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

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.
Amarasingham, R, Patel, PC, et al. (2013). Allocating scarce resources in real-time to reduce heart failure readmissions: a prospective, controlled study. *BMJ Quality & Safety*. 22: 998-1005.
Amarasingham, R, Patzer, RE, et al. (2014). Implementing electronic health care predictive analytics: considerations and challenges. *Health Affairs*. 33: 1148-1154.
Anonymous (2002). Risks and benefits of estrogen plus progestin in healthy postmenopausal women - principal results from the Women's Health Initiative randomized controlled trial. *Journal of the American Medical Association*. 288: 321-333.
Anonymous (2014). IDC Reveals Worldwide Big Data and Analytics Predictions for 2015. Framingham, MA, International Data Corporation. http://bit.ly/IDCBigDataFutureScape2015
Anonymous (2015). Estimating the reproducibility of psychological science. *Science*. 349: aac4716. http://science.sciencemag.org/content/349/6251/aac4716
Anonymous (2016). The Cost of Sequencing a Human Genome. Bethesda, MD, National Human Genome Research Institute. http://www.genome.gov/sequencingcosts/
Anonymous (2016). Toward fairness in data sharing. *New England Journal of Medicine*. 375: 405-407.
Baker, M (2016). 1,500 scientists lift the lid on reproducibility. *Nature*. 533: 452-454.
Barocas, S and Selbst, AD (2015). Big data's disparate impact. *California Law Review*. 104: 2016. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2477899
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.
Begley, CG and Ellis, LM (2012). Raise standards for preclinical cancer research. *Nature*. 483: 531-533.
Begley, CG and Ioannidis, JPA (2015). Reproducibility in science - improving the standard for basic and preclinical research. *Circulation Research*. 116: 116-126.

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.
Bourne, PE, Lorsch, JR, et al. (2015). Sustaining the big-data ecosystem. *Nature*. 527: S16-S17.
Boyd, D and Crawford, K (2012). Critical Questions for Big Data. *Information, Communication & Society*. 15: 662-679.
Broberg, C, Sklenar, J, et al. (2015). Feasibility of using electronic medical record data for tracking quality indicators in adults with congenital heart disease. *Congenital Heart Disease*. 10: E268-E277.
Burwell, SM (2015). Setting value-based payment goals - HHS efforts to improve U.S. health care. *New England Journal of Medicine*. 372: 897-899.
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
Cho, I, Park, I, et al. (2013). Using EHR data to predict hospital-acquired pressure ulcers: A prospective study of a Bayesian Network model. *International Journal of Medical Informatics*. 82: 1059-1067.
Collins, FS and Varmus, H (2015). A new initiative on precision medicine. *New England Journal of Medicine*. 372: 793-795.
Davenport, TH and Patil, DJ (2012). Data Scientist: The Sexiest Job of the 21st Century. Harvard Business Review, October, 2012. http://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/
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.
DesRoches, CM, Painter, MW, et al. (2015). Health Information Technology in the United States 2015 - Transition to a Post-HITECH World. Princeton, NJ, Robert Wood Johnson Foundation. http://www.rwjf.org/en/library/research/2015/09/health-information-technology-in-the-united-states-2015.html
Donoho, D (2015). 50 years of Data Science. Princeton NJ, Tukey Centennial Workshop. https://dl.dropboxusercontent.com/u/23421017/50YearsDataScience.pdf
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.
Dwan, K, Gamble, C, et al. (2013). Systematic review of the empirical evidence of study publication bias and outcome reporting bias - an updated review. *PLoS ONE*. 8(7): e66844. http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0066844
Dzau, VJ, McClellan, M, et al. (2016). Vital Directions for Health and Health Care: an initiative of the National Academy of Medicine. *Journal of the American Medical Association*. 316: 711-712.
Eklund, A, Nichols, TD, et al. (2016). Cluster failure: why fMRI inferences for spatial extent have inflated false-positive rates. *Proceedings of the National Academy of Sciences*. 113: 7900–7905.
Erskine, AR, Karunakaran, P, et al. (2016). How Geisinger Health System Uses Big Data to Save Lives. *Harvard Business Review,* https://hbr.org/2016/12/how-geisinger-health-system-uses-big-data-to-save-lives

