Information Retrieval from Electronic Health Records for Patient Cohort Discovery

University of Missouri Informatics Institute
Columbia, MO
March 7, 2019

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

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

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.

Hanbury, A, Müller, H, et al. (2015). Evaluation-as-a-service: overview and outlook. *arXiv.org*: arXiv:1512.07454. https://arxiv.org/abs/1512.07454

Harman, D (2011). Information Retrieval Evaluation. San Rafael, CA, Morgan & Claypool.

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

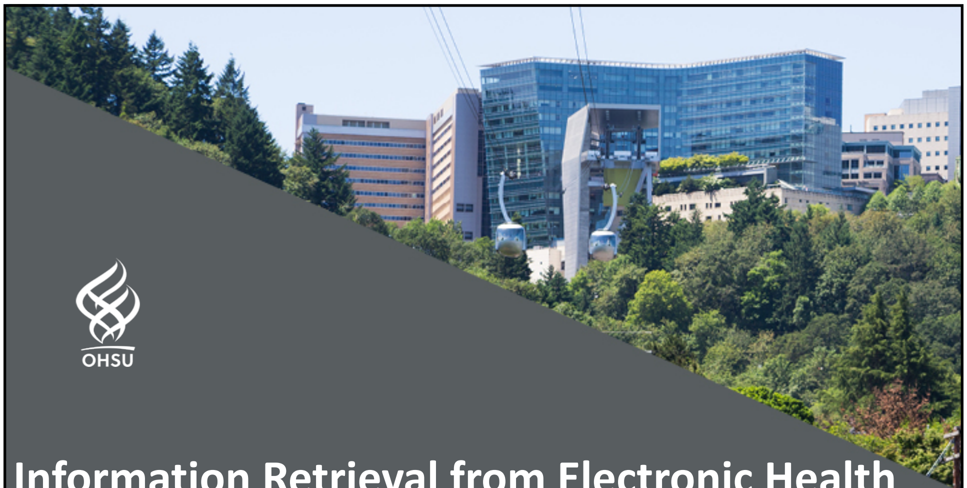
Hersh, W, Müller, H, et al. (2009). The ImageCLEFmed medical image retrieval task test collection. *Journal of Digital Imaging*. 22: 648-655.

Hersh, W and Voorhees, E (2009). TREC genomics special issue overview. *Information Retrieval*. 12: 1-15.

Hersh, WR (2009). Information Retrieval: A Health and Biomedical Perspective (3rd Edition). New York, NY, Springer.

Hersh, WR, Boone, KW, et al. (2018). Characteristics of the healthcare information technology workforce in the HITECH era: underestimated in size, still growing, and

adapting to advanced uses. *JAMIA Open*. 1: 188-194.
https://academic.oup.com/jamiaopen/article/1/2/188/5056817
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.
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
Müller, H, Clough, P, et al., Eds. (2010). ImageCLEF: Experimental Evaluation in Visual Information Retrieval. Heidelberg, Germany, Springer.
Myrick, KL and Ogburn, DF (2019). Percentage of office-based physicians using any electronic health record (EHR)/electronic medical record (EMR) system and physicians that have a certified EHR/EMR system, by specialty: National Electronic Health Records Survey, 2017. Hyattsville, MD, National Center for Health Statistics.
https://www.cdc.gov/nchs/data/nehrs/2017_NEHRS_Web_Table_EHR_Specialty.pdf
Roegiest, A and Cormack, GV (2016). An architecture for privacy-preserving and replicable high-recall retrieval experiments. *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, Pisa, Italy. 1085-1088.
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.
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 (2013). The TREC Medical Records Track. *Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics*, Washington, DC. 239-246.
Voorhees, EM and Harman, DK, Eds. (2005). TREC: Experiment and Evaluation in Information Retrieval. Cambridge, MA, MIT Press.
Voorhees, EM and Tong, RM (2011). Overview of the TREC 2011 Medical Records Track. *The Twentieth Text REtrieval Conference Proceedings (TREC 2011)*, Gaithersburg, MD. National Institute of Standards and Technology
Washington, V, DeSalvo, K, et al. (2017). The HITECH era and the path forward. *New England Journal of Medicine*. 377: 904-906.
Wu, S, Liu, S, et al. (2017). Intra-institutional EHR collections for patient-level information retrieval. *Journal of the American Society for Information Science & Technology*. 68: 2636-2648.
Zhu, D, Wu, ST, et al. (2014). Using large clinical corpora for query expansion in text-based cohort identification. *Journal of Biomedical Informatics*. 49: 275-281.



