Information Retrieval and Text Mining Evaluation Must Go Beyond "Users": Incorporating Real-World Context and Outcomes

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

Amini, I, Martinez, D, et al. (2016). Improving patient record search: a meta-data based approach. *Information Processing & Management*. 52: 258-272.
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 (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.
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.
Blumenthal, D (2011). Implementation of the federal health information technology initiative. *New England Journal of Medicine*. 365: 2426-2431.
Blumenthal, D (2011). Wiring the health system--origins and provisions of a new federal program. *New England Journal of Medicine*. 365: 2323-2329.
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.
Demner-Fushman, D, Abhyankar, S, et al. (2012). NLM at TREC 2012 Medical Records Track. *The Twenty-First Text REtrieval Conference Proceedings (TREC 2012)*, Gaithersburg, MD. National Institute for Standards and Technology
http://trec.nist.gov/pubs/trec21/papers/NLM.medical.final.pdf

Demner-Fushman, D, Abhyankar, S, et al. (2011). A knowledge-based approach to medical records retrieval. *The Twentieth Text REtrieval Conference Proceedings (TREC 2011)*, Gaithersburg, MD. National Institute for Standards and Technology
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
Egan, DE, Remde, JR, et al. (1989). Formative design-evaluation of Superbook. *ACM Transactions on Information Systems*. 7: 30-57.
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
Fidel, R and Soergel, D (1983). Factors affecting online bibliographic retrieval: a conceptual framework for research. *Journal of the American Society for Information Science*. 34: 163-180.
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
Hersh, W, Müller, H, et al. (2009). The ImageCLEFmed medical image retrieval task test collection. *Journal of Digital Imaging*. 22: 648-655.
Hersh, W and Voorhees, E (2009). TREC genomics special issue overview. *Information Retrieval*. 12: 1-15.
Hersh, WR (1994). Relevance and retrieval evaluation: perspectives from medicine. *Journal of the American Society for Information Science*. 45: 201-206.
Hersh, WR (2009). Information Retrieval: A Health and Biomedical Perspective (3rd Edition). New York, NY, Springer.
Hersh, WR, Crabtree, MK, et al. (2002). Factors associated with success for searching MEDLINE and applying evidence to answer clinical questions. *Journal of the American Medical Informatics Association*. 9: 283-293.
Hersh, WR and Greenes, RA (1990). SAPHIRE: an information retrieval environment featuring concept-matching, automatic indexing, and probabilistic retrieval. *Computers and Biomedical Research*. 23: 405-420.

Hersh, WR and Hickam, DH (1995). An evaluation of interactive Boolean and natural language searching with an on-line medical textbook. *Journal of the American Society for Information Science*. 46: 478-489.
Hersh, WR, Hickam, DH, et al. (1994). A performance and failure analysis of SAPHIRE with a MEDLINE test collection. *Journal of the American Medical Informatics Association*. 1: 51-60.
Hersh, WR, Müller, H, et al. (2006). Advancing biomedical image retrieval: development and analysis of a test collection. *Journal of the American Medical Informatics Association*. 13: 488-496.
Hersh, WR, Pentecost, J, et al. (1996). A task-oriented approach to information retrieval evaluation. *Journal of the American Society for Information Science*. 47: 50-56.
Ide, NC, Loane, RF, et al. (2007). Essie: a concept-based search engine for structured biomedical text. *Journal of the American Medical Informatics Association*. 14: 253-263.
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.
Joyner, MJ, Paneth, N, et al. (2016). What happens when underperforming big ideas in research become entrenched? *Journal of the American Medical Association*: Epub ahead of print.
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.
King, B, Wang, L, et al. (2011). Cengage Learning at TREC 2011 Medical Track. *The Twentieth Text REtrieval Conference Proceedings (TREC 2011)*, Gaithersburg, MD. National Institute for Standards and Technology
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
Lau, AY and Coiera, EW (2008). Impact of web searching and social feedback on consumer decision making: a prospective online experiment. *Journal of Medical Internet Research*. 10(1): e2. http://www.jmir.org/2008/1/e2/
Lau, AY, Kwok, TM, et al. (2011). How online crowds influence the way individual consumers answer health questions. *Applied Clinical Informatics*. 2: 177-189.
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
Martinez, D, Otegi, A, et al. (2014). Improving search over electronic health records using UMLS-based query expansion through random walks. *Journal of Biomedical Informatics*. 51: 100-106.

