OHSU

Applying Information Retrieval to the Electronic Health Record for Cohort Discovery

William Hersh, MD
Division of Informatics, Clinical Epidemiology, and Translational Data Science
Oregon Health & Science University
Portland, OR, USA
https://billhersh.info/presentations

Overview

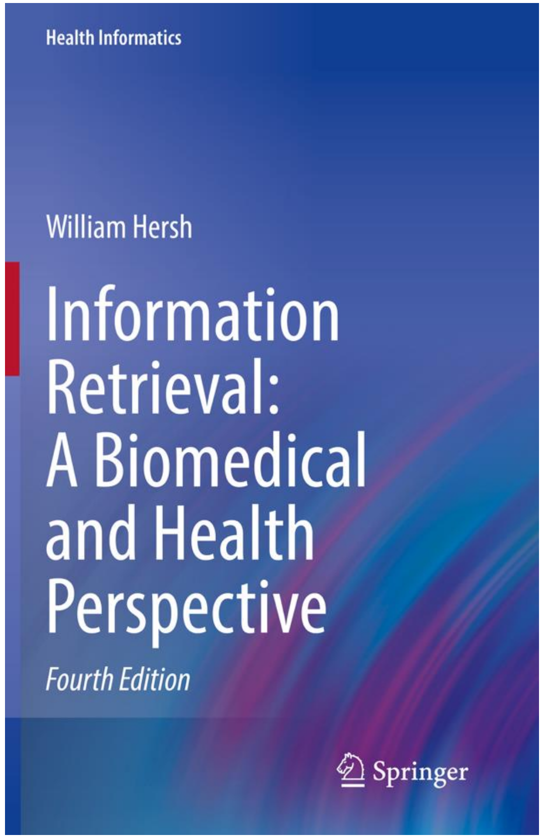
- Applying IR to the EHR
- Cohort discovery
- Challenges for EHR research

- This work funded by grant from
  - NLM 1R01LM011934

- With help from OHSU collaborators
  - Steven Bedrick
  - Jolie Kaner

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) (1996, 2002)
  - Participation/leadership of challenge evaluations, mainly Text Retrieval Conference (TREC) (Hersh, 2009; Voorhees, 2012; Roberts, 2016; Gupta, 2024)



(Hersh, 2020)

