Secondary Use of Clinical Data from Electronic Health Records: The TREC Medical Records Track

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

References Cited


Secondary Use of Clinical Data from Electronic Health Records: The TREC Medical Records Track

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

Overview

- Motivations for secondary use of clinical data
- Challenges for secondary use of clinical data
- Primer on information retrieval and related topics
- TREC Medical Records Track
- Conclusions and future directions
Motivations for secondary use of clinical data

- Many “secondary uses” or re-uses of electronic health record (EHR) data, including (Safran, 2007)
  - Personal health records (PHRs)
  - Clinical and translational research – generating hypotheses and facilitating research
  - Health information exchange (HIE)
  - Public health surveillance for emerging threats
  - Healthcare quality measurement and improvement

- Opportunities facilitated by growing incentives for “meaningful use” of EHRs in the HITECH Act (Blumenthal, 2011; Blumenthal, 2011), aiming toward the “learning healthcare system” (Friedman, 2010; Smith 2012)

- Successful demonstration that the phenotype in the EHR can be used with the genotype to replicate known associations as well as identify new ones, e.g., eMERGE (Kho, 2011; Denny, 2010)

Challenges for secondary use of clinical data

- EHR data does not automatically lead to knowledge
  - Data quality and accuracy is not a top priority for busy clinicians

- Little research, but problems identified
  - EHR data can be incorrect and incomplete, especially for longitudinal assessment (Berlin, 2011)
  - Much data is “locked” in text (Hripcsak, 2012)
  - Many steps in ICD-9 coding can lead to incorrectness or incompleteness (O’Malley, 2005)

- There are also important “provocations” about use of “big data” for research (Boyd, 2011)
Challenges (cont.)

• Many data “idiosyncrasies” (Weiner, 2011)
  – “Left censoring”: First instance of disease in record may not be when first manifested
  – “Right censoring”: Data source may not cover long enough time interval
  – Data might not be captured from other clinical (other hospitals or health systems) or non-clinical (OTC drugs) settings
  – Bias in testing or treatment
  – Institutional or personal variation in practice or documentation styles
  – Inconsistent use of coding or standards

Data in EHRs can be incomplete

• Claims data failed to identify more than half of patients with prognostically important cardiac conditions prior to admission for catheterization (Jollis, 1993)
• In Texas academic hospital, billing data alone only identified 22.7% and 52.2% respectively of patients with breast and endometrial cancer, increasing to 59.1% and 88.6% with a machine learning algorithm (Bernstam, 2010)
• At Columbia University Medical Center, 48.9% of patients with ICD-9 code for pancreatic cancers did not have corresponding disease documentation in pathology reports, with many data elements incompletely documented (Botsis, 2010)
Patients also get care at multiple sites

- Study of 3.7M patients in Massachusetts found 31% visited 2 or more hospitals over 5 years (57% of all visits) and 1% visited 5 or more hospitals (10% of all visits) (Bourgeois, 2010)
- Study of 2.8M emergency department (ED) patients in Indiana found 40% of patients had data at multiple institutions, with all 81 EDs sharing patients in a completely connected network (Finnell, 2011)

Primer on information retrieval (IR) and related topics

- Information retrieval
- Evaluation
- Challenge evaluations
Information retrieval (Hersh, 2009)

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

Elements of IR systems

- Retrieval
  - Boolean
  - Natural language

- Metadata

- Indexing
  - Words
  - Terms
  - Attributes

- Queries

- Search engine

- Content
Evaluation of IR systems

• 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
  \[ R = \frac{\text{#retrieved and relevant documents}}{\text{#relevant documents in collection}} \]
  
  - Usually use relative recall when not all relevant documents known, where denominator is number of known relevant documents in collection

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

Example of recall and precision

- Database
- Relevant
- Retrieved and relevant
- Retrieved

\[ R = \frac{30}{50} = 0.6 = 60\% \]
\[ P = \frac{30}{100} = 0.3 = 30\% \]
Some measures can be combined into a single aggregated measure

- Mean average precision (MAP) is mean of average precision for each topic (Harman, 2005)
  - Average precision is average of precision at each point of recall (relevant document retrieved)
  - Despite name, emphasizes recall
- Bpref accounts for when relevance information is significantly incomplete (Buckley, 2004)
- Normal discounted cumulative gain (NDCG) allows for graded relevance judgments (Jarvelin, 2002)
- MAP and NCDG can be “inferred” when there are incomplete judgments (Yilmaz, 2008)

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 of items to be retrieved – usually want 25-30 for “stability” (Buckley, 2000)
  - 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
Challenge evaluations (cont.)

