Applying Information Retrieval to the Electronic Health Record for Cohort Discovery and Rare Disease Detection

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

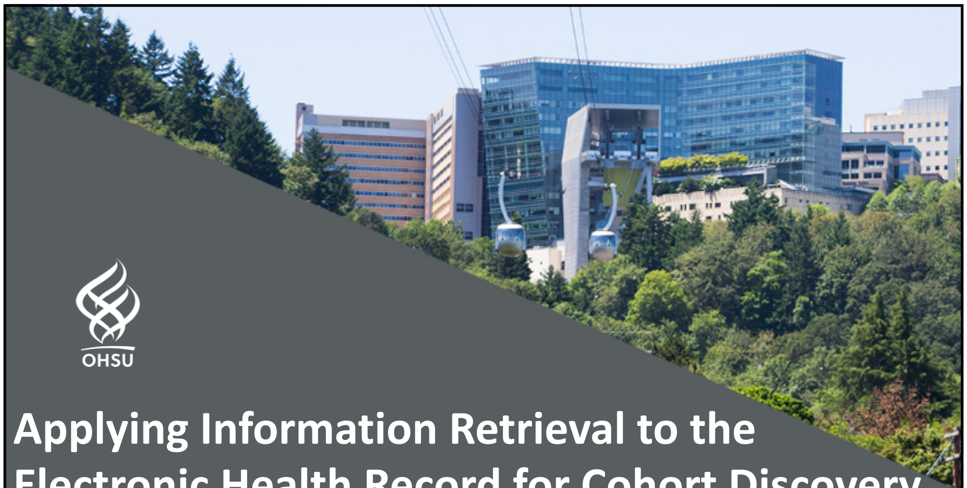
Distinguished Lecture Series, UCLA Biomedical Data Science Program
February 13, 2020

References

Buckley, C and Voorhees, EM (2005). Retrieval System Evaluation. TREC: Experiment and Evaluation in Information Retrieval. E. Voorhees and D. Harman. Cambridge, MA, MIT Press: 53-75.
Chamberlin, SR, Bedrick, SD, et al. (2019). Evaluation of patient-level retrieval from electronic health record data for a cohort discovery task. *medRxiv*: 19005280. https://www.medrxiv.org/content/10.1101/19005280v2
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.
Halamka, JD (2020). A New Model for Sharing Insights While Protecting Privacy. Dispatch from the Digital Health Frontier. http://geekdoctor.blogspot.com/2020/01/a-new-model-for-sharing-insights-while.html
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.
Harman, DK (2005). The TREC Ad Hoc Experiments. TREC: Experiment and Evaluation in Information Retrieval. E. Voorhees and D. Harman. Cambridge, MA, MIT Press: 79-98.
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, 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 and Hickam, DH (1998). How well do physicians use electronic information retrieval systems? A framework for investigation and review of the literature. *Journal of the American Medical Association*. 280: 1347-1352.
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.
Jarvelin, K and Kekalainen, J (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*. 20: 422-446.
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
Matheny, M, Israni, ST, et al., Eds. (2019). Artificial Intelligence in Health Care: The Hope, the Hype, the Promise, the Peril. Washington, DC, National Academy of Medicine.
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.
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.
Sardh, E, Harper, P, et al. (2019). Phase 1 trial of an RNA interference therapy for acute intermittent porphyria. *New England Journal of Medicine*. 380: 549-558.
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 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
Wang, Y, Wen, A, et al. (2019). Test collections for electronic health record-based clinical information retrieval. *JAMIA Open*. 2: 360-368. https://academic.oup.com/jamiaopen/advance-article/doi/10.1093/jamiaopen/ooz016/5510566
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.



Applying Information Retrieval to the Electronic Health Record for Cohort Discovery and Rare Disease Detection

William Hersh, MD
Oregon Health & Science University

Distinguished Lecture Series, UCLA Biomedical Data Science Program
February 13, 2020

Overview

- Applying IR to the EHR
- Use cases
  - Cohort discovery
  - Rare disease detection
- Challenges for EHR research

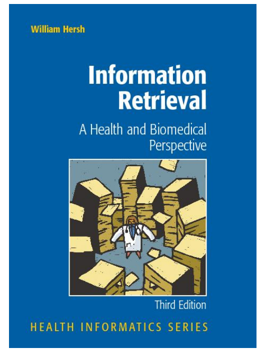
- This work funded by grants from
  - NLM 1R01LM011934
  - Alnylam Pharmaceuticals