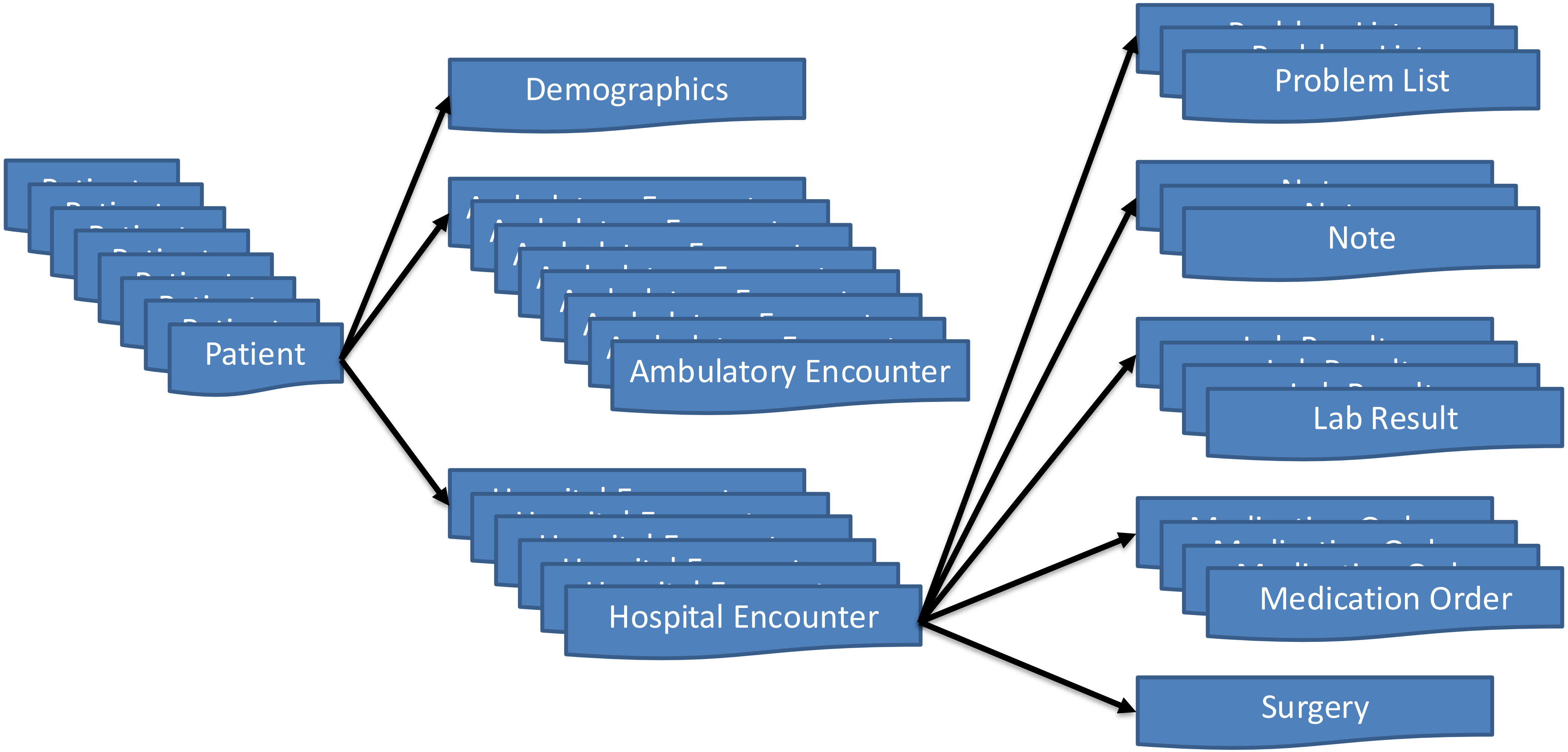
Applying IR to the EHR

- Increased 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, 2012; Voorhees, 2013)
- Awarding in 2014 of NIH R01 to (former) OHSU faculty Stephen Wu to explore methods in parallel with Mayo Clinic (Wu, 2017; Wang, 2019; Chamberlin, 2020)

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
  - For ranked output (Harman, 2011)
    - Mean average precision (MAP)
    - B-Pref – used when relevance judgments incomplete (Buckley, 2004)
    - Cumulative gain and other "inferred" measures (Järvelin, 2022)
- Usual approach in test collections is to average results across topics for each "run"

Electronic health record (EHR) is (typically) not a single document



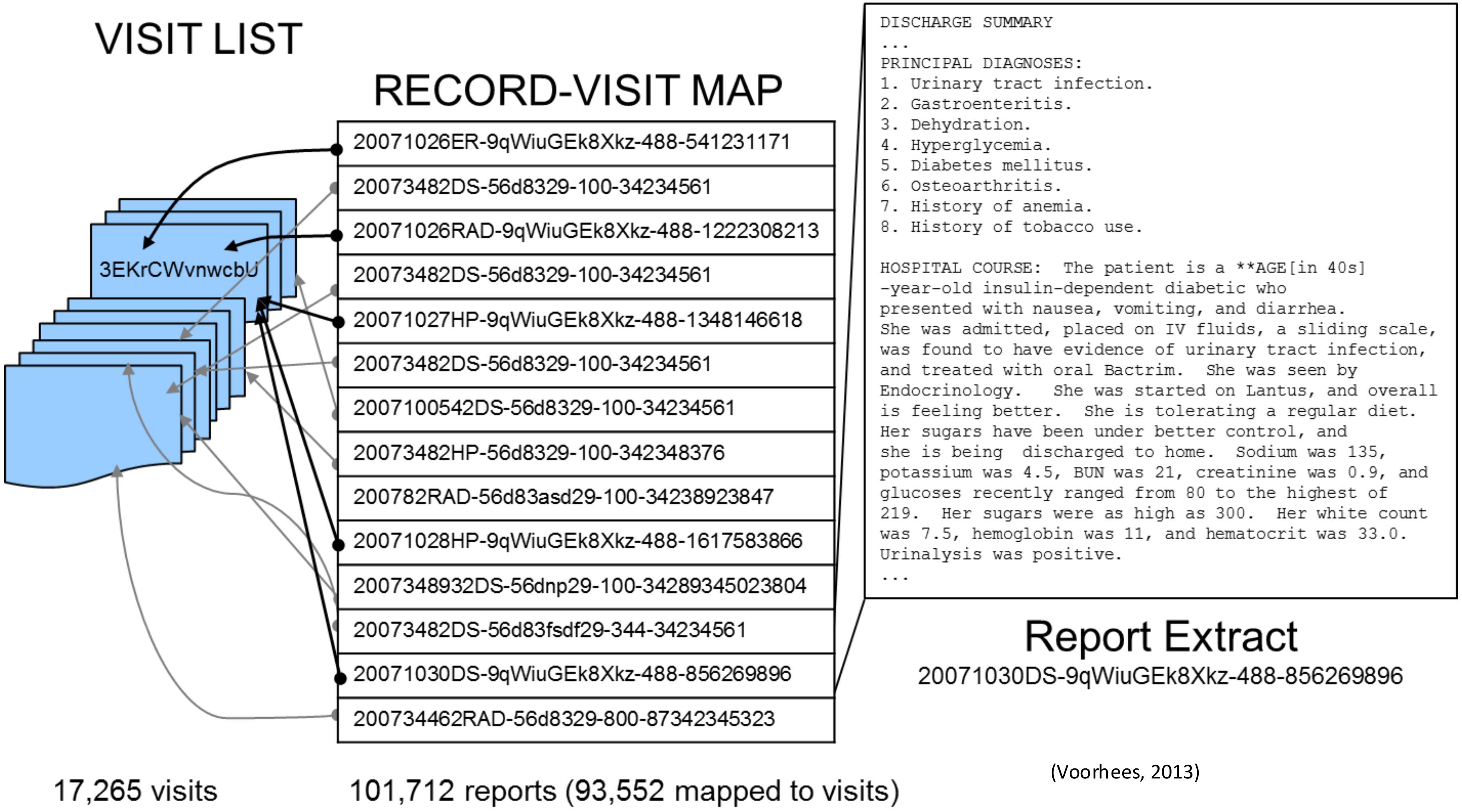
Cohort discovery

- Widely offered service by most academic medical centers to identify cohorts to retrieve for clinical study recruitment but little formal evaluation of approaches
- Early work – TREC Medical Records Track, 2011-2012
- Follow-on collaboration funded by NLM R01

TREC Medical Records Track (Voorhees, 2012; Voorhees, 2013)