McKibbon, KA, Lokker, C, et al. (2013). Net improvement of correct answers to therapy questions after PubMed searches: pre/post comparison. *Journal of Medical Internet Research*. 15: e243. http://www.jmir.org/2013/11/e243/
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.
Müller, H, Clough, P, et al., Eds. (2010). ImageCLEF: Experimental Evaluation in Visual Information Retrieval. Heidelberg, Germany, Springer.
Mynatt, BT, Leventhal, LM, et al. (1992). Hypertext or book: which is better for answering questions? *Proceedings of Computer-Human Interface 92*. 19-25.
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.
Roberts, K, Simpson, M, et al. (2016). State-of-the-art in biomedical literature retrieval for clinical cases: a survey of the TREC 2014 CDS track. *Information Retrieval Journal*. 19: 113-148.
Safran, C, Bloomrosen, M, et al. (2007). Toward a national framework for the secondary use of health data: an American Medical Informatics Association white paper. *Journal of the American Medical Informatics Association*. 14: 1-9.
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.
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.
vanDeursen, AJ (2012). Internet skill-related problems in accessing online health information. *International Journal of Medical Informatics*. 81: 61-72.
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
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.
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.
Westbrook, JI, Coiera, EW, et al. (2005). Do online information retrieval systems help experienced clinicians answer clinical questions? *Journal of the American Medical Informatics Association*. 12: 315-321.
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

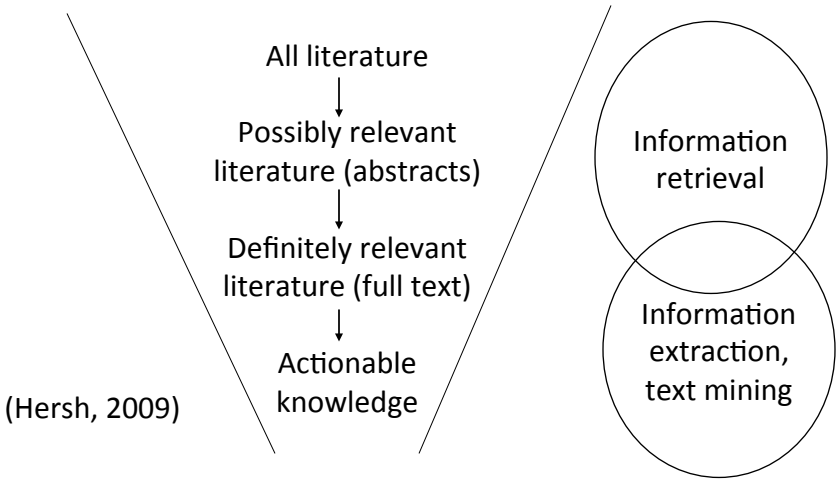
Information Retrieval and Text Mining Evaluation Must Go Beyond "Users": Incorporating Real-World Context and Outcomes

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

Information Retrieval and Text Mining Evaluation Must Go Beyond "Users"

- Personal history of domain-specific information retrieval evaluation
- Challenges for biomedical information retrieval (IR) and text mining
- Implications for development and evaluation going forward

IR and text mining in context of biomedical knowledge management



(Hersh, 2009)

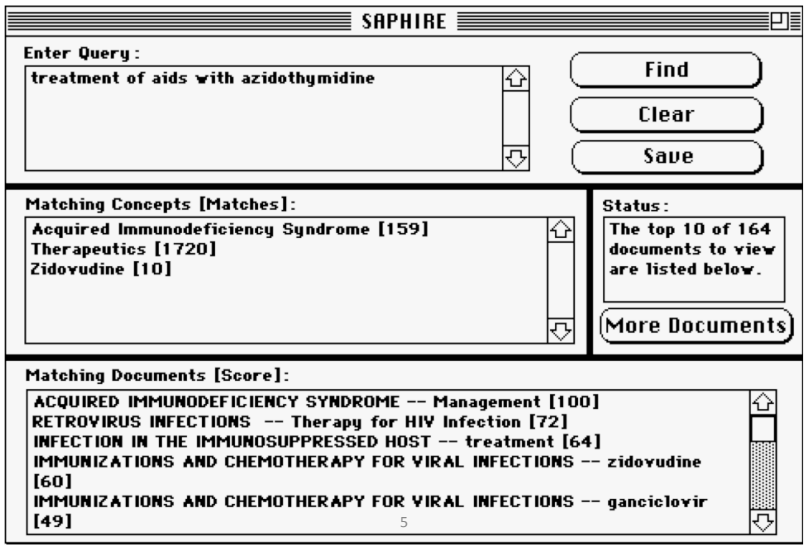
3

Personal journey in IR evaluation in health and biomedical domain

- SAPHIRE
- Toward task-oriented evaluation
- Factors association with successful searching
- Domain-specific retrieval evaluation

