Background and Disclosures

- Professor of Biostatistics and Computational Biology, Dana-Farber Cancer Institute
- Professor of Computational Biology and Bioinformatics, Harvard School of Public Health
- Many other academic titles
- Numerous advisory boards
- Co-Founder of GenoSpace, a Precision Genomic Medicine Software Company
Every revolution in science — from Copernican heliocentric model to the rise of statistical and quantum mechanics, from Darwin’s theory of evolution and natural selection to the theory of the gene — has been driven by one and only one thing: access to data.

—John Quackenbush
Driving Trends in Big Data
The Cloud

Image from https://www.data-hive.com/academy/primer.php
The iPhone (and other smart phones)

New Sources of Health and Medical Data

- Drug Research
- Social Media
- Patient Records
- Genomics
- Test Results
- Claims Data
- Home Monitoring
- Mobile Apps
It is estimated that by 2015, the average hospital will generate 665TB of data annually.

Medical imaging archives are increasing by 20-40% annually.

Today, 80% of data is unstructured such as images, video, and notes.
Transforming Medicine?

• New technologies from surveillance and exposure to genomics and imaging, from electronic health data to survey-based longitudinal studies are providing unprecedented data that have opened new avenues of investigation, transforming biomedical research into an information science.

• The challenge is to bring this information together with other information to better address fundamental problems, including a wide range of problems in health and biomedical research.

• In patient care, we want to increase efficiency, improve outcomes, and decrease cost.
Transforming Medicine?

• This is the era of Big Data in biomedical research with increases in the Three V’s: Volume, Velocity, and Variety.

  • Volume refers to the vast quantities of data we can now amass. We will soon generate as much data in one minute as we did between the beginning of history to 2008.

  • Velocity refers to the speed at which we generate data and speaks to the fact that nearly every medical instrument is networked and transmitting massive data in real time.

  • Variety refers to the complex, multifactorial data that we can now collect on each patient.
The Need for Data-Driven Precision Medicine
## US Cancer Prevalence Estimates
### 2010 and 2020

<table>
<thead>
<tr>
<th>Site</th>
<th>2010</th>
<th>2020</th>
<th>% change</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breast</td>
<td>3461</td>
<td>4538</td>
<td>31</td>
</tr>
<tr>
<td>Prostate</td>
<td>2311</td>
<td>3265</td>
<td>41</td>
</tr>
<tr>
<td>Colorectal</td>
<td>1216</td>
<td>1517</td>
<td>25</td>
</tr>
<tr>
<td>Melanoma</td>
<td>1225</td>
<td>1714</td>
<td>40</td>
</tr>
<tr>
<td>Lymphoma</td>
<td>639</td>
<td>812</td>
<td>27</td>
</tr>
<tr>
<td>Uterus</td>
<td>588</td>
<td>672</td>
<td>15</td>
</tr>
<tr>
<td>Bladder</td>
<td>514</td>
<td>629</td>
<td>22</td>
</tr>
<tr>
<td>Lung</td>
<td>374</td>
<td>457</td>
<td>22</td>
</tr>
<tr>
<td>Kidney</td>
<td>308</td>
<td>426</td>
<td>38</td>
</tr>
<tr>
<td>Leukemia</td>
<td>263</td>
<td>240</td>
<td>29</td>
</tr>
<tr>
<td><strong>All Sites</strong></td>
<td><strong>13,772</strong></td>
<td><strong>18,071</strong></td>
<td><strong>32</strong></td>
</tr>
</tbody>
</table>

Estimates of U.S. National Expenditures for Cancer Care 2010 – 2020

$124 billion and projected to rise to $207 billion (66% increase) by 2020

Ini. = within 1 year of Dx; Con = continuing; Last = last year
Biomarkers, Disease Subtyping and Targeted Therapy

Her-2+ (Herceptin) (Perjeta)
EML4-ALK (Xalkori)
K-ras (Erbitux) (Vectibix)
BRAF-V600 (Zelboraf)
CFTR-G551 (Kalydeco)

Companion Diagnostics: the Right Rx for the Right Disease (Subtype)
Non-Responders to Oncology Therapeutics Are Highly Prevalent and Very Costly

Disease Progression and Precision Care

- Birth
- Natural History of Disease
- Environment + Lifestyle
- Genetic Risk
- Early Detection
- Patient Stratification
- Disease Staging
- Treatment Options
- Clinical Care
- Outcomes
- Quality Of Life
- Death
- Biomarkers
The Data Must Be Useful and Usable And Accurate
Costs of Generating Data Have Plummeted

Cost per Genome

-$10,000,000$

-$1,000,000$

-$100,000$

-$10,000$

-$1,000$

Sep-01 Apr-02 Nov-02 Jun-03 Aug-04 Mar-05 Oct-05 May-06 Dec-06 Jul-07 Feb-08 Sep-08 Apr-09 Nov-09 Jun-10 Jan-11 Aug-11 Mar-12 Oct-12 May-13 Dec-13 Jul-14
Precision Medicine Demands Simplicity

