# Challenges for Information Retrieval and Text Mining in Biomedicine: Imperatives for Systems and Their Evaluation

William Hersh
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

Anonymous (2014). Connecting Health and Care for the Nation: A 10-Year Vision to Achieve an Interoperable Health IT Infrastructure. Washington, DC, Department of Health and Human Services. http://www.healthit.gov/sites/default/files/ONC10yearInteroperabilityConceptPaper.pdf
Anonymous (2015). Estimating the reproducibility of psychological science. *Science*. 349: aac4716. http://science.sciencemag.org/content/349/6251/aac4716
Anonymous (2016). Result and Artifact Review and Badging. New York, NY, Association of Computing Machinery. http://www.acm.org/publications/policies/artifact-review-badging/
Baker, M (2016). 1,500 scientists lift the lid on reproducibility. *Nature*. 533: 452-454.
Barkhuysen, P, deGrauw, W, et al. (2014). Is the quality of data in an electronic medical record sufficient for assessing the quality of primary care? *Journal of the American Medical Informatics Association*. 21: 692-698.
Bayley, KB, Belnap, T, et al. (2013). Challenges in using electronic health record data for CER: experience of four learning organizations. *Medical Care*. 51: S80-S86.
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.
Brennan, L, Watson, M, et al. (2012). The importance of knowing context of hospital episode statistics when reconfiguring the NHS. *British Medical Journal*. 344: e2432. http://www.bmj.com/content/344/bmj.e2432
Broberg, CS, Mitchell, J, et al. (2015). Electronic medical record integration with a database for adult congenital heart disease: early experience and progress in automating multicenter data collection. *International Journal of Cardiology*. 196: 178-182.

Cerrato, P (2012). IBM Watson Finally Graduates Medical School. Information Week, October 23, 2012. http://www.informationweek.com/healthcare/clinical-systems/ibm-watson-finally-graduates-medical-sch/240009562
Curfman, GD, Morrissey, S, et al. (2005). Expression of concern: Bombardier et al., "Comparison of upper gastrointestinal toxicity of rofecoxib and naproxen in patients with rheumatoid arthritis". *New England Journal of Medicine*. 353: 2318-2319.
D'Amore, JD, Mandel, JC, et al. (2014). Are Meaningful Use Stage 2 certified EHRs ready for interoperability? Findings from the SMART C-CDA Collaborative. *Journal of the American Medical Informatics Association*. 21: 1060-1068.
deLusignan, S, Chan, T, et al. (2005). The roles of policy and professionalism in the protection of processed clinical data: a literature review. *International Journal of Medical Informatics*. 76: 261-268.
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
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.
Ferrucci, D, Brown, E, et al. (2010). Building Watson: an overview of the DeepQA Project. *AI Magazine*. 31(3): 59-79. http://www.aaai.org/ojs/index.php/aimagazine/article/view/2303
Ferrucci, D, Levas, A, et al. (2012). Watson: beyond Jeopardy! *Artificial Intelligence*. 199-200: 93-105.
Ferrucci, DA (2012). Introduction to "This is Watson". *IBM Journal of Research and Development*. 56(3/4): 1. http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6177724
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.
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
Haug, C (2013). The downside of open-access publishing. *New England Journal of Medicine*. 368: 791-793.
Head, ML, Holman, L, et al. (2015). The extent and consequences of p-hacking in science. *PLoS Biology*. 13: e1002106. http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002106
Henry, J, Pylypchuk, Y, et al. (2016). Adoption of Electronic Health Record Systems among U.S. Non-Federal Acute Care Hospitals: 2008-2015. Washington, DC, Department of Health and Human Services. http://dashboard.healthit.gov/evaluations/data-briefs/non-federal-acute-care-hospital-ehr-adoption-2008-2015.php
Hersh, WR (2009). Information Retrieval: A Health and Biomedical Perspective (3rd Edition). New York, NY, Springer.
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.

