Grand Challenges in using Big Data in Healthcare

DEAN F. SITTIG, PHD
Agenda

- Define and discuss Big Data
- Review Grand Challenges associated with Big Data
- Q & A re: Big Data and Grand Challenge
The Big Data Use Continuum...

Discovery of new knowledge

What could happen?

Examples

1. Risk stratification/patient identification for integrated-care programs
2. Risk-adjusted benchmark/simulation of hospital productivity
3. Identification of patients with negative drug–drug interactions
4. Identification of patients with potential diseases ("patient finder")
5. Evaluation of clinical pathways
6. Evaluation of drug efficacy based on real-world data
7. Performance evaluation of integrated-care programs and contracts
8. Identification of inappropriate medication
9. Systematic reporting of misuse of drugs
10. Systematic identification of obsolete-drug usage
11. Personal health records
12. Community-wide EHR with environment and activity

Modified from McKinsey Big Data Value Demonstration Team
Utah Population Database (UPDB) used to find breast cancer gene

- Genealogical data linked to state cancer records & death certificates:
  - Mormon genealogy – 1 million records; 180,000 families
  - Utah Cancer Registry – 125,904 records
  - Utah death certificates – 500,000 records

- 117,407 individuals with cancer, 41,940 of whom linked to the UPDB genealogy records (35.7%)

- Identified families in which breast and ovarian cancer occurred more frequently – led to further testing and isolation of the BRCA gene
Removal of Vioxx from market...

- Kaiser Permanente - integrated managed care organization providing health care to more than 6 million residents in CA
- Population varies with respect to age, education, income, and ethnicity
- KP maintains computer records of outpatient & ER visits, admissions, medical procedures, laboratory testing, and outpatient drugs
- Mortality status, including cause of death is updated
Removal of Vioxx from market...

- Cohort of NSAID-treated patients Jan 1, 1999 - Dec 31, 2001
- Identified all individuals age 18–84 years who filled at least 1 prescription for Vioxx.
- Patients with no diagnoses of cancer, renal failure, liver failure, severe respiratory disease, organ transplantation, or HIV/AIDS
- Followed up cohort until end of study, acute MI, or death
Removal of Vioxx from market...

- 1,394,764 people contributed 2,302,029 person-years
- 8,143 patients (0.6%) with severe coronary heart disease
- For high-dose Vioxx, odds ratio was 3.58 (p=0.016)
  for standard-dose Vioxx 1.47 (p=0.054)
- Conclusion: Vioxx increases the risk of serious coronary heart disease
- 4 years after withdrawal from market, Merck paid $4.85 billion to settle 27,000 lawsuits
Definition of “Big Data”

- **Volume**: amount of information to be processed
- **Velocity**: rate at which the data must be processed to keep up with the inflow
- **Variety**: types, dimensions, and time scales that must be managed
- **Veracity**: uncertainty in the data’s accuracy and trustworthiness
Where does Big Data come from?

**Activity (claims) & cost data**
- Owners: payors and providers
- Examples: BC/BS or CMS claims; retail pharmacies

**Clinical data**
- Owners: providers
- Examples: EHRs, Health Information Exchanges (HIEs); VA clinical data warehouse

**Pharmaceutical R&D data**
- Owners: Pharma, academia
- Examples: Clinical trials, DNA screening data

**Patient behavior and sentiment data**
- Owners: consumers, outside of healthcare
- Examples: FitBit, patient satis, Twitter, Chat groups

Need to integrate data across sources for major opportunities

Modified from McKinsey Big Data Value Demonstration Team
Definition of “Big Data”

- **Volume**: amount of information to be processed
- **Velocity**: rate at which the data must be processed to keep up with the inflow
- **Variety**: types, dimensions, and time scales that must be managed
- **Veracity**: uncertainty in the data’s accuracy and trustworthiness
**Data Velocity**

- Real-time waveform analysis
  - Heart rate analysis requires processing approximately 500 data points per second
  - Could become a problem if you are simultaneously monitoring 1 million patients via a new iPhone app

- Real-time image processing – looking for abnormalities

- Quantifiable self – fitbit – activity tracking – 1.25 million units

- Genetic testing
Definition of “Big Data”

- **Volume**: amount of information to be processed
- **Velocity**: rate at which the data must be processed to keep up with the inflow
- **Variety**: types, dimensions, and time scales that must be managed
- **Veracity**: uncertainty in the data’s accuracy and trustworthiness
Data Variety

- Free-text data – clinical notes
- Image data – x-rays, CTs, MRIs, microscopy
- Audio files – heart sounds
- Waveform data – EKG, Blood pressure, oximeter, inhaled gases
- Geographic data – patient location – environmental exposure
- Genetic data – human genome, bacterial genome
Definition of “Big Data”

- **Volume**: amount of information to be processed
- **Velocity**: rate at which the data must be processed to keep up with the inflow
- **Variety**: types, dimensions, and time scales that must be managed
- **Veracity**: uncertainty in the data’s accuracy and trustworthiness
Data Veracity

- Data artifacts – EKG lead failure, BP line blockage,
- Assessing data authenticity – where did it come from
- Free-text – misspellings,
- Wrong patient ID errors
Grand Challenges...