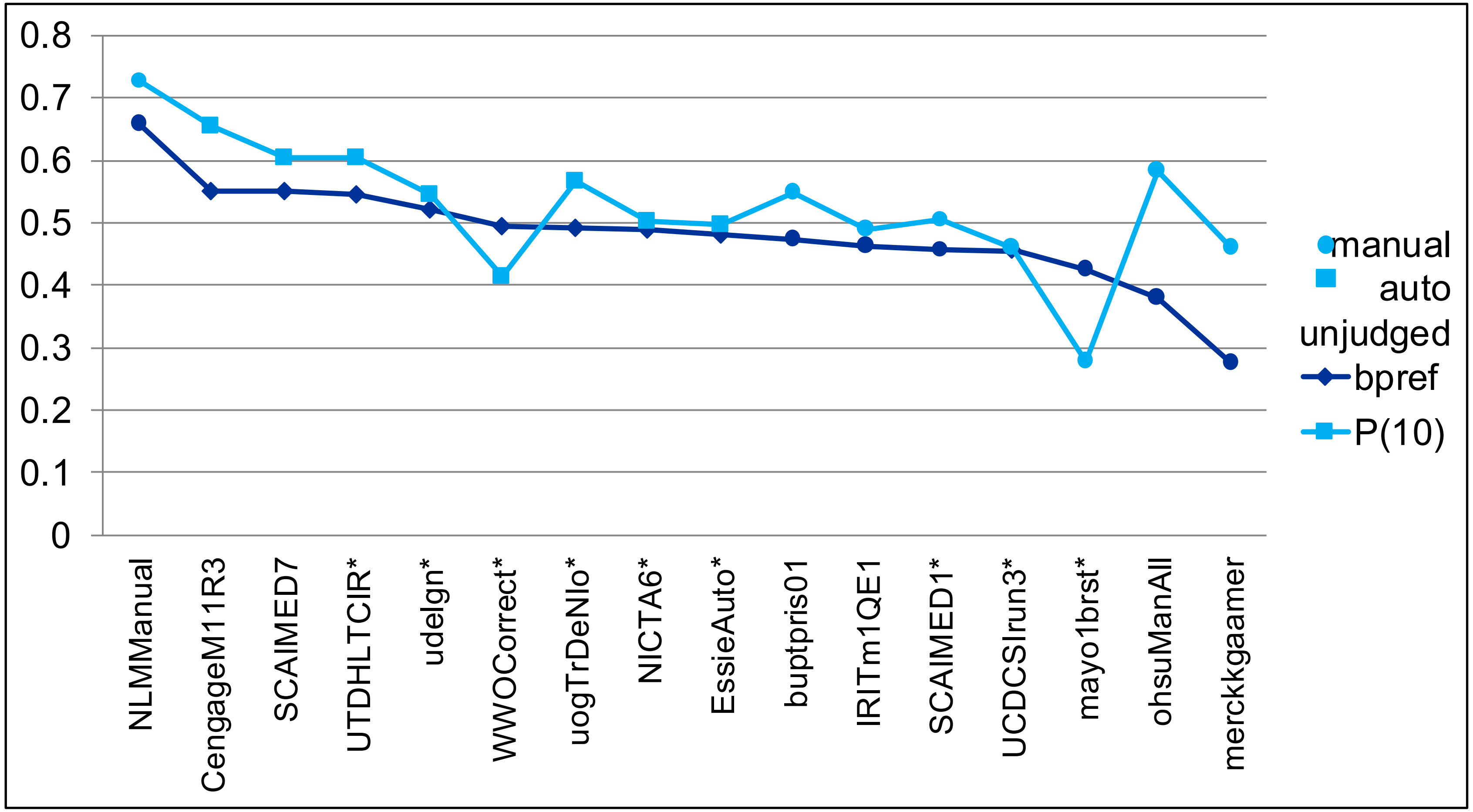
- Task – identify patients who are possible candidates for clinical studies/trials
- Documents – de-identified patient records from University of Pittsburgh Medical Center (UPMC)
- Topics – 35 clinical study topics from IOM key areas for comparative effectiveness research
- Relevance judgments – patients "relevant" to topics, judged by OHSU informatics students who were also physicians