Evans, RS, Benuzillo, J, et al. (2016). Automated identification and predictive tools to help identify high-risk heart failure patients: pilot evaluation. *Journal of the American Medical Informatics Association*: Epub ahead of print.
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.
Fung, K (2014). Google Flu Trends' Failure Shows Good Data > Big Data. Harvard Business Review, March 25, 2014. https://hbr.org/2014/03/google-flu-trends-failure-shows-good-data-big-data/
Geifman, N and Butte, AJ (2016). Do cancer clinical trial populations truly represent cancer patients? A comparison of open clinical trials to the Cancer Genome Atlas. *Pacific Symposium on Biocomputing*, Kohala Coast, HI. 309-320. http://www.worldscientific.com/doi/10.1142/9789814749411_0029
Gilbert, P, Rutland, MD, et al. (2013). Redesigning the work of case management: testing a predictive model for readmission. *American Journal of Managed Care*. 19(11 Spec No. 10): eS19-eSP25. http://www.ajmc.com/publications/issue/2013/2013-11-vol19-sp/redesigning-the-work-of-case-management-testing-a-predictive-model-for-readmission
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.
Gold, M and McLaughlin, C (2016). Assessing HITECH implementation and lessons: 5 years later. *Milbank Quarterly*. 94: 654-687.
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
Haug, C (2013). The downside of open-access publishing. *New England Journal of Medicine*. 368: 791-793.
Hebert, C, Shivade, C, et al. (2014). Diagnosis-specific readmission risk prediction using electronic health data: a retrospective cohort study. *BMC Medical Informatics & Decision Making*. 14: 65. http://www.biomedcentral.com/1472-6947/14/65
Hersh, W (2013). What is a Thinking Informatician to Think of IBM's Watson? Informatics Professor. http://informaticsprofessor.blogspot.com/2013/06/what-is-thinking-informatician-to-think.html
Hersh, W (2016). Generalizability and Reproducibility of Scientific Literature and the Limits to Machine Learning. Informatics Professor.
Hersh, WR (2015). What is the Difference (If Any) Between Informatics and Data Science? Informatics Professor. http://informaticsprofessor.blogspot.com/2015/07/what-is-difference-if-any-between.html
Hersh, WR, Weiner, MG, et al. (2013). Caveats for the use of operational electronic health record data in comparative effectiveness research. *Medical Care*. 51(Suppl 3): S30-S37.
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.

