Caveats and Recommendations for Use of Operational Electronic Health Record Data for Research and Quality Measurement

William Hersh, MD, FACP, FACMI
Diplomate, Clinical Informatics, American Board of Preventive Medicine
Professor and Chair
Department of Medical Informatics & Clinical Epidemiology
Oregon Health & Science University
Portland, OR, USA
Email: hersh@ohsu.edu
Web: www.billhersh.info
Blog: http://informaticsprofessor.blogspot.com

References

Amarasingham, R, Patel, PC, et al. (2013). Allocating scarce resources in real-time to reduce heart failure readmissions: a prospective, controlled study. *BMJ Quality & Safety*. 22: 998-1005.
Angrisano, C, Farrell, D, et al. (2007). Accounting for the Cost of Health Care in the United States. Washington, DC, McKinsey & Company. http://www.mckinsey.com/mgi/rp/healthcare/accounting_cost_healthcare.asp
Anonymous (2010). The State of Health Care Quality: 2010. Washington, DC, National Committee for Quality Assurance. http://www.ncqa.org/tabid/836/Default.aspx
Anonymous (2011). Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease. Washington, DC, National Academies Press.
Barkhuysen, P, deGrauw, W, et al. (2014). Is the quality of data in an electronic medical record sufficient for assessing the quality of primary care? *Journal of the American Medical Informatics Association*. 21: 692-698.
Blumenthal, D (2011). Implementation of the federal health information technology initiative. *New England Journal of Medicine*. 365: 2426-2431.
Blumenthal, D (2011). Wiring the health system--origins and provisions of a new federal program. *New England Journal of Medicine*. 365: 2323-2329.
Brill, S (2013). Bitter Pill: Why Medical Bills Are Killing Us. Time, April 4, 2013. http://healthland.time.com/2013/02/20/bitter-pill-why-medical-bills-are-killing-us/
Buntin, MB, Burke, MF, et al. (2011). The benefits of health information technology: a review of the recent literature shows predominantly positive results. *Health Affairs*. 30: 464-471.
Charles, D, Gabriel, M, et al. (2014). Adoption of Electronic Health Record Systems among U.S. Non-federal Acute Care Hospitals: 2008-2013. Washington, DC, Department of Health and Human Services. http://www.healthit.gov/sites/default/files/oncdatabrief16.pdf
Chaudhry, B, Wang, J, et al. (2006). Systematic review: impact of health information technology on quality, efficiency, and costs of medical care. *Annals of Internal Medicine*. 144: 742-752.

Classen, DC, Resar, R, et al. (2011). 'Global trigger tool' shows that adverse events in hospitals may be ten times greater than previously measured. *Health Affairs*. 30: 4581-4589.
Collins, FS, Hudson, KL, et al. (2014). PCORnet: turning a dream into reality. *Journal of the American Medical Informatics Association*. 21: 576-577.
Collins, FS and Varmus, H (2015). A new initiative on precision medicine. *New England Journal of Medicine*: Epub ahead of print.
Dahabreh, IJ and Kent, DM (2014). Can the learning health care system be educated with observational data? *Journal of the American Medical Association*. 312: 129-130.
Danaei, G, Rodríguez, LA, et al. (2011). Observational data for comparative effectiveness research: An emulation of randomised trials of statins and primary prevention of coronary heart disease. *Statistical Methods in Medical Research*. 22: 70-96.
deLusignan, S and vanWeel, C (2005). The use of routinely collected computer data for research in primary care: opportunities and challenges. *Family Practice*. 23: 253-263.
Demner-Fushman, D, Abhyankar, S, et al. (2012). NLM at TREC 2012 Medical Records Track. *The Twenty-First Text REtrieval Conference Proceedings (TREC 2012)*, Gaithersburg, MD. National Institute for Standards and Technology
http://trec.nist.gov/pubs/trec21/papers/NLM.medical.final.pdf
Denny, JC (2012). Mining Electronic Health Records in the Genomics Era. PLOS Computational Biology: Translational Bioinformatics. M. Kann and F. Lewitter.
Denny, JC, Bastarache, L, et al. (2013). Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nature Biotechnology*. 31: 1102-1111.
Diamond, CC, Mostashari, F, et al. (2009). Collecting and sharing data for population health: a new paradigm. *Health Affairs*. 28: 454-466.
Edinger, T, Cohen, AM, et al. (2012). Barriers to retrieving patient information from electronic health record data: failure analysis from the TREC Medical Records Track. *AMIA 2012 Annual Symposium*, Chicago, IL. 180-188.
Fleurence, RL, Curtis, LH, et al. (2014). Launching PCORnet, a national patient-centered clinical research network. *Journal of the American Medical Informatics Association*. 21: 578-582.
Galvez, E (2014). Interoperability Roadmap: Early Draft Material - Joint HITPC and HITSC Meeting. Washington, DC, Department of Health and Human Services.
http://www.healthit.gov/facas/sites/faca/files/Joint_HIT_DRAFT Nationwide Interoperability Roadmap Material FACA v2_2014-10-15.pdf
Gildersleeve, R and Cooper, P (2013). Development of an automated, real time surveillance tool for predicting readmissions at a community hospital. *Applied Clinical Informatics*. 4: 153-169.
Goldzweig, CL, Towfigh, A, et al. (2009). Costs and benefits of health information technology: new trends from the literature. *Health Affairs*. 28: w282-w293.
Gottesman, O, Kuivaniemi, H, et al. (2013). The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. *Genetics in Medicine*. 15: 761-771.
Hebert, C, Shivade, C, et al. (2014). Diagnosis-specific readmission risk prediction using electronic health data: a retrospective cohort study. *BMC Medical Informatics & Decision Making*. 14: 65. http://www.biomedcentral.com/1472-6947/14/65