Test collection structure

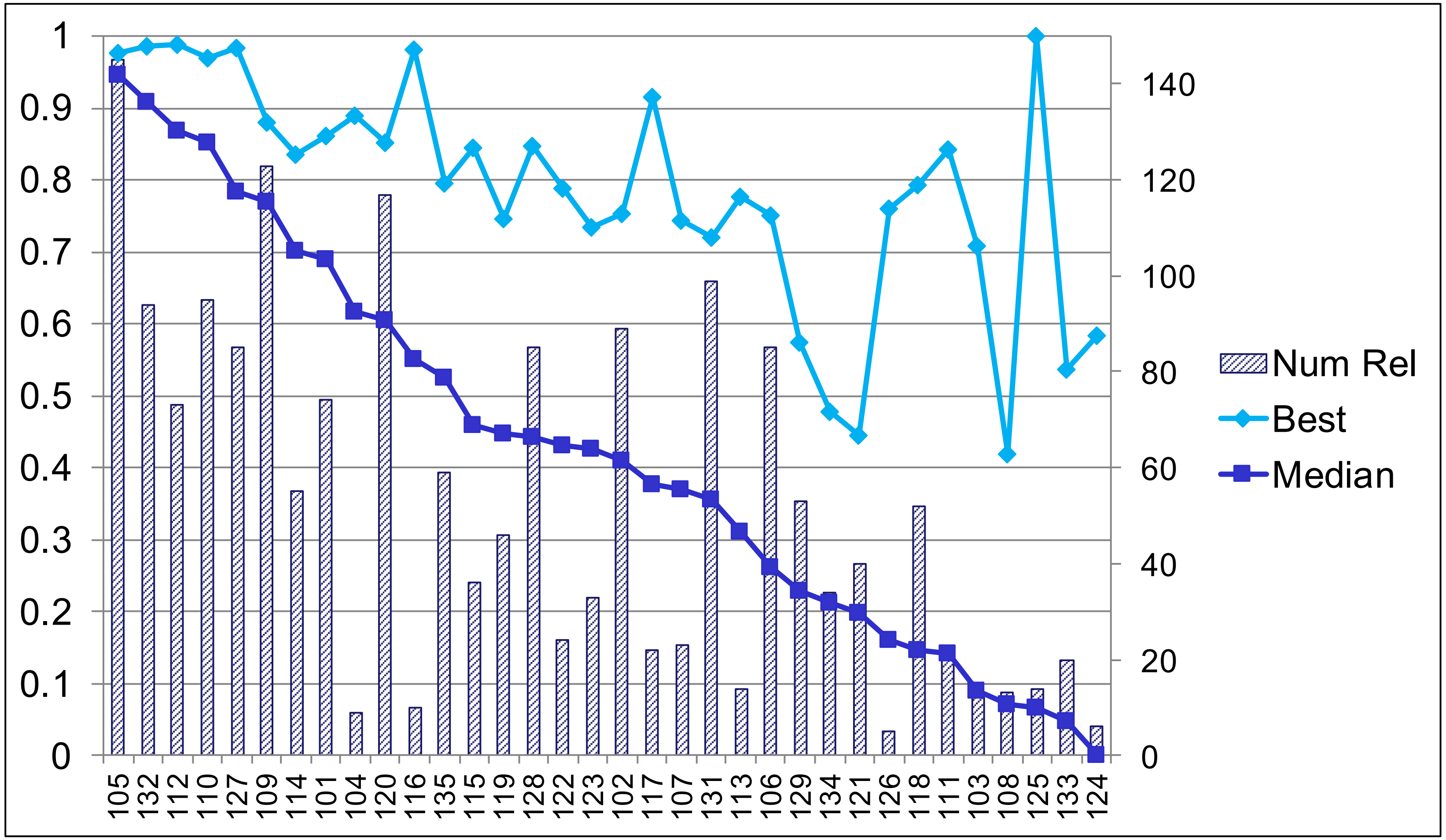


(Voorhees, 2013)

