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

Anonymous (2009). Initial National Priorities for Comparative Effectiveness Research. Washington, DC, Institute of Medicine. http://www.iom.edu/Reports/2009/ComparativeEffectivenessResearchPriorities.aspx.

Bedrick, S., Ambert, K., et al. (2011). Identifying Patients for Clinical Studies from Electronic Health Records: TREC Medical Records Track at OHSU. *The Twentieth Text REtrieval Conference Proceedings (TREC 2011)*, Gaithersburg, MD. National Institute for Standards and Technology.

Benin, A., Fenick, A., et al. (2011). How good are the data? Feasible approach to validation of metrics of quality derived from an outpatient electronic health record. *American Journal of Medical Quality,* 26: 441-451.

Benin, A., Vitkauskas, G., et al. (2005). Validity of using an electronic medical record for assessing quality of care in an outpatient setting. *Medical Care,* 43: 691-698.

Berlin, J. and Stang, P. (2011). *Clinical Data Sets That Need to Be Mined*, 104-114, in Olsen, L., Grossman, C. and McGinnis, J., eds. *Learning What Works: Infrastructure Required for Comparative Effectiveness Research*. Washington, DC. National Academies Press.

Bernstam, E., Herskovic, J., et al. (2010). Oncology research using electronic medical record data. *Journal of Clinical Oncology,* 28: suppl; abstr e16501.

Blumenthal, D. (2011a). Implementation of the federal health information technology initiative. *New England Journal of Medicine,* 365: 2426-2431.

Blumenthal, D. (2011b). Wiring the health system--origins and provisions of a new federal program. *New England Journal of Medicine,* 365: 2323-2329.

Botsis, T., Hartvigsen, G., et al. (2010). Secondary use of EHR: data quality issues and informatics opportunities. *AMIA Summits on Translational Science Proceedings*, San Francisco, CA.

Boyd, D. and Crawford, K. (2011). Six Provocations for Big Data. Cambridge, MA, Microsoft Research. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1926431.

Buckley, C. and Voorhees, E. (2000). Evaluating evaluation measure stability. *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Athens, Greece. ACM Press. 33-40.

Buckley, C. and Voorhees, E. (2004). Retrieval evaluation with incomplete information. *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Sheffield, England. ACM Press. 25-32.

Demner-Fushman, D., Abhyankar, S., et al. (2011). A knowledge-based approach to medical records retrieval. *The Twentieth Text REtrieval Conference Proceedings (TREC 2011)*, Gaithersburg, MD. National Institute for Standards and Technology.

Friedman, C., Wong, A., et al. (2010). Achieving a nationwide learning health system. *Science Translational Medicine,* 2(57): 57cm29.

Harman, D. (2005). *The TREC Ad Hoc Experiments*, 79-98, in Voorhees, E. and Harman, D., eds. *TREC: Experiment and Evaluation in Information Retrieval*. Cambridge, MA. MIT Press.

Hersh, W. (2009). *Information Retrieval: A Health and Biomedical Perspective (3rd Edition)*. New York, NY. Springer.

Hersh, W., Müller, H., et al. (2009). The ImageCLEFmed medical image retrieval task test collection. *Journal of Digital Imaging,* 22: 648-655.

Hersh, W. and Voorhees, E. (2009). TREC genomics special issue overview. *Information Retrieval,* 12: 1-15.

Ide, N., Loane, R., et al. (2007). Essie: a concept-based search engine for structured biomedical text. *Journal of the American Medical Informatics Association,* 14: 253-263.

Jollis, J., Ancukiewicz, M., et al. (1993). Discordance of databases designed for claims payment versus clinical information systems: implications for outcomes research. *Annals of Internal Medicine,* 119: 844-850.

King, B., Wang, L., et al. (2011). Cengage Learning at TREC 2011 Medical Track. *The Twentieth Text REtrieval Conference Proceedings (TREC 2011)*, Gaithersburg, MD. National Institute for Standards and Technology.

Müller, H., Clough, P., et al., eds. (2010). *ImageCLEF: Experimental Evaluation in Visual Information Retrieval*. Heidelberg, Germany. Springer.

O'Malley, K., Cook, K., et al. (2005). Measuring diagnoses: ICD code accuracy. *Health Services Research,* 40: 1620-1639.

Rhodes, E., Laffel, L., et al. (2007). Accuracy of administrative coding for type 2 diabetes in children, adolescents, and young adults. *Diabetes Care,* 30: 141-143.

Safran, C., Bloomrosen, M., et al. (2007). Toward a national framework for the secondary use of health data: an American Medical Informatics Association white paper. *Journal of the American Medical Informatics Association,* 14: 1-9.

Voorhees, E. and Harman, D., eds. (2005). *TREC: Experiment and Evaluation in Information Retrieval*. Cambridge, MA. MIT Press.