Hersh, W (2004). Health care information technology: progress and barriers. *Journal of the American Medical Association*. 292: 2273-2274.
Hersh, WR, Cimino, JJ, et al. (2013). Recommendations for the use of operational electronic health record data in comparative effectiveness research. *eGEMs (Generating Evidence & Methods to improve patient outcomes)*. 1: 14. http://repository.academyhealth.org/egems/vol1/iss1/14/
Hersh, WR, Weiner, MG, et al. (2013). Caveats for the use of operational electronic health record data in comparative effectiveness research. *Medical Care*. 51(Suppl 3): S30-S37.
Horner, P and Basu, A (2012). Analytics & the future of healthcare. Analytics, January/February 2012. http://www.analytics-magazine.org/januaryfebruary-2012/503-analytics-a-the-future-of-healthcare
Hsiao, CJ and Hing, E (2014). Use and Characteristics of Electronic Health Record Systems Among Office-based Physician Practices: United States, 2001–2013. Hyattsville, MD, National Center for Health Statistics, Centers for Disease Control and Prevention. http://www.cdc.gov/nchs/data/databriefs/db143.htm
Hussey, PS, Wertheimer, S, et al. (2013). The association between health care quality and cost: a systematic review. *Annals of Internal Medicine*. 158: 27-34.
Ide, NC, Loane, RF, et al. (2007). Essie: a concept-based search engine for structured biomedical text. *Journal of the American Medical Informatics Association*. 14: 253-263.
Jamoom, E and Hing, E (2015). Progress With Electronic Health Record Adoption Among Emergency and Outpatient Departments: United States, 2006–2011. Hyattsville, MD National Center for Health Statistics, Centers for Disease Control and Prevention. http://www.cdc.gov/nchs/data/databriefs/db187.htm
Jones, SS, Rudin, RS, et al. (2014). Health information technology: an updated systematic review with a focus on meaningful use. *Annals of Internal Medicine*. 160: 48-54.
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
Kohn, LT, Corrigan, JM, et al., Eds. (2000). To Err Is Human: Building a Safer Health System. Washington, DC, National Academies Press.
MacKenzie, SL, Wyatt, MC, et al. (2012). Practices and perspectives on building integrated data repositories: results from a 2010 CTSA survey. *Journal of the American Medical Informatics Association*. 19(e1): e119-e124.
McGlynn, EA, Asch, SM, et al. (2003). The quality of health care delivered to adults in the United States. *New England Journal of Medicine*. 348: 2635-2645.
Miller, DR, Safford, MM, et al. (2004). Who has diabetes? Best estimates of diabetes prevalence in the Department of Veterans Affairs based on computerized patient data. *Diabetes Care*. 27(Suppl 2): B10-21.
Newton, KM, Peissig, PL, et al. (2013). Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network. *Journal of the American Medical Informatics Association*. 20(e1): e147-154.
Rea, S, Pathak, J, et al. (2012). Building a robust, scalable and standards-driven infrastructure for secondary use of EHR data: The SHARPn project. *Journal of Biomedical Informatics*. 45: 763-771.
Richesson, RL and Andrews, JE, Eds. (2012). Clinical Research Informatics. New York, NY, Springer.

