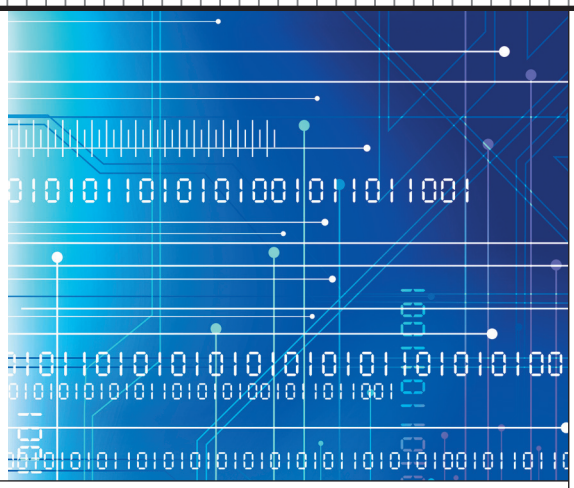
Mining Electronic Health Records

Naren Ramakrishnan, *Virginia Tech*
David A. Hanauer, *University of Michigan*
Benjamin J. Keller, *Eastern Michigan University*



Initial efforts to mine electronic health records are unlikely to yield many Eureka insights, but there are many opportunities for improving the delivery, efficiency, and effectiveness of healthcare.

With President Obama's Health Information Technology for Economic and Clinical Health (HITECH) initiatives making their way into policy, there has been considerable interest in the US about the promise of electronic health records. Some organizations are ahead of the curve and have been using electronic records for decades, while others are still in the planning phases.

A major goal of the new initiatives is to encourage the development of a digital infrastructure for providers and patients so that care can be delivered more effectively and efficiently. If successful, we may someday have true data interoperability among healthcare providers and, ultimately, improved patient outcomes.

As interest in digital medical records has grown, so too has debate about their nomenclature.

Some groups interchangeably refer to electronic medical records and electronic health records, whereas others make a clear distinction. The Healthcare Information and Management Systems Society defines an EMR system as a clinical information system, owned and operated by a healthcare delivery organization, that serves as the legal record of a patient encounter; EHR has a broader scope, containing data from multiple organizations as well as patient input. The increasingly popular EHR concept is closely associated with health information exchanges (HIEs), which are emerging regionally for transferring data between disparate healthcare systems.

Similarly, patients themselves own and maintain their personal health records, which are, at least theoretically, available to any provider with patient consent. The PHR concept has support from large technology companies, including Microsoft and Google.

The confusion over what to call digital medical records hints at a more fundamental problem: the lack of data standards.

LACK OF DATA STANDARDS

The lack of standardization in how medical information is coded and stored is a major difficulty that is hindering automation. The most readily available coded data constitute the ICD-9 (International Classification of Diseases, v9) and CPT-4 (Current Procedural Terminology, v4).

Maintained by the World Health Organization, ICD was initially developed to better track diseases in a standardized manner but, at least in the US, has come to be used primarily to support billing and financial auditing rather than for clinical care. CPT codes were developed by the American Medical Association to document procedures and laboratory tests; like ICD codes, although they were not designed for billing, they are now used almost exclusively for that purpose.

The US government mandates that by 2013 all healthcare providers must switch from ICD-9, which contains some 21,000 codes, to ICD-10, which has more than 150,000. This will create significant implementation issues and also has implications for research and longitudinal data mining as the codes that have been used historically will change.

Other code sets are increasingly popular for clinical care. Developed by the College of American Pathologists, SNOMED-CT (Systematized Nomenclature of Medicine—Clinical Terms) has distinct advantages over ICD codes for capturing discrete patient-care diagnoses.