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, Friedman, C, et al. (1995). Unlocking clinical data from narrative reports: a study of natural language processing. *Annals of Internal Medicine*. 122: 681-688.
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.
Jüni, P, Rutjes, AWS, et al. (2002). Are selective COX 2 inhibitors superior to traditional non steroidal anti-inflammatory drugs? *British Medical Journal*. 324: 1287-1288.
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.
Kris, MG, Gucalp, A, et al. (2015). Assessing the performance of Watson for oncology, a decision support system, using actual contemporary clinical cases. *ASCO Annual Meeting*, Chicago, IL http://meetinglibrary.asco.org/content/150420-156
Lohr, S (2012). The Future of High-Tech Health Care — and the Challenge. New york, NY. New York Times. February 13, 2012. http://bits.blogs.nytimes.com/2012/02/13/the-future-of-high-tech-health-care-and-the-challenge/
Markoff, J (2011). Computer Wins on 'Jeopardy!': Trivial, It's Not. New York, NY. New York Times. February 16, 2011. http://www.nytimes.com/2011/02/17/science/17jeopardy-watson.html
Merali, Z (2010). Computational science: ...Error. *Nature*. 467: 775-777.
Miller, DR, Safford, MM, et al. (2004). Who has diabetes? Best estimates of diabetes prevalence in the Department of Veterans Affairs based on computerized patient data. *Diabetes Care*. 27(Suppl 2): B10-21.
Moher, D and Moher, E (2016). Stop predatory publishers now: act collaboratively. *Annals of Internal Medicine*. 164: 616-617.
Osborn, R, Moulds, D, et al. (2015). Primary care physicians in ten countries report challenges caring for patients with complex health needs. *Health Affairs*. 34: 2104-2112.
Parsons, A, McCullough, C, et al. (2012). Validity of electronic health record-derived quality measurement for performance monitoring. *Journal of the American Medical Informatics Association*. 19: 604-609.
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.

Richesson, RL, Rusincovitch, SA, et al. (2013). A comparison of phenotype definitions for diabetes mellitus. *Journal of the American Medical Informatics Association*. 20(e2): e319-e326.
Roebuck, C (2012). The importance of knowing context of hospital episode statistics when reconfiguring the NHS. *British Medical Journal*: Rapid Response. http://www.bmj.com/content/344/bmj.e2432/rr/578977
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
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.
Seiler, KP, Bodycombe, NE, et al. (2011). Master data management: getting your house in order. *Combinatorial Chemistry & High Throughput Screening*. 14: 749-756.
Stanfill, MH, Williams, M, et al. (2010). A systematic literature review of automated clinical coding and classification systems. *Journal of the American Medical Informatics Association*. 17: 646-651.
Sterling, TD (1959). Publication decisions and their possible effects on inferences drawn from tests of significance - or vice versa. *Journal of the American Statistical Association*. 54: 30-34.
Turner, EH, Knoepflmacher, D, et al. (2012). Publication bias in antipsychotic trials: an analysis of efficacy comparing the published literature to the US Food and Drug Administration Database. *PLoS Medicine*. 9(3) http://www.plosmedicine.org/article/info%3Adoi%2F10.1371%2Fjournal.pmed.1001189
Turner, EH, Matthews, AM, et al. (2008). Selective publication of antidepressant trials and its influence on apparent efficacy. *New England Journal of Medicine*. 358: 252-260.
Voorhees, EM (2005). Question Answering in TREC. TREC - Experiment and Evaluation in Information Retrieval. E. Voorhees and D. Harman. Cambridge, MA, MIT Press: 233-257.
Wei, WQ, Leibson, CL, et al. (2013). The absence of longitudinal data limits the accuracy of high-throughput clinical phenotyping for identifying type 2 diabetes mellitus subjects. *International Journal of Medical Informatics*. 82: 239-247.
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.
Williams, S (2013). Absolute versus relative risk – making sense of media stories. Cancer Research UK. http://scienceblog.cancerresearchuk.org/2013/03/15/absolute-versus-relative-risk-making-sense-of-media-stories/
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
Zhang, Z and Sun, J (2010). Interval censoring. *Statistical Methods in Medical Research*. 19: 53-70.