- Fundamental scientific, technologic, or social problems
- Solutions require significant improvement in scientific knowledge, technical capabilities, or regulations
- Solutions should significantly improve access to data, quality of analyses, and predictions
- Solutions should be achievable within a decade
- Created to educate and inspire researchers, developers, funders, and policy-makers
Order of presentation...

- Not most important first...
- Not hardest first...
- Not easiest first...
- Not ones most likely to be solved first...
- Rather...Challenges likely to be faced when exploring a real-world healthcare problem
Creating, identifying, collecting, or gaining access to truly “big data” sets

- Most existing single site healthcare data sets not that big!
- High-resolution digital images
- Free-text data
- Opportunities for real “big data”:
  - genetic testing
  - “quantifiable-self”
  - Creation of large, region- or even nationwide clinical data sets
Creation of an overarching regulatory and policy framework

- NSA, Facebook, and Apple: privacy & confidentiality of patient-related data

- Specific issues:
  - Data sharing
  - Patient consent
  - Data stewardship
  - Maintenance of individual’s privacy

- Multiple IRBs claim jurisdiction of data and analysis

- HIPAA only permits research with informed consent or waiver

- Analyses often must be completed using methods that are privacy-preserving
  - Easy to design queries that expose a single person
  - Fuzzing - removal or modification of the least significant digits from a number
  - Aggregation - conversion of an individual values to a range
  - Difficult to implement correctly - trivial to defeat, or impossible to answer certain questions
Linking patients across organizations

- Requires combining data gathered by multiple organizations
- Necessary to achieve longitudinal data resource and to fill in the gaps in care
- Difficult because patients have very few, if any, unique, permanent, identifying characteristics:
  - date of birth, gender, names, addresses, and most gov’t issued IDs (driver’s license numbers, passports, SSNs)
- Rely on combinations of items to reduce the probability of incorrect links
  - often one or more items are missing, or have recently changed
- Current accuracy estimates: 90-95% correct matches
  - In DB with 10 million patients - 500,000 – 1 million incorrect
Dealing with “opt-in” or “opt-out” biases

- Types of patients that are in the database
  - VA health system has nationwide clinical data repository
  - Contains clinical data from all VA patients seen in the last 10 years
  - 84.3% of these patients are male and most are over 65 years of age

- Oversample a subset of the patients to create an artificial sample that more closely reflects the composition of nation as a whole

- Other clinical data resources have similar degrees of bias:
  - Kaiser Permanente has huge data warehouses containing all information from their patients
  - Relatively younger and healthier patients
  - Members get insurance from employer: young, healthy, working
Organ donors in European Union

% of Adult Population Volunteering to be on Organ Donor

- Denmark: 4.25%
- Netherlands: 27.5%
- United Kingdom: 17.17%
- Germany: 12%
- Austria: 99.98%
- Belgium: 98%
- France: 99.91%
- Hungary: 99.97%
- Poland: 99.5%
- Portugal: 99.64%
- Sweden: 85.9%
Organ donors in European Union

Opt-in

Opt-out
Dealing with missing data

- Often one or more data points describing patients missing
- Missing data can occur for a variety of reasons:
  - Failure to collect, enter, understand the data
  - MU criteria require 80% of patients to have at least 1 problem
  - % of diabetic patients with documented problem range 65-95%
- Impute values of missing data based on distribution of observed variables
  - Such assumptions are untestable
Overcoming missing data?

- Clinical trial data to answer the research question is prospectively collected for every participant
- Big Data - participants chosen based on eligibility criteria AND availability of sufficient data for extraction
- Reviewed # days with medication orders and # days with laboratory results for each patient for the year preceding the provision of anesthetic services
- Estimated count of days with laboratory results for ASA 4 (most sick) was 5.05 times the count of days with laboratory results for ASA 1 (least sick)
- On the other hand...
- If a patient comes into ED with a severe problem and dies, care team spends little time documenting symptoms so in the EHR, the patient appears to be healthy...other than the death!
Limitations of retrospective, observational study designs

- Analysis of big data limits one to retrospective study designs

- Inherent limitations:
  - Before-after studies must deal with overarching secular trends that often overwhelm small differences due to study variable(s)
  - Case-control and retrospective cohort designs deal with unknowns that account for why some patients received a particular test or treatment

