Information Retrieval Evaluation in the Ubiquitous Search Era: A View from the Biomedical/Health Domain

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

References

Anonymous (2012). From Screen to Script: The Doctor's Digital Path to Treatment. New York, NY, Manhattan Research; Google. http://www.thinkwithgoogle.com/insights/library/studies/the-doctors-digital-path-to-treatment/
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
Davies, K (2006). Search and Deploy. Bio-IT World, October 16, 2006. http://www.bio-itworld.com/issues/2006/oct/biogen-idec/
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
Edinger, T, Cohen, AM, et al. (2012). Barriers to retrieving patient information from electronic health record data: failure analysis from the TREC Medical Records Track. *AMIA 2012 Annual Symposium*, Chicago, IL. 180-188.
Egan, DE, Remde, JR, et al. (1989). Formative design-evaluation of Superbook. *ACM Transactions on Information Systems*. 7: 30-57.
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.
Friedman, CP, Wong, AK, et al. (2010). Achieving a nationwide learning health system. *Science Translational Medicine*. 2(57): 57cm29. http://stm.sciencemag.org/content/2/57/57cm29.full
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, Cimino, JJ, et al. (2013). Recommendations for the use of operational electronic health record data in comparative effectiveness research. *eGEMs (Generating Evidence & Methods to improve patient outcomes)*. 1: 14. http://repository.academyhealth.org/egems/vol1/iss1/14/
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.
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.
Insel, TR, Volkow, ND, et al. (2003). Neuroscience networks: data-sharing in an information age. *PLoS Biology*. 1: E17.
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
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.
Marchionini, G (1992). Interfaces for end-user information seeking. *Journal of the American Society for Information Science*. 43: 156-163.
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
McKibbon, KA and Fridsma, DB (2006). Effectiveness of clinician-selected electronic information resources for answering primary care physicians' information needs. *Journal of the American Medical Informatics Association*. 13: 653-659.
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/
Meadow, CT, Boyce, BR, et al. (2007). Text Information Retrieval Systems, Third Edition. San Diego, CA, Academic Press.
Metzger, J and Rhoads, J (2012). Summary of Key Provisions in Final Rule for Stage 2 HITECH Meaningful Use. Falls Church, VA, Computer Sciences Corp. http://assets1.csc.com/health_services/downloads/CSC_Key_Provisions_of_Final_Rule_for_Stage_2.pdf

Müller, H, Clough, P, et al., Eds. (2010). ImageCLEF: Experimental Evaluation in Visual Information Retrieval. Heidelberg, Germany, Springer.
Mulrow, CD, Cook, DJ, et al. (1997). Systematic reviews: critical links in the great chain of evidence. *Annals of Internal Medicine*. 126: 389-391.
Mynatt, BT, Leventhal, LM, et al. (1992). Hypertext or book: which is better for answering questions? *Proceedings of Computer-Human Interface 92*. 19-25.
Purcell, K, Brenner, J, et al. (2012). Search Engine Use 2012. Washington, DC, Pew Internet & American Life Project. http://www.pewinternet.org/Reports/2012/Search-Engine-Use-2012.aspx
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.
Smith, M, Saunders, R, et al. (2012). Best Care at Lower Cost: The Path to Continuously Learning Health Care in America. Washington, DC, National Academies Press.
Taylor, A (2012). A study of the information search behaviour of the millennial generation. *Information Research*. 17(1) http://informationr.net/ir/17-1/paper508.html
Thiele, RH, Poiro, NC, et al. (2010). Speed, accuracy, and confidence in Google, Ovid, PubMed, and UpToDate: results of a randomised trial. *Postgraduate Medical Journal*. 86: 459-465.
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
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.

Information Retrieval Evaluation in the Ubiquitous Search Era: View from the Biomedical/Health Domain

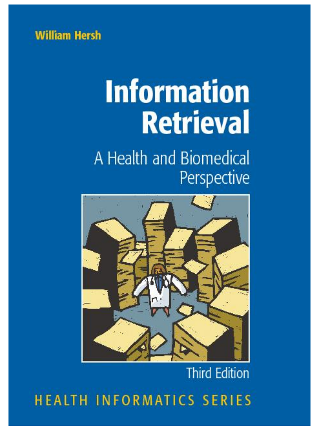
William Hersh, MD
Professor and Chair
Department of Medical Informatics & Clinical Epidemiology
School of Medicine
Oregon Health & Science University
Email: hersh@ohsu.edu
Web: www.billhersh.info
Blog: http://informaticsprofessor.blogspot.com

