Big Data Science and Analytics in Health and Biomedicine

William Hersh, MD, FACP, FACMI
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

References

Adams, J and Klein, J (2011). Business Intelligence and Analytics in Health Care - A Primer. Washington, DC, The Advisory Board Company. http://www.advisory.com/Research/IT-Strategy-Council/Research-Notes/2011/Business-Intelligence-and-Analytics-in-Health-Care

Afzal, Z, Engelkes, M, et al. (2013). Automatic generation of case-detection algorithms to identify children with asthma from large electronic health record databases. *Pharmacoepidemiology and Drug Safety*. 22: 826-833.

Aggarwal, CC and Zhai, C, Eds. (2012). Mining Text Data. New York, NY, Springer.

Alvarez, CA, Clark, CA, et al. (2013). Predicting out of intensive care unit cardiopulmonary arrest or death using electronic medical record data. *BMC Medical Informatics & Decision Making*. 13: 28. http://www.biomedcentral.com/1472-6947/13/28

Amarasingham, R, Moore, BJ, et al. (2010). An automated model to identify heart failure patients at risk for 30-day readmission or death using electronic medical record data. *Medical Care*. 48: 981-988.

Anonymous (2011). Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease. Washington, DC, National Academies Press.

Anonymous (2012). The value of analytics in healthcare - From insights to outcomes. Somers, NY, IBM Global Services. http://www-935.ibm.com/services/us/gbs/thoughtleadership/ibv-healthcare-analytics.html

Anonymous (2013). Readmissions Reduction Program. Washington, DC, Center for Medicare and Medicaid Services. http://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/AcuteInpatientPPS/Readmissions-Reduction-Program.html

Anonymous (2015). Big Data Now: 2014 Edition. Sebastopol, CA, O'Reilly Media.

Ashley, EA (2015). The Precision Medicine Initiative - a new national effort. *Journal of the American Medical Association*. 313: 2119-2120.

Bates, DW, Saria, S, et al. (2014). Big data in health care: using analytics to identify and manage high-risk and high-cost patients. *Health Affairs*. 33: 1123-1131.

Bellazzi, R and Zupan, B (2008). Predictive data mining in clinical medicine: current issues and guidelines. *International Journal of Medical Informatics*. 77: 81-97.

Bourgeois, FC, Olson, KL, et al. (2010). Patients treated at multiple acute health care facilities: quantifying information fragmentation. *Archives of Internal Medicine*. 170: 1989-1995.

Buneman, P and Davidson, SB (2010). Data provenance – the foundation of data quality. Pittsburgh, PA, Carnegie Mellon University Software Engineering Institute. http://www.sei.cmu.edu/measurement/research/upload/Davidson.pdf

Burke, J (2013). Health Analytics: Gaining the Insights to Transform Health Care. Hoboken, NJ, Wiley.

Charlson, M, Wells, MT, et al. (2014). The Charlson comorbidity index can be used prospectively to identify patients who will incur high future costs. *PLoS ONE*. 9(12): e112479. http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0112479
Coiera, E, Wang, Y, et al. (2014). Predicting the cumulative risk of death during hospitalization by modeling weekend, weekday and diurnal mortality risks. *BMC Health Services Research*. 14: 226. http://www.biomedcentral.com/1472-6963/14/226
Collins, FS and Varmus, H (2015). A new initiative on precision medicine. *New England Journal of Medicine*. 372: 793-795.
Crown, WH (2015). Potential application of machine learning in health outcomes research and some statistical cautions. *Value Health*. 18: 137-140.
Davenport, TH (2012). Analytical integration in healthcare. Analytics, January/February 2012. http://www.analytics-magazine.org/januaryfebruary-2012/504-analytical-integration-in-healthcare
Davenport, TH and Harris, JG (2007). Competing on Analytics: The New Science of Winning. Cambridge, MA, Harvard Business School Press.
deLusignan, S and vanWeel, C (2005). The use of routinely collected computer data for research in primary care: opportunities and challenges. *Family Practice*. 23: 253-263.
Dhar, V (2013). Data science and prediction. *Communications of the ACM*. 56(12): 64-73.
Diamond, CC, Mostashari, F, et al. (2009). Collecting and sharing data for population health: a new paradigm. *Health Affairs*. 28: 454-466.
Donzé, J, Aujesky, D, et al. (2013). Potentially avoidable 30-day hospital readmissions in medical patients: derivation and validation of a prediction model. *JAMA Internal Medicine*. 173: 632-638.
Finnell, JT, Overhage, JM, et al. (2011). All health care is not local: an evaluation of the distribution of emergency department care delivered in Indiana. *AMIA Annual Symposium Proceedings*, Washington, DC. 409-416.
FitzHenry, F, Murff, HJ, et al. (2013). Exploring the frontier of electronic health record surveillance: the case of postoperative complications. *Medical Care*. 51: 509-516.
Flach, P (2012). Machine Learning: The Art and Science of Algorithms that Make Sense of Data. Cambridge, England, Cambridge University Press.
Gardner, E (2013). The HIT Approach to Big Data. Health Data Management, March 1, 2013. http://www.healthdatamanagement.com/issues/21_3/The-HIT-Approach-to-Big-Data-Anayltics-45735-1.html
Gensinger, RA, Ed. (2014). Analytics in Healthcare: An Introduction. Chicago, IL, Healthcare Information Management Systems Society.
Gildersleeve, R and Cooper, P (2013). Development of an automated, real time surveillance tool for predicting readmissions at a community hospital. *Applied Clinical Informatics*. 4: 153-169.
Gottlieb, A, Stein, GY, et al. (2013). A method for inferring medical diagnoses from patient similarities. *BMC Medicine*. 11: 194. http://www.biomedcentral.com/1741-7015/11/194
Grus, J (2015). Data Science From Scratch. Sebastopol, CA, O'Reilly Media.
Halamka, J (2013). The "Post EHR" Era. Life as a Healthcare CIO, February 12, 2013. http://geekdoctor.blogspot.com/2013/02/the-post-ehr-era.html
Hamburg, MA and Collins, FS (2010). The path to personalized medicine. *New England Journal of Medicine*. 363: 301-304.
Hersh, W (2012). From Implementation to Analytics: The Future Work of Informatics. Informatics Professor, April 1, 2012. http://informaticsprofessor.blogspot.com/2012/04/from-implementation-to-analytics-future.html
Hersh, WR (2007). The full spectrum of biomedical informatics education at Oregon Health & Science University. *Methods of Information in Medicine*. 46: 80-83.

