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

References

Balwani, M., Sardh, E., Ventura, P., Peiró, P.A., Rees, D.C., Stölzel, U., Bissell, D.M., Bonkovsky, H.L., Windyga, J., Anderson, K.E., Parker, C., Silver, S.M., Keel, S.B., Wang, J.-D., Stein, P.E., Harper, P., Vassiliou, D., Wang, B., Phillips, J., Ivanova, A., Langendonk, J.G., Kauppinen, R., Minder, E., Horie, Y., Penz, C., Chen, J., Liu, S., Ko, J.J., Sweetser, M.T., Garg, P., Vaishnaw, A., Kim, J.B., Simon, A.R., Gouya, L., ENVISION Investigators, 2020. Phase 3 Trial of RNAi Therapeutic Givosiran for Acute Intermittent Porphyria. N Engl J Med 382, 2289–2301. https://doi.org/10.1056/NEJMoa1913147

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

Cohen, A.M., Chamberlin, S., Deloughery, T., Nguyen, M., Bedrick, S., Meninger, S., Ko, J.J., Amin, J.J., Wei, A.J., Hersh, W., 2020. Detecting rare diseases in electronic health records using machine learning and knowledge engineering: Case study of acute hepatic porphyria. PLoS ONE 15, e0235574. https://doi.org/10.1371/journal.pone.0235574

Demner-Fushman, D., Abhyankar, S., Jimeno-Yepes, A., Loane, R., Lang, F., Mork, J.G., Ide, N., Aronson, A.R., 2012. NLM at TREC 2012 Medical Records Track, in: The Twenty-First Text REtrieval Conference (TREC 2012) Proceedings.

Demner-Fushman, D., Abhyankar, S., Jimeno-Yepes, A., Loane, R., Rance, B., Lang, F., Ide, N., Apostolova, E., Aronson, A.R., 2011. A Knowledge-Based Approach to Medical Records Retrieval, in: The Twentieth Text REtrieval Conference (TREC 2011) Proceedings.

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.

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

Haendel, M., Vasilevsky, N., Unni, D., Bologa, C., Harris, N., Rehm, H., Hamosh, A., Baynam, G., Groza, T., McMurry, J., Dawkins, H., Rath, A., Thaxon, C., Bocci, G., Joachimiak, M.P., Köhler, S., Robinson, P.N., Mungall, C., Oprea, T.I., 2020. How many rare diseases are there? Nat Rev Drug Discov 19, 77–78. https://doi.org/10.1038/d41573-019-00180-y

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

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<50::AID-ASI5>3.0.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., 2022. Health Informatics: Practical Guide, 8th Edition. Lulu.com.

Hersh, W.R., Cohen, A.M., Nguyen, M.M., Bensching, K.L., Deloughery, T.G., 2022. Clinical study applying machine learning to detect a rare disease: results and lessons learned. JAMIA Open 5, ooac053. https://doi.org/10.1093/jamiaopen/ooac053

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

Ide, N.C., Loane, R.F., Demner-Fushman, D., 2007. Essie: a concept-based search engine for structured biomedical text. J Am Med Inform Assoc 14, 253–263. https://doi.org/10.1197/jamia.M2233

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

King, B., Wang, L., Provalov, I., Zhou, J., 2011. Cengage Learning at TREC 2011 Medical Track, in: The Twentieth Text REtrieval Conference (TREC 2011) Proceedings.