Overview

- Role of IR in health and biomedicine
- Personal journey: IR evaluation in health and biomedicine
  - Early work
  - Task-oriented evaluation
  - Use case-driven batch evaluation
- Future directions and recommendations

The world of IR has changed

- Evolution of my book
  - In first edition (1996), last chapter devoted to "special topic" of the Internet and Web
- Most people have used a search engine
  - And have strong opinions about them
- Previous concern of access to information (e.g., Gregor Mendel) has given way to *information overload*, *data smog*, and *information chaos*
- 91% of US Internet users (73% of US adults) have used a search engine (Purcell, 2012)



OREGON HEALTH & SCIENCE UNIVERSITY

3

IR and online access firmly planted in health and biomedicine

- Biology is now defined as an "information science" (Insel, 2003)
- Pharmaceutical companies compete for informatics/library talent (Davies, 2006)
- Search for health information by clinicians, researchers, and patients/consumers is ubiquitous (Purcell, 2012; Google/Manhattan Research, 2012)
  - It's even part of "meaningful use" rule for electronic health record adoption! (Metzger, 2012)

OREGON HEALTH & SCIENCE UNIVERSITY

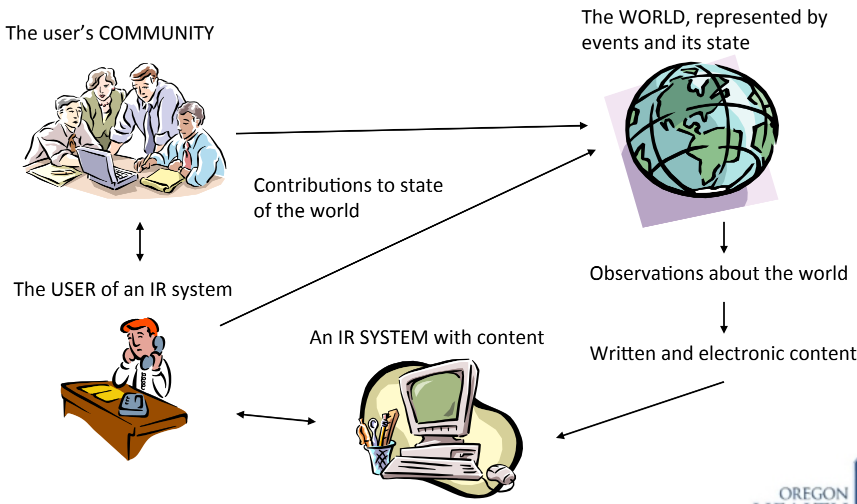
4

Popular IR-related icons permeate our lives

Models show us that information "retrieval" is more than just searching

- Information universe
- User interaction
- Medical decision-making
- Knowledge management

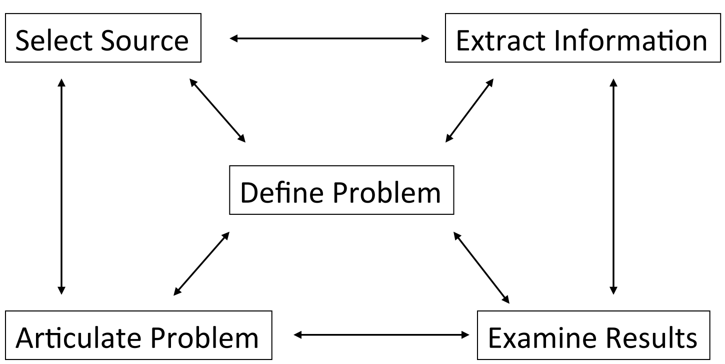
Information universe (Meadow, 2007)



