guidance; machine understanding of human behavior, emotional, and physiological states; and the need to respond appropriately to unintended stimuli. Numerous machine perception, cognition, and communication challenges also remain such as image-guided intervention, speech and language understanding, two-hand-like manipulative dexterity, and learning systems that adapt to an individual's long-term change of state. Solutions likely lie at the intersection of new and ongoing research in computer science, materials, psychology, and neuroscience.

The societal pressure to mitigate the healthcare crisis presents an unprecedented opportunity for computing, information science, and engineering. Whereas the pursuit of understanding the pathogenesis of disease will be accelerated with new algorithms and increasingly powerful computation and data architectures, we look to other computation-enabled means to provide additional avenues to the pursuit of quality of life. Multidisciplinary approaches are required to engineer a privacy-maintaining information infrastructure with secure, real-time access to unprecedented amounts of heterogeneous health, medical, and treatment data. New generations of algorithms must be developed to utilize the resulting global resource of population-based evidence for assisted discovery, knowledge creation, and even individual point-of-care decisions. Analytics based on modeling phenomena ranging from the physiology of humans to their social interactions are required to optimize therapies ranging from molecular medicine to behavioral interventions.

Such advances in human-centered computing in combination with standardization and commercialization of unobtrusive sensing and robotics will trigger a disruptive change in healthcare and wellbeing by empowering individuals to more directly participate. Finally, partnerships among academic, industrial, and governmental bodies are required to enable these computer science innovations and realize their deployment in order to help transform healthcare.

References

1. G. Anderson and P. Markovich, *Multinational Comparisons of Health Systems Data*, The Commonwealth Fund, 2009.
2. T.K. Landauer, *The Trouble With Computers: Usefulness, Usability, and Productivity*, MIT Press, 1995.
3. President's Council of Advisors on Science and Technology (PCAST), *Realizing the Full Potential of Health Information Technology To Improve Healthcare for Americans: The Path Forward*, Executive Office of the President, 2010.
4. Inst. of Medicine, *Digital Infrastructure for the Learning Health System: The Foundation for Continuous Improvement in Health and Health Care*: *Workshop Series Summary*, Nat'l Academies Press, 2011.
5. L. Northrop et al., *Ultra-Large-Scale Systems: The Software Challenge of the Future*, Software Eng. Inst., Carnegie Mellon Univ., 2006.
6. A. Bandura, "Self-Efficacy in Health Functioning," *Cambridge Handbook of Psychology, Health and Medicine*, 2nd ed., S. Ayers et al., eds., Cambridge Univ. Press, 2007.
7. C. Chen et al., "The Kaiser Permanente Electronic Health Record: Transforming and Streamlining Modalities of Care," *Health Affairs*, vol. 28, no. 2, 2009, pp. 323–333.

Howard Wactlar is vice provost for research computing, associate dean in the School of Computer Science, and alumni research professor of computer science at Carnegie Mellon University. Contact him at wactlar@cmu.edu.

Misha Pavel is a professor in the Department of Biomedical Engineering at the Oregon Health and Science University. Contact him at pavel@bme.ogi.edu.

Will Barkis is a AAAS science and technology policy fellow at the American Association for the Advancement of Science. Contact him at wbarkis@gmail.com.

Opportunities and Challenges in Association and Episode Discovery from Electronic Health Records

David A. Hanauer, *University of Michigan Medical School*
Kai Zheng, *University of Michigan School of Public Health and School of Information*
Naren Ramakrishnan, *Virginia Tech*
Benjamin J. Keller, *Eastern Michigan University*

As healthcare practices, both small and large, move from traditional paper-based patient charts to electronic health records (EHRs), new opportunities are emerging for secondary uses of data captured as part of routine care. Such opportunities include not only traditional research methodologies involving relatively small cohorts of selected patients, but also large-scale data mining analyses encompassing hundreds of thousands or even millions of patients at once.

Performing these nontraditional analyses has required novel computational approaches, sometimes borrowing from techniques originally developed in other fields such as genomics and network theory. Additionally, to interpret such large volumes of data in a meaningful

way often requires interactive visual analytics that were developed, and have demonstrated enormous value, in other disciplines. Along with these new and exciting possibilities created with the application of computational sciences to clinical data, new issues have also emerged that still have yet to be adequately addressed.

Composition of an EHR

EHRs are quite variable in their design and structure, but most of them contain certain core components including diagnoses, procedures, medications, progress notes, assessments, and plans. Sometimes these data elements are coded and other times (especially in the case of documentation of clinicians' conclusions and reasoning underlying the conclusions) are recorded in a free-text, narrative format often created by dictation and transcription. Free text does not lend itself well to computation but medical natural language processing (NLP) algorithms can help extract predefined fields that can then be used for computational analyses.