Voorhees, E. and Tong, R. (2011). Overview of the TREC 2011 Medical Records Track. *The Twentieth Text REtrieval Conference Proceedings (TREC 2011)*, Gaithersburg, MD. National Institute for Standards and Technology.

Wei, W., Leibson, C., et al. (2012). Impact of data fragmentation across healthcare centers on the accuracy of a high-throughput clinical phenotyping algorithm for specifying subjects with type 2 diabetes mellitus. *Journal of the American Medical Informatics Association,* 19: 219-224.

Weiner, M. (2011). Evidence Generation Using Data-Centric, Prospective, Outcomes Research Methodologies. San Francisco, CA, Presentation at AMIA Clinical Research Informatics Summit.

Wright, A., Pang, J., et al. (2012). Improving completeness of electronic problem lists through clinical decision support: a randomized, controlled trial. *Journal of the American Medical Informatics Association*: Epub ahead of print.

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)

3

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)

4

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

OREGON HEALTH & SCIENCE UNIVERSITY

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)

OREGON HEALTH & SCIENCE UNIVERSITY

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

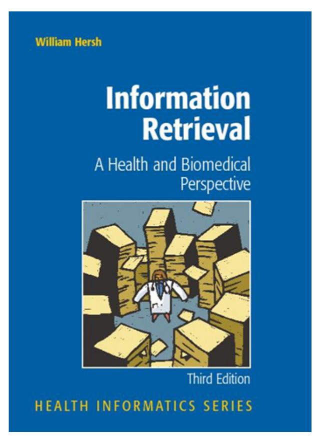
- Study of 3.7M patients in Massachusetts found 31% visited 2 or more hospitals over 5 years (57% of all visits) and 1% visited 5 or more hospitals (10% of all visits) (Bourgeois, 2010)
- Study of 2.8M emergency department (ED) patients in Indiana found 40% of patients had data at multiple institutions,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

Primer on information retrieval (IR) and related topics

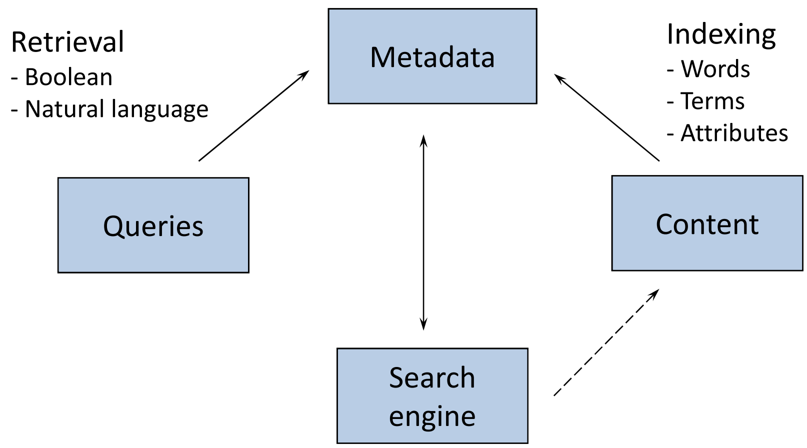
- Information retrieval
- Evaluation
- Challenge evaluations

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



11

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 measures
    - 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.

12

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

13

Recall and precision

- Recall

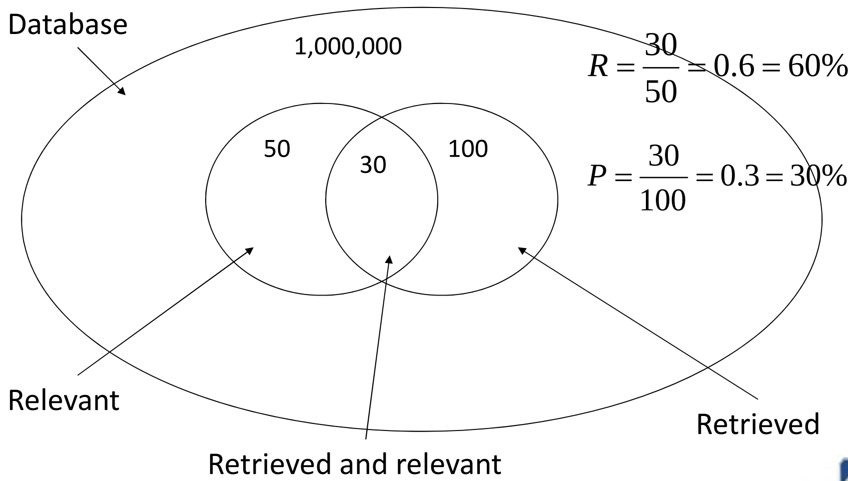
$$R = \frac{\#\,retrieved\ and\ relevant\ documents}{\#\,relevant\ documents\ in\ collection}$$