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., Alam, T., Bedrick, S., Demner-Fushman, D., Lo, K., Soboroff, I., Voorhees, E., Wang, L.L., Hersh, W.R., 2021. Searching for scientific evidence in a pandemic: An overview of TREC-COVID. J Biomed Inform 121, 103865. https://doi.org/10.1016/j.jbi.2021.103865

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

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

Zhu, D., Wu, S., Carterette, B., Liu, H., 2014. Using large clinical corpora for query expansion in text-based cohort identification. J Biomed Inform 49, 275–281. https://doi.org/10.1016/j.jbi.2014.03.010

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

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

1

Overview

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

2

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) (1996, 2002)
  - Participation/leadership of challenge evaluations, mainly TREC (2009; Voorhees, 2012; Roberts, 2016; Roberts, 2021)
- Overview textbook: Hersh W, *Information Retrieval: A Biomedical & Health Perspective, 4th Edition* (Springer, 2020)

Health Informatics

William Hersh

Information Retrieval: A Biomedical and Health Perspective

Fourth Edition

Springer
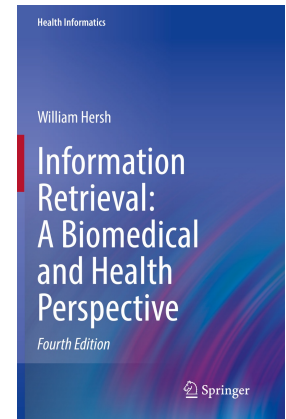
3

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, 2012; Voorhees, 2013)
- 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, 2020)
- Added rare disease surveillance task in collaboration with Alnylam Pharmaceuticals

4

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)

5

Electronic health record (EHR) structure
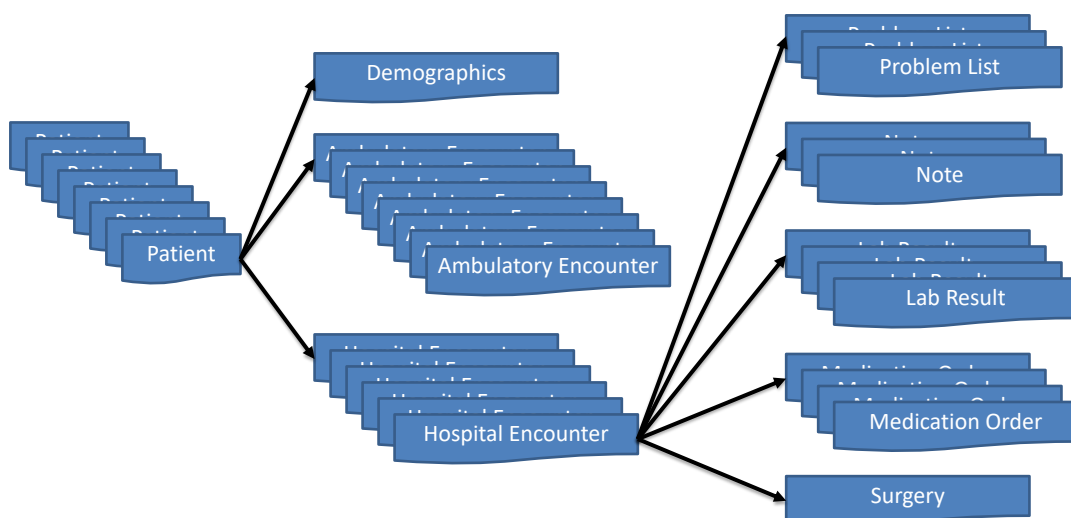
6

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 funded by NLM R01

TREC Medical Records Track

- Appealing task given HITECH investment and EHR adoption
  - 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