Evaluation results for top runs – 2011



But as commonly seen in IR, wide variation across topics



“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

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

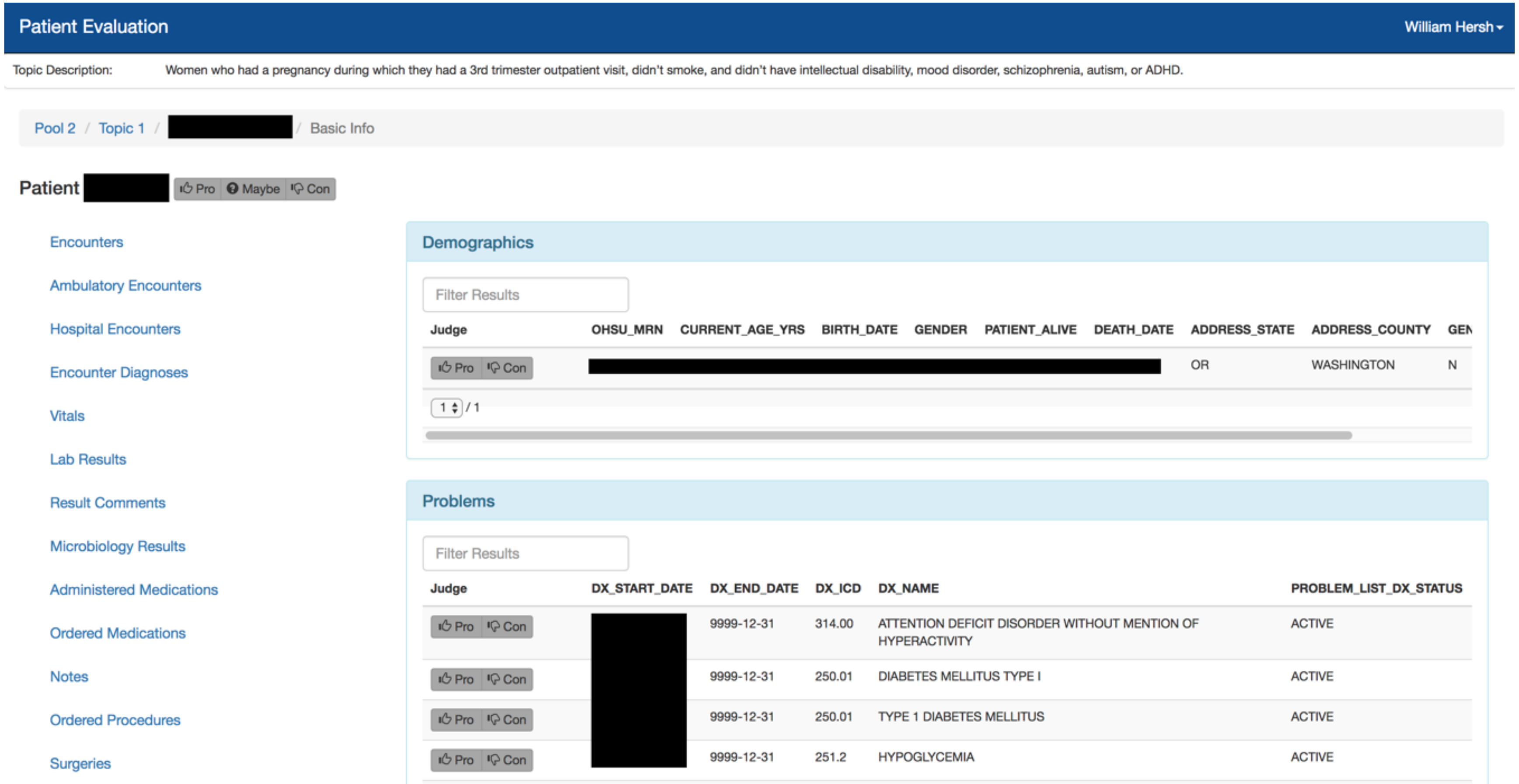
Extending cohort discovery work

- Mutli-site collaboration of OHSU, Mayo Clinic, and University of Texas Houston Health Science Center
- 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
- Initial methods (Wu, 2017) and results (Chamberlin, 2020)

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

Judgments from Patient Relevance Assessment Interface (PRAI)



Topic examples – summary and full

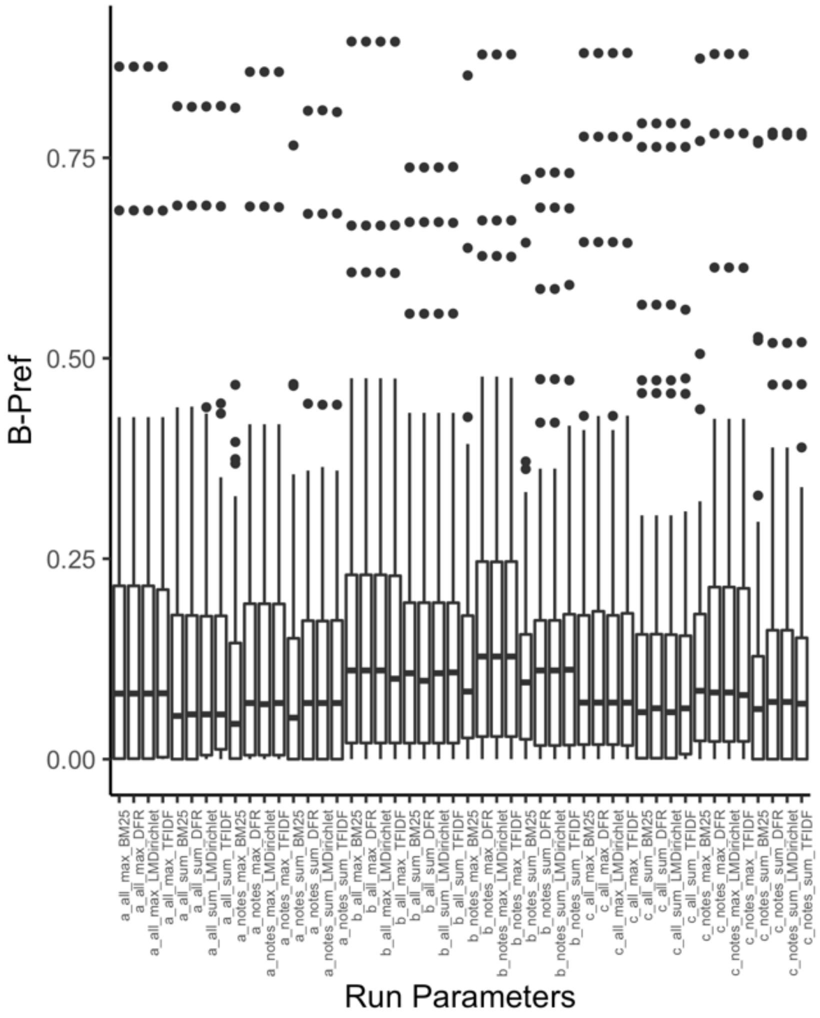
Adults with IBD who haven't had GI surgery	Adults with inflammatory bowel disease who haven't had surgery involving the small intestine, colon, rectum, or anus.
Adults with a Vitamin D lab result	Adults with a lab result for 25-hydroxy Vitamin D collected between May 15 and October 15.
Postherpetic neuralgia 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.