- With help from OHSU collaborators
  - Steven Bedrick
  - Steven Chamberlin
  - Aaron Cohen

2

Information retrieval (IR, aka, search)

- We all do it – Google, PubMed, etc.
- As academics, we evaluate it – personal journey from
  - Knowledge-based information (1990, 1994, 1998)
  - Studies of users (mostly physicians) (1995, 2002)
  - Participation/leadership of challenge evaluations, mainly TREC (2009; Voorhees, 2012; Roberts, 2016)
- Forthcoming 4th edition of *Information Retrieval: A Biomedical & Health Perspective* (Springer, 2020)



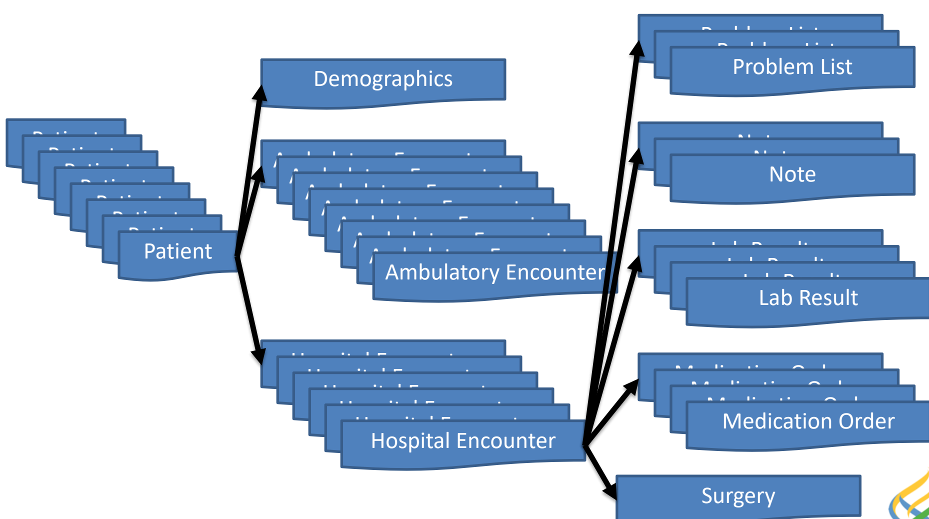
Applying IR to the EHR

- Growing availability of data with incentives for electronic health record (EHR) adoption in HITECH Act of 2009
- With availability of EHR data, first effort was cohort discovery task of TREC Medical Records Track (Voorhees, 2011; Voorhees, 2012)
- Awarding in 2014 of NIH R01 to (former) OHSU faculty Stephen Wu to explore methods in parallel with Mayo Clinic (Zhu, 2014; Wu, 2017; Wang, 2019; Chamberlin, 2019)
- Added rare disease surveillance task in collaboration with Alnylam Pharmaceuticals

IR system evaluation based on test collections of "documents"

- Recall $R = \frac{\#\,retrieved\ and\ relevant\ documents}{\#\,relevant\ documents\ in\,collection}$
- Precision $P = \frac{\#\,retrieved\ and\ relevant\ documents}{\#\,retrieved\ documents}$
- Aggregate measures
  - F – combining and (optional) weighting of R and P
- Measures for ranked output (Harman, 2011)
  - Mean average precision (MAP) (Harman, 2005)
  - B-Pref – used when relevance judgments incomplete (Buckley, 2004)
  - Others – normalized distributed cumulative gain (NDCG), inferred measures (Jarvelin, 2002)

Electronic health record (EHR) structure



Uses cases for EHR retrieval

- Cohort discovery
  - Can we “retrieve” cohorts of patients who are candidates for specific clinical studies?
- Rare disease detection
  - Can we discover patients who may be candidates for diagnosis and treatment of rare disease?