9

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
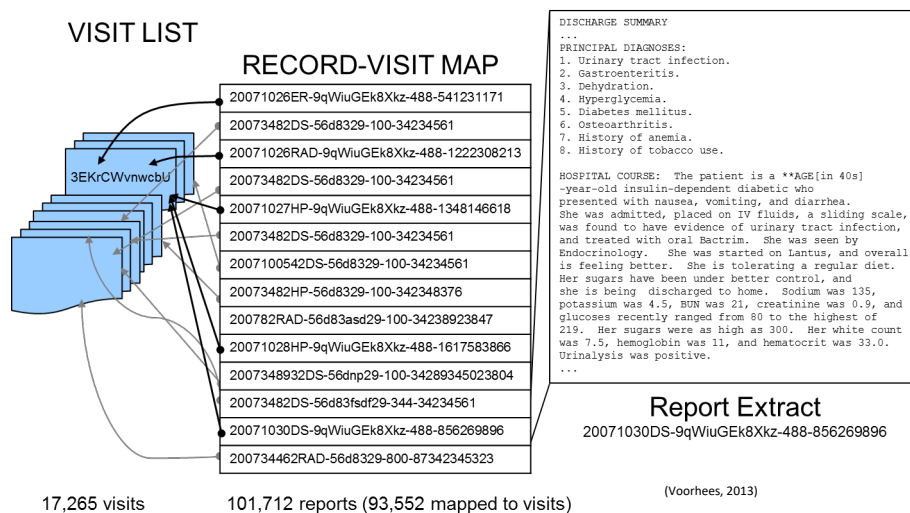
10

Test collection structure

VISIT LIST

RECORD-VISIT MAP

3EKrCWvnwcbU

20071026ER-9qWiuGEk8Xkz-488-541231171
20073482DS-56d8329-100-34234561
20071026RAD-9qWiuGEk8Xkz-488-1222308213
20073482DS-56d8329-100-34234561
20071027HP-9qWiuGEk8Xkz-488-1348146618
20073482DS-56d8329-100-34234561
2007100542DS-56d8329-100-34234561
20073482HP-56d8329-100-342348376
200782RAD-56d83asd29-100-34238923847
20071028HP-9qWiuGEk8Xkz-488-1617583866
2007348932DS-56dnp29-100-34289345023804
20073482DS-56d83fsdf29-344-34234561
20071030DS-9qWiuGEk8Xkz-488-856269896
200734462RAD-56d8329-800-87342345323

```
DISCHARGE SUMMARY
...
PRINCIPAL DIAGNOSES:
1. Urinary tract infection.
2. Gastroenteritis.
3. Dehydration.
4. Hyperglycemia.
5. Diabetes mellitus.
6. Osteoarthritis.
7. History of anemia.
8. History of tobacco use.

HOSPITAL COURSE:  The patient is a **AGE[in 40s]
-year-old insulin-dependent diabetic who
presented with nausea, vomiting, and diarrhea.
She was admitted, placed on IV fluids, a sliding scale,
was found to have evidence of urinary tract infection,
and treated with oral Bactrim.  She was seen by
Endocrinology.   She was started on Lantus, and overall
is feeling better.  She is tolerating a regular diet.
Her sugars have been under better control, and
she is being  discharged to home.  Sodium was 135,
potassium was 4.5, BUN was 21, creatinine was 0.9, and
glucoses recently ranged from 80 to the highest of
219.  Her sugars were as high as 300.  Her white count
was 7.5, hemoglobin was 11, and hematocrit was 33.0.
Urinalysis was positive.
...
```

Report Extract
20071030DS-9qWiuGEk8Xkz-488-856269896

17,265 visits   101,712 reports (93,552 mapped to visits)   (Voorhees, 2013)