Richesson, RL, Rusincovitch, SA, et al. (2013). A comparison of phenotype definitions for diabetes mellitus. *Journal of the American Medical Informatics Association*. 20(e2): e319-e326.
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.
Schoen, C, Osborn, R, et al. (2009). A survey of primary care physicians in eleven countries, 2009: perspectives on care, costs, and experiences. *Health Affairs*. 28: w1171-1183.
Smith, M, Saunders, R, et al. (2012). Best Care at Lower Cost: The Path to Continuously Learning Health Care in America. Washington, DC, National Academies Press.
Smith, PC, Araya-Guerra, R, et al. (2005). Missing clinical information during primary care visits. *Journal of the American Medical Association*. 293: 565-571.
Tannen, RL, Weiner, MG, et al. (2008). Replicated studies of two randomized trials of angiotensin-converting enzyme inhibitors: further empiric validation of the 'prior event rate ratio' to adjust for unmeasured confounding by indication. *Pharmacoepidemiology and Drug Safety*. 17: 671-685.
Tannen, RL, Weiner, MG, et al. (2009). Use of primary care electronic medical record database in drug efficacy research on cardiovascular outcomes: comparison of database and randomised controlled trial findings. *British Medical Journal*. 338: b81. http://www.bmj.com/cgi/content/full/338/jan27_1/b81
Tannen, RL, Weiner, MG, et al. (2007). A simulation using data from a primary care practice database closely replicated the Women's Health Initiative trial. *Journal of Clinical Epidemiology*. 60: 686-695.
VanDenBos, J, Rustagi, K, et al. (2011). The $17.1 billion problem: the annual cost of measurable medical errors. *Health Affairs*. 30: 596-603.
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. National Institute of Standards and Technology http://trec.nist.gov/pubs/trec21/papers/MED12OVERVIEW.pdf
Voorhees, EM and Harman, DK, Eds. (2005). TREC: Experiment and Evaluation in Information Retrieval. Cambridge, MA, MIT Press.
Wei, WQ, Leibson, CL, et al. (2013). The absence of longitudinal data limits the accuracy of high-throughput clinical phenotyping for identifying type 2 diabetes mellitus subjects. *International Journal of Medical Informatics*. 82: 239-247.
Weiner, MG, Barnhart, K, et al. (2008). Hormone therapy and coronary heart disease in young women. *Menopause*. 15: 86-93.
Xu, H, Aldrich, MC, et al. (2014). Validating drug repurposing signals using electronic health records: a case study of metformin associated with reduced cancer mortality. *Journal of the American Medical Informatics Association*: Epub ahead of print.
Zhang, Z and Sun, J (2010). Interval censoring. *Statistical Methods in Medical Research*. 19: 53-70.

Caveats and Recommendations for Use of Operational Electronic Health Record Data for Research and Quality Measurement

William Hersh, MD, FACP, FACMI
Diplomate, Clinical Informatics, ABPM
Professor and Chair
Department of Medical Informatics & Clinical Epidemiology
Oregon Health & Science University
Portland, OR, USA
Email: hersh@ohsu.edu
Web: www.billhersh.info
Blog: http://informaticsprofessor.blogspot.com

1

Outline

- Information-related problems and solutions in healthcare
- Opportunities for secondary use or re-use of clinical data for research and other purposes
- Caveats of operational clinical data
- Role of informatics research

