Healthcare Data Analytics

What is Biomedical & Health Informatics?
William Hersh, MD
Copyright 2018
Oregon Health & Science University

Healthcare data analytics

- Rationale
- Definitions
- Applications
- Results
- Challenges
Rationale

- Although focus in recent years has been on EHR implementation and meaningful use (MU), informatics work in future will shift to putting the data and information to work improving quality and lowering cost of healthcare (Hersh, 2012; MIT Critical Data, 2016)
- As quantity and complexity of healthcare data grow through EHR capture, genomics, and other sources, number of facts per clinical decision will increase, requiring increasing help for decision-makers (Stead, 2011)
- Physicians and others can leverage this data to improve care (Sniderman, 2015; Parikh, 2016; Rumsfeld, 2016)

Definitions

- Both a buzz-word and an important emerging area
- Davenport (2007) – “the extensive use of data, statistical and quantitative analysis, explanatory and predictive models, and fact-based management to drive decisions and actions”
- IBM (2012) – “the systematic use of data and related business insights developed through applied analytical disciplines (e.g. statistical, contextual, quantitative, predictive, cognitive, other [including emerging] models) to drive fact-based decision making for planning, management, measurement and learning”
Levels of analytics (Adams, 2011)

<table>
<thead>
<tr>
<th>Degree of Competitive Advantage and Complexity</th>
<th>Optimization</th>
<th>How can we achieve the best outcomes?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predictive modeling</td>
<td>Identifying high-risk patients</td>
<td>What will happen next if...?</td>
</tr>
<tr>
<td>Forecasting</td>
<td>Public health issues</td>
<td>What if these trends continue?</td>
</tr>
<tr>
<td>Simulation</td>
<td>Business processes</td>
<td>What could happen if...?</td>
</tr>
<tr>
<td>Alerts</td>
<td>Infection outbreaks</td>
<td>When are actions needed?</td>
</tr>
<tr>
<td>Query/drill-down</td>
<td>&quot;What and why&quot;</td>
<td>What exactly is the problem?</td>
</tr>
<tr>
<td>Ad hoc reporting</td>
<td>Out-of-range metrics</td>
<td>How many, how often, where?</td>
</tr>
<tr>
<td>Standard reporting</td>
<td>Key metrics</td>
<td>What happened?</td>
</tr>
</tbody>
</table>

Related terms

- **Big Data** – 4 Vs of (Zikopolous, 2011; O’Reilly, 2015)
  - Volume
  - Velocity
  - Variety
  - Veracity
- **Data mining** – processing and modeling of data to discover previously unknown patterns or relationships (Bellazzi, 2008; Zaki, 2014)
- **Text mining** – applying data mining to unstructured textual data (Aggarwal, 2012)
Related terms (cont.)

• Machine learning – area of computer science focused on systems and algorithms that learn from data (Flach, 2012; Crown, 2015)
  – Deep learning – use of computational models with deep layers requiring substantial processing (LeCun, 2015)
• Data science – “the science of learning from data; it studies the methods involved in the analysis and processing of data and proposes technology to improve methods in an evidence-based manner” (Donoho, 2015)
  – “A person who is better at statistics than any software engineer and better at software engineering than any statistician” (or substitute informatician for software engineer? – Hersh, 2016)

More related terms

• Data provenance – origin and trustworthiness (Buneman, 2010)
• Business intelligence – use of data to obtain timely, valuable insights into business and clinical data (Adams, 2011)
• Precision (IOM, 2011; Collins, 2015; Ashley, 2015), personalized (Hamburg, 2010), or computational medicine (Winslow, 2012)
• Analytics pipeline - adapted from Kumar (2013) for healthcare (Hersh, 2014)
Analytics is well-employed outside of healthcare

- Amazon and Netflix recommend books and movies with great precision
- Many sports teams, such as the Oakland Athletics and New England Patriots, have used “moneyball” to select players, plays, strategies, etc. (Lewis, 2004; Davenport, 2007)
- Twitter volume and other linkages can predict stock market prices (Ruiz, 2012)
- US 2012 election showed value of using data: re-election of President Obama (Scherer, 2012) and predictive ability of Nate Silver (Salant, 2012)
- Individual traits such as sexual orientation, political affiliation, personality types, and ethnicity can be discerned from Facebook “likes” with high accuracy (Kosinski, 2013)
- “Internet advertising” (Smith, 2014), aiming to solve “Wanamaker dilemma” (O’Reilly, 2012)
- Government (e.g., National Security Agency in US) tracking of email, phone calls, and other digital trails (Levy, 2014)
- Detecting credit card fraud (Ryoo, 2015)

What about analytics in healthcare?