Information Retrieval from Electronic Health Records for Patient Cohort Discovery

William Hersh, MD
Professor and Chair
Department of Medical Informatics & Clinical Epidemiology
School of Medicine
Oregon Health & Science University
Portland, OR, USA

Overview

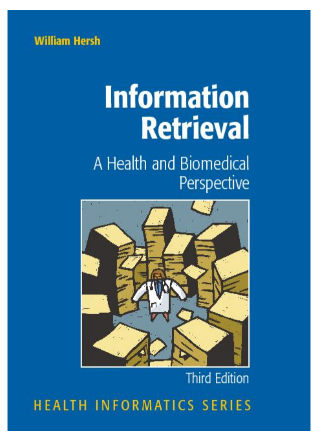
- Re-use of clinical data
- Primer on information retrieval (IR) and challenge evaluations
- TREC Medical Records Track
- Information retrieval from electronic health record (EHR) for patient cohort discovery
- Future work

Re-use of clinical data

- Many re-uses (secondary use) of EHR data, including (Safran, 2007)
  - Clinical and translational research – generating hypotheses and facilitating research
  - Public health surveillance for emerging threats
  - Healthcare quality measurement and improvement
- Opportunities facilitated by widespread adoption of EHRs (Washington, 2017)
  - 96% of hospitals (Henry, 2016), 86% of physicians (Myrick, 2019)

Information retrieval – IR (Hersh, 2009)

- Aka, "search"
- Focus on indexing and retrieval of information
- Historically centered on text in documents, but increasingly associated with many types of content
- www.irbook.info



Evaluation of IR systems (Harman, 2011)

- System-oriented – how well system performs
  - Historically focused on relevance-based measures
    - Recall and precision – proportions of relevant documents retrieved
  - When documents ranked, can combine both in a single measure
    - Mean average precision (MAP)
    - Normal discounted cumulative gain (NDCG)
    - Binary preference (Bpref)
- User-oriented – how well user performs with system
  - e.g., performing task, user satisfaction, etc.

System-oriented IR evaluation

- Historically assessed with *test collections*, which consist of
  - Content – fixed yet realistic collections of documents, images, etc.
  - Topics – statements of information need that can be fashioned into queries entered into retrieval systems
  - Relevance judgments – by expert humans for which content items should be retrieved for which topics
- Evaluation consists of *runs* using a specific IR approach with output for each topic measured and averaged across topics

Recall and precision

- Recall

$$R = \frac{\#\,retrieved\ and\ relevant\ documents}{\#\,relevant\ documents\ in\, collection}$$

  - Usually use *relative recall* when not all relevant documents known, where denominator is number of known relevant documents in collection
- Precision

$$P = \frac{\#\,retrieved\ and\ relevant\ documents}{\#\,retrieved\ documents}$$

Challenge evaluations

- A common approach in computer science, not limited to IR
- Develop a common task, data set, evaluation metrics, etc., ideally aiming for real-world size and representation for data, tasks, etc.
- In case of IR, this usually means
  - Test collection of content items
  - Topics to be retrieved
  - Runs from participating groups with retrieval for each topic
  - Relevance judgments of which content items are relevant to which topics – judged items derived from submitted runs