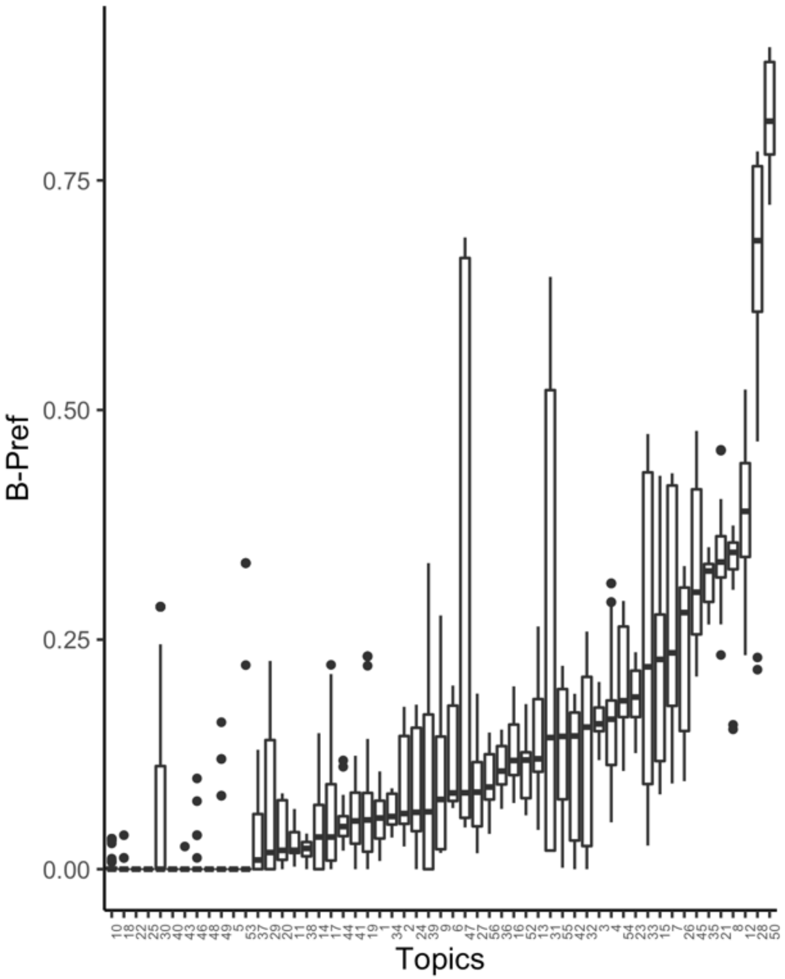
A. **Adults 18-64 years old with rheumatoid arthritis who have had a lab test for cyclic citrullinated peptide IgG antibody with a result greater than 40 units.**

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.**

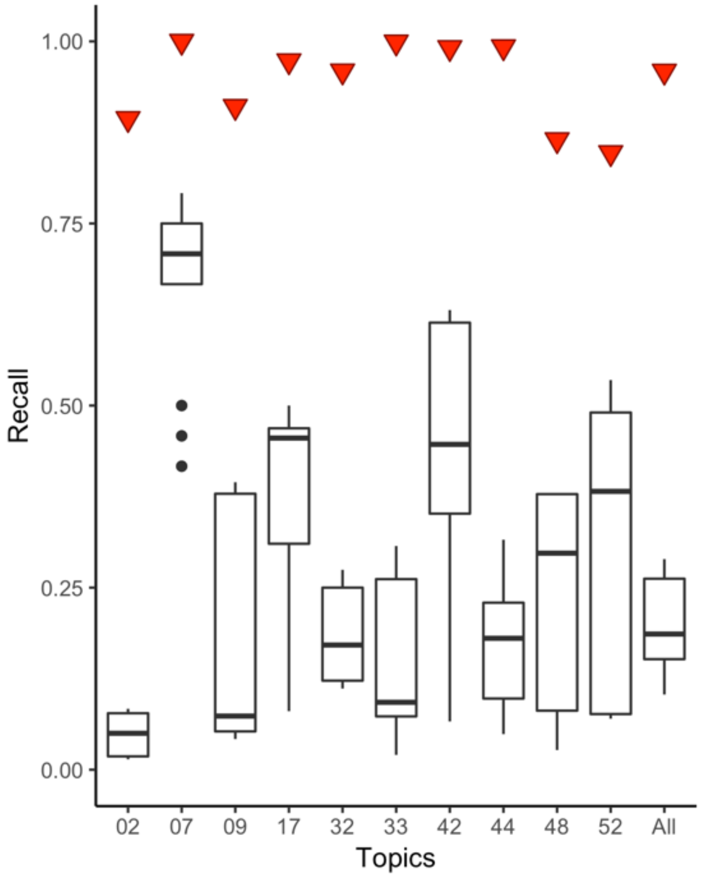
C. **Adults 18-64 years old with rheumatoid arthritis and lab result for positive anti-CCP IgG > 40 units.**
   - I. **Demographics inclusion**
     - a. **Age: 18-64 years**
   - II. **Diagnosis inclusion**
     - a. **Rheumatoid arthritis (ICD-9): 714.0**
   - III. **Lab inclusion**
     - a. **Cyclic citrullinated peptide IgG antibody (anti-CCP IgG): > 40 units**