• Typical flow of events in an IR challenge evaluation

Release of document collection to participating groups → Experimental runs and submission of results → Relevance judgments → Analysis of results

• In IR, challenge evaluation results usually show wide variation between topics and between systems
  – Should be viewed as relative, not absolute performance
  – Averages can obscure variations

Some well-known challenge evaluations in IR

• Text Retrieval Conference (TREC, trec.nist.gov; Voorhees, 2005) – sponsored by National Institute for Standards and Technology (NIST)
  – Many “tracks” of interest, such as routing/filtering, Web searching, question-answering, etc.
  – Non-medical, with exception of Genomics Track (Hersh, 2009)

• Cross-Language Evaluation Forum (CLEF, www.clef-campaign.org)
  – Focus on retrieval across languages, European-based
  – Additional focus on image retrieval, which includes medical image retrieval tasks (Hersh, 2009; Müller, 2010)

• Both operate on annual cycle of test collection release, experiments, and analysis of results
TREC Medical Records Track

- Appealing task given societal value and leveraging HITECH investment
  - NIST involved in HITECH in various ways
- Has always been easier with knowledge-based content than patient-specific data due to a variety of reasons
  - 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
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)

Wide variations in number of documents per visit

23 visits > 100 reports; max report size 415
Topic development and relevance assessments

- Task – Identify patients who are possible candidates for clinical studies/trials
  - Had to be done at “visit” level due to de-identification of records
- 2011 topics derived from 100 top critical medical research priorities in comparative effectiveness research (IOM, 2009)
- Topic development done as IR course student project
  - Selected 35 topics from 54 assessed for appropriateness for data and with at least some relevant “visits”
- Relevance judgments by OHSU informatics students who were physicians

Sample topics from 2011

- Patients taking atypical antipsychotics without a diagnosis of schizophrenia or bipolar depression
- Patients treated for lower extremity chronic wound
- Patients with atrial fibrillation treated with ablation
- Elderly patients with ventilator-associated pneumonia
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

Evaluation results for top runs ...
... 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 disectomy

- 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 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>

Topic development and relevance assessments for 2012 track

- Task – same as 2011
- Topic development same as 2011, but topics derived from
  - Unused 46 top critical medical research priorities in comparative effectiveness research (IOM, 2009) – 16
  - Meaningful use Stage 1 quality measures – 12
  - OHSUMED test collection literature retrieval topics recast for this task – 22
- Relevance judgments by OHSU and other BMI students who were physicians
  - 25 physicians judged 1-9 full topics each
Participation in 2012

- Runs consisted of ranked list of up to 1000 visits per topic for each of 50 topics
  - Automatic – no human intervention from input of topic statement to output of ranked list
  - Manual – everything else
- Up to 4 runs per participating group
- More judging resources than 2011 allowed more relevance judgments
  - For each topic, pooled top 15 from all runs and 25% of all documents ranked 16-100 by any run
- 88 runs submitted from 24 groups
  - 82 automatic
  - 6 manual

Preliminary results for 2012 – more details at conference Nov. 7-9
What approaches did (and did not) work?

- Best results in 2011 and 2012 obtained from NLM group (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 did provided small or inconsistent value for this task in 2011 (and probably 2012)
  - Document focusing, term expansion, etc.

Conclusions and future directions

- Growing amount of EHR data provides potential benefit for the learning healthcare system
  - Many challenges to use of EHR data exist – incompleteness and incorrectness
- TREC Medical Records Track extended IR challenge evaluation approach to a patient selection triage task
  - Initial results show mixed success for different methods – common with a new IR task
  - Best results so far from expert-constructed Boolean queries
  - IR techniques known to work well with news and literature documents do not work well for this task – new automated approaches required
- Future work also requires development of new test collections, which will be challenging not only due to resources but also privacy concerns
  - Do we need patient consent for data use?