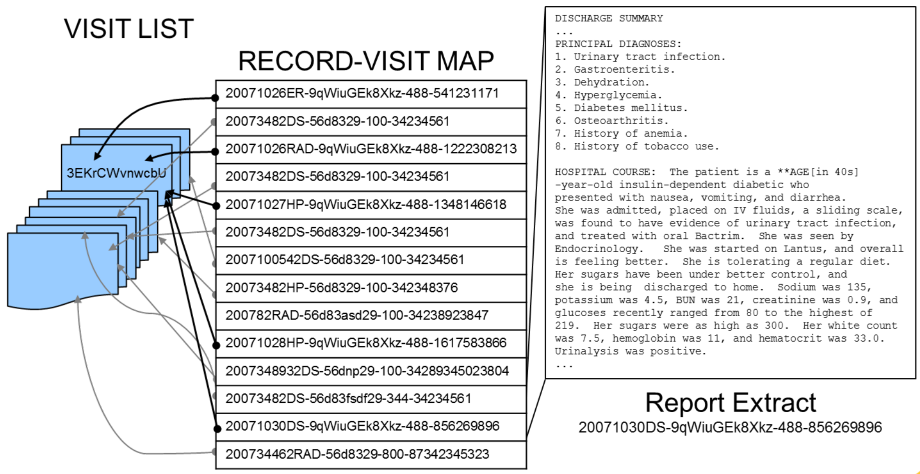
Some well-known challenge evaluations in IR

- Text Retrieval Conference (TREC, http://trec.nist.gov; Voorhees, 2005) – sponsored by National Institute for Standards and Technology (NIST), started in 1992
  - Mostly non-biomedical; first domain-specific track was Genomics Track (Hersh, 2009)
- Conferences and Labs of the Evaluation Forum (CLEF, www.clef-initiative.eu)
  - Focus on retrieval across languages, but also on image retrieval, including medical image retrieval tasks – www.imageclef.org (Hersh, 2009; Müller, 2010)
- TREC has inspired other challenge evaluations, e.g., n2c2/i2b2 NLP Shared Task, https://n2c2.dbmi.hms.harvard.edu/

TREC Medical Records Track (TRECMed)

- Focused on use case of identifying patients for possible recruitment into clinical studies
  - Task to "retrieve" patients who fit search criteria
- Used de-identified EHR data from University of Pittsburgh Medical Center (UPMC)
- Ran for two years (Voorhees, 2011; Voorhees, 2012; Voorhees, 2013)

Test collection



17,265 visits　　101,712 reports (93,552 mapped to visits)

(Voorhees, 2013)

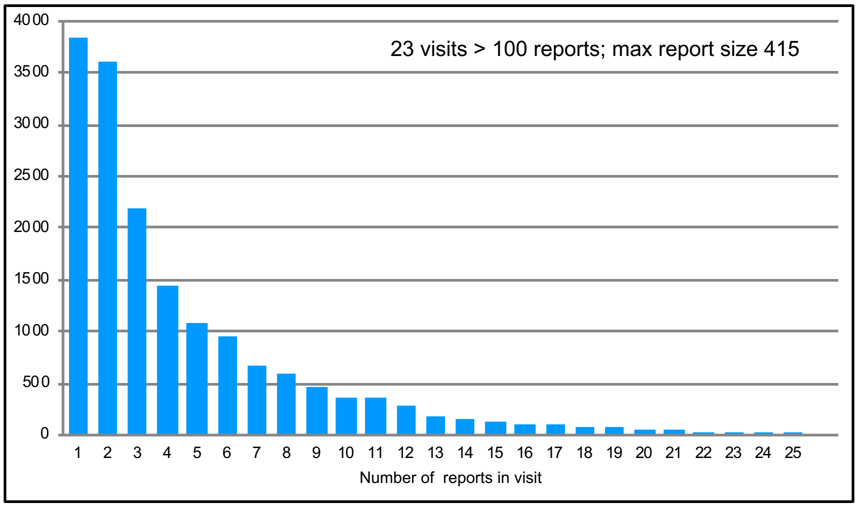
11

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)

12

Wide variations in number of documents per visit



13 (Voorhees, 2013)