# Challenges for Information Retrieval and Text Mining in Biomedicine: Imperatives for Systems and Their Evaluation

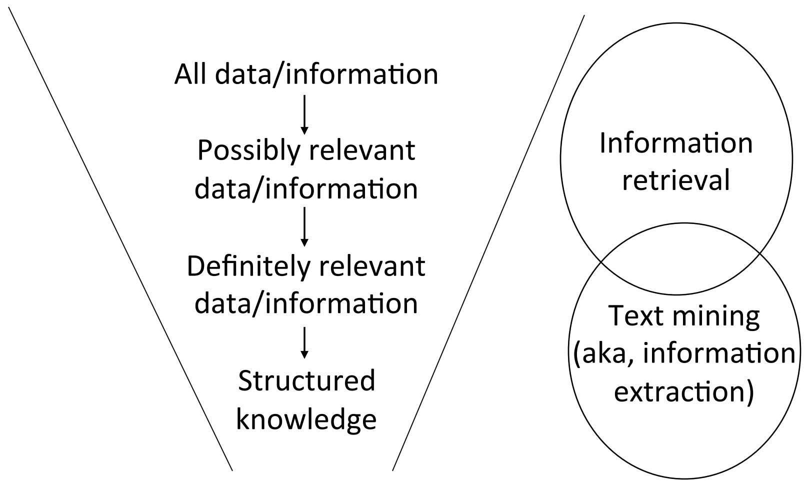
William Hersh
Professor and Chair
Department of Medical Informatics & Clinical Epidemiology
Oregon Health & Science University
Portland, OR, USA
Email: hersh@ohsu.edu
Web: www.billhersh.info
Blog: http://informaticsprofessor.blogspot.com
Twitter: @williamhersh



## Challenges for Information Retrieval and Text Mining in Biomedicine

- Definitions
- Rationale
- Challenges
- Implications

## Information retrieval and text mining (Hersh, 2009 – revised)




3

## Information retrieval (IR) and text mining must be driven by

- Appropriate use cases
- Understanding of the content and challenges of the two major types of data/information
  – Patient-specific
  – Knowledge-based
- Realistic evaluation

4

# Patient-specific data/information

- Data/information about patients, historically based in the medical record (electronic health record, EHR)
- But also growing amounts from personal health records (PHRs), wearable devices and sensors, social media, etc.
- Some of this data may be highly private

5

# Knowledge-based data/information

- The knowledge base of biomedicine and health
- Origin usually from scientific studies published in literature but many derived works in reviews, guidelines, textbooks, compendia, and Web sites

6

## What are the important use cases?

- Patient-specific
  - Clinical decision support
  - Precision medicine – more precise clinical measurements, including genomics, biomarkers, etc.
  - "Re-use" of data for research, quality measurement and improvement
- Knowledge-based
  - Connecting clinicians, patients, and others with knowledge to inform health and healthcare
  - "Mining" the literature for associations, question-answering, and other tasks

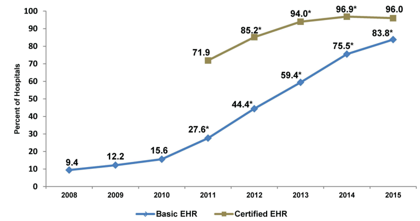


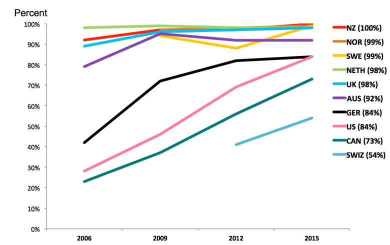
## Why is research in IR and text mining methods important?

