So Just What Is “Big Data”?

James E. Tcheng, MD, FACC, FSCAI
Disclosures

• James E. Tcheng, MD, FACC, FSCAI

Affiliations / Financial Relationships / Other RWI

• ACC – Chair, Informatics and Health IT Task Force
• Philips Healthcare – consultant (compensated)
• Epic Systems – consultant (not compensated)
• MedStreaming – consultant (not compensated)
• GE Healthcare – consultant (not compensated)
• Lumedx – consultant (not compensated)
A megabyte (MB) is to a gigabyte (GB) as a terabyte (TB) is to a:

1. Exabyte (EB)
2. Petabyte (PB)
3. Yottabyte (YB)
4. Zettabyte (ZB)
A typical large tertiary care hospital generates ~100 TB of data per year. The printed (text) content of the Library of Congress is estimated to be:

1. 10 TB
2. 100 TB
3. 1000 TB (=1 PB)
4. 10,000 TB (=10 PB)
One gram of DNA can theoretically hold how much data?

1. $\frac{1}{2}$ Terabyte (TB)
2. $\frac{1}{2}$ Petabyte (PB)
3. $\frac{1}{2}$ Exabyte (EB)
4. $\frac{1}{2}$ Zettabyte (ZB)
Big Data Defined

• Volume. Volume. Volume …
  – megabyte, gigabyte, terabyte, petabyte, exabyte (=10^9 GB),
    zettabyte, yottabyte …
• Velocity … faster and faster!
• Variety … both structured data and unstructured information (e.g., analog text / audio / video)
• Variability … quality / veracity, quantity, timing
• Value … targets operational insight (e.g., individual patient care)

• Library of Congress: ~10 TB of text (20 PB of audio, video)
• Duke Heart Center: ~30TB clinical data / yr
  = Facebook in 1 day; CERN in 1 second
• Google: processes 24PB of data / day
• “All words ever spoken by humans”: 5 EB of text
• 1 gram of DNA: 455 EB (~1/2 ZB)
• 2015 worldwide IP traffic: 966 EB (61% Internet video)
Endless Variety

Genomic

Other ‘Omics

Imaging

Phenotypic

Exposure

Clinical
Huge Variability

95% of the world’s data is unstructured
  – Text, images, video, voice, etc.
  – Most healthcare data is unstructured

New data types are emerging
  – Messaging, social media, sensor data
Peak Break-Up Times
According to Facebook status updates

Likelihood of a Break Up per Day

- Spring Break
- Valentines Day
- April Fools Day
- Mondays
- Summer Holiday
- Before Winter Holidays
- Christmas

David McCandless & Lee Byron
InformationIsBeautiful.net / LeeBryon.com

Source: searches for “we broke up because”
taken from the infographic ultrabook
The Visual Miscellaneum
So Why Big Data?

- Big volume ... isn’t new (stock exchanges)
- Real time velocity ... isn’t new (railway management)
- Variety and variability of data ... is increasing (and increasingly difficult to manage)
- Requires changing the data ownership paradigm
- VALUE ... the potential for insight and understanding
Turning Big Data into Value

‘Data-fication’ of the World
- Documentation
- Events
- Procedures
- Billing
- Images
- Registries
- Social Media
- ‘Omics
- Sensors
- Etc.

Analyzing Big Data:
- Natural language processing
- Text analytics
- Information extraction
- Data mining
- Predictive modelling
- Inferential analysis
- Comparative effectiveness
- Etc.

Visualizing Big Data:
- Infographics
- Advanced data visualization
- Interactive data
- Contextual modelling
- Etc.

Value

Volume
Velocity
Variety
Variability
Every 10% ↑ in guidelines adherence → 10% ↓ in mortality
OR 0.90 (0.84-0.97, \( P < .001 \))

Peterson ED et al, *JAMA* 2006; 295:1912-1920
<table>
<thead>
<tr>
<th>Drug</th>
<th>Efficacy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anti-Depressants</td>
<td>62 %</td>
</tr>
<tr>
<td>Asthma</td>
<td>60 %</td>
</tr>
<tr>
<td>Diabetes</td>
<td>57 %</td>
</tr>
<tr>
<td>Arthritis</td>
<td>50 %</td>
</tr>
<tr>
<td>Alzheimer</td>
<td>30 %</td>
</tr>
<tr>
<td>Cancer</td>
<td>25 %</td>
</tr>
</tbody>
</table>

*Drug does not work*
Big Data in Healthcare

• 6 Key High-Risk, High-Cost Use Cases
  – high-cost patients (identify early, intensively manage)
  – high-probability readmissions (match resources to risk)
  – triage (assessment in context → posterior probabilities)
  – clinical decompensation (↑signal : noise, address alarm fatigue)
  – adverse events (e.g. AKI, infection, adverse drug events)
  – treatment optimization for complex, multi-organ disease (multisite longitudinal registries – e.g., PCORnet)

Bates DW et al., Health Affairs (2014) 33;1123
6/09 – 3/10
826 million tweets
146 million geo-located
Across 1300 counties
Eval for stress & hostility:
health, attractiveness,
job, curse words
+CDC, US Census data

STRONGER correlation than 10 classic predictors, incl. education, SES, smoking, hypertension (r=.42 vs .36, p=.049)

Eichstaedt JC et al., Psychological Science, 2015;26:159
Implications – Big Data in Cardiology

• What analytics approach?
  – Start at the end – and perfection is the enemy of good

• What measurement sources?
  – Cardiology already data-rich, including non-traditional data

• How to make predictions actionable?
  – Not just reductionist CDS – think heuristic medical decision making

• Similarity of development cohort to use case?
  – Model (clinical trial) populations versus registries versus reality

• Tailoring the intervention to the individual
  – Actualization, delivery can be high-cost
Big Trends in Health IT

• We are falling well short of achieving “meaningful” use of HIT
  – EHR certification criteria, etc., are stifling innovation

• We must move from closed systems to open software architectures
  – enabling Big Data analytics – not just document transport

• Big Data – vision of an interoperable health data infrastructure
  – Using patient-reported, ‘omic, embedded / wireless sensors, open sources
  – New discovery likely dependent on shared access to international health data

JASON Report, AHRQ 2014                      10 Year HIT Infrastructure, ONC 2014