Topic development and relevance assessments for TRECMed 2011

- Task – Identify patients who are possible candidates for clinical studies/trials
- Topics derived from 100 top critical medical research priorities in comparative effectiveness research (IOM, 2009)
  - Selected 35 topics assessed for appropriateness for data and with at least some relevant "visits"
- Relevance judgments by physicians who were OHSU informatics students

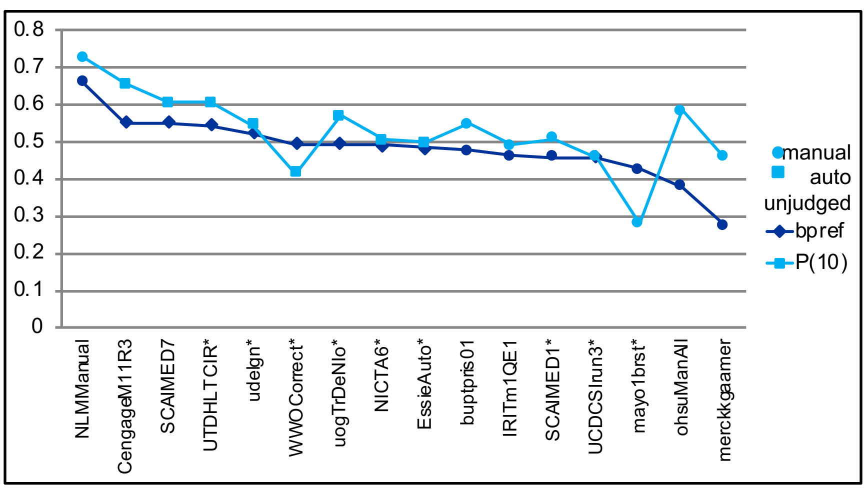
14

Participation in 2011

- Runs consisted of ranked list of up to 1000 visits per topic for each of 35 topics
  - Automatic – no human intervention from input of topic statement to output of ranked list
  - Manual – everything else
- Up to 8 runs per participating group
- Subset of retrieved visits contributed to judgment sets
  - Because resources for judging limited, could only judge relatively small sample of visits, necessitating use of BPref for primary evaluation measure
- 127 runs submitted from 29 groups
  - 109 automatic
  - 18 manual

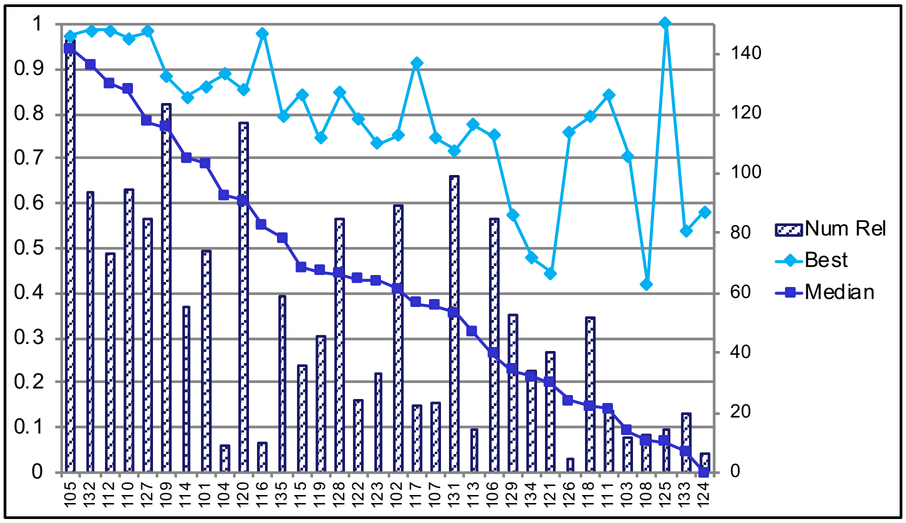
15

Evaluation results for top runs …



16

… BUT, wide variation among topics



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

Failure analysis for TRECMed 2011 topics (Edinger, 2012)