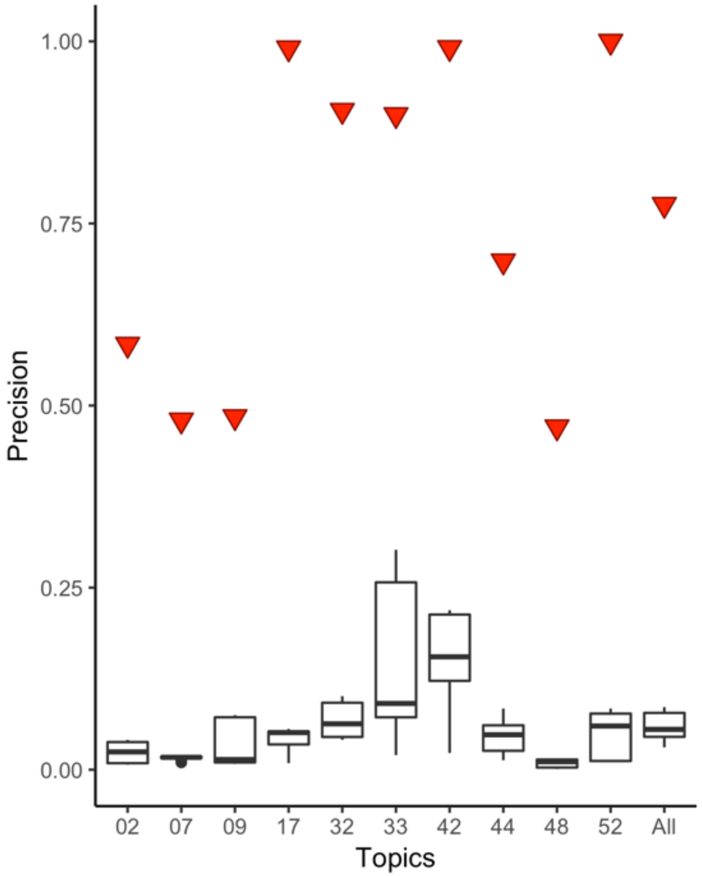
Simple word-based query performance not optimal for most topics





Reformulating as Boolean queries with additional relevance judgments for 10 topics





Improved relative recall and precision

Results and future work …

- Focus on 13 topics due to resource constraints
- To facilitate comparative methods and results without sharing data, standardizing all data, code, and methods using Observational Medical Outcomes Partnership (OMOP) format (Yates, 2025)
- Current results show best performance with Boolean over word-based methods
- Role for LLM-based methods (Chen, 2025; Layne, 2025)?
  - Most evaluated is TrialGPT
    - Applies state-of-art LLMs to retrieval, matching, and ranking – comparable performance to human experts (Jin, 2024)
    - Identified twice as many candidates while maintaining specificity (Syed, 2026)

OHSU

Challenges for EHR retrieval work

- Need for large and realistic data sets
  - Scalability of methods
  - 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?
  - Controlled access to identified data (Guerrero, 2019)
  - Evaluation as a service (Roegiest, 2016; Hopfgartner, 2018; Fröbe, 2023)

Questions?

William Hersh, MD
Professor
Division of Informatics, Clinical Epidemiology, and Translational Data Science
Department of Medicine
School of Medicine
Oregon Health & Science University
Portland, OR, USA
Email – hersh@ohsu.edu
Web – https://billhersh.info/

References Cited

Buckley, C., Voorhees, E.M., 2004. Retrieval evaluation with incomplete information, in: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR ’04. Association for Computing Machinery, New York, NY, USA, pp. 25–32. https://doi.org/10.1145/1008992.1009000

Chamberlin, S.R., Bedrick, S.D., Cohen, A.M., Wang, Y., Wen, A., Liu, S., Liu, H., Hersh, W.R., 2020. Evaluation of patient-level retrieval from electronic health record data for a cohort discovery task. JAMIA Open 3, 395–404. https://doi.org/10.1093/jamiaopen/ooaa026

Chen, H., Li, X., He, X., Chen, A., McGill, J., Webber, E.C., Xu, H., Liu, M., Bian, J., 2025. Enhancing Patient-Trial Matching With Large Language Models: A Scoping Review of Emerging Applications and Approaches. JCO Clin Cancer Inform 9, e2500071. https://doi.org/10.1200/CCI-25-00071

Edinger, T., Cohen, A.M., Bedrick, S., Ambert, K., Hersh, W., 2012. Barriers to retrieving patient information from electronic health record data: failure analysis from the TREC Medical Records Track. AMIA Annu Symp Proc 2012, 180–188.

Fröbe, M., Reimer, J.H., MacAvaney, S., Deckers, N., Reich, S., Bevendorff, J., Stein, B., Hagen, M., Potthast, M., 2023. The Information Retrieval Experiment Platform, in: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR ’23. Association for Computing Machinery, New York, NY, USA, pp. 2826–2836. https://doi.org/10.1145/3539618.3591888

Guerrero, S.C., Sridhar, S., Edmonds, C., Solis, C.F., Zhang, J., McPherson, D.D., Bernstam, E.V., 2019. Access to Routinely Collected Clinical Data for Research: A Process Implemented at an Academic Medical Center. Clin Transl Sci 12, 231–235. https://doi.org/10.1111/cts.12614

Gupta, D., Demner-Fushman, D., Hersh, W., Bedrick, S., Roberts, K., 2024. Overview of TREC 2024 Biomedical Generative Retrieval (BioGen) Track. https://doi.org/10.48550/arXiv.2411.18069

Harman, D., Marchionini, G., 2011. Information Retrieval Evaluation, 1st edition. ed. Morgan & Claypool Publishers, San Rafael, Calif.

Hersh, W., 2020. Information Retrieval: A Biomedical and Health Perspective, 4th ed, Health Informatics. Springer International Publishing.

Hersh, W., Hickam, D., 1994. Use of a multi-application computer workstation in a clinical setting. Bull Med Libr Assoc 82, 382–389.