Among coded data, there is a multitude of controlled medical vocabularies in use, such as the International Classification of Diseases version 9 (ICD-9) and Current Procedural Terminology version 4 (CPT-4) codes, that are commonly used for billing and reimbursement. Although clinicians are increasingly required to code their findings (such as diagnoses) and actions (such as medication prescriptions) as part of their clinical care, a fundamental conflict exists between expressivity allowed by narrative documentation and computability enabled by coded data, which has led to numerous usability issues and significant user resistance.[1]

A New Kind of Research

Traditional clinical research is usually conducted with clearly defined patient populations selected based on rigorous inclusion and exclusion criteria; data for such studies are hence collected uniformly according to a predefined, rigid study protocol.

In the new computational paradigm, it has become possible to use vast amounts of data captured as part of routine patient care practice, not originally intended for research. The compromises in the uniformity of data can be compensated for, at least in theory, by much larger cohorts of patients that have their conditions, treatments, and outcomes recorded in EHR systems. Thus, the hope is that even with a decreased signal-to-noise ratio or greater variation, the likelihood of making new discoveries is comparable to traditional approaches.

Another advantage of applying computational methods to analyze clinical data is that there are no predisposed biases about what can or should be discovered, and therefore multiple hypotheses can be tested, the significance of which can be algorithmically assessed across the entire set of patients.

Associations in Diagnoses

A few years ago, we applied an algorithm for discovering associations among gene-expression data to the highly variable free-text diagnoses of more than 325,000 patients in our EHR system at the University of Michigan.[2] The dataset consisted of 1.5 million diagnoses that included about 20,000 distinct free-text diagnoses, each occurring in five or more patients. Hypertension, the most common diagnosis, appeared more than 58,000 times. The 3,500 most frequently appearing terms were mapped to one another to reduce the free-text variability so that, for example, T1DM was made equivalent to type 1 diabetes mellitus.

The investigation was based on software originally developed for gene-expression signature analysis. In our case, each patient's collection of diagnoses were analogous to a gene-expression signature. Odds ratios and p-values were computed for every diagnosis pair using an all-versus-all association analysis approach. We used Fisher's exact test to determine significance.

Results were visualized using network diagrams to help identify meaningful associations. Of the nearly half-million highly significant associations discovered, many were known (see Figure 2), which provided confirmation that the approach was working. Other associations were recently reported in the literature, suggesting that the approach might be useful for hypothesis generation. Such associations included a history of smoking and amyotrophic lateral sclerosis (Lou Gehrig's disease) as well as fibromyalgia and hypothyroidism. Some associations were novel, but they might share as yet undescribed biological underpinnings, such as those between granuloma annulare (GA) and osteoarthritis (OA), as well as between pyloric stenosis and ventricular septal defect. Both GA and OA have been treated with niacin, although no common pathway is currently described. Lastly, some associations were unusual with no plausible explanation (such as a possible connection between cat bites and depression).

Temporal Episodes in EHRs

More recently, we have taken into account temporal relations between events to better elucidate patterns of disease progression. An example of a known short-time-scale event is the development of a rash a few days

following the administration of certain antibiotics in patients with infectious mononucleosis, whereas a longer-term event is the development of cancer decades after exposure to radiation from computed tomography (CT, or CAT) scanners.

In this study,[3] we retrieved the longitudinal medical records of 1.6 million patients that contained nearly 100 million coded ICD-9 and CPT-4 diagnoses and procedures. The record with the longest time span involved a patient's medical encounters (for example, ambulatory visits or inpatient admissions) over 22 years. This suggested to us the potential for identifying patterns of sequences of codes that might be contained in the data.