4

Concept-based IR using UMLS Metathesaurus (Hersh, 1990)



5

Set out to evaluate SAPHIRE and IR in biomedicine

- Concept-based approach did not impart value over word indexing and searching (Hersh, JAMIA, 1994)
- Experience of several evaluations led to concern with use of recall/precision (Hersh, JASIS, 1994)
  - How much difference is meaningful?
  - How valid is batch evaluation for understand how well user will search?

6

Led to "task-oriented" evaluation approaches

- Motivated by Egan (1989) and Mynatt (1992)
- Major task in medicine: answering questions
- How can we evaluate systems in interactive use for answering questions?
- Undertook parallel approaches in
  - Medicine – Using electronic textbooks and bibliographic databases
  - General news – TREC Interactive Track

Medical textbook – Boolean vs. natural language (1995)

- Searching medical textbook (*Scientific American Medicine*) with Boolean and natural language interfaces
  - Medical students answering ten short-answer questions
  - Randomized to one interface or other, asked to search on questions they rated lowest confidence before searching
  - Pre-searching correctness very low (1.7/10)
  - Correctness improved markedly with searching (4.0/5)
  - When incorrect with searching, document with correct answer retrieved two-thirds of time and viewed half of time

MEDLINE – Boolean vs. natural language (1996)

- Searching MEDLINE with Ovid (Boolean) and Knowledge Finder (natural language)
  - Medical students answering yes/no clinical questions
  - 37.5% answered correctly before searching
  - 85.4% answered correctly after searching
  - No difference across systems in time taken, relevant articles retrieved, or user satisfaction

9

Factors associated with successful searching (Hersh, 2002)

- Medical and nurse practitioner (NP) students success of using a retrieval system to answer clinical questions
  - Had to provide not only answer but level of evidence supporting it
    - Yes with good evidence
    - Indeterminate evidence
    - No with good evidence
- Look at factors associated with success
  - Based on model of factors associated with successful use of retrieval systems (Fidel, 1983) adapted to this setting
  - Dependent variable was correctness of answer

10

Major categories and some factors in the model

- Associated answering question correctly with independent variables
  - Answers – correct before searching, certainty, time
  - Demographic – age, gender, school
  - Computer experience – general, searching, specific MEDLINE features
  - Cognitive – set of factors shown in past to be associated with successful computer and/or retrieval system use
  - Search mechanics – sets retrieved, references viewed
  - User satisfaction – from Questionnaire for User Interface Satisfaction (QUIS)
  - Retrieval – recall, precision

Results

- 66 participants, 45 medical and 21 NP students
  - NP students all female, medical students evenly divided
  - NP students older, with more computer use but less searching and EBM experience
  - Medical students scored higher on cognitive tests, especially of spatial visualization

With searching, medical students increased rate of correctness to 51.6% but NP students remained virtually unchanged at 34.7%, i.e., searching did not help NP students

Prior to searching, rate of correctness (32.1%) about equal to chance for both groups, i.e., equal to chance

No difference in recall or precision for correct answering or student type, i.e., it did not impact correct answering

		Post-Search	
		Incorrect	Correct
Pre-Search	Incorrect M N	41 (20%) 27 (12%) 14 (14%)	63 (19%) 45 (20%) 18 (18%)
	Correct M		

Variable	Incorrect	Correct	p value
Recall	18%	18%	.61
Precision	28%	29%	.99

Variable	All	Medical	NP
Recall	18%	18%	20%
Precision	29%	30%	26%

OHSU

Work followed on by others

- Clinicians
  - Physicians and nurse consultants searching full-text and MEDLINE resource – both improved answering with searching (Westbrook, 2005)
  - Physicians had modest improvement in answering with searching; no difference between Pubmed and Clinical Queries (McKibbon, 2013)
- Others
  - Lau (2008, 2011) – college students searching PubMed, MedlinePLUS, and others
    - Correct answering 61.2% before searching and 82.0% after
    - Confidence not associated with correctness
  - Van Duersen (2012) – older and less educated searchers had poorer search skills
    - Younger searchers more likely to use nonrelevant search results and unreliable sources