  - Usually use *relative recall* when not all relevant documents known, where denominator is number of known relevant documents in collection
- Precision

$$P = \frac{\#\,retrieved\ and\ relevant\ documents}{\#\,retrieved\ documents}$$

14

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)
- 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

19

OREGON HEALTH & SCIENCE UNIVERSITY

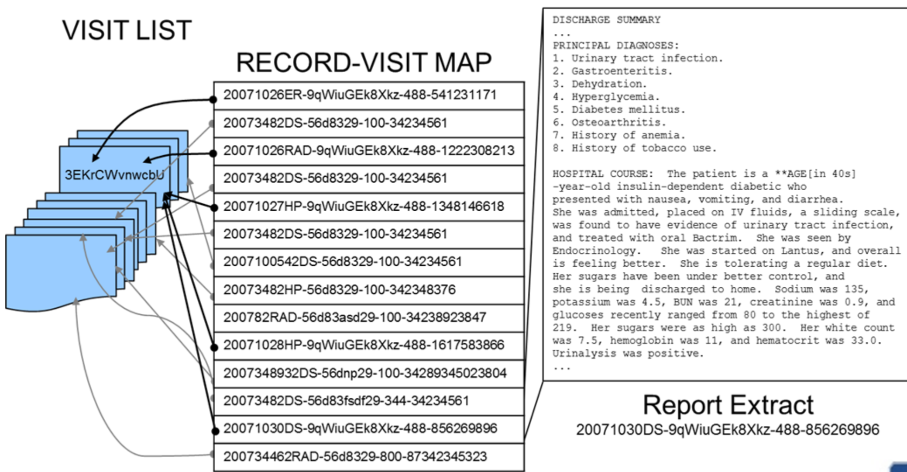
TREC Medical Records Track

- 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)

20

OREGON HEALTH & SCIENCE UNIVERSITY

Test collection



17,265 visits

101,712 reports (93,552 mapped to visits)

(Courtesy, Ellen Voorhees, NIST)