2

Many problems in healthcare have information-related solutions

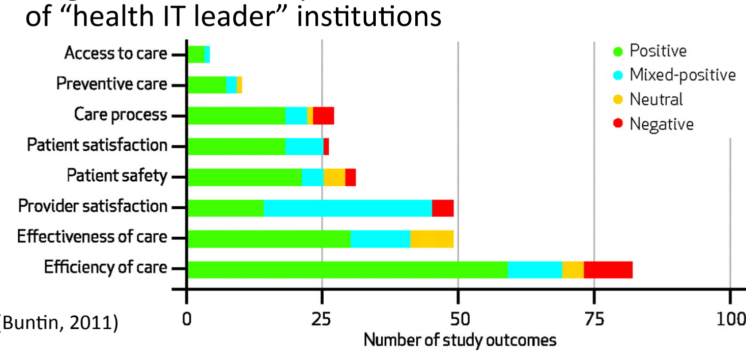
- Quality – not as good as it could be (McGlynn, 2003; Schoen, 2009; NCQA, 2010)
- Safety – errors cause morbidity and mortality; many preventable (Kohn, 2000; Classen, 2011; van den Bos, 2011; Smith 2012)
- Cost – rising costs not sustainable; US spends more but gets less (Angrisano, 2007; Brill, 2013)
- Inaccessible information – missing information frequent in primary care (Smith, 2005)

OREGON HEALTH & SCIENCE UNIVERSITY

3

Growing evidence that information interventions are part of solution

- Systematic reviews (Chaudhry, 2006; Goldzweig, 2009; Buntin, 2011; Jones, 2014) have identified benefits in a variety of areas, although
  - Quality of many studies could be better
  - Large number of early studies came from a small number of "health IT leader" institutions



(Buntin, 2011)

What are the major challenges in getting where we want? (Hersh, 2004)

Health Care Information Technology
Progress and Barriers

William Hersh, MD

IN THE 3 DECADES SINCE THE TERM "MEDICAL INFORMATics" was first used, individuals working at the intersection of information technology (IT) and medicine have developed and evaluated computer applications aim-

in this issue of *JAMA*, Slack demonstrates the value that patient-physician e-mail can have in improving patient care, and also catalogs the incomplete but encouraging underlying evidence.[11] As with many applications of IT, the technology can improve the existing situation but also empower clinicians and patients to think more fundamentally about how inno-

- Cost
- Technical challenges
- Interoperability
- Privacy and confidentiality
- Workforce

care IT.[8,9] It is no exaggeration to declare that the years ahead portend the "decade of health information technology."[10]

Informatics is poised to have a major impact in patient-clinician communication. In the Clinical Crossroads article

See also p 2255.

ment. The rest goes to those who typically do not pay for

Author Affiliation: Department of Medical Informatics & Clinical Epidemiology, Oregon Health & Science University, Portland.
Corresponding Author: William Hersh, MD, Department of Medical Informatics & Clinical Epidemiology, Oregon Health & Science University School of Medicine, 3181 SW Sam Jackson Park Rd, BICC, Portland, OR 97201-3098 (hersh@ohsu.edu).

(Reprinted) JAMA, November 10, 2004—Vol 292, No. 18 2273

OREGON HEALTH & SCIENCE UNIVERSITY

US has made substantial investment in health information technology (HIT)

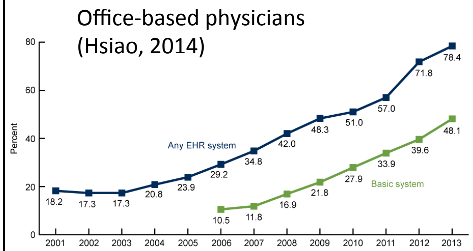
"To improve the quality of our health care while lowering its cost, we will make the immediate investments necessary to ensure that within five years, all of America's medical records are computerized … It just won't save billions of dollars and thousands of jobs – it will save lives by reducing the deadly but preventable medical errors that pervade our health care system."
January 5, 2009

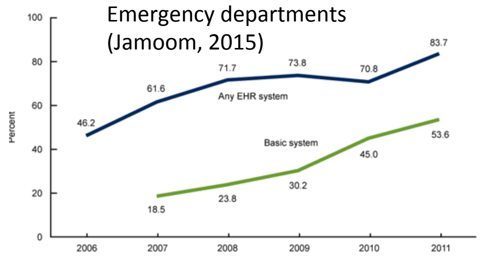
Health Information Technology for Economic and Clinical Health (HITECH) Act of the American Recovery and Reinvestment Act (ARRA) (Blumenthal, 2011)

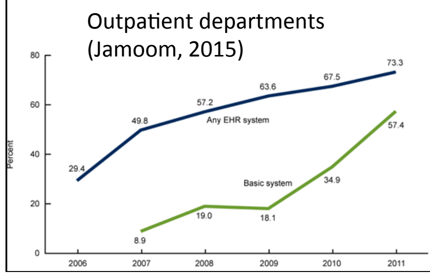
- Incentives for electronic health record (EHR) adoption by physicians and hospitals (up to $27B)
- Direct grants administered by federal agencies ($2B, including $118M for workforce development)