7

OREGON HEALTH & SCIENCE UNIVERSITY

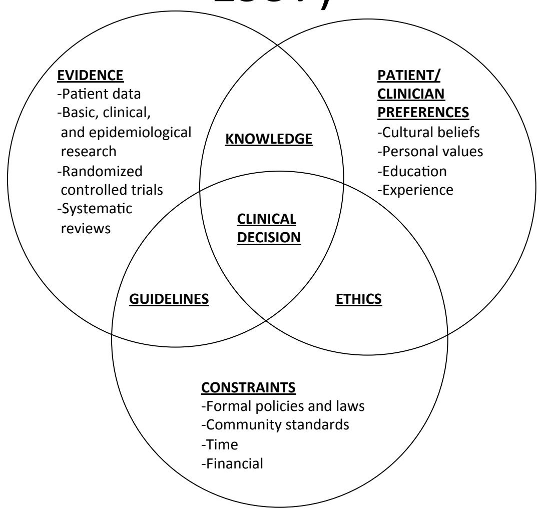
User interaction (Marchionini, 1992)



8

OREGON HEALTH & SCIENCE UNIVERSITY

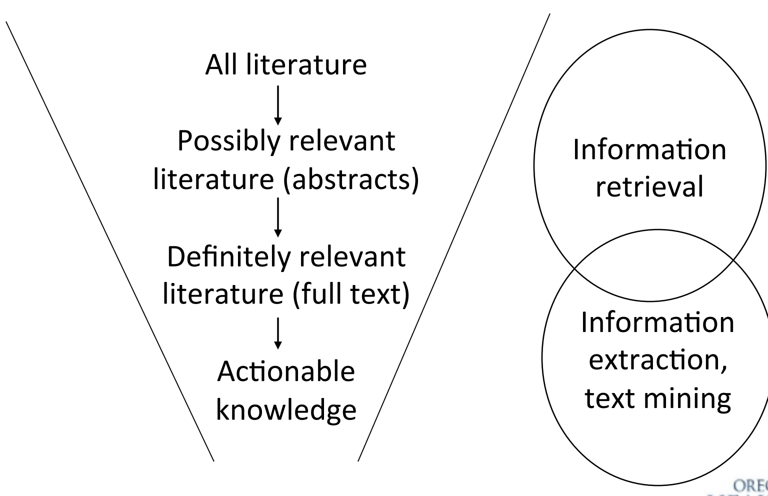
Medical decision-making (Mulrow, 1997)



OREGON HEALTH & SCIENCE UNIVERSITY

9

IR in context of biomedical knowledge management (Hersh, 2009)



OREGON HEALTH & SCIENCE UNIVERSITY

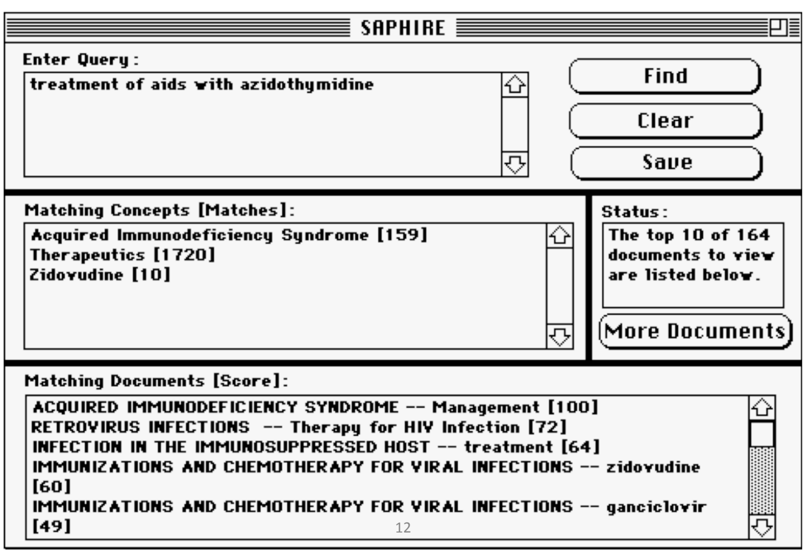
10

Personal journey in IR evaluation in health and biomedical domain

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

OREGON HEALTH & SCIENCE UNIVERSITY

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



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?

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 bibliographic databases and electronic textbooks
  - 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

OREGON HEALTH & SCIENCE UNIVERSITY

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

OREGON HEALTH & SCIENCE UNIVERSITY

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

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

OREGON HEALTH & SCIENCE UNIVERSITY

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
- Prior to searching, rate of correctness (32.1%) about equal to chance for both groups
  - Rating of certainly low for both groups
- With searching, medical students increased rate of correctness to 51.6% but NP students remained virtually unchanged at 34.7%
  - NP student difference was likely due to judging evidence

OREGON HEALTH & SCIENCE UNIVERSITY

Results (cont.)