11
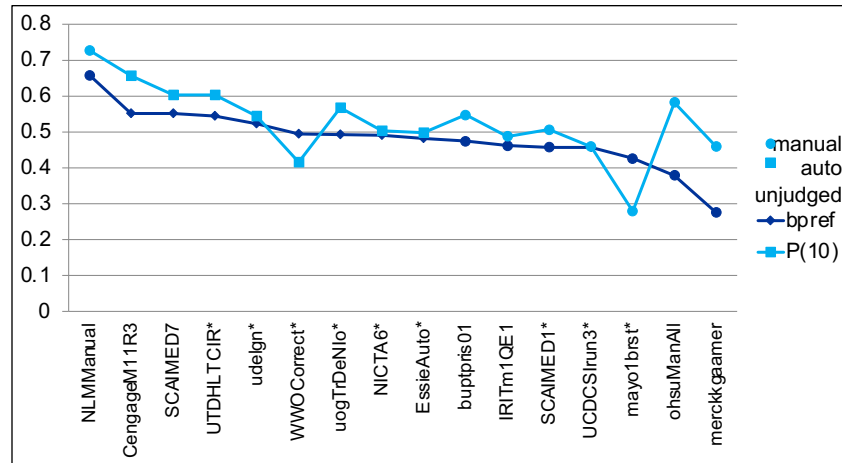
Some issues for test collection

- De-identified to remove protected health information (PHI), e.g., age number → range
- De-identification precluded linkage of same patient across different visits (encounters)
- UPMC only authorized use for TREC 2011 and TREC 2012 but no longer available
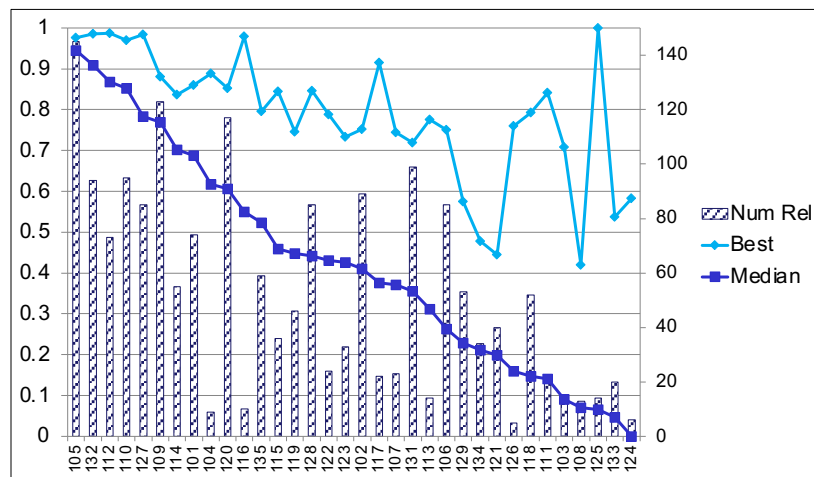
12

Evaluation results for top runs – 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
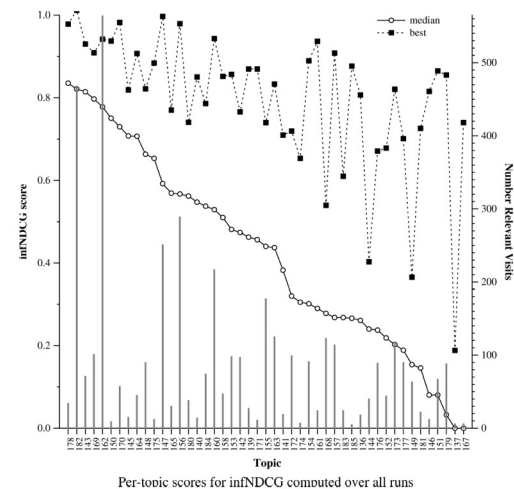
What approaches did (and did not) work?

- Best results in 2011 and 2012 from NLM group (Demner-Fushman, 2011; Demner-Fushman, 2012)
  - 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, 2020)

19

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

20

Judgments from Patient Relevance Assessment Interface (PRAI)

Patient Evaluation · William Hersh

Topic Description: Women who had a pregnancy during which they had a 3rd trimester outpatient visit, didn't smoke, and didn't have intellectual disability, mood disorder, schizophrenia, autism, or ADHD.