- Motivated by the challenges in the following slides
- The methods to achieve those use cases still need improvement, led by research and evaluation
- Countering hype – especially that sold to scientists, administrators, clinical leaders, and others

# Challenges for patient-specific data/ information

- Since 2010, the growth in EHR use in the US (Henry, 2016) and many other countries (Osborn, 2015) has ushered in a new era of digital data that goes beyond the EHR
- But re-using clinical data for purposes beyond documentation has many challenges






# Caveats for the Use of Operational Electronic Health Record Data in Comparative Effectiveness Research

*William R. Hersh, MD,* Mark G. Weiner, MD,† Peter J. Embi, MD, MS,‡ Judith R. Logan, MD, MS,**
*Philip R.O. Payne, PhD,‡ Elmer V. Bernstam, MD, MSE,§ Harold P. Lehmann, MD, PhD,‖*
*George Hripcsak, MD, MS,¶ Timothy H. Hartzog, MD, MS,# James J. Cimino, MD,***
*and Joel H. Saltz, MD, PhD††*

Operational clinical data may be
- Inaccurate
- Incomplete
- Transformed in ways that undermine meaning
- Unrecoverable for re-use
- Of unknown provenance
- Of insufficient granularity
- Incompatible with research protocols

records, clinical data, coded (claims) data, biomedical informatics

(*Med Care* 2013;51: S30–S37)

Health Information Technology (ONC) through the Strategic Health IT Advanced Research Projects (SHARP) Program,

# Inaccurate

- Documentation not always a top priority for busy clinicians (de Lusignan, 2005)
- Data entry errors in a recent analysis in the English National Health Service (NHS) – yearly hospital statistics showed approximately (Brennan, 2012)
  - 20,000 adults attending pediatric outpatient services
  - 17,000 males admitted to obstetrical inpatient services – mainly due to male newborns (Roebuck, 2012)
  - 8,000 males admitted to gynecology inpatient services

# Inaccurate (cont.)

- Analysis of EHR systems of four known national leaders assessed use of data for studies on treatment of hypertension and found five categories of reasons why data were problematic (Bayley, 2013)
  - Missing
  - Erroneous
  - Un-interpretable
  - Inconsistent
  - Inaccessible in text notes

# Incomplete

- Not every diagnosis is recorded at every visit; absence of evidence is not always evidence of absence, an example of a concern known by statisticians as *censoring* (Zhang, 2010)
- Makes seemingly simple tasks such as identifying diabetic patients challenging (Miller, 2004; Wei, 2013; Richesson, 2013)
- Undermine ability to automate quality measurement
  - Measures under-reported based on under-capture of data due to variation in clinical workflow and documentation practices (Parsons, 2012)
  - Correct when present but not infrequently missing in primary care EHRs (Barkhuysen, 2014)

OHSU

# Incomplete (cont.)

- Studies of health information exchange (HIE)
  - Study of 3.7 million patients in Massachusetts found 31% visited two or more hospitals over five years (57% of all visits) and 1% visited five or more hospitals (10% of all visits) (Bourgeois, 2010)
  - Analysis of 2.8 million emergency department patients in Indiana found 40% had data at multiple institutions (Finnell, 2011)



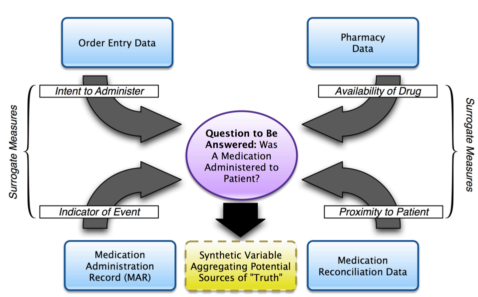
OHSU

# Unrecoverable for research

- Despite adoption of EHRs, major problem now is lack of interoperability due to incomplete adherence to standards (ONC, 2014 and many, many others)
- Many clinical data are "locked" in narrative text reports (Hripcsak, 1995; Hripcsak, 2012), including summaries of care (D'Amore, 2012)
- State of the art for performance of NLP has improved dramatically over the last couple decades, but is still far from perfect (Stanfill, 2010)
- Electronic records of patients at academic medical centers not easy to combine for aggregation (Broberg, 2015)

