Big Data Is Not Enough: People and Systems Are Needed to Benefit Health and Biomedicine

William Hersh, MD
Professor and Chair
Department of Medical Informatics & Clinical Epidemiology
Oregon Health & Science University
Portland, OR, USA
Email: hersh@ohsu.edu
Web: www.billhersh.info
Blog: http://informaticsprofessor.blogspot.com
Twitter: @williamhersh

References


Charlson, M., Wells, M., Ullman, R., King, F., & Shmukler, C. (2014). The Charlson comorbidity index can be used prospectively to identify patients who will incur high future costs. *PLoS ONE, 9*(12), e112479.


Ioannidis, J. (2005b). Why most published research findings are false. *PLoS Medicine, 2*(8), e124.


Big Data Is Not Enough: People and Systems Are Needed to Benefit Health and Biomedicine

William Hersh, MD
Professor and Chair
Department of Medical Informatics & Clinical Epidemiology
Oregon Health & Science University
Portland, OR, USA
Email: hersh@ohsu.edu
Web: www.billhersh.info
Blog: http://informaticsprofessor.blogspot.com
Twitter: @williamhersh

Big Data is not enough

- Many use cases for Big Data
- Growing quantity of data available at decreasing cost
- Much demonstration of predictive ability; less so of value
- Many caveats for different types of biomedical data
- Effective solutions require people and systems
Many use cases for Big Data in medicine (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

Growing quantity at increasingly lower cost of data

- Last half-decade has seen dramatic growth in adoption of electronic health record (EHR) by hospitals (96%) and physicians (83%) (DesRoches, 2015; Gold, 2016)
- Cost of genome sequencing has fallen faster than Moore’s Law (NHGRI, 2016)
- Proliferation of other data sources
  - Imaging
  - Wearables
  - Web and social media
Important data-related initiatives from US government

- Sync for Science (White, 2016) – http://syncfor.science
- Vital Directions for Health and Health Care (Dzau, 2016)
- Cancer Moonshot (Singer, 2016) – https://www.cancer.gov/research/key-initiatives/moonshot-cancer-initiative
- 21st Century Cures (Kesselheim, 2017)

Rationale

- Growing quantity and complexity of healthcare data through EHR capture, genomics, and other sources require more decision support (Stead, 2011)
- With shift of payment from “volume to value,” healthcare organizations will need to manage information better to deliver better care (Horner, 2012; Burwell, 2015)
- 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
Ever-growing number of studies demonstrating predictive ability

- Using EHR data to predict patients at risk for readmission (Amarasingham, 2010; Donzé, 2013; Gildersleeve, 2013; Hebert, 2014; Shadmi, 2015)
- Identifying patients who might be eligible for participation in clinical studies (Voorhees, 2012)
- Detecting postoperative complications (FitzHenry, 2013; Tien, 2015)
- Detecting potential delays in cancer diagnosis (Murphy, 2014)
- Predicting future patient costs (Charlson, 2014)

Predictive studies (cont.)

- Optimizing primary care physician panel size (Rajkomar, 2016)
- Real-time alerting of mortality risk and prolonged hospitalization from EHR data (Khurana, 2016)
- Elucidating treatment pathways for common diseases (Hripcsak, 2016)
- NLP-based case-finding algorithm of HIE data increased detection of diabetes cases (Zheng, 2016)
- The list goes on and on ...
BUT, studies demonstrating improved patient outcomes are fewer

- Readmission tool applied with case management reduced 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)
- Tool to reduce delay in cancer diagnosis led to earlier diagnosis for colorectal and prostate cancer (Murphy, 2015)

Newer studies of outcomes

- 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)
- 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
  - Early detection and treatment of sepsis
  - Monitoring and control of surgery costs and outcomes
- In cohort of children with cerebral palsy, implementation of a learning health system led to (Lowes, 2016)
  - 43% reduced hospital days
  - 30% reduction in emergency department visits
  - 210% reduction in healthcare costs
Some challenges for analytical use of clinical (EHR) data