Hersh, WR (2014). Healthcare Data Analytics. Health Informatics: Practical Guide for Healthcare and Information Technology Professionals, Sixth Edition. R. Hoyt and A. Yoshihashi. Pensacola, FL, Lulu.com: 62-75.
Horner, P and Basu, A (2012). Analytics & the future of healthcare. Analytics, January/February 2012. http://www.analytics-magazine.org/januaryfebruary-2012/503-analytics-a-the-future-of-healthcare
Hripcsak, G and Albers, DJ (2012). Next-generation phenotyping of electronic health records. *Journal of the American Medical Informatics Association*. 20: 117-121.
Kellermann, AL and Jones, SS (2013). What will it take to achieve the as-yet-unfulfilled promises of health information technology? *Health Affairs*. 32: 63-68.
Köpcke, F, Lubgan, D, et al. (2013). Evaluating predictive modeling algorithms to assess patient eligibility for clinical trials from routine data. *BMC Medical Informatics & Decision Making*. 13: 134. http://www.biomedcentral.com/1472-6947/13/134
Kosinski, M, Stillwell, D, et al. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*. 110: 5802-5805.
Krumholz, HM (2014). Big data and new knowledge in medicine: the thinking, training, and tools needed for a learning health system. *Health Affairs*. 33: 1163-1170.
Kumar, A, Niu, F, et al. (2013). Hazy: making it easier to build and maintain big-data analytics. *Communications of the ACM*. 56(3): 40-49.
Levy, S (2014). How the NSA Almost Killed the Internet. Wired, January 7, 2014. http://www.wired.com/threatlevel/2014/01/how-the-us-almost-killed-the-internet/
Lewis, M (2004). Moneyball: The Art of Winning an Unfair Game. New York, NY, W. W. Norton & Company.
Longhurst, CA, Harrington, RA, et al. (2014). A 'green button' for using aggregate patient data at the point of care. *Health Affairs*. 33: 1229-1235.
Makam, AN, Nguyen, OK, et al. (2013). Identifying patients with diabetes and the earliest date of diagnosis in real time: an electronic health record case-finding algorithm. *BMC Medical Informatics & Decision Making*. 13: 81. http://www.biomedcentral.com/1472-6947/13/81
Manor-Shulman, O, Beyene, J, et al. (2008). Quantifying the volume of documented clinical information in critical illness. *Journal of Critical Care*. 23: 245-250.
Marconi, K and Lehmann, H, Eds. (2014). Big Data and Health Analytics. Boca Raton, FL, CRC Press.
Mathias, JS, Agrawal, A, et al. (2013). Development of a 5 year life expectancy index in older adults using predictive mining of electronic health record data. *Journal of the American Medical Informatics Association*. 20(e1): e118-e124.
Murphy, DR, Laxmisan, A, et al. (2014). Electronic health record-based triggers to detect potential delays in cancer diagnosis. *BMJ Quality & Safety*. 23: 8-16.
Nijhawan, AE, Clark, C, et al. (2012). An electronic medical record-based model to predict 30-day risk of readmission and death among HIV-infected inpatients. *Journal of Acquired Immunodeficiency Syndrome*. 61: 349-358.
O'Reilly, T, Loukides, M, et al. (2012). How Data Science Is Transforming Health Care - Solving the Wanamaker Dilemma. Sebastopol, CA, O'Reilly Media.
Ruiz, EJ, Hristidis, V, et al. (2012). Correlating financial tme series with micro-blogging activity. *Fifth ACM International Conference on Web Search & Data Mining*, Seattle, WA http://www.cs.ucr.edu/~vagelis/publications/wsdm2012-microblog-financial.pdf
Salant, JD and Curtis, L (2012). Nate Silver-Led Statistics Men Crush Pundits in Election. Bloomberg, November 7, 2012. http://www.bloomberg.com/news/2012-11-07/nate-silver-led-statistics-men-crush-pundits-in-election.html
Scherer, M (2012). Inside the Secret World of the Data Crunchers Who Helped Obama Win. Time, November 7, 2012. http://swampland.time.com/2012/11/07/inside-the-secret-world-of-quants-and-data-crunchers-who-helped-obama-win/