Pool 2 / Topic 1 / ▮ / Basic Info

Patient ▮ Pro ❓ Maybe 👎 Con

Encounters
Ambulatory Encounters
Hospital Encounters
Encounter Diagnoses
Vitals
Lab Results
Result Comments
Microbiology Results
Administered Medications
Ordered Medications
Notes
Ordered Procedures
Surgeries

Demographics

Filter Results

Judge	OHSU_MRN	CURRENT_AGE_YRS	BIRTH_DATE	GENDER	PATIENT_ALIVE	DEATH_DATE	ADDRESS_STATE	ADDRESS_COUNTY	GEN
👍 Pro 👎 Con							OR	WASHINGTON	N

1 / 1

Problems

Filter Results

Judge	DX_START_DATE	DX_END_DATE	DX_ICD	DX_NAME	PROBLEM_LIST_DX_STATUS
👍 Pro 👎 Con		9999-12-31	314.00	ATTENTION DEFICIT DISORDER WITHOUT MENTION OF HYPERACTIVITY	ACTIVE
👍 Pro 👎 Con		9999-12-31	250.01	DIABETES MELLITUS TYPE I	ACTIVE
👍 Pro 👎 Con		9999-12-31	250.01	TYPE 1 DIABETES MELLITUS	ACTIVE
👍 Pro 👎 Con		9999-12-31	251.2	HYPOGLYCEMIA	ACTIVE

21

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

22

## 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)

23

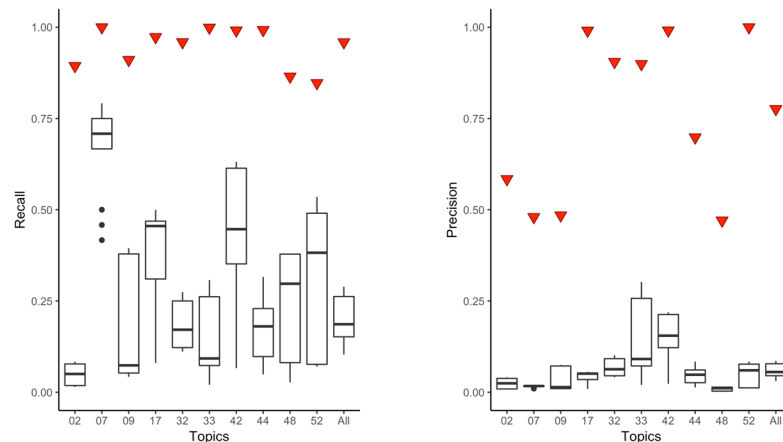## Word-based query performance not optimal for most topics
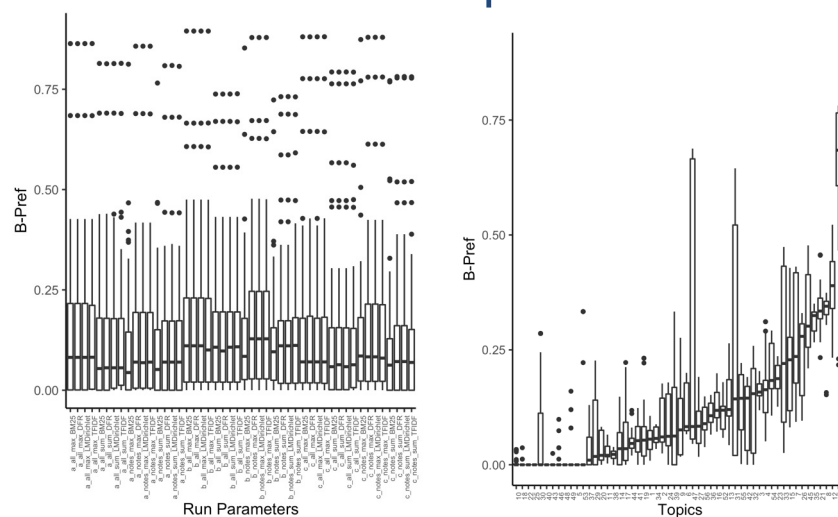
24

Reformulating as Boolean queries with additional relevance judgments for 10 topics

Improved relative recall and precision

25

Continued work

- For renewed R01, added third site, University of Texas Health Science Center Houston
- To facilitate comparative methods and results without sharing data, standardizing all data, code, and methods
  - Including new relevance judgment systems
- Adopting Observational Medical Outcomes Partnership (OMOP) format for data
  - Emerging standard for EHR data in research settings