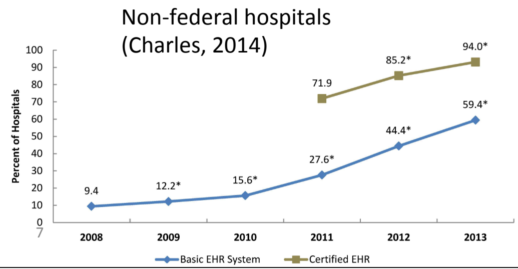
OREGON HEALTH & SCIENCE UNIVERSITY

Which has led to significant EHR adoption in the US









Providing opportunities for “secondary use” or “re-use” of clinical data

- (Safran, 2007; SHARPn, Rea, 2012)
- Using data to improve care delivery
- Healthcare quality measurement and improvement
- Clinical and translational research
- Public health surveillance
- Implementing the learning health system

OREGON HEALTH & SCIENCE UNIVERSITY

Using data to improve healthcare

- With shift of payment from "volume to value," healthcare organizations will need to manage information better to provide better care (Diamond, 2009; Horner, 2012)
- Predictive analytics is use of data to anticipate poor outcomes or increased resource use – applied by many to problem of early hospital re-admission (e.g., Gildersleeve, 2013; Amarasingham, 2013; Herbert, 2014)
- A requirement for "precision medicine" (IOM, 2011; Collins, 2015)
- Also can be used to measure quality of care delivered to make it more "accountable" (Hussey, 2013; Barkhuysen, 2014)

9

OREGON HEALTH & SCIENCE UNIVERSITY

Clinical and translational research

- Many roles for clinical research informatics (Richesson, 2012)
  - Led in part by activities of NIH Clinical and Translational Science Award (CTSA) Program (Mackenzie, 2012)
- One of largest and most productive efforts has been eMERGE Network – connecting genotype-phenotype (Gottesman, 2013; Newton, 2013)
  - http://emerge.mc.vanderbilt.edu
  - Has used EHR data to identify genomic variants associated with various phenotypes (Denny, 2012; Denny, 2013)

10

OREGON HEALTH & SCIENCE UNIVERSITY

Clinical and translational research (cont.)

- Other successes include replication of clinical studies, e.g.,
  - Randomized controlled trials (RCT)
    - Women's Health Initiative (Tannen, 2007; Weiner, 2008)
    - Other cardiovascular diseases (Tannen, 2008; Tannen, 2009) and value of statin drugs in primary prevention of coronary heart disease (Danaei, 2011)
  - Observational studies
    - Metformin and reduced cancer mortality rate (Xu, 2014)
- Much potential for using propensity scores with observational studies as complement to RCTs
  - Often but not always obtain same results as RCTs (Dahabreh, 2014)

Caveats for the Use of Operational Electronic Health Record Data in Comparative Effectiveness Research

William R. Hersh, MD, Mark G. Weiner, MD,† Peter J. Embi, MD, MS,‡ Judith R. Logan, MD, MS,* Philip R.O. Payne, PhD,‡ Elmer V. Bernstam, MD, MSE,§ Harold P. Lehmann, MD, PhD,|| George Hripcsak, MD, MS,¶ Timothy H. Hartzog, MD, MS,# James J. Cimino, MD,** and Joel H. Saltz, MD, PhD††*

Operational clinical data may be (Medical Care, 2013):
- Inaccurate
- Incomplete
- Transformed in ways that undermine meaning
- Unrecoverable for research
- Of unknown provenance
- Of insufficient granularity
- Incompatible with research protocols

(*Med Care* 2013;00: 000–000)

Health Information Technology (ONC) through the Strategic Health IT Advanced Research Projects (SHARP) Program,

Caveats of clinical data

- Documentation not always a top priority for busy clinicians (de Lusignan, 2005)
- Not every diagnosis is recorded at every visit; absence of evidence is not always evidence of absence, an example of a concern known by statisticians as *censoring* (Zhang, 2010)
- Makes seemingly simple tasks such as identifying diabetic patients challenging (Miller, 2004; Wei, 2013; Richesson, 2013)

