Challenges of Big Data in Health

Toshiba – JH Collaboration
Scott L. Zeger
Professor and Director
Johns Hopkins inHealth Initiative
• 40 year old man, no family history, tests “positive” in a cancer screening test
• What is his cancer state?

Data from population of “similar” people

<table>
<thead>
<tr>
<th>Test result</th>
<th>True cancer status</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Positive</td>
<td>15</td>
<td>985</td>
</tr>
<tr>
<td>Negative</td>
<td>5</td>
<td>8,995</td>
</tr>
<tr>
<td>Total</td>
<td>20</td>
<td>9,980</td>
</tr>
</tbody>
</table>
“Big Data” – large, complex data sets that challenge current data science knowledge/skills that hold out the promise of hidden truths → “data mining” for “gold”
Boole - a - Bayes

George Boole 1815-1864

Thomas Bayes 1701-1761
Radar to Land Aircraft Safely at BWI

Initial Radar Measurements (radar echoes)

Filtered (smoothed) "track" updates

Smoothed track updates

Symbol of type of aircraft

Indicates expected future direction and speed

Display of altitude

Data:
- Track No. 1234
- Position (x,y,z)
- Speed and direction
- Acceleration
- Identification

© 2014 Hopkins inHealth
Track State Analogy

- location
- velocity
- acceleration
- accuracy
- uncertainties

- genotype
- phenotype
- environment/behavior
- biomarkers/labs/images
- identification
**Biohealth Pilot Projects**

- Cancer screening
- Cardiovascular disease diagnosis and treatment
- Genomics of cystic fibrosis
- Telomere biology and chronic diseases
- OncoSpace in Radiation Oncology
- Management of autoimmune diseases
- Myostatin in sarcopenia

**Population Health Demonstration**

- Cancer screening and early diagnosis
- Cardiovascular disease
- Age-related cognitive loss
- Obesity and Diabetes
- Children’s asthma prevention and control

**Methodology Cores**

- Health measurement
- Bioethics
- Big Data and software solutions
- Statistical design and analysis
- Behavior change and dissemination
- Finance – organization models

© 2014 Hopkins inHealth
A data system designed to individualize radiation therapy

Todd McNutt, Kim Evans, Joe Moore, Harry Quon, Joseph Herman, Andrew Sharabi, Wuyang Yang, John Wong, Theodore DeWeese

Disclosure:
Funding from Elekta and Philips
Sample Automated Radiation Plan

Original

Automated plan

30% reduction in dose to parotids

Auto plan
Original Plan
Dot: right
No-dot: left
Model Framework for Infections (I) and Pathogen Measurements (M)

- **Blood infection, \( I_B \)**
  - Blood culture measure, \( M_{B,CX} \)

- **Lung aspirate culture and PCR measure*, \( M_{L,CX} \) and \( M_{L,PCR} \)

- **Pleural fluid culture and PCR measure*, \( M_{PE,CX} \) and \( M_{PE,PCR} \)

- **Blood LytA PCR measure, \( M_{B,PCR} \)**

- **NP measure, \( M_{NP,PCR} \)**

- **NP measure, \( M_{np,cx} \)**

- **Sputum measure, \( M_{SP,PCR} \)**

- **Sputum measure, \( M_{SP,cx} \)**

- **Gastric aspirate*, \( M_{GA,TB,cx} \)**

- **A/C antibodies measure, \( M_{SERA} \)**

* collected on a subset of cases

§ TB testing only

- Measurement of pathogen/evidence of infection
- Pathogen direction of travel
- Potential non-pneumonia causes of pathogen detection
- Sample collected in both cases and controls

© 2014 Hopkins inHealth
Covariates for adjustment (X)

Unobserved Lung Infection (I)

Clinical Pneumonia (Y)

Population Pathogen Distribution

Infection Measurements (M)

Prob(M|Case) Prob(M|Control)

Unknown regression coefficient \( \gamma \)
All that glitters….
Beware the Texas Sharpshooter
Congratulations to Toshiba
Welcome to Johns Hopkins!

Thank You