- With shift of payment from “volume to value,” healthcare organizations will need to manage information better to deliver better care (Diamond, 2009; Horner, 2012; Burwell, 2015)
  - To realize this, they must achieve “analytic integration” (Davenport, 2012)
- New care delivery models (e.g., accountable care organizations) will require better access to data (e.g., health information exchange, HIE)
  - Halamka (2013): ACO = HIE + analytics
- Recent overviews (Gensinger, 2014; Marconi, 2014; Reddy, 2015)
Applications of analytics in healthcare

- Early application – identifying patients at risk for hospital readmission within 30 days of discharge
- Centers for Medicare and Medicaid Services (CMS) Readmissions Reduction Program penalizes hospitals for excessive numbers of readmissions (2013)
- Several studies have used EHR data to predict patients at risk for readmission (Amarasingham, 2010; Donzé, 2013; Gildersleeve, 2013; Hebert, 2014; Shadmi, 2015; Tabak, 2015)
- Many other applications that identify or predict using EHR data; fewer actually applied to achieve improved outcomes

Other applications of analytics – identifying cases

- Identifying patients who might be eligible for participation in clinical studies (Voorhees, 2012; Wu, 2017)
- Identification of children with asthma (Afzal, 2013)
- Detecting postoperative complications (FitzHenry, 2013; Tien, 2015)
- Identifying patients with diabetes
  - Including earliest date of diagnosis (Makam, 2013)
  - Using ontology-based algorithm (Rahimi, 2014)
  - Using NLP-based case-finding algorithm of HIE data (Zheng, 2016)
- Improving on ICD-9 to identify patients with hepatocellular carcinoma (Sada, 2016)
- Automated phenotyping using EHR data and machine learning (Halpern, 2016)
Other applications of analytics – predicting utilization and outcomes

- Predicting
  - Healthcare utilization (Haas, 2013; Charlson, 2014)
  - Primary care panel size (Rajkomar, 2016)
- Predicting risk of suicide better than clinicians (Tran, 2014)
- Predicting mortality with EHR data
  - Risk-standardized mortality after acute MI (McNamara, 2015)
  - In ICU, augmented with patient similarity data (Lee, 2015)
  - During hospitalization (Khurana, 2016)
  - One-year death risk when hospitalized (van Walraven, 2017)
- Cancer
  - Detecting potential delays in cancer diagnosis (Murphy, 2014)
  - Predicting trajectory of disease using NLP (Jensen, 2017)
- Improving prediction of conventional approaches
  - Severity of illness in ICU based on usual physiological models (Lee, 2017)
  - Cardiovascular event risk prediction (Weng, 2017)

Most important use cases for data analytics (Bates, 2014)

- High-cost patients – looking for ways to intervene early
- Readmissions – preventing
- Triage – appropriate level of care
- Decompensation – when patient’s condition worsens
- Adverse events – awareness
- Treatment optimization – especially for diseases affecting multiple organ systems
Requirements for data analytics in healthcare

- Infrastructure (Amarasingham, 2014)
  - Stakeholder engagement
  - Human subjects research protection
  - Protection of patient privacy
  - Data assurance and quality
  - Interoperability of health information systems
  - Transparency
  - Sustainability
- New models of thinking and training (Krumholz, 2014)
- New tools, e.g., “green button” to help clinicians aggregate data in local EHR (Longhurst, 2014)

Does application of analytics improve patient outcomes?

- Readmission tool applied to case management approach helped reduce readmissions (Gilbert, 2013)
- Bayesian network model embedded in EHR to predict hospital-acquired pressure ulcers led to tenfold reduction in ulcers and one-third reduction in intensive care unit length of stay (Cho, 2013)
- Readmission risk tool intervention reduced risk of readmission for patients with congestive heart failure but not those with acute myocardial infarction or pneumonia (Amarasingham, 2013)
- Use of EHR-based acuity score allowed intervention that reduced in-hospital mortality from 1.9% to 1.3% (Rothman, 2015)
- Randomized controlled trial of tool to reduce delay in cancer diagnosis led to earlier diagnosis for colorectal and prostate cancer (Murphy, 2015)
- Use of predictive report based on NLP tool reduced time in discharge planning meetings and 30-day all-cause mortality although not cost or readmissions (Evans, 2016)
Studies of outcomes (cont.)