OREGON HEALTH & SCIENCE UNIVERSITY

"Idiosyncrasies" of clinical data (Hersh, 2013)

- "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

Overcoming the caveats: recommendations for EHR data use

- (Hersh, 2013)
- Assessing and using data
- Adaptation of “best evidence” approaches to use of operational data
- Need for standards and interoperability
- Appropriate use of informatics expertise

AcademyHealth

eGEMs

Home > EDM_PUBLICATIONS > eGEMs > Vol. 1 (2013) > Iss. 1

Recommendations for the Use of Operational Electronic Health Record Data in Comparative Effectiveness Research

Download

William R. Hersh, Oregon Health & Science University — Follow
James Cimino, National Institutes of Health Clinical Center — Follow
Philip R.O. Payne, The Ohio State University — Follow
Peter Embi, The Ohio State University — Follow
Judith Logan, Oregon Health & Science University — Follow
Mark Weiner, AstraZeneca — Follow
Elmer V. Bernstam, The University of Texas Health Science Center at Houston — Follow
Harold Lehmann, Johns Hopkins University — Follow
George Hripcsak, Columbia University — Follow
Timothy Hartzog, Medical University of South Carolina — Follow
Joel Saltz, Emory University — Follow

Included in Health Information Technology Commons, Health Services Research Commons

SHARE

Altmetric 6

Abstract

There is an increasing amount of clinical data in operational electronic health record (EHR) systems. Such data provide substantial opportunities for their re-use for many purposes, including comparative effectiveness research (CER). In a previous paper, we

15

A big challenge is interoperable data

INFORMATICS PROFESSOR

THIS BLOG MAINTAINS THE THOUGHTS ON VARIOUS TOPICS RELATED TO BIOMEDICAL AND HEALTH INFORMATICS BY DR. WILLIAM HERSH, PROFESSOR AND CHAIR, DEPARTMENT OF MEDICAL INFORMATICS & CLINICAL EPIDEMIOLOGY, OREGON HEALTH & SCIENCE UNIVERSITY.

WEDNESDAY, MAY 15, 2013

Universal EHR? No. Universal Data Access? Yes.

A recent blog posting calls for a "universal EMR" for the entire healthcare system. The author prov laments how lack of access to the c impedes optimal clinical care. I wo clinical research, and public health well.

However, I do not agree that a "uni this problem. Instead, I would adv

WILLIAM HERSH

INFORMATICS PROFESSOR

THIS BLOG MAINTAINS THE THOUGHTS ON VARIOUS TOPICS RELATED TO BIOMEDICAL AND HEALTH INFORMATICS BY DR. WILLIAM HERSH, PROFESSOR AND CHAIR, DEPARTMENT OF MEDICAL INFORMATICS & CLINICAL EPIDEMIOLOGY, OREGON HEALTH & SCIENCE UNIVERSITY.

SATURDAY, SEPTEMBER 6, 2014

Unscrambling Eggs and the Need for Comprehensive Data Standards and Interoperability

Two local informatics-related happenings recently provided teac moments demonstrating why a comprehensive approach to stan and interoperability is so critical for realizing the value of health Fortunately, the Office of the National Coordinator for Health IT (ONC) has prioritized interoperability among its activities movir forward, and other emerging work on standards provides hope t problems I will described that occurred locally (and I know occu

INFORMATICS PROFESSOR

THIS BLOG MAINTAINS THE THOUGHTS ON VARIOUS TOPICS RELATED TO BIOMEDICAL AND HEALTH INFORMATICS BY DR. WILLIAM HERSH, PROFESSOR AND CHAIR, DEPARTMENT OF MEDICAL INFORMATICS & CLINICAL EPIDEMIOLOGY, OREGON HEALTH & SCIENCE UNIVERSITY.