			Post-Search	
			Incorrect	Correct
Pre-Search	Incorrect		133 (41%)	87 (27%)
		M	81 (36%)	70 (31%)
		N	52 (52%)	17 (17%)
	Correct		41 (13%)	63 (19%)
		M	27 (12%)	45 (20%)
		N	14 (14%)	18 (18%)

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%

OREGON HEALTH & SCIENCE UNIVERSITY

Work followed on by others

- Physicians and nurse consultants searching full-text and MEDLINE resource – both improved with searching (Westbrook, 2005)
- Physicians using self-chosen resource improved minimally (McKibbon, 2006)
- Physician searching improved more with textbook than Google or MEDLINE (Thiele, 2010)
- Physicians had modest improvement with searching, no difference between Pubmed and Clinical Queries (McKibbon, 2013)

OREGON HEALTH & SCIENCE UNIVERSITY

Including study of non-clinicians

- 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 have lower search skills although younger searchers more likely to use nonrelevant search results and unreliable sources
- Taylor (2012) – same attributes of younger ("millenial generation") searchers seen in general

OREGON HEALTH & SCIENCE UNIVERSITY

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

TREC Medical Records Track

- Adapting IR techniques to medical records
- 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 "secondary use" of clinical data (Safran, 2007)
- Opportunities facilitated by growing incentives for "meaningful use" of EHRs in the HITECH Act (Blumenthal, 2011; Blumenthal, 2011), aiming toward the "learning healthcare system" (Friedman, 2010; Smith 2012)

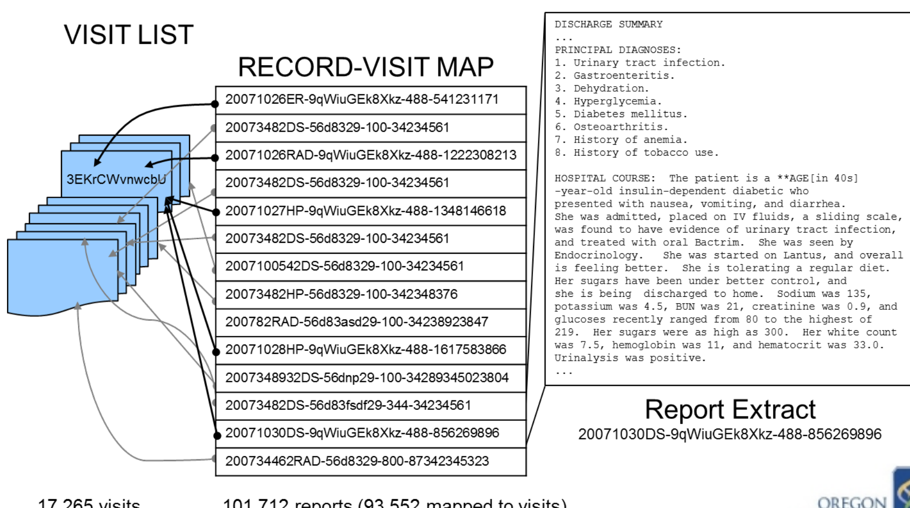
Challenges for secondary use of clinical data

- EHR data does not automatically lead to knowledge (Hersh, 2013; Hersh, 2013)
  - Data quality and accuracy is not a top priority for busy clinicians
  - Patients get care in many places, so record may be incomplete
  - Data provenance often a concern; where does data come from?
  - Best evidence for medical tests and treatments comes from experiments, i.e., evidence-based medicine

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 (Voorhees, 2012)

Test collection



(Courtesy, Ellen Voorhees, NIST)

Some issues for test collection

- De-identified to remove protected health information (PHI), e.g., age number → range
- De-identification precludes linkage of same patient across different visits (encounters)
- UPMC only authorized use for TREC 2011 and TREC 2012 but nothing else, including any other research (unless approved by UPMC)

OREGON HEALTH & SCIENCE UNIVERSITY

Easy and hard topics

- Easiest – best median bpref
  - 105: Patients with dementia
  - 132: Patients admitted for surgery of the cervical spine for fusion or discectomy
- Hardest – worst best bpref and worst median bpref
  - 108: Patients treated for vascular claudication surgically
  - 124: Patients who present to the hospital with episodes of acute loss of vision secondary to glaucoma
- Large differences between best and median bpref
  - 125: Patients co-infected with Hepatitis C and HIV
  - 103: Hospitalized patients treated for methicillin-resistant Staphylococcus aureus (MRSA) endocarditis
  - 111: Patients with chronic back pain who receive an intraspinal pain-medicine pump