- Data quality and accuracy is not a top priority for busy clinicians (de Lusignan, 2005)
- Data quantity can be overwhelming – average pediatric ICU patient generates 1348 information items per 24 hours (Manor-Shulman, 2008)
- Patients get care at different institutions (Bourgeois, 2010; Finnell, 2011)
- Much data is “locked” in text (Hripcsak, 2012)
- EHRs of academic medical centers not easy to combine for aggregation (Broberg, 2015)

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

Information from scientific publications can also be problematic

- Science, driven by experimentation, is the best source of truth, but just because something is written in a journal article does not mean it is true
  - Winner’s curse (Ioannidis, 2005; Young, 2008) leads to publication bias (Dwan, 2013)
  - Reproducibility (Begley, 2012; Science, 2015; Begley, 2015; Baker, 2016)
  - Clinical trials may not be representative of patient populations (Weng, 2014; Prieto-Centurion, 2014; Geifman, 2016)
  - Use of surrogate endpoints may distort efficacy (Kim, 2015)
  - Reversal (Ioannidis, 2005; Prasad, 2013; Prasad, 2015)
  - Erroneous information in reference materials (Randhawa, 2015)
  - Outright fraud not infrequent (RetractionWatch.com), may be driven by predatory publishing (Haug, 2013; Moher, 2016)
Results can be misleading, conflicting, or hyped

- Observational studies can mislead us, e.g., Women’s Health Initiative (JAMA, 2002)
- Observational studies do not discern cause and effect, e.g., diet and cancer (Schoenfeld, 2013)
- Hype about new technologies not yet fully assessed, e.g., IBM Watson – much promise but much hype (Hersh, 2013; Hersh, 2016; Schank, 2016)

Biomedical researchers are not necessarily good software engineers

- Many scientific researchers write code but are not always well-versed in best practices of testing and error detection (Merali, 2010)
- Scientists have history of relying on incorrect data or models (Sainani, 2011)
- They may also not be good about selection of best software packages for their work (Joppa, 2013)
- 3000 of 40,000 studies using fMRI may have false-positive results due to faulty algorithms and bugs (Eklund, 2016)
Should there be more sharing of scientific data? Yes, but ...

- Came to fore with ICMJE guidelines (Taichman, 2016) and NEJM “research parasites” editorial (Longo, 2016)
  - Pro: fairness to funders (taxpayers) and subjects (patients)
  - Con: researchers who carried out the heavy work need period of embargo and protection from misuse of their data (ICIFTDS, 2016); costs of curating and organizing 27K clinical trials per year; amount of actual use modest (Strom, 2016)
- Informatics issues: need for attention to standards (Kush, 2014); workflows, patient engagement (Tennenbaum, 2016)

Other concerns

- Boyd (2012) – critical questions for Big Data
  - Big Data changes the definition of knowledge
  - Claims to objectivity and accuracy are misleading
  - Bigger data are not always better data
  - Taken out of context, Big Data loses its meaning
  - Just because it is accessible does not make it ethical
  - Limited access to Big Data creates new digital divides
- Fung (2014) – Big Data is OCCAM
  - Observational
  - Lacking Controls
  - Seemingly Complete
  - Adapted
  - Merged
- Big Data not neutral; reflects our values and priorities (Richards, 2014; Barocas, 2015)
Big Data requires more than the data; also takes people

• Data scientists – the “sexiest profession of the 21st century” (Davenport, 2012)
• 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

Big Data also requires systems

• 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 users of data (Krumholz, 2014)
Some axes to grind

• Is data science really new or different?
  – Statisticians (Donoho, 2016) and informaticians (Hersh, 2015) have been doing some of this for a long time

• Will Big Data transform medicine?
  – In some areas, but need more demonstration of value than ability to predict

• How can we optimize its use?
  – Research focused on its applications and their outcomes
  – Don’t oversell it, especially to clinicians

Much promise for Big Data in Health and Biomedicine, 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 and systems to apply and disseminate