FRIDAY, FEBRUARY 6, 2015

The Conundrum of Structured vs. Unstructured Data

As in all complex endeavors, the push for a healthcare system underpinned by structured and interoperable electronic health record (EHR) data has turned out to be more complicated than we might have anticipated when acceleration of EHR adoption was begun about a decade ago. This does not mean that anyone was right or wrong; it just shows the inherent complexities of trying to solve the real problems that motivate data-related problems in healthcare. These healthcare problems have been well-documented over the past couple decades by

WILLIAM HERSH

http://www.billhersh.info/

Challenges to EHRs have spurred focus on interoperability

- Office of National Coordinator for Health IT (ONC) developing interoperability road map for 10-year path forward (Galvez, 2014)
- Emerging approaches include
  - RESTful architectures for efficient client-server interaction
  - OAuth2 for Internet-based security
  - Standard application programming interface (API) for query/retrieval of data
    - Need for both documents and discrete data
    - Emerging standard is Fast Health Interoperability Resources (FHIR) – http://wiki.hl7.org/index.php?title=FHIR_for_Clinical_Users

17

OREGON HEALTH & SCIENCE UNIVERSITY

Also need to develop clinical data research networks

- Established
  - HMO Research Network – facilitates clinical research
    - www.hmoresearchnetwork.org
  - FDA Mini-Sentinel Network – safety surveillance
    - www.mini-sentinel.org
- New
  - PCORnet – www.pcornet.org
    - Clinical data research networks (CDRNs) – 11 networks aggregating data on >1M patients each
      - (Fleurence, 2014; Collins, 2014; and other papers in JAMIA special issue)
    - Common Data Model for subset of data

18

OREGON HEALTH & SCIENCE UNIVERSITY

Critical role for informatics research

- Many areas of need, e.g.,
  - Data standards and interoperability
  - Data science
  - People and organizational issues
  - Information retrieval and text mining
- Currently funded by many different entities
  - NLM – basic research in informatics, especially clinical informatics
  - Other NIH institutes – tend to focus on bioinformatics
  - AHRQ and PCORI – applied clinical informatics

OREGON HEALTH & SCIENCE UNIVERSITY

An instance of research: cohort discovery

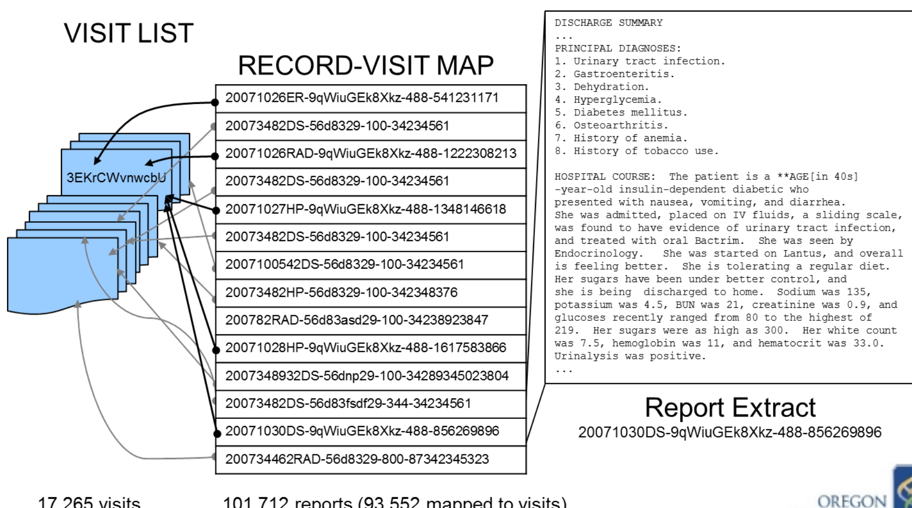
- Adapting information retrieval techniques to medical records
- Use case somewhat different from usual information retrieval: want to retrieve records and data within them to identify patients who might be candidates for clinical studies
- Another goal: working with large quantity of data, i.e., not few hundred documents typical to natural language processing studies

OREGON HEALTH & SCIENCE UNIVERSITY

Challenges for informatics research with medical records

- 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)
- Part of Text Retrieval Conference (TREC), an annual challenge evaluation sponsored by National Institute for Standards and Technology (NIST) (Voorhees, 2005)
- Medical Records Track launched in 2011, repeated in 2012 (Voorhees, 2012)

Test collection



(Courtesy, Ellen Voorhees, NIST)

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)