- Lead to bias and confounding
  - Matching, stratification and adjustment (e.g. through regression) are available, but only control for confounders that are measured
  - Important confounders are either unmeasurable or simply unmeasured in the available dataset.
**Hormone replacement therapy (HRT) was common**

- In 1990’s 38% of U.S. women aged 50 to 74 years were using hormone therapy
- Based on analysis of large retrospective data sets: it could help prevent cardiovascular disease while causing no harm
- Recommendation was retracted based on large prospective, RCTs showed an increase in cardiovascular events following use of HRT medications.
- Explanations for differences have been proposed:
  - Retrospective data sets may have over-sampled patients with higher socioeconomic status
  - Retrospective studies adjusted for either alcohol use or exercise, both known to be more common in women using HRT, showed no benefit from HRT.
Separating spurious correlations from true causation

- Confusing one or more events that happen during the same time period with those that cause each other
  - Fall in the number of pirates over the last 400 years often associated with global warming

- Try to understand potential cause and effect relationship using known physical or biological principles

- Most big data discoveries have been confirmations of known or suspected relationships
US spending on science, space, and technology correlates with Suicides by hanging, strangulation and suffocation

Correlation: 99%  
Sources: US OMB & CDC  
tylervigen.com

<table>
<thead>
<tr>
<th>Year</th>
<th>US spending on science, space, and technology (in millions of today's dollars)</th>
<th>Suicides by hanging, strangulation and suffocation (in deaths)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1999</td>
<td>18,079</td>
<td>5,427</td>
</tr>
<tr>
<td>2000</td>
<td>18,594</td>
<td>5,688</td>
</tr>
<tr>
<td>2001</td>
<td>19,753</td>
<td>6,198</td>
</tr>
<tr>
<td>2002</td>
<td>20,734</td>
<td>6,462</td>
</tr>
<tr>
<td>2003</td>
<td>20,831</td>
<td>6,635</td>
</tr>
<tr>
<td>2004</td>
<td>23,029</td>
<td>7,336</td>
</tr>
<tr>
<td>2005</td>
<td>23,597</td>
<td>7,248</td>
</tr>
<tr>
<td>2006</td>
<td>23,584</td>
<td>7,491</td>
</tr>
<tr>
<td>2007</td>
<td>25,525</td>
<td>8,161</td>
</tr>
<tr>
<td>2008</td>
<td>27,731</td>
<td>8,578</td>
</tr>
<tr>
<td>2009</td>
<td>29,449</td>
<td>9,000</td>
</tr>
</tbody>
</table>

Correlation: 0.992082

Permalink - Mark as interesting (5,147) - Not interesting (2,370)

http://www.tylervigen.com/
In General...You can’t “prove” causation using retrospective data

- Violates Koch’s Postulates from 1905 – Used to identify the bacteria responsible for Tuberculosis

- Violates most of Sir Austin Bradford Hill’s Criteria for Causation – used to help health services researchers determine that “poverty leads to poor health outcomes”

- On the other hand, you can often:
  - Gather enough evidence to generate equipoise, to make a prospective, randomized trial acceptable
<table>
<thead>
<tr>
<th>Missing data</th>
<th>Prospective Cohort Study</th>
<th>Retrospective Big Data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Can require that blood pressure be measured in all patients at time of enrollment</td>
<td>Can only use patients for whom blood pressure was measured – these may not be representative of all patients</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variability in accuracy</th>
<th>Prospective Cohort Study</th>
<th>Retrospective Big Data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Can enforce protocol to have nurse measure BP three times after patient has been sitting still for at least five minutes</td>
<td>BPs measured in a variety of different situations</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Patient representativeness</th>
<th>Prospective Cohort Study</th>
<th>Retrospective Big Data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Can define the inclusion criteria for the cohort and, with sufficient resources, can likely build a cohort that is representative of the target population</td>
<td>Must use data that’s available – if large enough, can use weighting to make the sample resemble the target population</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Loss to followup</th>
<th>Prospective Cohort Study</th>
<th>Retrospective Big Data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Significant risk, although have the option to trace patients, offer incentives, etc. to maximize followup</td>
<td>Significant risk, particularly in an open health system. With HIE or federation, might be able to track data as patient moves across providers</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cost</th>
<th>Prospective Cohort Study</th>
<th>Retrospective Big Data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Very expensive, especially if followed prospectively</td>
<td>Less expensive</td>
</tr>
</tbody>
</table>
Summary

- Big data has **big potential** in healthcare
- Current healthcare-related data is “**not that big**”
- Many **challenges to overcome** if we are to continue the strong traditions of scientific research including: ethics, precision, validity, and explanatory power
Thank you...

Contact Information...

Dean Sittig – Dean.F.Sittig@uth.tmc.edu
@DeanSittig