26

Rare disease detection

- Over 1200 known rare disorders that affect < 1 in 200K patients worldwide, many under-diagnosed (https://rarediseases.org/; Haendel, 2020)
- Acute Intermittent Porphyria (AIP, aka Acute Hepatic Porphyria)
  - Rare genetic disease of heme biosynthesis – variable penetrance
  - Incidence 1 per 100K in population
  - Often undiagnosed for long time
  - Significant morbidity and effect on quality of life
  - "Neurovisceral" symptoms common with other diseases
    - Abdominal pain
    - Nausea and vomiting
    - Psychiatric changes
  - Diagnosed by inexpensive urine porphobilinogen test
  - New highly effective (and highly expensive) treatment available – RNA-silencing molecule givosiran (Balwani, 2020)

27

27

Can we detect rare diseases earlier using population-based techniques with EHR data?

- Funding from Alnylam Pharmaceuticals
- 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
- Preparation for machine learning
  - Positive training cases from ICD-10-CM E80.21 (47) with manual review to verify (30)
  - Negative training cases were the rest
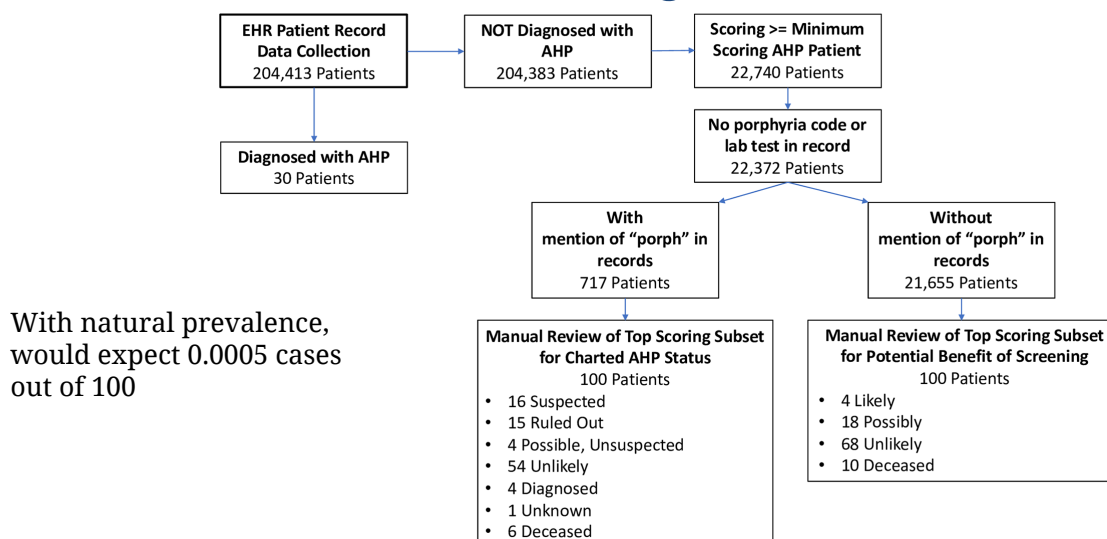
28

28

Machine learning approach (Cohen, 2020)

- Parsed EHR record into features – scored by frequency of appearance, labeled features by the EHR source document
- 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
- Applied trained model back to full data set – ranked patients by margin distance

Aimed to identify patients with symptoms but no consideration of diagnosis of AIP

| EHR Patient Record Data Collection 204,413 Patients | → | NOT Diagnosed with AHP 204,383 Patients | → | Scoring >= Minimum Scoring AHP Patient 22,740 Patients |

Diagnosed with AHP
30 Patients

No porphyria code or lab test in record
22,372 Patients

| With mention of "porph" in records 717 Patients | Without mention of "porph" in records 21,655 Patients |