# Of unknown provenance and insufficient granularity

- Provenance – knowing where your data come from (Seiler, 2011)




- Granularity – knowing what your data mean
  - Diagnostic codes assigned for billing purposes may be generalized to a broad class of diagnosis due to regulatory and documentation requirements
  - For example, patient with set of complex cytogenetic and morphologic indicators of a pre-leukemic state may be described as having "myelodysplastic syndromes (MDS)" for billing purposes, but this is insufficient for other purposes, including research

16

# Many data "idiosyncrasies" between clinical practice and research protocols

- "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

# Challenges for knowledge-based data/ information

- Methodological challenges
- Publication bias and the "winner's curse"
- Reproducibility
- Misconduct
- Hype

# Methodological challenges

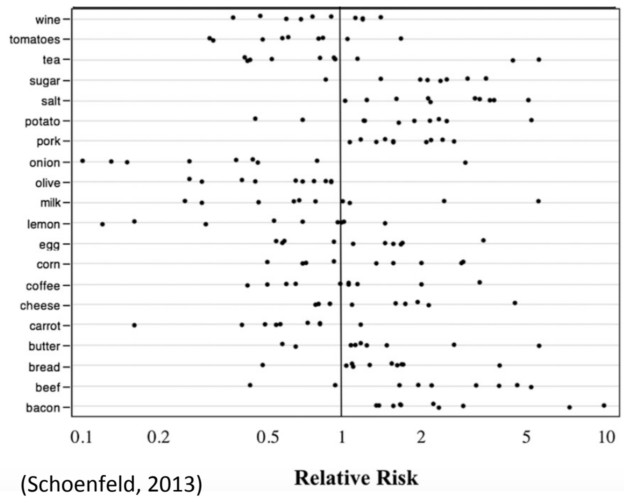
- IR and text mining may be better at finding knowledge but humans are (for now) better at appraising it
- Critical appraisal is needed because there are many limitations to current medical studies, even with gold-standard randomized controlled trials



# Problems with RCTs

- Experimental studies are the best approach for discerning cause and effect, but have limitations, e.g.
  - Samples may not represent populations (Weng, 2014; Prieto-Centurion, 2014; Geifman, 2016)
  - "Medical reversal" of earlier results not uncommon (Prasad, 2013; Prasad, 2015)
  - Surrogate measures may not be associated with desired clinical outcomes (Kim, 2015; Prasad, 2015)
  - Like many other studies, temptations for p-hacking (Head, 2015)
  - Differences between relative and absolute risk (Williams, 2013)

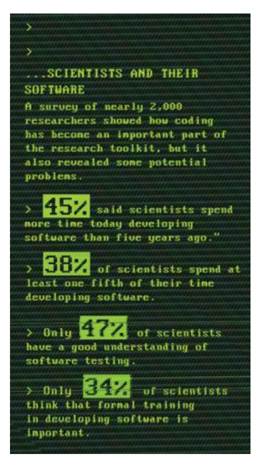
# Observation studies have challenges as well, e.g., what causes cancer




(Schoenfeld, 2013)

# 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)




22

# Publication bias and the "winner's curse"

- Publication bias is a long-known problem, not limited to biomedicine (Sterling, 1959; Dwan, 2013)
- As a result, what is reported in the scientific literature may not reflect the totality of knowledge, but instead representing the "winner's curse" of results that have been positive and thus more likely to be published (Ionnaidis, 2005; Young, 2008)
- Initial positive results not infrequently later overturned (Ionnaidis, 2005)

23

OHSU

# Discrepancies between FDA reporting and published literature

- Selective publication of antidepressant trials (Turner, 2008) – studies with positive results more likely to be published (37 of 38) than those with negative results (22 of 36 not published, 11 of 36 published in way to convey positive results)
- Similar picture with antipsychotic drugs (Turner, 2012)
- FDA data also led to discovery of studies of COX-2 inhibitors (Vioxx and Celebrex) with altered study design and omission of results that led to obfuscation of cardiac complications (Jüni, 2002; Curfman, 2005)

