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


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.
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.


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
Information retrieval (IR, aka, search)

- We all do it – Google, PubMed, etc.
- As academics, we evaluate it – personal journey from
  - Participation/leadership of challenge evaluations, mainly TREC (2009; Voorhees, 2012; Roberts, 2016)

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**
  \[ R = \frac{\# \text{retrieved and relevant documents}}{\# \text{relevant documents in collection}} \]

- **Precision**
  \[ P = \frac{\# \text{retrieved and relevant documents}}{\# \text{retrieved documents}} \]

- **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**

- Patient
  - Demographics
  - Ambulatory Encounter
  - Hospital Encounter
- Problem List
- Note
- Lab Result
- Medication Order
- Surgery
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
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)

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)

<table>
<thead>
<tr>
<th>Reasons for Incorrect Retrieval</th>
<th>Number of Visits</th>
<th>Number of Topics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Visits Judged Not Relevant</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Topic terms mentioned as future possibility</td>
<td>16</td>
<td>9</td>
</tr>
<tr>
<td>Topic symptom/condition/procedure done in the past</td>
<td>22</td>
<td>9</td>
</tr>
<tr>
<td>All topic criteria present but not in the time/sequence specified by the topic description</td>
<td>19</td>
<td>6</td>
</tr>
<tr>
<td>Most, but not all, required topic criteria present</td>
<td>17</td>
<td>8</td>
</tr>
<tr>
<td>Topic terms denied or ruled out</td>
<td>19</td>
<td>10</td>
</tr>
<tr>
<td>Notes contain very similar term confused with topic term</td>
<td>13</td>
<td>11</td>
</tr>
<tr>
<td>Non-relevant reference in record to topic terms</td>
<td>37</td>
<td>18</td>
</tr>
<tr>
<td>Topic terms not present—unclear why record was ranked highly</td>
<td>14</td>
<td>8</td>
</tr>
<tr>
<td>Topic present—record is relevant—disagree with expert judgment</td>
<td>25</td>
<td>11</td>
</tr>
<tr>
<td>Visits Judged Relevant</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Topic not present—record is not relevant—disagree with expert judgment</td>
<td>44</td>
<td>21</td>
</tr>
<tr>
<td>Topic present in record but overlooked in search</td>
<td>103</td>
<td>27</td>
</tr>
<tr>
<td>Visit notes used a synonym or lexical variant for topic terms</td>
<td>22</td>
<td>10</td>
</tr>
<tr>
<td>Topic terms not named in notes and must be inferred</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Topic terms present in diagnosis list but not visit notes</td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>
Evaluation results from 2012 were comparable (Voorhees, 2012)

<table>
<thead>
<tr>
<th>Run</th>
<th>indNDCG</th>
<th>P(10)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NLMManual*</td>
<td>0.680</td>
<td>0.749</td>
</tr>
<tr>
<td>udelSUM</td>
<td>0.578</td>
<td>0.592</td>
</tr>
<tr>
<td>semnamed2</td>
<td>0.547</td>
<td>0.557</td>
</tr>
<tr>
<td>ohsuManBool*</td>
<td>0.526</td>
<td>0.611</td>
</tr>
<tr>
<td>atigoo1</td>
<td>0.524</td>
<td>0.519</td>
</tr>
<tr>
<td>UDinfoMed123</td>
<td>0.517</td>
<td>0.528</td>
</tr>
<tr>
<td>uogTrMCovQrd</td>
<td>0.500</td>
<td>0.553</td>
</tr>
<tr>
<td>NICTAUBC4</td>
<td>0.487</td>
<td>0.517</td>
</tr>
</tbody>
</table>

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)

---

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

<table>
<thead>
<tr>
<th>Type</th>
<th>Patients</th>
<th>Encounters</th>
<th>Records</th>
<th>Average</th>
<th>Median</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Administered Meds</td>
<td>47,208</td>
<td>1,25,631</td>
<td>6,497,157</td>
<td>51.634</td>
<td>6</td>
<td>-</td>
</tr>
<tr>
<td>Ambulatory Encounters</td>
<td>98,965</td>
<td>3,760,205</td>
<td>3,760,205</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Current Meds</td>
<td>92,783</td>
<td>-</td>
<td>31,997,402</td>
<td>344,863</td>
<td>64</td>
<td>20,102</td>
</tr>
<tr>
<td>Demographics</td>
<td>99,965</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Encounter Attributes</td>
<td>99,965</td>
<td>6,273,137</td>
<td>6,273,137</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Encounter Diagnoses</td>
<td>99,965</td>
<td>3,725,603</td>
<td>18,170,896</td>
<td>4.777</td>
<td>4</td>
<td>107</td>
</tr>
<tr>
<td>Notes</td>
<td>99,968</td>
<td>3,491,659</td>
<td>10,111,930</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Hospital Encounters</td>
<td>73,303</td>
<td>466,252</td>
<td>466,252</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Lab-Results</td>
<td>83,435</td>
<td>753,461</td>
<td>20,186,748</td>
<td>27,523</td>
<td>12</td>
<td>19488</td>
</tr>
<tr>
<td>Microbiology Results</td>
<td>27,515</td>
<td>65,373</td>
<td>296,548</td>
<td>4.536</td>
<td>1</td>
<td>268</td>
</tr>
<tr>
<td>Medications Ordered</td>
<td>94,089</td>
<td>1,388,086</td>
<td>5,336,506</td>
<td>3.845</td>
<td>1</td>
<td>1551</td>
</tr>
<tr>
<td>Procedures Ordered</td>
<td>96,514</td>
<td>1,880,309</td>
<td>7,229,854</td>
<td>3.845</td>
<td>1</td>
<td>6681</td>
</tr>
<tr>
<td>Problem List</td>
<td>90,722</td>
<td>-</td>
<td>761,260</td>
<td>8.391</td>
<td>6</td>
<td>182</td>
</tr>
<tr>
<td>Result Comments</td>
<td>72,716</td>
<td>468,814</td>
<td>916,534</td>
<td>1.925</td>
<td>1</td>
<td>691</td>
</tr>
<tr>
<td>Surgeries</td>
<td>18,640</td>
<td>29,895</td>
<td>51,889</td>
<td>1.067</td>
<td>1</td>
<td>41</td>
</tr>
<tr>
<td>Vitals</td>
<td>99,098</td>
<td>1,362,431</td>
<td>6,647,115</td>
<td>4.879</td>
<td>2</td>
<td>6387</td>
</tr>
</tbody>
</table>
Judgments from Patient Relevance Assessment Interface (PRAI)

To p i c e x a m p l e s – 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?

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
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)

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
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