Shadmi, E, Flaks-Manov, N, et al. (2015). Predicting 30-day readmissions with preadmission electronic health record data. *Medical Care*. 53: 283-289.
Singal, AG, Rahimi, RS, et al. (2013). An automated model using electronic medical record data identifies patients with cirrhosis at high risk for readmission. *Clinical Gastroenterology and Hepatology*. 11: 1335-1341.
Smith, M (2014). Targeted: How Technology Is Revolutionizing Advertising and the Way Companies Reach Consumers. Washington, DC, AMACOM.
Stead, WW, Searle, JR, et al. (2011). Biomedical informatics: changing what physicians need to know and how they learn. *Academic Medicine*. 86: 429-434.
Tai-Seale, M, Wilson, CJ, et al. (2013). Leveraging electronic health records to develop measurements for processes of care. *Health Services Research*. 49: 628-644.
Voorhees, E and Hersh, W (2012). Overview of the TREC 2012 Medical Records Track. *The Twenty-First Text REtrieval Conference Proceedings (TREC 2012)*, Gaithersburg, MD. National Institute of Standards and Technology http://trec.nist.gov/pubs/trec21/papers/MED12OVERVIEW.pdf
Winslow, RL, Trayanova, N, et al. (2012). Computational medicine: translating models to clinical care. *Science Translational Medicine*. 4: 158rv111.
http://stm.sciencemag.org/content/4/158/158rv11.short
Zaki, MJ and Meira, W (2014). Data Mining and Analysis: Fundamental Concepts and Algorithms. Cambridge, England, Cambridge University Press.
Zikopoulos, P, Eaton, C, et al. (2011). Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data. New York, NY, McGraw-Hill.

Big Data Science and Analytics in Health and Biomedicine

William Hersh, MD, FACP, FACMI
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

OREGON HEALTH & SCIENCE UNIVERSITY

1

Big Data Science and Analytics in Healthcare

- Rationale
- Definitions
- Applications
- Results
- Challenges
- Workforce
- Further study

OREGON HEALTH & SCIENCE UNIVERSITY

2

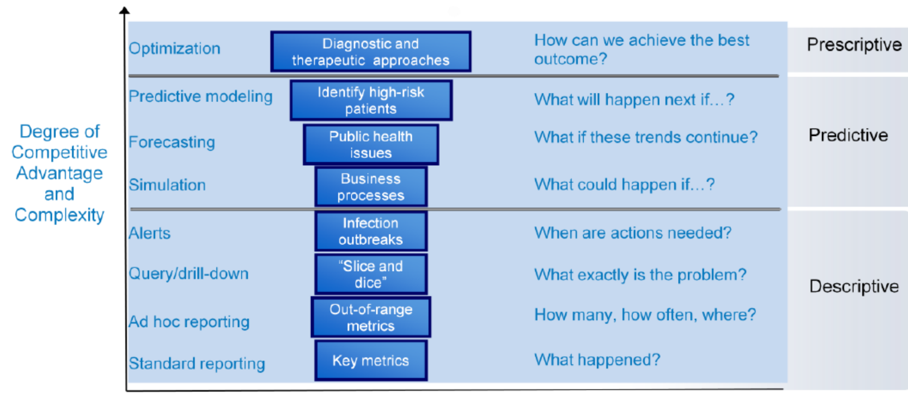
Rationale

- Although focus in recent years has been on electronic health record (EHR) implementation and "meaningful use," informatics work in the future will shift to putting the data and information to good use (Hersh, 2012)
- As the quantity <u>and</u> complexity of healthcare data grow through EHR capture, genomics, and other sources, the number of facts per clinical decision will increase, requiring increasing help for decision-makers (Stead, 2011)

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)



5

Related terms

- Machine learning – area of computer science focused on systems and algorithms that learn from data (Flach, 2012; Crown, 2015)
- 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)
- Big data – data of growing volume, velocity, variety, and veracity (Zikopolous, 2011; O'Reilly, 2015)
  - e.g., ~9 petabytes of data of Kaiser-Permanente (Gardner, 2013)

6

Related terms (cont.)