One such advantage is the ability to support "postcoordination" of

0018-9162/10/$26.00 © 2010 IEEE Published by the IEEE Computer Society

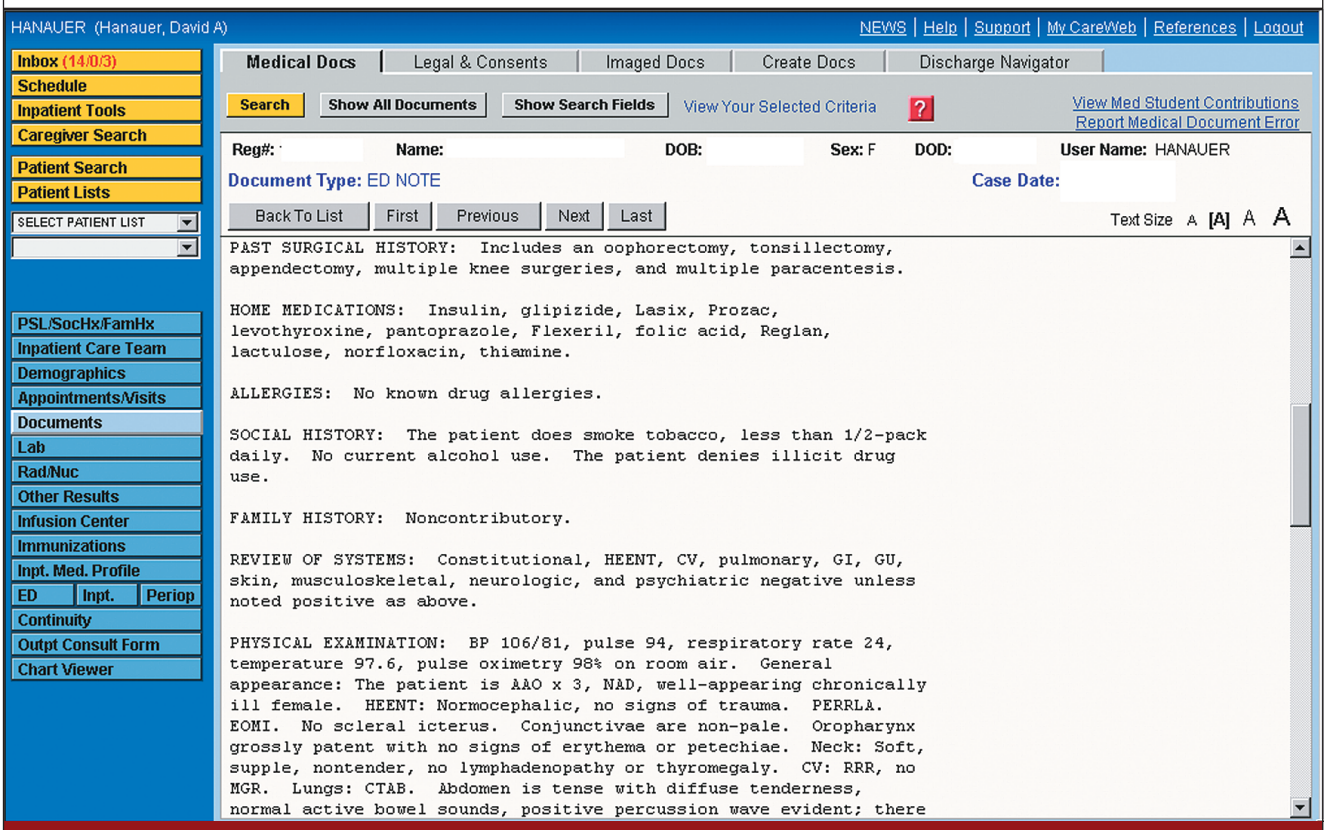


Figure 1. Free text data in medical records often contains crucial information but is not easily extractable for mining. (All identifiable information in this figure has been removed or changed.)

concepts, allowing richer and more detailed clinical descriptions with a more manageable set of codes. For example, the concept of "acute left-sided chest pain" could be created using codes for "left," "side," and "acute onset" (all qualifier values) as well as "chest" (body structure) and "pain" (finding). However, the complexity of such a system comes from its flexibility, as the same concept could also be coded using "acute onset" and "left-sided chest pain" (clinical finding).

Thus, the ability to map between similar concepts will be essential and, like all coded data, the utility for data mining will depend heavily on the initial coding's quality. The use of SNOMED-CT in clinical medicine has yet to gain widespread acceptance but likely will increase over time. It is possible to map between code sets using resources such as the Unified Medical Language System (UMLS), but doing this right is a challenge in itself.

One problem with the use of medical codes for research is that, at least for some codes, the presence of a coded diagnosis does not necessarily mean that the coding is accurate. For example, one study found that of young patients with an ICD-9 code for type 2 diabetes, only 16 percent truly had this disorder based on a chart review (E.T. Rhodes et al., "Accuracy of Administrative Coding for Type 2 Diabetes in Children, Adolescents, and Young Adults," *Diabetes Care*, Jan. 2007, pp. 141-143).

While this is an extreme example, it highlights the risks in blindly using coded data for research without understanding the data's provenance and the reasons for the code assignments. Because providers are typically required to provide ICD codes for insurance and billing purposes, they might assign codes even before making a final diagnosis.