Cohort discovery

- Widely offered service by most academic medical centers but little formal evaluation of approaches
- Early work – TREC Medical Records Track, 2011-2012
- Follow-on collaboration with Mayo Clinic

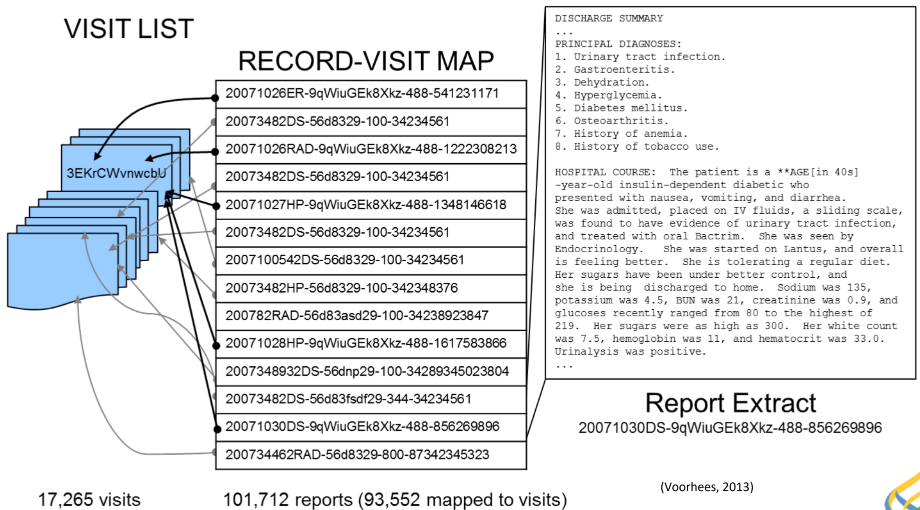
TREC Medical Records Track

- Appealing task given HITECH investment
  - NIST involved in HITECH in various ways
- More challenging with patient-specific data due to
  - 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 for EHR retrieval

- Task
  - Identify patients who are possible candidates for clinical studies/trials
- "Documents"
  - At "visit" level due to de-identification of records
- "Topics"
  - Selected 35 clinical study topics from IOM key areas for comparative effectiveness research
- "Relevance judgments"
  - Patients "relevant" to topics, judged by OHSU informatics students who were also physicians

Test collection structure

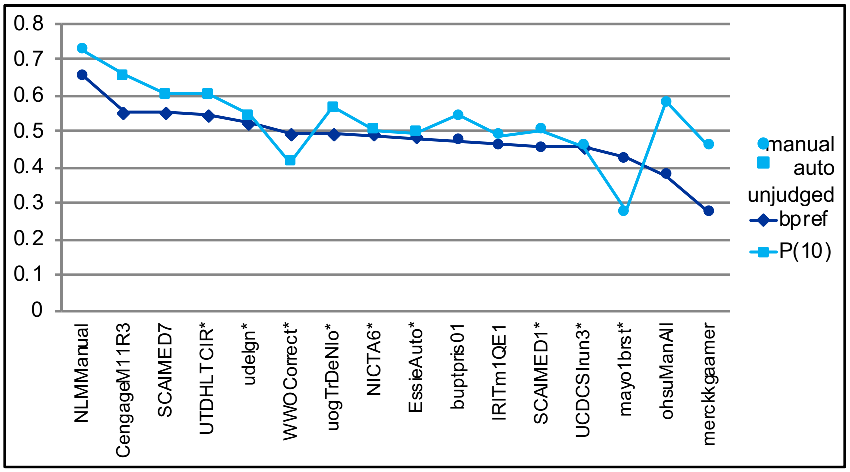


(Voorhees, 2013)

Some issues for test collection

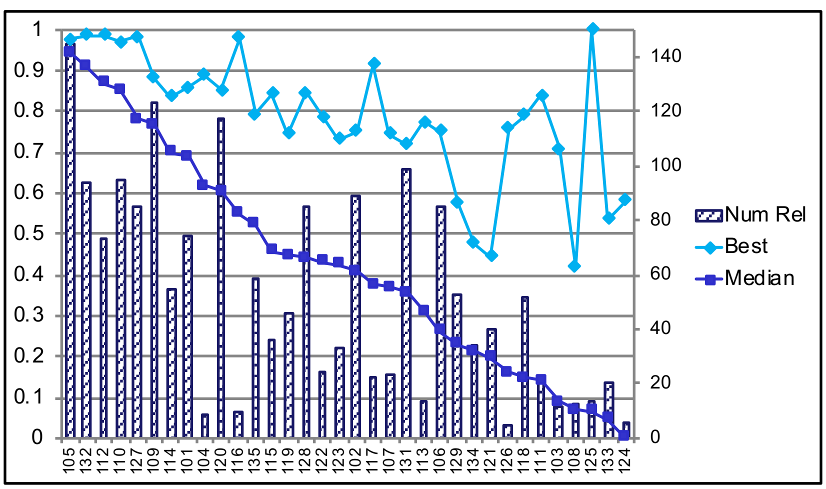
- De-identified to remove protected health information (PHI), e.g., age number → range
- De-identification precludes linkage of same patient across different visits (encounters)
- UPMC only authorized use for TREC 2011 and TREC 2012 but no longer available