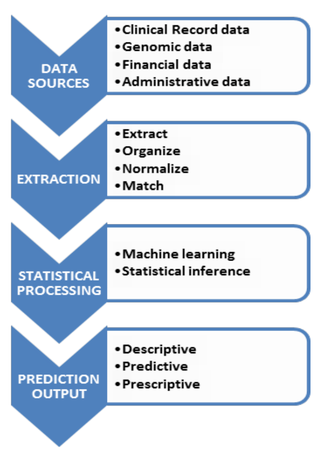
- Data science – distinguished from statistics by understanding of varying types and how to manipulate and leverage (Dhar, 2013; Grus, 2015)
- Data provenance – origin and trustworthiness (Buneman, 2010)
- Business intelligence – use of data to obtain timely, valuable insights into business and clinical data (Adams, 2011)
- Personalized (Hamburg, 2010), precision (IOM, 2011; Collins, 2015; Ashley, 2015), or computational medicine (Winslow, 2012)

7

OREGON HEALTH & SCIENCE UNIVERSITY

Analytics pipeline

- Adapted from Kumar (2013) for healthcare (Hersh, 2014)



8

OREGON HEALTH & SCIENCE UNIVERSITY

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)
- 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" is a growing area (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)

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)
  - 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 (Burke, 2013; Gensinger, 2014; Marconi, 2014)

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; Shadmi, 2015)

OREGON HEALTH & SCIENCE UNIVERSITY

Applications of analytics – identifying other clinical situations

- Predicting 30-day risk of readmission and death among HIV-infected inpatients (Nijhawan, 2012)
- Identification of children with asthma (Afzal, 2013)
- Detecting postoperative complications (FitzHenry, 2013)
- Measuring processes of care (Tai-Seale, 2013)
- Determining five-year life expectancy (Mathias, 2013)
- Detecting potential delays in cancer diagnosis (Murphy, 2014)
- Identifying patients with cirrhosis at high risk for readmission (Singal, 2013)
- Predicting out of intensive care unit cardiopulmonary arrest or death (Alvarez, 2013)
- Predicting hospital death by day or time of day (Coiera, 2014)
- Predicting future patient costs (Charlson, 2014)

OREGON HEALTH & SCIENCE UNIVERSITY

Applications of analytics – patient identification and diagnosis

- Identifying patients who might be eligible for participation in clinical studies (Voorhees, 2012)
- Determining eligibility for clinical trials (Köpcke, 2013)
- Identifying patients with diabetes and the earliest date of diagnosis (Makam, 2013)
- Predicting diagnosis in new patients (Gottlieb, 2013)

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

OREGON HEALTH & SCIENCE UNIVERSITY

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)

OREGON HEALTH & SCIENCE UNIVERSITY

Challenges for analytical use of clinical data

- Data quality and accuracy is not a top priority for busy clinicians (de Lusignan, 2005)
- Patients get care at different places (Bourgeois, 2010; Finnell, 2011)
- Standards and interoperability – mature approaches but lack of widespread adoption (Kellermann, 2013)
- Much data is "locked" in text (Hripcsak, 2012)
- Average pediatric ICU patient generates 1348 information items per 24 hours (Manor-Shulman, 2008)

OREGON HEALTH & SCIENCE UNIVERSITY

How can I learn more in Oregon? Study informatics?

- Many educational opportunities at a variety of levels, mostly graduate
  - http://www.amia.org/informatics-academic-training-programs
- OHSU program one of largest and well-established (Hersh, 2007)
  - http://www.ohsu.edu/informatics-education
  - Graduate level programs at Certificate, Master's, and PhD levels
  - "Building block" approach allows courses to be carried forward to higher levels

17

OREGON HEALTH & SCIENCE UNIVERSITY

Conclusions

- There are plentiful opportunities for data analytics in healthcare
- We must be cognizant of caveats of using operational clinical data
- We must implement best practices for using such data
- There are many career opportunities in healthcare data analytics

18

OREGON HEALTH & SCIENCE UNIVERSITY