Such inconsistencies in how data are recorded can quickly add up to serious misinterpretation. The experiences of "e-Patient Dave" deBronkart (www.patientdave.blogspot.com) have become legendary. DeBronkart was a cancer survivor, but when he opened an account at Google Health and transferred his medical record information into it, the system incorrectly concluded that the cancer had spread to other parts of his body and that he had had a stroke.

RESEARCH CHALLENGES

Given such broad-based issues in medical data interpretation, what challenges do data mining researchers face?

Data incompleteness

In general, many computer scientists and informaticians assume that most medical records are readily "minable" but, contrary to this belief, much of the relevant data is "locked up" in free text documents. Figure 1 shows an example of a free text report

from the University of Michigan's CareWeb system. Such documents have details that likely will never be in the record's coded portions. Thus, for example, data mining might not detect the association between asthma exacerbation and smoking if asthma exacerbation is coded but smoking is only described in clinical documents. Researchers have used natural-language-processing techniques to detect smoking history in medical records, but such techniques are not yet widely available and require significant expertise to implement.

Similarly, image data, unless processed into a computable form, will also likely be ignored—other than perhaps as a diagnosis code attached to the radiologists' findings. The amount of available information depends on the incentives to code and how well codification integrates with clinicians' workflows and thought processes.

Free text dictations are, and for some time will remain, one of the best ways to communicate clinical information between healthcare providers. In fact, even when an EHR offers data-coding options, many clinicians still prefer to manually document aspects of a patient's care.

Such data incompleteness leads to an inherent bias in any inferences drawn from coded data. In particular, we are more likely to mine procedural/policy biases within the organization rather than patterns involving medical outcomes.

Information integration

For more computer-readable data sources—laboratory findings, genomic tests, numerical and other physiological data—information integration is critical for both research and clinical care. HIEs could eventually resolve such issues, but today even a single provider faces problems as many large institutions have multiple systems from different vendors to fulfill the various care functions—cardiovascular, ophthalmological, and so on.

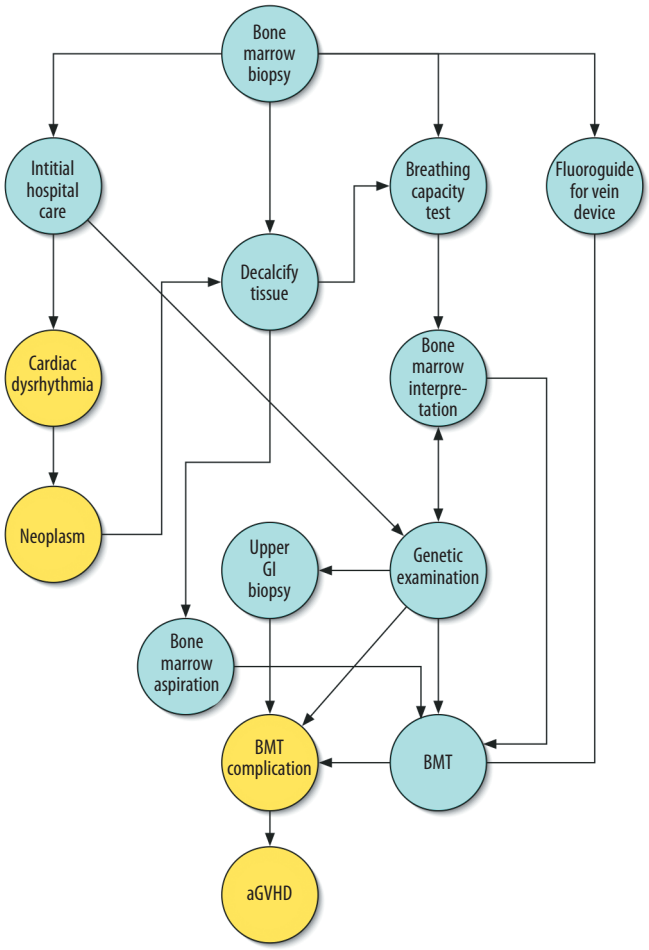


Figure 2. Network diagram of diagnosis and procedure codes mined from EHRs.

Sequential modeling

Much of a patient's interaction with a provider or hospital is temporal in nature, and there is a sequentiality to how symptoms and diagnoses develop and how milestones and procedures correlate with them. Hence, being able to mine sequential patterns or model temporal characteristics is likely to be central in many data mining efforts.

Figure 2 depicts a network diagram mined from thousands of EHRs that summarizes sequential patterns of diagnosis (yellow nodes) and procedure (blue nodes) codes. Arrows represent the flow of time. These patterns can often tell a story about a group of patients—in this case, a bone marrow biopsy (presumably due to a concern about cancer), followed by various tests and procedures, and ultimately a bone marrow transplant (BMT). The patients in this series all subsequently developed a major BMT complication—acute graft versus host disease (aGVHD).