Hripcsak, G, Ryan, PB, et al. (2016). Characterizing treatment pathways at scale using the OHDSI network. *Proceedings of the National Academy of Sciences*. 113: 7329-7336.
Ioannidis, JP (2005). Contradicted and initially stronger effects in highly cited clinical research. *Journal of the American Medical Association*. 294: 218-228.
Ioannidis, JP (2005). Why most published research findings are false. *PLoS Medicine*. 2(8): e124. http://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.0020124
Joppa, LN, McInerny, G, et al. (2013). Troubling trends in scientific software use. *Science*. 340: 814-815.
Kesselheim, AS and Avorn, J (2017). New "21st Century Cures" legislation: speed and ease vs science. *Journal of the American Medical Association*: Epub ahead of print.
Khurana, HS, Groves, RH, et al. (2016). Real-time automated sampling of electronic medical records predicts hospital mortality. *American Journal of Medicine*. 129: 688-698.
Kim, C and Prasad, V (2015). Strength of validation for surrogate end points used in the US Food and Drug Administration's approval of oncology drugs. *Mayo Clinic Proceedings*: Epub ahead of print.
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.
Kush, R and Goldman, M (2014). Fostering responsible data sharing through standards. *New England Journal of Medicine*. 370: 2163-2165.
Longo, DL and Drazen, JM (2016). Data sharing. *New England Journal of Medicine*. 374: 276-277.
Lowes, LP, Noritz, GH, et al. (2016). 'Learn From Every Patient': implementation and early results of a learning health system. *Developmental Medicine & Child Neurology*. 59: 183-191.
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.
Manyika, J, Chui, M, et al. (2011). Big data: The next frontier for innovation, competition, and productivity, McKinsey Global Institute. http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation
Merali, Z (2010). Computational science: ...Error. *Nature*. 467: 775-777.
Moher, D and Moher, E (2016). Stop predatory publishers now: act collaboratively. *Annals of Internal Medicine*. 164: 616-617.
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.
Murphy, DR, Wu, L, et al. (2015). Electronic trigger-based intervention to reduce delays in diagnostic evaluation for cancer: a cluster randomized controlled trial. *Journal of Clinical Oncology*. 33: 3560-3567.
Prasad, V, Kim, C, et al. (2015). The strength of association between surrogate end points and survival in oncology: a systematic review of trial-level meta-analyses. *JAMA Internal Medicine*. 175: 1389-1398.
Prasad, V, Vandross, A, et al. (2013). A decade of reversal: an analysis of 146 contradicted medical practices. *Mayo Clinic Proceedings*. 88: 790-798.
Prasad, VK and Cifu, AS (2015). Ending Medical Reversal: Improving Outcomes, Saving Lives. Baltimore, MD, Johns Hopkins University Press.

Prieto-Centurion, V, Rolle, AJ, et al. (2014). Multicenter study comparing case definitions used to identify patients with chronic obstructive pulmonary disease. *American Journal of Respiratory and Critical Care Medicine*. 190: 989-995.
Rajkomar, A, Yim, JW, et al. (2016). Weighting primary care patient panel size: a novel electronic health record-derived measure using machine learning. *JMIR Medical Informatics*. 4(4): e29. http://medinform.jmir.org/2016/4/e29/
Randhawa, AS, Babalola, O, et al. (2016). A collaborative assessment among 11 pharmaceutical companies of misinformation in commonly used online drug information compendia. *Annals of Pharmacotherapy*. 50: 352-359.
Richards, NM and King, JH (2014). Big data ethics. *Wake Forest Law Review*. 49: 393-432. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2384174
Rothman, M, Rimar, J, et al. (2015). Mortality reduction associated with proactive use of EMR-based acuity score by an RN team at an urban hospital. *BMJ Quality & Safety*. 24: 734-735. http://qualitysafety.bmj.com/content/24/11/734.abstract
Sainani, K (2011). Error! – What Biomedical Computing Can Learn From Its Mistakes. Biomedical Computation Review, September 1, 2011. http://biomedicalcomputationreview.org/content/error-%E2%80%93-what-biomedical-computing-can-learn-its-mistakes
Schank, R (2016). The fraudulent claims made by IBM about Watson and AI. They are not doing "cognitive computing" no matter how many times they say they are. Roger Schank. http://www.rogerschank.com/fraudulent-claims-made-by-IBM-about-Watson-and-AI
Schoenfeld, JD and Ioannidis, JPA (2013). Is everything we eat associated with cancer? A systematic cookbook review. *American Journal of Clinical Nutrition*. 97: 127-134.
Shadmi, E, Flaks-Manov, N, et al. (2015). Predicting 30-day readmissions with preadmission electronic health record data. *Medical Care*. 53: 283-289.
Singer, DS, Jacks, T, et al. (2016). A U.S. "Cancer Moonshot" to accelerate cancer research. *Science*. 353: 1105-1106.
Stead, WW, Searle, JR, et al. (2011). Biomedical informatics: changing what physicians need to know and how they learn. *Academic Medicine*. 86: 429-434.
Strom, BL, Buyse, ME, et al. (2016). Data sharing — is the juice worth the squeeze? *New England Journal of Medicine*. 375: 1608-1609.
Taichman, DB, Backus, J, et al. (2016). Sharing clinical trial data: a proposal from the International Committee of Medical Journal Editors. *New England Journal of Medicine*. 374: 384-386.
Tenenbaum, JD, Avillach, P, et al. (2016). An informatics research agenda to support precision medicine: seven key areas. *Journal of the American Medical Informatics Association*. 23: 791-795.
Tien, M, Kashyap, R, et al. (2015). Retrospective derivation and validation of an automated electronic search algorithm to identify post operative cardiovascular and thromboembolic complications. *Applied Clinical Informatics*. 6: 565-576.
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