Evaluation results for top runs (Voorhees, 2011)



13

But as commonly seen in IR, wide variation across topics



14

Easy and hard topics

- Easiest – best median B-Pref
  - 105: Patients with dementia
  - 132: Patients admitted for surgery of the cervical spine for fusion or discectomy
- Hardest – worst best B-Pref and worst median B-Pref
  - 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 B-Pref
  - 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

15

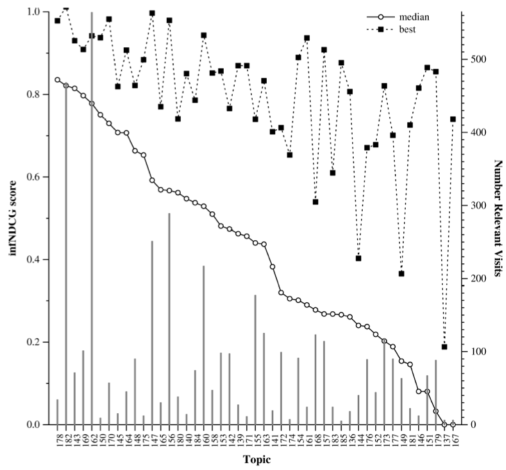
Failure analysis for 2011 topics (Edinger, 2012)

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

16

Evaluation results from 2012 were comparable (Voorhees, 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 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
- Approaches commonly successful in general IR provided small or inconsistent value for this task
  - Document focusing, term expansion, etc.

Extending cohort discovery work

- Mayo Clinic-OHSU collaboration
  - Hongfang Liu, Mayo Clinic, Co-PI
  - Stephen Wu, OHSU, Co-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 TREC Medical Record Track collection (Zhu, 2014)
- Methods (Wu, 2017) and results (Chamberlin, 2019 – medRxiv 19005280)

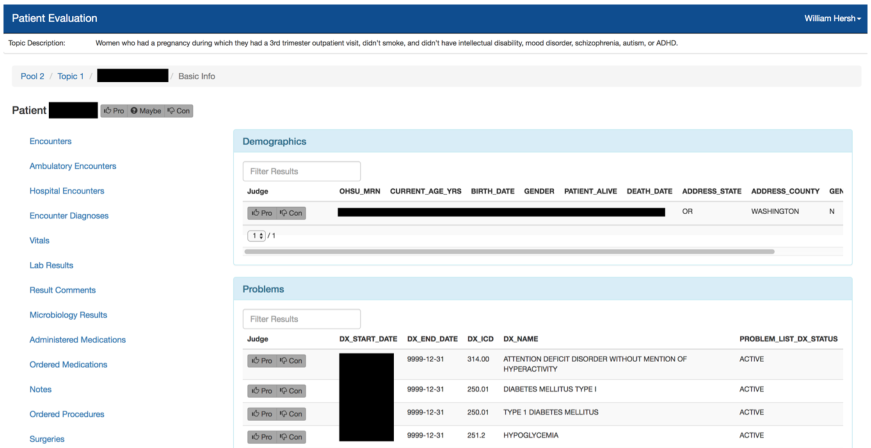
19

Original EHR data – 100K OHSU patients having ≥3 visits

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

20

Judgments from Patient Relevance Assessment Interface (PRAI)



Topic examples – summary and full

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.
3rd trimester prenatal visit with midwife or Ob/Gyn	Women who had a pregnancy with a 3rd trimester outpatient prenatal visit with an obstetrician and gynecologist or midwife.

A. Adults 18-64 years old with rheumatoid arthritis who have had a lab test for cyclic citrullinated peptide IgG antibody with a result greater than 40 units.