Interactive exploration

Some analysis tasks can be easily supported when there is a focused question or pared-down dataset that contains the patients of interest. Other broad-based types of exploration fall into the true discovery category. Methods are needed to interactively explore data—including interrogation, manipulation, and visualization—and to verify novel patterns.

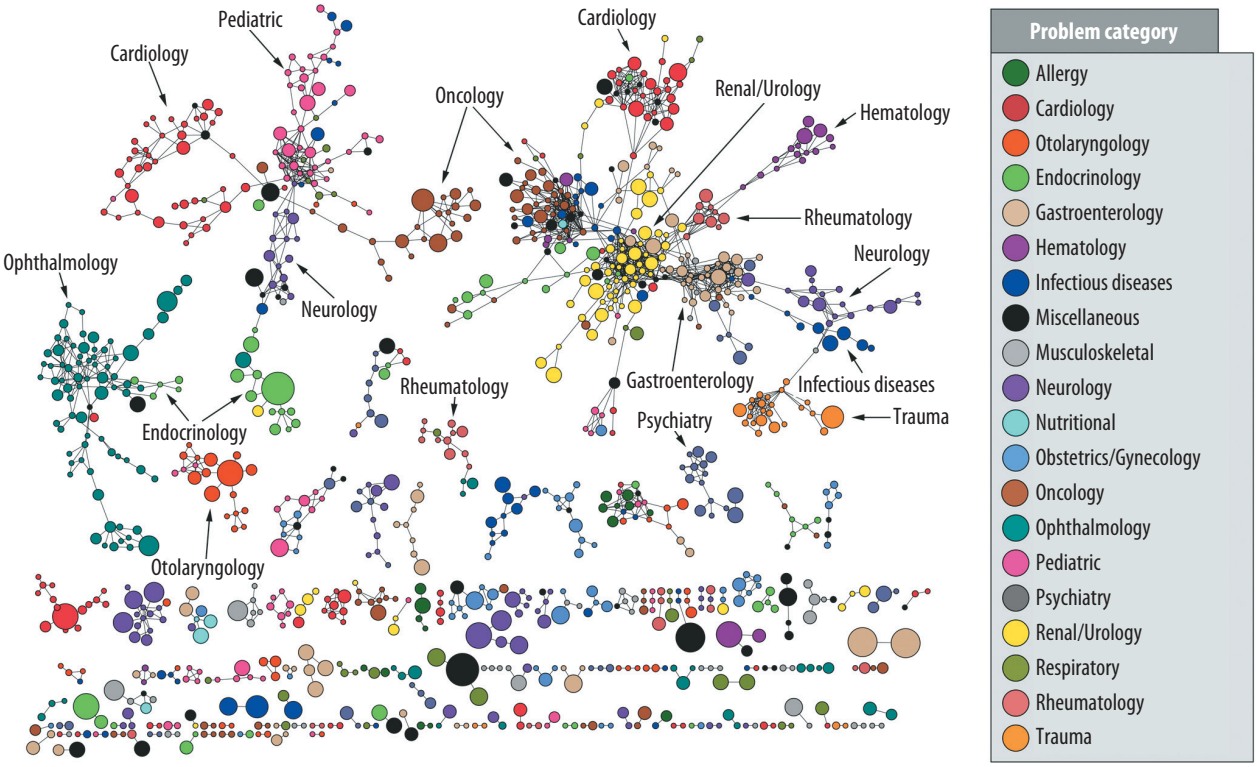


Figure 3. Network diagram from an association analysis of more than 300,000 patients based on their problems as entered into an EHR. Only the most significant associations are shown. The color of the nodes represents high-level disease categories and demonstrates how many problems of the same category cluster together and how they relate to problems from other categories.

Some general data mining issues are relevant in the EHR context as well. For instance, we may not want to find well-known or common patterns, and it takes more work to sift through less-prominent patterns to determine the more statistically significant or clinically meaningful ones. New measures of interest to the clinician must be defined. Similarly, there are many ways to segment data: do we study records by age, race, gender, lab values, or other, completely different measures?

Verification and validation

There is often no universal source of truth with which to compare data mining results. Thus, unlike a growing body of literature in the "–omics" community, we cannot readily screen an algorithm's findings against a known clinical data source to see what inferences are known and unknown. Comparing datasets from two large medical centers might yield interesting differences, but medical interpretation of inferences must be done very carefully.