Weng, C, Li, Y, et al. (2014). A distribution-based method for assessing the differences between clinical trial target populations and patient populations in electronic health records. *Applied Clinical Informatics*. 5: 463-479.
White, J, Briggs, J, et al. (2016). NIH and ONC Launch the Sync for Science (S4S) Pilot: Enabling Individual Health Data Access and Donation. Health IT Buzz. https://www.healthit.gov/buzz-blog/health-innovation/nih-and-onc-launch-the-sync-for-science-pilot/
Young, NS, Ioannidis, JP, et al. (2008). Why current publication practices may distort science. *PLoS Medicine*. 5(10): e201. http://www.plosmedicine.org/article/info%3Adoi%2F10.1371%2Fjournal.pmed.0050201
Zheng, L, Wang, Y, et al. (2016). Web-based real-time case finding for the population health management of patients with diabetes mellitus: a prospective validation of the natural language processing–based algorithm with statewide electronic medical records. *JMIR Medical Informatics*. 4(4): e37.

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

Thanks for the invite to a "sister" city!

KEEP PORTLAND WIRED!

KEEP PORTLAND BEERED!

Big Data is not enough

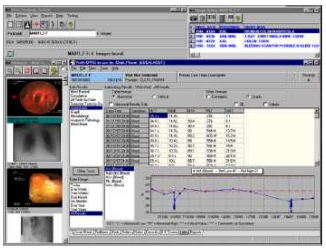
- Many use cases for Big Data
- Growing quantity of data available at decreasing cost
- Much demonstration of predictive ability; less so of value
- Many caveats for different types of biomedical data
- Effective solutions require people and systems

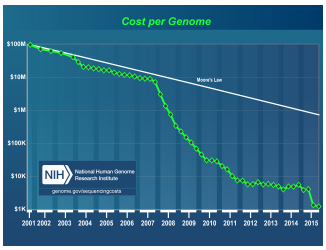
Many use cases for Big Data in medicine (Bates, 2014)

- 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





5

Important data-related initiatives from US government

- Big Data to Knowledge (BD2K) (Bourne, 2015) – https://datascience.nih.gov
- Sync for Science (White, 2016) – http://syncfor.science
- Vital Directions for Health and Health Care (Dzau, 2016)
- Precision Medicine Initiative (Collins, 2015) – https://www.nih.gov/research-training/allofus-research-program
- Cancer Moonshot (Singer, 2016) – https://www.cancer.gov/research/key-initiatives/moonshot-cancer-initiative
- 21st Century Cures (Kesselheim, 2017)

6

OHSU

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

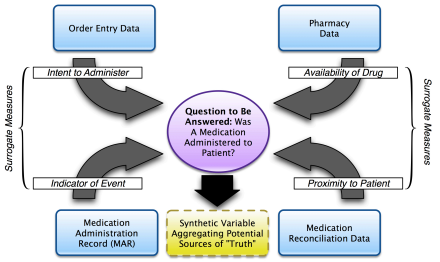
- Use of predictive report based on NLP tool reduced time in discharge planning meetings and 30-day all-cause mortality although not cost or readmissions (Evans, 2016)
- Development and use of a universal data architecture at Geisinger has led to successes in (Erskine, 2016)
  - Closing loop on appropriate 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



