria”</td>
<td>8/36 (22)</td>
<td>7.8 (1.7–50)</td>
</tr>
<tr>
<td>Present &gt;60 days</td>
<td>“Urine protein”</td>
<td>7/23 (30)</td>
<td>14.7 (3.3–96)</td>
</tr>
<tr>
<td>Pancreatitis</td>
<td>“Pancreatitis,” “Lipase”</td>
<td>5/8 (63)</td>
<td>11.8 (3.8–27)</td>
</tr>
<tr>
<td>Antiphospholipid antibodies</td>
<td>“Aspirin”</td>
<td>6/51 (12)</td>
<td>1.0 (0.3–3.7)</td>
</tr>
</tbody>
</table>

*In all cases, the sentences surrounding the keywords were manually reviewed to determine their relevance to our patient. Pancreatitis was defined as an elevated lipase level (twice the upper limit of normal) coexisting with abdominal pain. We used the word “aspirin” as a proxy for antiphospholipid antibodies, since it is standard practice at our institution to give all patients with these antibodies aspirin; if “aspirin” was found in the chart, than antiphospholipid-antibody status was confirmed by investigating the laboratory results.*

Evidence-Based Medicine in the EMR Era
Jennifer Frankovich, M.D., Christopher A. Longhurst, M.D., and Scott M. Sutherland, M.D. N Engl J Med 2011; 365:1758-1759
Take Home Points

• Escalating the value chain
  – Data ➔ Information ➔ Evidence/knowledge

• Big Data
  – 5 Vs: **Value**>Veracity>Volume>Variety>Velocity
  – Information **value** is key
    • Messy clinical/history data requires transform to information
    • Big informative data: imaging / genomic

“Computational precision medicine: Data science for healing humanity - one person at a time”
  - Lussier Group