Easy and hard topics

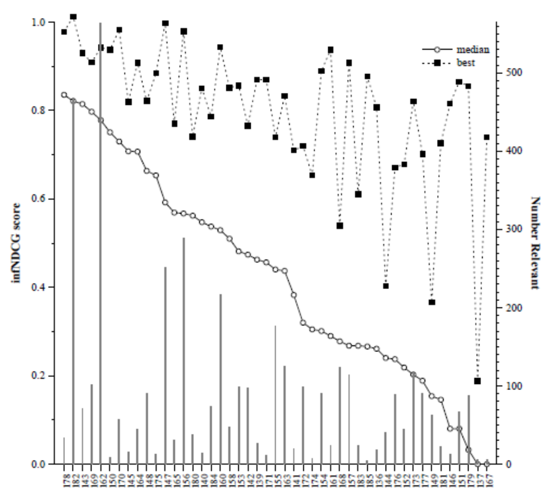
- Easiest – best median bpref
  - 105: Patients with dementia
  - 132: Patients admitted for surgery of the cervical spine for fusion or discectomy
- Hardest – worst best bpref and worst median bpref
  - 108: Patients treated for vascular claudication surgically
  - 124: Patients who present to the hospital 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

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

Results for 2012

Run	infNDCG	infAP	P(10)
NLMManual*	0.680	0.366	0.749
udelSUM	0.578	0.286	0.592
sennamed2	0.547	0.275	0.557
ohsuManBool*	0.526	0.250	0.611
atigeo1	0.524	0.224	0.519
UDinfoMed123	0.517	0.236	0.528
uogTrMConQRd	0.509	0.231	0.553
NICTAUBC4	0.487	0.216	0.517



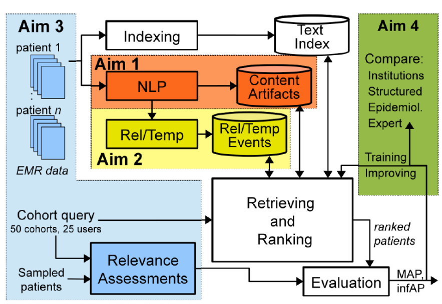
Per-topic scores for infNDCG computed over all runs

What approaches did (and did not) work?

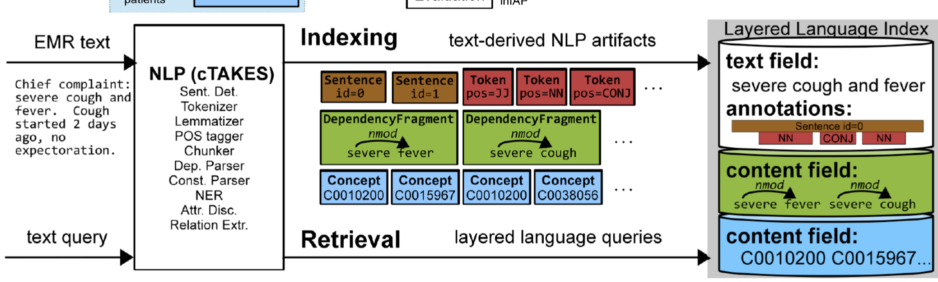
- Best results in 2011 and 2012 obtained from NLM group (Demner-Fushman, 2011)
  - Top results from manually constructed queries using Essie domain-specific search engine (Ide, 2007)
  - Other automated processes fared less well, e.g., creation of PICO frames, negation, term expansion, etc.
- Best automated 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 provided small or inconsistent value
  - Document focusing, term expansion, etc.

OREGON HEALTH & SCIENCE UNIVERSITY

Continued work – Mayo Clinic-OHSU collaboration



Led by:
- Hongfang Liu – Mayo
- Stephen Wu – OHSU



Conclusions

- There are plentiful opportunities for secondary use or re-use of clinical data
- We must be cognizant of caveats of using operational clinical data
- We must implement best practices for using such data
- We need consensus on approaches to standards and interoperability
- Research to discover best methods and useful results is critical

29

For more information

- Bill Hersh
  - http://www.billhersh.info
- Informatics Professor blog
  - http://informaticsprofessor.blogspot.com
- OHSU Department of Medical Informatics & Clinical Epidemiology (DMICE)
  - http://www.ohsu.edu/informatics
  - http://www.youtube.com/watch?v=T-74duDDvwU
  - http://oninformatics.com
- What is Biomedical and Health Informatics?
  - http://www.billhersh.info/whatis
- Office of the National Coordinator for Health IT (ONC)
  - http://healthit.hhs.gov
- American Medical Informatics Association (AMIA)
  - http://www.amia.org
- National Library of Medicine (NLM)
  - http://www.nlm.nih.gov

30