OHSU

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

OHSU

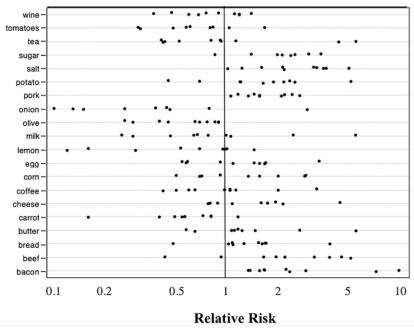
Information from scientific publications can also be problematic

- 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)

15

Results can be misleading, conflicting, or hyped

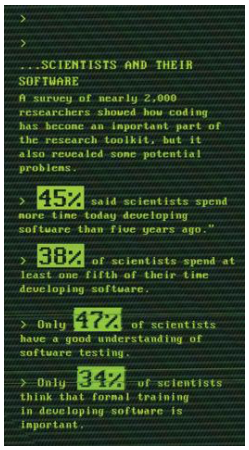
- Observational studies can mislead us, e.g., Women's Health Initiative (JAMA, 2002)
- Observational studies do not discern cause and effect, e.g., diet and cancer (Schoenfeld, 2013)
- Hype about new technologies not yet fully assessed, e.g., IBM Watson – much promise but much hype (Hersh, 2013; Hersh,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



16

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)



17

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)

18

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
  - **O**bservational
  - Lacking **C**ontrols
  - Seemingly **C**omplete
  - **A**dapted
  - **M**erged
- 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

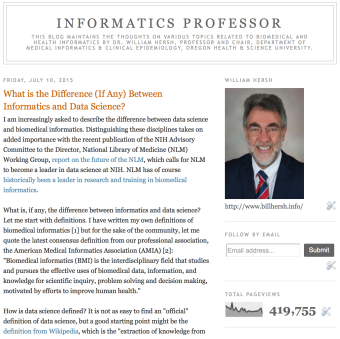
Big Data also requires systems

- 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)

OHSU

Some axes to grind

- Is data science really new or different?
  - Statisticans (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



INFORMATICS PROFESSOR

THIS BLOG MAINTAINS THE THOUGHTS ON VARIOUS TOPICS RELATED TO BIOMEDICAL AND HEALTH INFORMATICS BY DR. WILLIAM HERSH, PROFESSOR AND CHAIR, DEPARTMENT OF MEDICAL INFORMATICS & CLINICAL EPIDEMIOLOGY, OREGON HEALTH & SCIENCE UNIVERSITY.

FRIDAY, JULY 10, 2015

What is the Difference (If Any) Between Informatics and Data Science?

I am increasingly asked to describe the difference between data science and biomedical informatics. Distinguishing these disciplines takes on added importance with the recent publication of the NIH Advisory Committee to the Director, National Library of Medicine (NLM) Working Group, report on the future of the NLM, which calls for NLM to become a leader in data science at NIH. NLM has of course historically been a leader in research and training in biomedical informatics.

What is, if any, the difference between informatics and data science? Let me start with definitions. I have written my own definitions of biomedical informatics [1] but for the sake of the community, let me quote the latest consensus definition from our professional association, the American Medical Informatics Association (AMIA) [2]: "Biomedical informatics (BMI) is the interdisciplinary field that studies and pursues the effective uses of biomedical data, information, and knowledge for scientific inquiry, problem solving and decision making, motivated by efforts to improve human health."

How is data science defined? It is not as easy to find an "official" definition of data science, but a good starting point might be the definition from Wikipedia, which is the "extraction of knowledge from

WILLIAM HERSH

http://www.billhersh.info/

FOLLOW BY EMAIL

Email address... Submit

TOTAL PAGEVIEWS

419,755

OHSU

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