Hersh, W., Pentecost, J., Hickam, D., 1996. A task-oriented approach to information retrieval evaluation. Journal of the American Society for Information Science 47, 50–56. https://doi.org/10.1002/(SICI)1097-4571(199601)47:1%253C50::AID-ASI5%253E3.0.CO;2-1

Hersh, W., Voorhees, E., 2009. TREC genomics special issue overview. Inf Retrieval 12, 1–15. https://doi.org/10.1007/s10791-008-9076-6

Hersh, W.R., Crabtree, M.K., Hickam, D.H., Sacherek, L., Friedman, C.P., Tidmarsh, P., Mosbaek, C., Kraemer, D., 2002. Factors associated with success in searching MEDLINE and applying evidence to answer clinical questions. J Am Med Inform Assoc 9, 283–293. https://doi.org/10.1197/jamia.m0996

Hersh, W.R., Greenes, R.A., 1990. SAPHIRE--an information retrieval system featuring concept matching, automatic indexing, probabilistic retrieval, and hierarchical relationships. Comput. Biomed. Res. 23, 410–425. https://doi.org/10.1016/0010-4809(90)90031-7
Hersh, W.R., Hickam, D.H., 1998. How well do physicians use electronic information retrieval systems? A framework for investigation and systematic review. JAMA 280, 1347–1352. https://doi.org/10.1001/jama.280.15.1347
Hopfgartner, F., Hanbury, A., Müller, H., Eggel, I., Balog, K., Brodt, T., Cormack, G.V., Lin, J., Kalpathy-Cramer, J., Kando, N., Kato, M.P., Krithara, A., Gollub, T., Potthast, M., Viegas, E., Mercer, S., 2018. Evaluation-as-a-Service for the Computational Sciences: Overview and Outlook. J. Data and Information Quality 10, 15:1-15:32. https://doi.org/10.1145/3239570
Järvelin, K., Kekäläinen, J., 2002. Cumulated gain-based evaluation of IR techniques. ACM Trans. Inf. Syst. 20, 422–446. https://doi.org/10.1145/582415.582418
Jin, Q., Wang, Z., Floudas, C.S., Chen, F., Gong, C., Bracken-Clarke, D., Xue, E., Yang, Y., Sun, J., Lu, Z., 2024. Matching patients to clinical trials with large language models. Nat Commun 15, 9074. https://doi.org/10.1038/s41467-024-53081-z
Layne, E., Olivas, C., Hershenhouse, J., Ganjavi, C., Cei, F., Gill, I., Cacciamani, G.E., 2025. Large language models for automating clinical trial matching. Curr Opin Urol 35, 250–258. https://doi.org/10.1097/MOU.0000000000001281
Rajpurkar, P., Chen, E., Banerjee, O., Topol, E.J., 2022. AI in health and medicine. Nat Med 1–8. https://doi.org/10.1038/s41591-021-01614-0
Roberts, K., Simpson, M., Demner-Fushman, D., Voorhees, E., Hersh, W., 2016. State-of-the-art in biomedical literature retrieval for clinical cases: a survey of the TREC 2014 CDS track. Inf Retrieval J 19, 113–148. https://doi.org/10.1007/s10791-015-9259-x
Roegiest, A., Cormack, G.V., 2016. An Architecture for Privacy-Preserving and Replicable High-Recall Retrieval Experiments, in: Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '16. Association for Computing Machinery, New York, NY, USA, pp. 1085–1088. https://doi.org/10.1145/2911451.2911456
Safran, C., Bloomrosen, M., Hammond, W.E., Labkoff, S., Markel-Fox, S., Tang, P.C., Detmer, D.E., Expert Panel, null, 2007. Toward a national framework for the secondary use of health data: an American Medical Informatics Association White Paper. J Am Med Inform Assoc 14, 1–9. https://doi.org/10.1197/jamia.M2273
Syed, M., Hamidi, M., Bikkanuri, M., Dierschke, N.A., Katragadda, H.V., Zozus, M., Teixeira, A.L., 2026. Translating evidence into practice: adapting TrialGPT for real-world clinical trial eligibility screening. J Am Med Inform Assoc 33, 909–913. https://doi.org/10.1093/jamia/ocag006
Voorhees, E.M., 2013. The TREC Medical Records Track, in: Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics, BCB'13. Association for Computing Machinery, New York, NY, USA, pp. 239–246. https://doi.org/10.1145/2506583.2506624
Voorhees, E.M., Harman, D.K. (Eds.), 2005. TREC: Experiment and Evaluation in Information Retrieval. The MIT Press, Cambridge, Mass.
Voorhees, E.M., Hersh, W., 2012. Overview of the TREC 2012 Medical Records Track, in: The Twenty-First Text REtrieval Conference (TREC 2012) Proceedings.

Wang, Y., Wen, A., Liu, S., Hersh, W., Bedrick, S., Liu, H., 2019. Test collections for electronic health record-based clinical information retrieval. Jamia Open 2, 360–368. https://doi.org/10.1093/jamiaopen/ooz016

Wu, S., Liu, S., Wang, Y., Timmons, T., Uppili, H., Bedrick, S., Hersh, W., Liu, H., 2017. Intrainstitutional EHR collections for patient-level information retrieval. Journal of the Association for Information Science and Technology 68, 2636–2648. https://doi.org/10.1002/asi.23884

Yates, A., Benton, E., Humes, I., Lawhead, M., Harrelson, H., Bentley, I., Mahmood, R., Hersh, W., Bedrick, S., 2025. The OMOP Annotator: A Database Agnostic Tool for Reviewing and Augmenting the Patient Record – OHDSI, in: ODHSI 2025 Global Collaborator Showcase. Presented at the ODHSI 2025 Global Symposium.