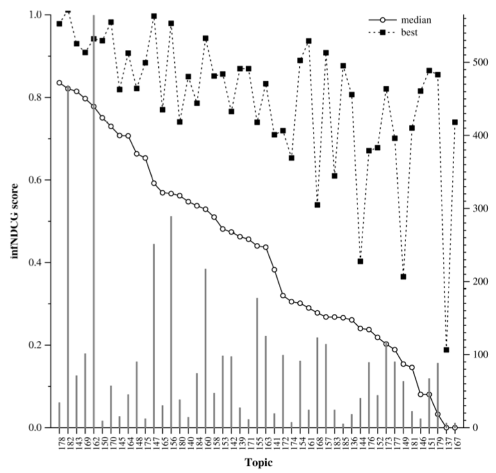
Reasons for Incorrect Retrieval	Number of Visits	Number of Topics
Visits Judged Not Relevant		
Topic terms mentioned as future possibility	16	9
Topic symptom/condition/procedure done in the past	22	9
All topic criteria present but not in the time/sequence specified by the topic description	19	6
Most, 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

Participation in TRECMed 2012

- Task and document collection same as 2011
- Developed 50 new topics
- More judging resources than 2011 allowed for more relevance judgments
  - For each topic, pooled top 15 from all runs and 25% of all visits ranked 16-100 by any run
- 88 runs submitted from 24 groups
  - 82 automatic
  - 6 manual

Evaluation results from 2012

Run	infNDCG	P(10)
NLMManual*	0.680	0.749
udelSUM	0.578	0.592
sennamed2	0.547	0.557
ohsuManBool*	0.526	0.611
atigeo1	0.524	0.519
UDinfoMed123	0.517	0.528
uogTrMConQRd	0.509	0.553
NICTAUBC4	0.487	0.517



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; 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 for this task in 2011 and 2012
  - Document focusing, term expansion, etc.

OHSU

Next step: use of bigger and better data

- Mayo Clinic-OHSU collaboration (Wu, 2017)
  - Original PI, Stephen Wu, Mayo → OHSU
- Funded by NLM R01
  - Hongfang Liu, Mayo Clinic, Co-PI
  - Steven Bedrick, OHSU, Site PI
  - William Hersh, OHSU, Co-I
- Aimed to add natural language processing (NLP) and language modeling (LM) to base IR methods on large amounts of unmodified (not de-identified) text from EHR
  - Preliminary data showed improvement over baseline IR techniques with TRECMed collection (Zhu, 2014)

Collections of patient data

- OHSU
  - Extraction of patients from Oregon Clinical and Translational Research Institute (OCTRI) Research Data Warehouse (RDW) having inpatient or outpatient encounters in primary care departments (Internal Medicine, Family Medicine, or Pediatrics) with
    - 3 or more encounters
    - 5 or more text entries
    - Between 1/1/2009 and 12/31/2013
  - Stored on (highly!) secure server
- Mayo using comparable approach and quantities

OHSU data model and statistics

Type	Patients	Encounters	Records	Average	Median	Max
Administered Meds	47,208	125,831	6,497,157	51.634	6	-
Ambulatory Encounters	99,965	3,760,205	3,760,205	-	-	-
Current Meds	92,783	-	31,997,402	344.863	64	20,102
Demographics	99,965	-				
Encounter Attributes	99,965	6,273,137	6,273,137			
Encounter Diagnoses	99,938	3,725,603	18,170,896	4.877	4	107
Notes	99,868	3,491,659	10,111,930			
Hospital Encounters	73,303	466,252	466,252			
Lab Results	83,435	733,461	20,186,748	27.523	12	19488
Microbiology Results	27,515	65,373	296548	4.536	1	268
Medications Ordered	94,089	1,388,086	5,336,506	3.845	1	1551
Procedures Ordered	98,514	1,880,309	7,229,854	3.845	1	6681
Problem List	90,722	-	761,260	8.391	6	182
Result Comments	72,716	468,814	916,554	1.955	1	691
Surgeries	18,640	29,895	31,889	1.067	1	41
Vitals	99,098	1,362,431	6,647,115	4.879	2	6387