- Development and use of a universal data architecture at Geisinger has led to successes in (Erskine, 2016)
  - Closing loop on appropriate treatment and lack of follow-up (Graf, 2014)
  - Early detection and treatment of sepsis
  - Monitoring and control of surgery costs and outcomes
- In children with cerebral palsy, implementation of a learning health system led to (Lowes, 2017)
  - 43% reduced hospital days
  - 30% reduction in emergency department visits
  - 210% reduction in healthcare costs

Some challenges for analytical use of clinical data

- Data quality and accuracy is not a top priority for busy clinicians (de Lusignan, 2005)
- Average pediatric ICU patient generates 1348 information items per 24 hours (Manor-Shulman, 2008)
- Patients get care at different places (Bourgeois, 2010; Finnell, 2011)
- Much data is “locked” in text (Hripcsak, 2012)
- Electronic records of patients at academic medical centers not easy to combine for aggregation (Broberg, 2015)
- Even structured data not usable purely automated for clinical score calculation (Aakre, 2017)
Caveats for use of operational EHR data (Hersh, 2013) – may be

- Inaccurate
- Incomplete
- Transformed in ways that undermine meaning
- Unrecoverable
- Of unknown provenance
- Of insufficient granularity
- Incompatible with research protocols

Many “idiosyncrasies” of clinical data (Hersh, 2013)

- “Left censoring” – First instance of disease in record may not be when first manifested
- “Right censoring” – Data source may not cover long enough time interval
- Data might not be captured from other clinical (other hospitals or health systems) or non-clinical (OTC drugs) settings
- Bias in testing or treatment
- Institutional or personal variation in practice or documentation styles
- Inconsistent use of coding or standards
Recommendations for use of operational EHR data (Hersh, 2013)

<table>
<thead>
<tr>
<th>Recommendation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apply an Evidence-Based Approach</td>
<td>Ask an answerable question, find the best EHR data (“evidence”), appraise the data, apply evidence to question</td>
</tr>
<tr>
<td>Evaluate and Manage Data</td>
<td>Assess availability, completeness, quality (validity), and transformability of data</td>
</tr>
<tr>
<td>Create Tools for Data Management</td>
<td>Create software (especially pipelines) for data aggregation, validation and transformation</td>
</tr>
<tr>
<td>Determine Metrics for Data Assessment</td>
<td>Determine whether a particular site’s data are “research grade”</td>
</tr>
<tr>
<td>Develop Methods for Comparative Validation</td>
<td>Develop tools that support analysis of multi-site data collections</td>
</tr>
<tr>
<td>Develop a Methodology Knowledge Base</td>
<td>Develop a data catalogue that relates data elements to recommended transformations</td>
</tr>
<tr>
<td>Standardize Reporting Methods</td>
<td>Provide details of data sources, provenance and manipulation, to support comparison of data</td>
</tr>
<tr>
<td>Engage Informatics Expertise</td>
<td>Ensure validity of findings derived from data collected from disparate sources</td>
</tr>
<tr>
<td>Include an Informatics Research Agenda</td>
<td>Generate systematic studies of inherent biases in EHR and data collection methods, such as data entry user interfaces</td>
</tr>
</tbody>
</table>

Apply an evidence-based medicine approach (Hersh, 2013)?

- **Ask an answerable question**
  - Can question be answered by the data we have?
- **Find the best evidence**
  - In this case, best evidence is EHR data needed to answer the question
- **Critically appraise the evidence**
  - Does the data answer the question?
  - Are there confounders?
- **Apply it to the patient situation**
  - Can the data be applied to this setting?
Who does/should do analytics?

- Data scientists – the “sexiest profession of the 21st century” (Davenport, 2012)
- Key skill sets include
  - Machine learning, based upon a foundation of statistics (especially Bayesian), computer science (representation and manipulation of data), and knowledge of correlation and causation (modeling) (Dhar, 2013)
  - IBM – both “numerate” and business-oriented skills (Fraser, 2013)
  - NIH – big data researchers need training in quantitative sciences, domain expertise, ability to work in diverse teams, and understanding concepts of managing and sharing data (NIH, 2013)

How many are needed?

- McKinsey (Manyika, 2011) – need in US in all industries (not just healthcare) for
  - 140,000-190,000 individuals who have “deep analytical talent”
  - 1.5 million “data-savvy managers needed to take full advantage of big data”
- Similar analysis by IDC (2014) of need for 180,000 with “deep” talent and 5-fold around with skills in data management and interpretation
What skills are needed (Otero, 2014; Garfield, 2017)?

- Programming – especially with data-oriented tools, such as SQL and statistical packages
- Statistics – working knowledge to apply tools and techniques
- Domain knowledge
- Communication – ability to understand needs of people and organizations and articulate results back to them

- Is this informatics? Or a specialization of informatics? Or something totally different?

Much promise for data analytics, but need

- Other aspects of informatics
  – Robust EHRs and other clinical data sources
  – Standards and interoperability
  – Health information exchange
  – Usability of clinical systems
- Improved completeness and quality of data
- Research demonstrating how best applied to improve health and outcomes
- Human expertise to apply and disseminate