B. Adults 18-64 years old with rheumatoid arthritis and lab result for positive anti-CCP IgG > 40 units.

- **I. A 58-year-old female presents with morning stiffness and joint pain in her hands, especially her fingers, which improves after about 30 minutes, but doesn't remit fully. On examination she is found to have ulnar deviation, decreased grip strength, and joint tenderness over the MCP and PIP joints. She has a positive rheumatoid factor and is positive for anti-CCP Ab at 45 units.**

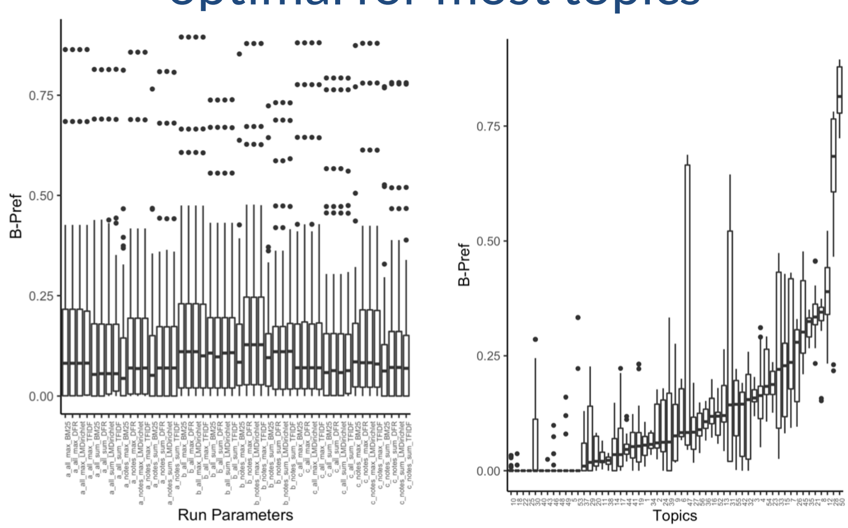
C. Adults 18-64 years old with rheumatoid arthritis and lab result for positive anti-CCP IgG > 40 units.

- **I. Demographics inclusion**
  - **a. Age: 18-64 years**
- **II. Diagnosis inclusion**
  - **a. Rheumatoid arthritis (ICD-9): 714.0**
- **III. Lab inclusion**
  - **a. Cyclic citrullinated peptide IgG antibody (anti-CCP IgG): > 40 units**

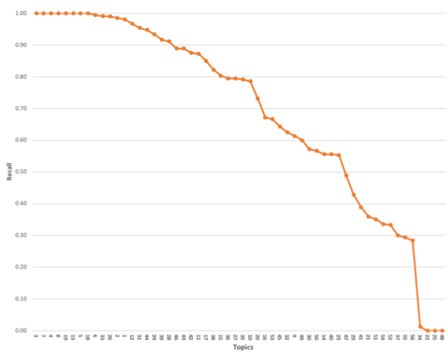
Initial approach used word-based queries with varying parameters

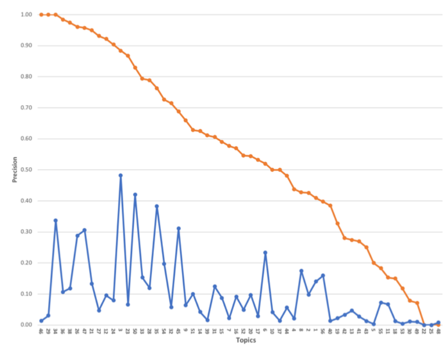
- Topic representation
  - A – summary statement
  - B – clinical case
  - C – detailed criteria
- Text subset
  - Just text notes
  - All of record
- Relevance aggregation
  - Sum of all retrieved
  - Max retrieved
- Retrieval ranking
  - BM25 (Roberston, 1994)
  - DFR (Amati, 2002)
  - LMDir (Zhai, 2004)
  - Lucene – aka, TFIDF (Salton, 1988)