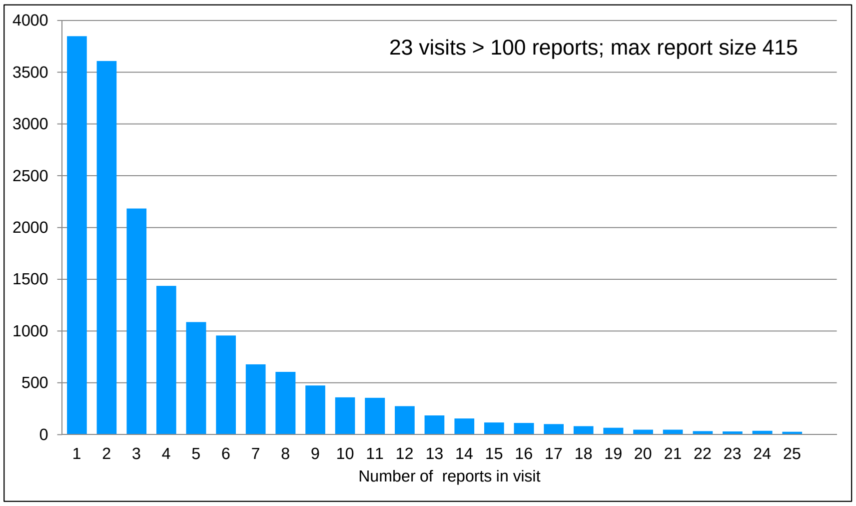
OREGON HEALTH & SCIENCE UNIVERSITY

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

OREGON HEALTH & SCIENCE UNIVERSITY

Wide variations in number of documents per visit



(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

OREGON HEALTH & SCIENCE UNIVERSITY

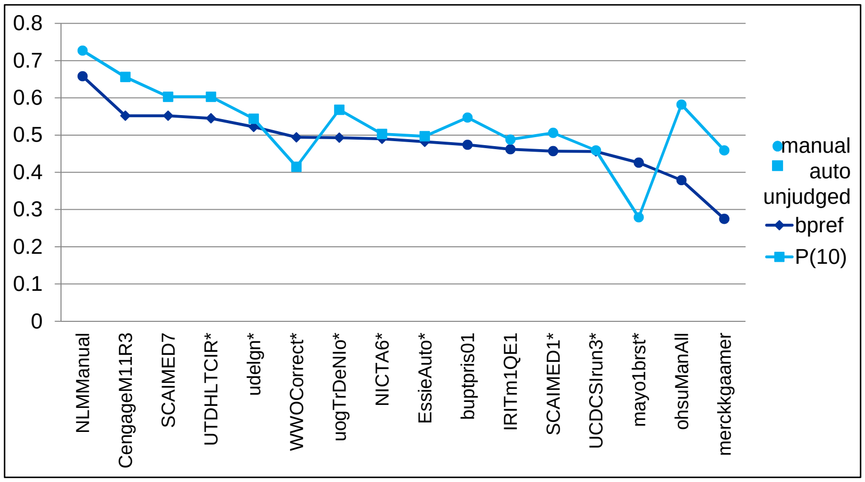
Sample topics

- 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

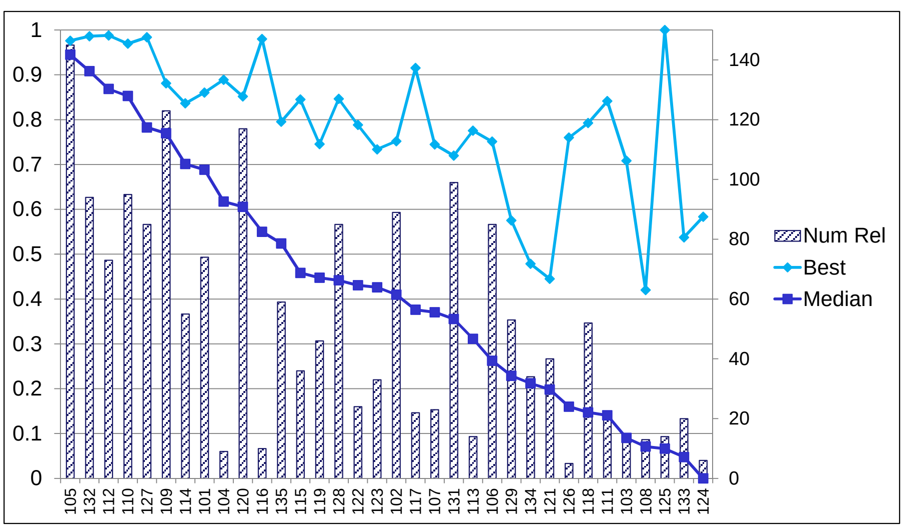
- Runs consisted of ranked list of up to 1000 visits per topic for each of 35 topics
  - Automatic – no human intervention from input of topic statement to output of ranked list
  - Manual – everything else
- Up to 8 runs per participating group
- Subset of retrieved visits contributed to judgment sets
  - 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 …



27

… BUT, wide variation among topics



28

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.

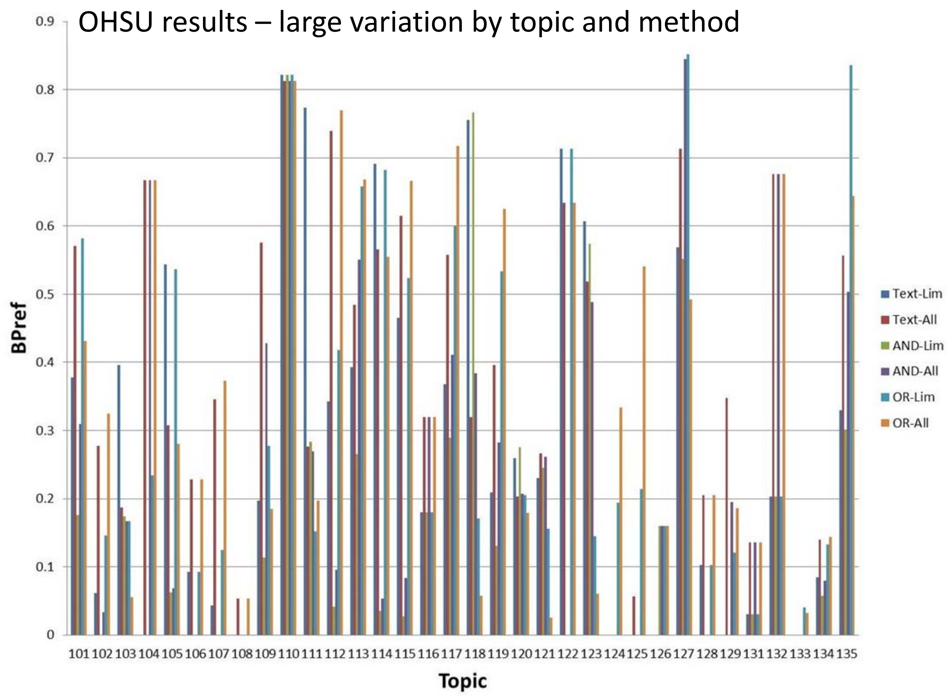
OHSU approach (Bedrick, 2011)

- 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

Run	BPref	P@10
Text-only – All	0.3751	0.5853
Text-only – High	0.2894	0.4824
Text AND ICD-9 – All	0.2497	0.4471
Text AND ICD-9 – High	0.1695	0.3235
Text OR ICD-9 – All	0.3657	0.4618
Text OR ICD-9 – High	0.3238	0.4206

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

OREGON HEALTH & SCIENCE UNIVERSITY