Big Data and the Research HARBOR (Helping Advance Research By Organizing Resources)

J. Kathleen Tracy, PhD
Director, Clinical and Translational Research Informatics Center (CTRIC)
Associate Director, Center for Health Informatics and Bioimaging (CHIB)
Universities, Innovation & Big Data

• Natural incubators for innovation
• Diverse disciplines
• Semi-common infrastructure
Big Data, Big Science

• Novel, creative approaches
• Knowledge integration across disparate fields
• Collaboration across multiple areas of expertise that cross major disciplinary, scientific, and administrative boundaries
## Opportunities and Challenges for Healthcare and Biomedical Research

<table>
<thead>
<tr>
<th>Opportunities</th>
<th>Challenges</th>
</tr>
</thead>
<tbody>
<tr>
<td>Greater access to more diverse data</td>
<td>The 4 V’s</td>
</tr>
<tr>
<td>Identification of previously unknown relationships (e.g., unintended effects of txs, differential response to txs for sub-groups of patients)</td>
<td>New methods</td>
</tr>
<tr>
<td>Identification of a greater array of targets for cost-savings</td>
<td>Analytic tools</td>
</tr>
<tr>
<td>Personalization of care and greater engagement with patients as partners</td>
<td>Protection of patient privacy</td>
</tr>
<tr>
<td>Building the next generation of scientists and care providers</td>
<td>Agility</td>
</tr>
</tbody>
</table>
Big Data: Opportunities and Challenges
What are some impacts of **Big Data**?

- After Haiti’s 2010 earthquake, Columbia University tracked the movements of 2 million refugees by the SIM cards in their cell phones and were able to determine where health risks would likely develop.
Our Unique Capacities

• Affiliation with UMMS
• Proximity and affiliation with Baltimore VA Hospital
• Diversity of patient populations
• Depth and breadth of scientific expertise of the University of Maryland System
The UM Research Harbor is an interactive, web-based platform that provides one-stop shopping for research support needs. Through a centralized hub, researchers and their staff at the University of Maryland can access our data warehouse, identify and access research support resources, tools and services, find experts, access regulatory support, learn about educational and training opportunities, and much more.

<table>
<thead>
<tr>
<th>Researchers Area</th>
<th>Community Area</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Warehouse</td>
<td>Community resources</td>
</tr>
<tr>
<td>Regulatory support (CICERO, WebRPR)</td>
<td>COMPASS Volunteer Registry</td>
</tr>
<tr>
<td>Research Support Services</td>
<td>COMPASS Community Partners Registry</td>
</tr>
<tr>
<td>Research Tools</td>
<td></td>
</tr>
<tr>
<td>Educational Resources</td>
<td></td>
</tr>
<tr>
<td>Specialty scientific “ports”</td>
<td></td>
</tr>
<tr>
<td>Personalized Medicine</td>
<td>SOM</td>
</tr>
<tr>
<td>-----------------------</td>
<td>--------------</td>
</tr>
<tr>
<td>BioBank</td>
<td>CICERO</td>
</tr>
<tr>
<td>Coeus</td>
<td>COMPASS</td>
</tr>
</tbody>
</table>

**Teleform**
**RedCap**
**OpenClinica**
**R-studio**
**NGS Exome, Genome, RNA-seq, ChIP-seq**

**Recruitment**
**Curated Catalog**
**Data Depot & Sharing**
**Secondary Sources**
**Research Compliance**
**Standardized Capture**
**Volunteer Registry**

**HARBOR**
HARBOR Data Warehouse

• Data catalog of data sources and related documentation for data that have been placed in Warehouse

• Provides ability to search for and compile identified (with IRB approval) and de-identified datasets from a variety of data sources
  – UMMS clinical data repository
  – Observational registries (e.g., cancer, anesthesiology, sports medicine)
  – COMPASS Volunteers
  – Geospatial data

• Provides tutorial information related to appropriate end use of data from the Warehouse
The purpose of the HARBOR is to expose data, connect collaborators, provide common tools/services, promote scientific innovation through use of common infrastructure.
Big Data, Research HARBOR & Innovation
Examples of *Big Data*

Healthcare insurance organizations collect billing, procedure and diagnostic code data from all providers in their networks on a daily, weekly, and monthly basis to assess costs, meet regulatory requirements, determining whether the organization qualifies for additional reimbursement based on more cost-effective standard of care. Insurance organizations also mine these data to better care for patients and contain costs.

- *Creating tailored reminders to female patients that have not received a Pap smear in the last 3 years.*

- *Creating an automated outbound calling system for girls age 9-26 who started the HPV vaccination series but did not complete all 3 doses.*
Big Data in Healthcare

• Rich information for health policy
  – How does spending differ by location? Patient mix?
  – What are the trends in disease prevalence?
  – What are the trends in treatment choices?
  – How do disease, treatments, outcomes, etc. vary from region to region?
    By gender? By type of insurance coverage? By provider?
  – Which providers are better/worse in quality and cost?

• Support for performance improvement
  – Transparent reporting of provider and payer results
  – Data can be used by providers to drive their QI efforts
Big Data in Healthcare (cont.)

• Leading causes of illness and hospitalization
• Rates of accidents, infections and cancer
• Geographic differences in incidence of diseases, such as diabetes or heart disease
• Ethnic, sex or socioeconomic variations in illness
• Most expensive diagnoses and procedures
• Role of prevention on illness and costs
• Unique genetic variations in response to treatments, outcomes
• Comparative effectiveness studies
Links to Other Data and Initiatives

• Quality – CMS, state reports, regional collaboratives
• State or National vital statistics – to assess mortality rates
• Hospital discharge datasets – for additional data detail and measures
• Health Information Exchanges – integrate claims and clinical (EMR) data
• Proprietary data marts that contain claims data for different payors, healthcare delivery systems, patient populations
In-Depth Example: Colectomy

• Average duration in OR = 182 minutes (Pandit & Carey, 2006)

• Data diversity and complexity
  – Patient-level data
  – Provider-level data
  – Facility-level data
  – Geospatial data
Whole Genome Sequences
MRI/CT OR Data
Labs
Nursing notes
Results/Outcomes
OR Data
Medications
Admissions
Hospital infections
Readmissions
Mortality
Physicians
Medicare history
Social history
Occupational history
Environmental data
Weather data
Neighborhood data
Geospatial data
Environmental data-home
Environmental data-work
Facility-level data
Patient level data
Patient history data
What to do?

• Use NLP to extract meaningful, standardized data elements
• Identify patterns and develop predictive algorithms
• Develop tools and methods to analyze highly diverse and complex data across multiple levels of measurement
Ex. of Big Data and Building a Comparative Effectiveness Trial

• Leverage HARBOR data (e.g., registries) for recruiting participant groups
  – Random selection

• Use of historical health system data rather than self-reported history to characterize participant history

• Use routine clinical data for monitoring at least secondary outcomes

“Do what you can, with what you have, where you are.” Theodore Roosevelt