Word-based query performance not optimal for most topics



Reformulating as Boolean queries led to better performance



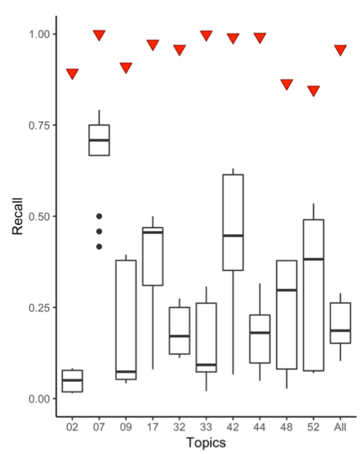


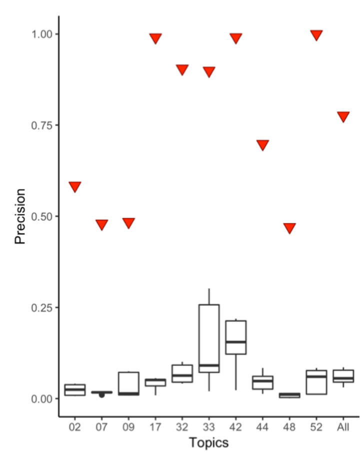
Good recall for many queries

Better precision for all queries

(Without additional relevance judgments)

Additional relevance judgments on 10 topics





Good relative recall and much improved precision

Rare disease detection

- Over 1200 known rare disorders that affect < 1 in 200K patients worldwide, many under-diagnosed
  - https://rarediseases.org/
- Acute Intermittent Porphyria (AIP)
  - Rare genetic disease of heme biosynthesis – variable penetrance
  - Incidence 1 per 100K in population
  - Often long lead time for diagnosis
  - Significant morbidity and effect on quality of life
  - "Neurovisceral" symptoms common with other diseases
    - Abdominal pain
    - Nausea and vomiting
    - Weakness,
    - Psychiatric changes
  - New treatments available, including RNA-silencing molecule (Sardh, 2019)
  - Diagnosed by inexpensive urine porphobilinogen test
- Can we detect rare diseases earlier using EHR data?

27

OHSU

Methods

- Expanded EHR data set to 200+ K patients
  - Updated base data set to 200K patients
    - Including from post-2015 era of ICD-10-CM coding
  - Enriched with 5,571 additional patients having "porph" in diagnoses, lab tests, and notes
    - 308 with ICD-9-CM 277.1 – Disorders of porphyrin metabolism
- Preparation for machine learning
  - Positive training cases from ICD-10-CM E80.21 with manual review to verify
  - Negative training cases were the rest

28

OHSU

Machine learning approach

- Parsed EHR record into features
  - Unigrams and bigrams of text
  - Labeled features by the EHR source document
  - Scored by frequency of appearance
- Univariate feature analysis
  - Manually choose features not directly tied to provider attributes or suspecting patient had porphyria, e.g., "DeLoughery" and "cimetidine"
- Trained on full dataset, with best performance using support vector machine (SVM) with radial basis function (RBF) kernel

Preliminary results from work in progress

- Applied trained model back to full data set – ranked patients by margin distance
- Manually reviewed top 100 ranked "negative cases" for potential for porphyria
- Found cases with no diagnosis explaining symptoms
  - Very Likely – 1
  - Likely – 3
  - Possible – 18
- Note with natural prevalence, would expect 0.0005 cases out of 100

Next steps

- Follow up clinical study on the 34 possible cases
  - Contacting primary care provider via email and inform that computer model suggests testing for AIP
- Explore new machine learning approaches to identify additional patients for possible diagnosis
- Apply methodology to other rare diseases

Conclusions

- Cohort retrieval
  - With large EHR collections, classic word-based EHRs do not work well; structured queries required
- Rare disease surveillance
  - Early results, but promise for using EHR to facilitate diagnoses
- For both, need robust and accessible data to advance research methods

Challenges for EHR retrieval work

- Need for large and realistic data sets
  - Scalability of methods, especially for rare diseases
  - More generalizable to real world
- Big challenge is patient privacy
  - Data not readily sharable
  - Leading to concerns about reproducibility
- Can we solve privacy problems?
  - Exhaustive de-identification, including of notes – is it possible?
  - De-identification with controlled access (Halamka, 2020)
  - Evaluation as a service (Hanbury, 2015; Roegiest, 2016)

OHSU

Opportunities going forward

- Upside value
  - Developing generalizable methods to achieve the value of "secondary use" of EHR data first envisioned by AMIA (Safran, 2007)
- Must move beyond predictions (Matheny, NAM, 2019)
  - Need actionable data that can improve health, care outcomes, care delivery, etc.
  - Done in ways that do not exacerbate bias and inequities

OHSU

Questions?

William Hersh, MD
Professor and Chair
Department of Medical Informatics & Clinical Epidemiology
School of Medicine
Oregon Health & Science University
Portland, OR, USA
http://www.ohsu.edu/informatics

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