<table>
<thead>
<tr>
<th>Mary Jones</th>
<th>RM12345</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disease: Colorectal Neoplasm</td>
<td></td>
</tr>
<tr>
<td>DOB: 5/31/53 (Age: 60y)</td>
<td></td>
</tr>
<tr>
<td>Sex: Female</td>
<td></td>
</tr>
<tr>
<td>Report Date: 08/01/2012</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Results</th>
<th>Interpretative Summary</th>
<th>Clinical Trials</th>
<th>References</th>
<th>Technical Details</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Therapeutic Associations</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Potential Therapeutic Response / Drug Class</td>
<td>Disease Association</td>
<td>Gene</td>
<td>Alteration</td>
<td></td>
</tr>
<tr>
<td>Panitumumab</td>
<td>Colonic Neoplasms</td>
<td>EGFR</td>
<td>p.L858R</td>
<td></td>
</tr>
<tr>
<td>Erlotinib</td>
<td>Colorectal Neoplasms</td>
<td>BRAF</td>
<td>p.V600E</td>
<td></td>
</tr>
<tr>
<td>Dasatinib</td>
<td>Colorectal Neoplasms</td>
<td>KRAS</td>
<td>p.G13</td>
<td></td>
</tr>
<tr>
<td>Ponatinib</td>
<td>Colonic Neoplasms</td>
<td>ABL1</td>
<td>p.E255K</td>
<td></td>
</tr>
</tbody>
</table>

**Additional Information**

**BRAF**

This gene encodes a protein belonging to the raf/mil family of serine/threonine protein kinases. This protein plays a role in regulating the MAP kinase/ERKs signaling pathway, which affects cell division, differentiation, and secretion. Mutations in this gene are associated with cardiofaciocutaneous syndrome, a disease characterized by heart defects, mental retardation, and a distinctive facial appearance. Mutations in this gene have also been associated with various cancers, including non-Hodgkin lymphoma, colorectal cancer, malignant melanoma, thyroid carcinoma, non-small cell lung carcinoma, and adenocarcinoma of lung. A pseudogene, which is located on chromosome X, has been identified for this gene. (provided by RefSeq Jul 2008)
Evidence: We Still Need Research

“Even within NCCN, certainly the majority of decision nodes that are enshrined in NCCN [guidelines] are not supported by high level evidence.”

Dr. Clifford Hudis
President, ASCO
Interview in Cancer Letter 22 Nov. 2013, 39

NCCN = National Cooperative Cancer Center Network
Evidence: We Still Need Research

“There is a lack of evidence showing the impact of guidelines on clinical practice and patient outcomes.”

Dr. G. H. Lyman
University of Washington School of Medicine
Medscape 11 April 2014

• response to McKesson Specialty Health press release that CMS is considering proposal from NCCN, US Oncology, and McKesson to use NCCN guidelines to control cost and promote more uniform medical practice
We Need to Invest in Big Data Research
National Research Council on Big Data

• National Research Council’s Committee on Massive Data Analysis concluded in their 2013 “Frontiers of Massive Data Analysis” report that the challenges associated with massive data go far beyond the technical aspects of data management (although those are not to be ignored.).

• The NRC consensus report noted the key element in meeting Big Data’s challenges was development of rigorous quantitative and statistical methods.

http://www.nap.edu/catalog.php?record_id=18374
In any discussion of massive data and inference, it is essential to be aware that it is quite possible to turn data into something resembling knowledge when actually it is not. Moreover, it can be quite difficult to know that this has happened.

http://www.nap.edu/catalog.php?record_id=18374
Key Challenges in Big Data

• **Preprocessing (Normalization) and Hot Spot Detection**
  • Need methods to compare measurements across sources and to rapidly identify salient features

• **Data Integration**
  • Need methods that can combine data from various sources where there are hidden correlations in the data

• **Reproducible Research**
  • Need to leverage the volume and velocity of the data to provide opportunities for validation of findings

• **Network Methods**
  • Need to move beyond correlations in studying relationships in data
Big Data Alone Are Not a Panacea

Number of people who died by becoming tangled in their bedsheets correlates with Total revenue generated by skiing facilities (US)

Spurious Correlations: http://www.tylervigen.com/
Key Questions to Answer about the use of Big Data in Health Care and Biomedical Research

• How do we construct informed consents to enable research?
• How do we balance protecting patient privacy with the need to store and access massive quantities of (fundamentally identifiable) data?
• How do we assure fair use of the appropriate data while addressing data ownership issues?
• How do we identify key “data consumer” user groups, define their potential use cases, and assure they have appropriate access to data?
• How do we distill information to provide appropriate information, in context, at the point of care?
• How do we navigate data security, cloud systems, and appropriate access in ways that are enabling rather than limiting?
• How do we scale as more and more quantitative data become available?
• What investments do we need to make in Big Data research to enable biomedical research and improve health care?