OREGON HEALTH & SCIENCE UNIVERSITY

Failure analysis for 2011 topics (Edinger, 2012)

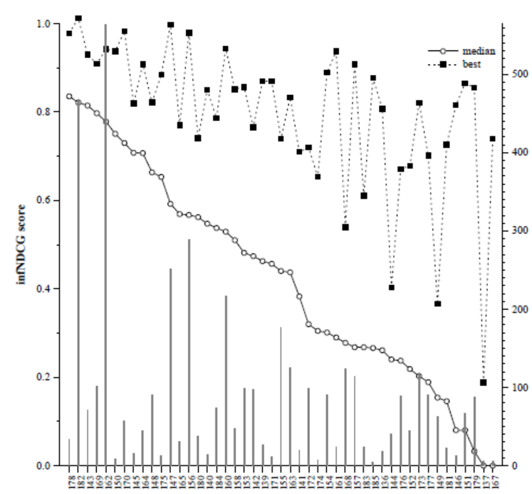
Reasons for Incorrect Retrieval	Number of Visits	Number of Topics
Visits Judged Not Relevant		
Topic terms mentioned as future possibility	16	9
Topic symptom/condition/procedure done in the past	22	9
All topic criteria present but not in the time/sequence specified by the topic description	19	6
Most, but not all, required topic criteria present	17	8
Topic terms denied or ruled out	19	10
Notes contain very similar term confused with topic term	13	11
Non-relevant reference in record to topic terms	37	18
Topic terms not present—unclear why record was ranked highly	14	8
Topic present—record is relevant—disagree with expert judgment	25	11
Visits Judged Relevant		
Topic not present—record is not relevant—disagree with expert judgment	44	21
Topic present in record but overlooked in search	103	27
Visit notes used a synonym or lexical variant for topic terms	22	10
Topic terms not named in notes and must be inferred	3	2
Topic terms present in diagnosis list but not visit notes	5	5

OREGON HEALTH & SCIENCE UNIVERSITY

33

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



34

Per-topic scores for infNDCG computed over all runs

What approaches did (and did not) work?

- Best results in 2011 and 2012 obtained from NLM group (Demner-Fushman, 2011)
  - Top results from manually constructed queries using Essie domain-specific search engine (Ide, 2007)
  - Other automated processes fared less well, e.g., creation of PICO frames, negation, term expansion, etc.
- Best automated results in 2011 obtained by Cengage (King, 2011)
  - Filtered by age, race, gender, admission status; terms expanded by UMLS Metathesaurus
- Benefits of approaches commonly successful in IR provided small or inconsistent value
  - Document focusing, term expansion, etc.

OREGON HEALTH & SCIENCE UNIVERSITY

Another (prominent) approach to question-answering: Watson

- IBM Watson evolved out of PIQUANT system used in TREC QA Track to new system, DeepQA (Ferrucci, 2010)
  - Additional (exhaustive) details in special issues of *IBM Journal of Research and Development* (Ferrucci, 2012)
- Achieved fame by beating humans at *Jeopardy!* television game (Markoff, 2011)
- Has turned to other domains, including healthcare
  - Has "graduated" medical school (Cerrato, 2012)
  - First results are in (Ferrucci, 2012)

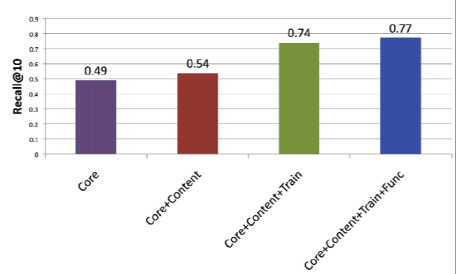
OREGON HEALTH & SCIENCE UNIVERSITY

Watson's first foray into 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)



Conclusions and future directions

- Evaluation must focus on real-world
  - Use cases
  - Collections and topics
- Use cases should focus on tasks of clinicians, researchers, and other specific roles
- Collections should reflect type and quantity of information appropriate to use cases

39

OREGON HEALTH & SCIENCE UNIVERSITY