OHSU

Back to batch evaluation: domain-specific IR

- TREC Genomics Track
- ImageCLEFmed
- TREC Medical Records Track

TREC Genomics Track (Hersh, 2009)

- Based on use case of exploding research in genomics and inability to biologists to know all that might impact work
- First TREC track devoted to "domain-specific" retrieval, with focus on IR systems for genomics researchers
- History
  - 2004-2005 – focus on ad hoc retrieval and document categorization
  - 2006-2007 – focus on passage retrieval and question-answering as means to improve document retrieval

Lessons learned (Hersh, 2009)

- Ad hoc retrieval
  - Modest benefit for techniques known to work well in general IR, e.g., stop word removal, stemming, weighting
  - Query term expansion, especially domain-specific and/or done by humans, helped most
- QA
  - Most consistent benefit from query expansion and paragraph-length passage retrieval
- For all experiments, big problem (as always) was lack of detailed description and use of low-performing baselines

Image retrieval – ImageCLEF medical image retrieval task

- Biomedical professionals increasingly use images for research, clinical care, and education, yet we know very little about how they find them
- Developed test collection and exploration of information needs motivating use of image retrieval systems (Hersh, 2006; Hersh, 2009; Müller, 2010)
- Started with ad hoc retrieval and added tasks
  - Modality detection
  - Case finding
- Overall conclusions: text yielded most consistent results with image features providing variable value

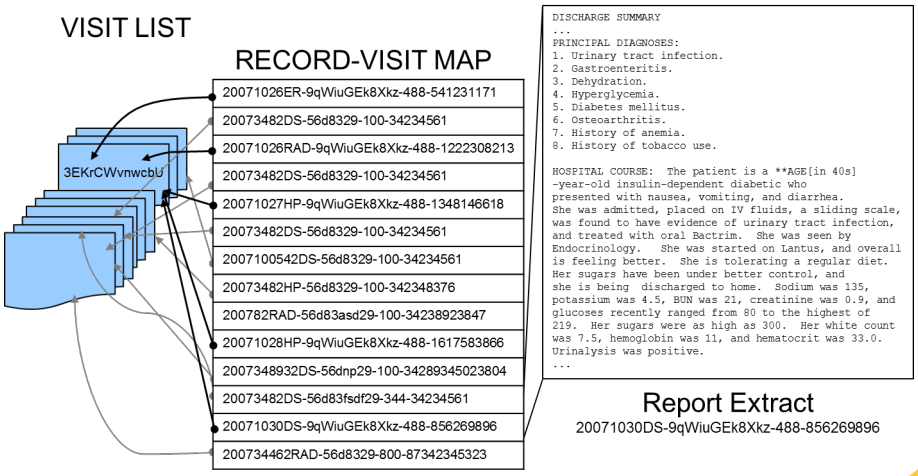
TREC Medical Records Track (Voorhees, 2012)

- Adapting IR techniques to electronic health records (EHRs)
- Use case somewhat different – want to retrieve records and data within them to identify patients who might be candidates for clinical studies
- Motivated by larger desire for "re-use" of clinical data (Safran, 2007)
- Opportunities facilitated by growing incentives for "meaningful use" of EHRs in the HITECH Act (Blumenthal, 2011; Blumenthal, 2011)

Challenges for informatics research with medical records

- Has always been easier with knowledge-based content than patient-specific data due to a variety of reasons
  - Privacy issues
  - Task issues
- Facilitated with development of large-scale, de-identified data set from University of Pittsburgh Medical Center (UPMC)
- Launched in 2011, repeated in 2012

Test collection



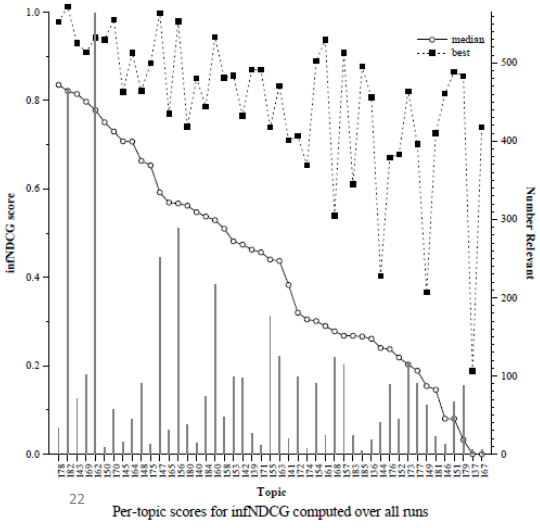
17,265 visits

101,712 reports (93,552 mapped to visits)

(Courtesy, Ellen Voorhees, NIST)

Results for 2012

Run	infNDCG	infAP	P(10)
NLMManual*	0.680	0.366	0.749
udelSUM	0.578	0.286	0.592
sennamed2	0.547	0.275	0.557
ohsuManBool*	0.526	0.250	0.611
atigeo1	0.524	0.224	0.519
UDinfoMed123	0.517	0.236	0.528
uogTrMConQRd	0.509	0.231	0.553
NICTAUBC4	0.487	0.216	0.517



Per-topic scores for infNDCG computed over all runs

Which approaches did (and did not) work?