With natural prevalence, would expect 0.0005 cases out of 100

Manual Review of Top Scoring Subset for Charted AHP Status
100 Patients
- 16 Suspected
- 15 Ruled Out
- 4 Possible, Unsuspected
- 54 Unlikely
- 4 Diagnosed
- 1 Unknown
- 6 Deceased

Manual Review of Top Scoring Subset for Potential Benefit of Screening
100 Patients
- 4 Likely
- 18 Possibly
- 68 Unlikely
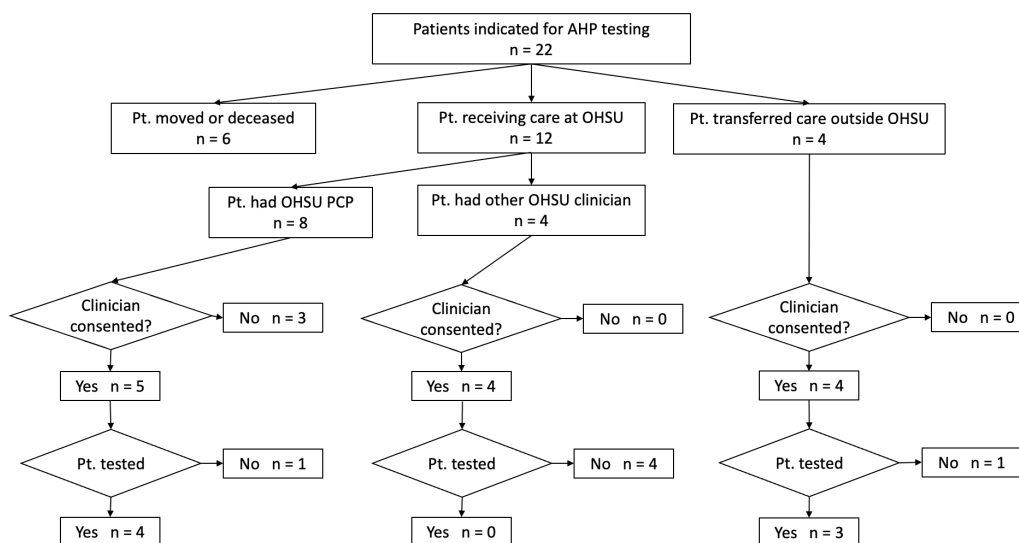- 10 Deceased

Clinical study (Hersh, 2022)

- Hindered by prolonged IRB process and COVID-19 pandemic, study was launched in late 2020
- IRB protocol required initial contact with primary care physician and, if they approved, offering the patient urine porphobilinogen testing
- Aimed to contact and enroll all 22 patients with AIP symptomatology but "unrecognized"

31

31

Clinician and patient participation

Patients indicated for AHP testing
n = 22

Pt. moved or deceased
n = 6

Pt. receiving care at OHSU
n = 12

Pt. transferred care outside OHSU
n = 4

Pt. had OHSU PCP
n = 8

Pt. had other OHSU clinician
n = 4

Clinician consented? — No n = 3 — Yes n = 5 — Pt. tested — No n = 1 — Yes n = 4

Clinician consented? — No n = 0 — Yes n = 4 — Pt. tested — No n = 4 — Yes n = 0

Clinician consented? — No n = 0 — Yes n = 4 — Pt. tested — No n = 1 — Yes n = 3

32

32

And the results showed...

- All 7 patients who came for testing had normal urine porphobilinogen

- Lessons learned
  - Identified patients who were reasonable candidates for testing
  - Two-step approval required for patients not under our care but complicated
  - Rare diseases are rare
  - For other diseases, testing may be expensive and/or harmful
  - (Don't attempt clinical validation studies of machine learning during a pandemic?)

33

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 (Guerrero, 2019)
  - Evaluation as a service (Roegiest, 2016; Hopfgartner, 2018)

34

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 (Rajpurkar, 2022)
  - Need actionable data that can improve health, care outcomes, care delivery, etc.
  - Done in ways that do not exacerbate bias and inequities

35

35

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

36

36

18