Privacy preservation

Researchers have extensively studied privacy in the EHR context due to the US Health Insurance Portability and Accountability Act. HIPAA provides many safeguards when health information is exchanged between parties. Institutional review boards also play a role in protecting patients when the data are used for research. New privacy models are continually emerging, and there must be greater integration between formal model development and the specific clinical contexts in which EHR data is processed and interpreted.

OPPORTUNITIES

Initial efforts to mine EHRs are unlikely to yield many Eureka insights that have thus far eluded clinicians and practitioners. However, there are many opportunities for improving the delivery, efficiency, and effectiveness of healthcare.

Operations management

Data mining can help to define predicted census reports at hospitals. Such reports are an important tool for estimating staffing needs for the upcoming day and are currently defined heuristically using factors such as day of week, time of year, elective surgeries for the next day, and the current hospital census. Using data mining algorithms such as episode discovery, researchers will be able to formally model census reports as mixture models of key episodes from historical data.

Preventive healthcare

While automation is permeating major urban hospitals, the bulk of US

patients are serviced through secondary clinics and community hospitals. Being able to mine patterns across these practices can aid in preventive healthcare. Hospital networks periodically provide data to primary-care providers and smaller practices to aid in their management of specific conditions and their patients' statistics relative to the rest of the network. Data mining algorithms can help improve these reports by revealing temporal-event patterns with both diagnostic and predictive purposes, such as what factors as an outpatient might lead to an inpatient hospitalization.

Chronic disease treatment and prevention

Specific chronic ailments such as diabetes, obesity, and cardiovascular disease are among the leading causes of death and disability in the US and can be better understood by mining preexisting health records. Analyzing data patterns in conjunction with patients' electronic health history can lead to more robust conclusions regarding effective treatments and help predict who may be at greatest risk for developing certain complications.

Association analysis

Common features among disparate patients—whether diagnoses, procedures, or even lab data—can be discovered using association analyses. University of Michigan researchers recently performed such an analysis on diagnoses of more than 300,000 patients (D.A. Hanauer, D.R. Rhodes, and A.M. Chinnaiyan, "Exploring Clinical Associations Using '-Omics' Based Enrichment Analyses," *PLoS ONE*, 13 Apr. 2009, e5203). The study found both recently reported and novel associations including those between osteoarthritis and granuloma annulare and between ventricular septal defects and pyloric stenosis, shown at a high level in Figure 3.

Population tracking

Sites like Google Flu Trends use aggregated search data to track flu activity; similarly, aggregated information across medical systems can help monitor the prevalence and spread of infections such as H1N1 ("swine flu") and other influenzas. Cooperation and interchange between different hospital and healthcare systems is crucial for this goal.

Side-effect modeling

Just as the Arrowsmith system used indirect information to link biomedical records (N.R. Smalheiser and D.R. Swanson, "Using ARROWSMITH: A Computer Assisted Approach to Formulating and Assessing Scientific Hypotheses, *Computer Methods and Programs in Biomedicine*, Nov. 1998, pp. 149-153), there are significant opportunities to infer interactions from medical records. A decade ago, researchers discovered a rare link between intussusception (blockage of the intestine) and administration of the rotavirus vaccine given to infants to prevent severe diarrhea, leading to the vaccine's withdrawal. With more automation comes the potential to infer such side effects more quickly than is currently possible.

Currently, health IT research is focused on "first order" issues such as health system integration, interoperability, reducing medical errors, and providing reliable support to healthcare providers within and across networks. Opportunities for mining and computer-aided decision making are nevertheless blossoming across all stages of the enterprise. New funding initiatives and renewed interest in this area should prove to be a strong impetus.

However, because electronic health data standards have not yet been fully developed or agreed upon, we may end up with an infrastructure comprised of too many noninterchangable and proprietary systems, resulting in a "tower of Babel" of such data. We must therefore be cautious in adopting EHRs too rapidly, as they could delay needed data standardization. Ultimately, high-quality, coded clinical information will surely allow data mining to prove its worth.

***Naren Ramakrishnan**, the AI Redux column editor, is a professor of computer science at Virginia Tech. Contact him at naren@cs.vt.edu.*

***David Hanauer** is a clinical assistant professor of pediatrics at the University of Michigan, Ann Arbor, where he is affiliated with the Department of Pediatrics, the Comprehensive Cancer Center, and the Center for Computational Medicine and Bioinformatics. Contact him at hanauer@umich.edu.*

***Benjamin Keller** is an associate professor of computer science at Eastern Michigan University. Contact him at bkeller@emich.edu.*