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


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)

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 is inaccurate and incomplete, especially for longitudinal assessment (Berlin, 2011)
  – Many steps in process of ICD-9 assignment can lead to inaccuracy (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 is incomplete

• Claims data failed to identify more than half of patients with prognostically important cardiac conditions prior to admission for catheterization (Jollis, 1993)
• Various approaches generated variable rate of retrieval of cases for quality measurement (Benin, 2005; Rhodes, 2007); algorithmic methods can lead to improvement (Benin, 2011)
• 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)
Data incomplete (cont.)

• 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)

• Alerting system to add 17 problems to patient problem lists accepted 41% of time (Wright, 2012)

• Data from two medical centers in a Minnesota were found to better predict Type 2 diabetes mellitus than single center (Wei, 2012)

Patients get care in multiple places

• 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 common (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**

- **Content**

- **Search engine**

**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) – mean of average precision across topics
    - Bpref – takes into account retrieved but unjudged documents
- **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

\[ R = \frac{30}{50} = 0.6 = 60\% \]

\[ P = \frac{30}{100} = 0.3 = 30\% \]

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)
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)
  - 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)
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 nothing else, including TREC 2012 or any other research
Wide variations in number of documents per visit

![Bar chart showing wide variations in number of documents per visit.](image)

23 visits > 100 reports; max report size 415

(Courtesy, Ellen Voorhees, NIST)

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
- Topics derived from 100 top critical medical research priorities in comparative effectiveness research (IOM, 2009)
- Topic development done as IR course student project
  - Selected topics appropriate for data and with at least some relevant “visits”
- Relevance judgments by OHSU BMI students who were physicians
Sample topics

• 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

• 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 not do complete judgments, necessitating use of BPref for 1° 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 discectomy
- 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

What approaches did (and did not) work?

- Best results obtained from NLM group (Demner-Fushman, 2011)
  - Top results from manually constructed queries using Essie domain-specific search engine (Ide, 2007) – BPref = 0.658
  - Other automated processes fared less well, e.g., creation of PICO frames, negation, term expansion, etc. – BPref = 0.4822
- Best automated results also obtained by Cengage (King, 2011)
  - Filtered by age, race, gender, admission status; terms expanded by UMLS Metathesaurus – BPref = 0.552
- Benefits of approaches commonly successful in IR did provided small or inconsistent value for this task
  - Document focusing, term expansion, etc.
OHSU approach (Bedrick, 2011)

- Manually constructed queries of text and ICD-9 codes, run against all and high-yield (discharge summary, emergency department) documents
- Visits ranked by top-ranking documents
- Text and ICD-9 combined by Boolean operators

<table>
<thead>
<tr>
<th>Run</th>
<th>BPref</th>
<th>P@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text-only – All</td>
<td>0.3751</td>
<td>0.5853</td>
</tr>
<tr>
<td>Text-only – High</td>
<td>0.2894</td>
<td>0.4824</td>
</tr>
<tr>
<td>Text AND ICD-9 – All</td>
<td>0.2497</td>
<td>0.4471</td>
</tr>
<tr>
<td>Text AND ICD-9 – High</td>
<td>0.1695</td>
<td>0.3235</td>
</tr>
<tr>
<td>Text OR ICD-9 – All</td>
<td>0.3657</td>
<td>0.4618</td>
</tr>
<tr>
<td>Text OR ICD-9 – High</td>
<td>0.3238</td>
<td>0.4206</td>
</tr>
</tbody>
</table>

Example query (topic #127)
Text: (diabetes mellitus) OR diabetic OR DM OR hypertension
AND (morbid obesity)
ICD-9: 278.01 AND (250.* OR 401.* OR 405.*)

OHSU results – large variation by topic and method
Conclusions and future directions

- Growing amount of EHR data provides potential benefit for learning healthcare system
  - Many challenges to use of EHR data exist
  - One potentially beneficial technique is understanding of data in clinical narrative text
- 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
- Future work can hopefully proceed from this and other data sets – if there is continued access to the test collection allowed