24

OHSU

# Reproducibility

- In recent years, another problem has been identified: inability to reproduce results (Begley, 2016)
  - ACM: "An experimental result is not fully established unless it can be independently reproduced" (2016)
- Documented in
  - Preclinical studies analyzed by pharmaceutical companies looking for promising drugs that might be candidates for commercial development (Begley, 2012)
  - Psychology research (Science, 2015)
- Recent survey of over 1500 scientists found over half agreed with statement: There is a "reproducibility crisis" in science (Baker, 2016)
  - 50-80% (depending on the field) reported unable to reproduce an experiment yet very few trying or able to publish about it





# Misconduct

- Many well-known cases, true scope of fraudulent science probably impossible to know because science operates on honor systems
- Documentation of many cases: Retractionwatch.com
- Predatory journals – fueled in part by open access movement (Haug, 2013; Moher, 2016)

# Hype

- Highest-profile system is IBM Watson
  – Developed out of TREC Question-Answering Track (Voorhees, 2005; Ferrucci, 2010)
  – Additional (exhaustive) details in special issue of IBM Journal of Research and Development (Ferrucci, 2012)
  – Beat humans at Jeopardy! (Markoff, 2011)
  – Now being applied to healthcare (Lohr, 2012); has "graduated" medical school (Cerrato, 2012)

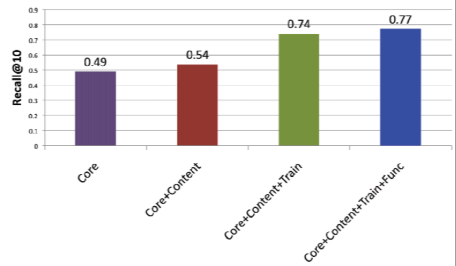


# Applying Watson to medicine (Ferrucci, 2012)

- Trained using several resources from internal medicine: *ACP Medicine*, *PIER*, *Merck Manual*, and *MKSAP*
- Concept adaptation process required
  – Named entity detection – e.g., disambiguation of terms and their senses
  – Measure recognition and interpretation – e.g., age or blood test value
  – Recognition of unary relations – e.g., elevated <test result>
- Trained with 5000 questions from *Doctor's Dilemma*, a competition like Jeopardy!, in which medical trainees participate and is run by the ACP each year
  – Sample question is, `Familial adenomatous polyposis is caused by mutations of this gene`, with the answer being, `APC Gene`
    - Googling the question gives the correct answer at the top of its ranking to this and two other sample questions listed

# Evaluation of Watson on internal medicine questions (Ferrucci, 2012)

- Evaluated on an additional 188 unseen questions
- Primary outcome measure was recall at 10 answers
  - How would Watson compare against other systems, such as Google or Pubmed, or using other measures, such as MRR?
- Future use case for Watson is applying system to data in EHR, ultimately aiming to serve as a clinical decision support system (Cerrato, 2012)
  - Performance so far falls "within evidence-based standards" (Kris, 2015)






# Implications for IR and text mining research

- The use cases driving IR and text mining in biomedicine are important
  - The future of clinical medicine needs these tools
- There are many challenges in developing and evaluating systems
  - But overcoming them is important
- The agenda for IR and text mining is identical to that of biomedical informatics generally, e.g.,
  - Standards and interoperability
  - Realistic and rigorous evaluation and reproducibility

## Some solutions we can pursue

- System development – should
  - Accommodate important use cases
  - Address challenges with data and information
- Evaluation
  - System-oriented studies fine for initial evaluation but must translate to focus on use cases, including studies of users and clinical outcomes
- Must not forget that biomedical informatics is a field that applies information solutions to real problems in health and healthcare

31

OHSU