- Best results in 2011 and 2012 obtained from NLM group (Demner-Fushman, 2011; Demner-Fushman, 2012)
  - Top results from manually constructed queries using Essie domain-specific search engine (Ide, 2007)
- Many approaches known to work in general IR fared less well, e.g., term expansion, document focusing, etc.
  - Other domain-specific approaches also did not show benefit, e.g., creation of PICO frames, negation
- Some success with
  - Results filtered by age, race, gender, admission status; terms expanded by UMLS Metathesaurus (King, 2011)
  - Expansion by concepts and relationships in UMLS Metathesaurus (Martinez, 2014)
  - Pseudorelevance feedback using ICD-9 codes (Amini, 2016)

TREC Clinical Decision Support Track (Roberts, 2016)

- www.trec-cds.org
- Ad hoc search of biomedical literature (PubMed Central Open Access Subset – 1.25M articles)
- Topics are patient descriptions in three information need categories
  - Diagnosis
  - Test
  - Treatment
- Currently in third year of operation
- Limitation: ad hoc searching of literature not a common activity of clinicians seeking answers to questions

Good searching is not enough – must take into account context of science

- 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 (RCTs)

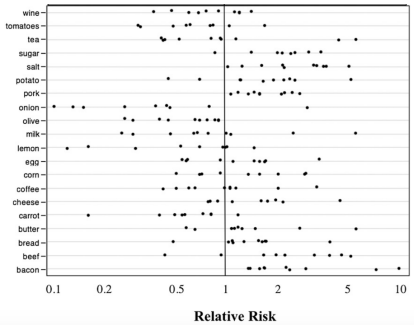
Problems with RCTs

- 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)

27

Problems with results of non-RCTs

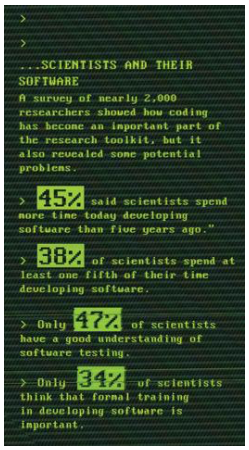
- Observational studies can mislead us, e.g., Women's Health Initiative (JAMA, 2002)
- Observational studies do not discern cause and effect, e.g., diet and cancer (Schoenfeld, 2013)
- New technologies and techniques not yet fully assessed, e.g., precision medicine and EHR usage (Joyner, 2016)



28

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)



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)

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)

Reproducibility

- In recent years, another problem has been identified: inability to reproduce results (Begley, 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)

33

Hype

- Example of high-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)

34

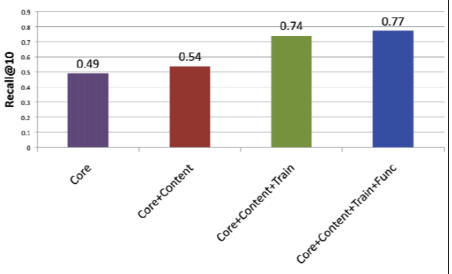
Applying Watson to medicine (Ferrucci, 2012)

- Concept adaptation process required
  - Named entity detection
  - Measure recognition and interpretation
  - Recognition of unary relations
- Trained using several resources from internal medicine: *ACP Medicine*, *PIER*, *Merck Manual*, and *MKSAP*
- Trained with 5000 questions from *Doctor's Dilemma*, a competition like *Jeopardy!* run by American College of Physicians (ACP) annually
  - Sample question, `Familial adenomatous polyposis is caused by mutations of this gene`, with answer, `APC Gene`
    - Googling the question gives the correct answer at the top of its ranking to this and two other sample questions listed

35

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?
- Awaiting further studies…



36

Why is the context of science important to IR and text mining?

- 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

37

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
    - Realistic use cases
    - Studies of users and incorporation of research and clinical outcomes

38