25

Topics

- From OHSU
  - Derived from clinical study data requests by researchers from the Oregon Clinical and Translational Research Institute (OCTRI), querying Research Data Warehouse (RDW) (29 topics)
- From Mayo
  - Phenotype KnowledgeBase (PheKB) (5 topics)
  - Healthcare quality measures from NQF (12 topics)
  - Rochester Epidemiology Project (REP) (8 topics)
  - Mayo RDW (2 topics)

26

Some topics

Adults with IBD who haven't had GI surgery	Adults with inflammatory bowel disease who haven't had surgery involving the small intestine, colon, rectum, or anus.
Adults with a Vitamin D lab result	Adults with a lab result for 25-hydroxy Vitamin D collected between May 15 and October 15.
Postherpetic neuralgia treated with topical and systemic medication	Adults with postherpetic neuralgia ever treated by concurrent use of topical and non-opioid systemic medications.
Children seen in ED with oral pain	Children who were seen in the emergency department with herpetic gingivostomatitis, herpangina or hand, foot, and mouth disease, tonsillitis, gingivitis, or ulceration (aphthae, stomatitis, or mucositis) not due to chemotherapy or radiation.
ACE inhibitor-induced cough	Adults who have used an ACE inhibitor and experienced ACE inhibitor-induced cough.

Evaluation across sites

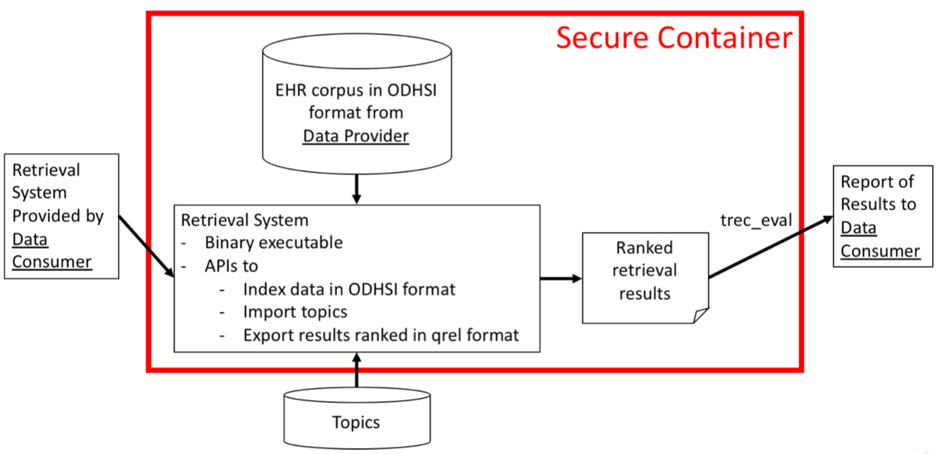
- Relevance assessments being done behind firewall at each site
  - Using Web-based relevance judging system based on one used for TREC –deployed at both sites
- Early results show techniques that worked in TRECMed less successful here
  - More complex and longitudinal data?
  - Requires queries on both structured and unstructured data?
- Following on with
  - Structured and more complex queries
  - Failure analysis

Future directions

- Starting collaborative project with a pharmaceutical company to detect EHR signals for a rare disease with common symptoms, porphyria
- Developing new methods to allow access to highly private data by other researchers (Hanbury, 2015; Roegiest, 2016)
  - Used for email spam detection, corporate repositories, etc.

29

Evaluation as a Service (EaaS)



30

Thank You!

William Hersh, MD
Professor and Chair
Department of Medical Informatics & Clinical Epidemiology
School of Medicine
Oregon Health & Science University
Portland, OR, USA

Email: hersh@ohsu.edu
Web: www.billhersh.info
Blog: http://informaticsprofessor.blogspot.com
Twitter: @williamhersh