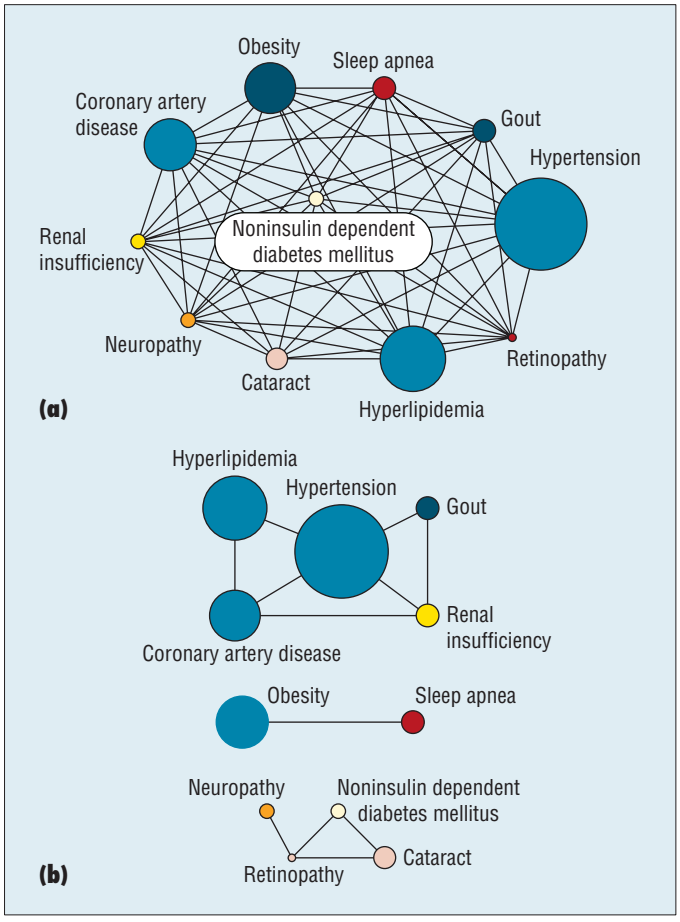

Figure 2. Visualizing meaningful associations. (a) A network diagram showing the most significant medical problems associated with noninsulin dependent (Type 2) diabetes mellitus. (b) When edges between less significantly associated nodes are removed, three separate network diagrams remain, with the most highly significant associations clearly demonstrated. For example, sleep apnea is most commonly associated with obesity.

A straightforward implementation of sequential pattern-finding algorithms did give us a level of data reduction. However, many of the patterns uncovered were permutations or near permutations of each other, representing different serializations of the same sets of codes. This led us to mine partial orders of codes where the objective was to compress alternative orderings of codes into a hierarchical structure.

Figure 3 shows an example of how a general diagnosis of hip pain is partitioned into subsequent diagnoses of variable order, ultimately leading to a hip replacement. The pattern encodes alternative orderings of progression. There is pelvic osteoarthrosis and a joint symptom followed by a femoral neck fracture. Both pelvic osteoarthrosis and a joint symptom occur before the hip replacement, but there is no specific ordering between these two stages other than both following an initial diagnosis of pelvic pain. Based on the existence of such patterns, it is possible to develop a "temporal process query" engine that can search for specific symptoms preceding and succeeding (or having no particular temporal order with) given conditions of interest. Furthermore, such a query engine can also use timeline-based visualizations to help practitioners determine historical diagnoses encountered in the past and prompt practitioners for necessary investigations.

Challenges

In our experience, mining large datasets of patient records stored in EHRs, in addition to handling the nearly infinite combinatorial possibilities for pattern detection, is approaching the limit of modern supercomputing technologies. Therefore, we found that a priori decisions had to be made with respect to pruning the analytic space either by constraining the time scales or the codes included in the analysis. This trade-off will inherently reduce our ability to discover certain associations, but it is necessary for the results to be manageable. Nevertheless, we expect that this approach can still help identify temporal patterns that might not otherwise be easily detected. Such correlations continue to be reported, including one recently announced by the US Food and Drug Administration related to the possible association between breast implants and the development of a rare cancer known as anaplastic large cell lymphoma (ALCL).[4]

These approaches will only show correlations and statistical associations, but they cannot determine causation. Thus, they are good for initial discovery and generating hypotheses, but further and more rigorous follow-up studies will almost always have to follow. The association between a cat bite and depression, for example, provides a perfect illustration of this conundrum.

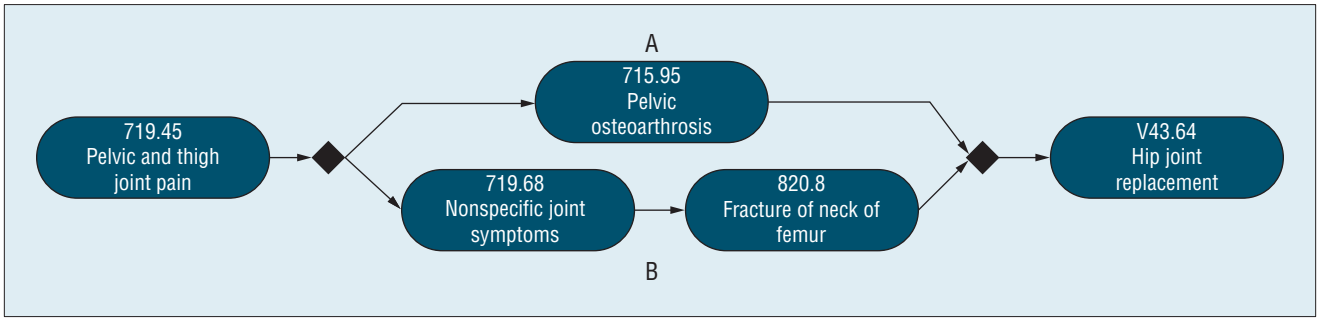

Figure 3. A five-sequence pattern demonstrating an initial diagnosis of pelvic pain followed by the variable sequence of (path A) osteoarthritis and (path B) nonspecific joint pains and subsequent femoral neck fracture, with a final common pathway converging on a hip-joint replacement. The variability is that path A sometimes precedes path B and vice versa. The codes in the diagram are International Classification of Disease version 9 (ICD-9).

Although it is certainly intriguing, the association raises multiple questions not only about cause and effect but also clinical significance. Do cats bite depressed people more than happy people? Do people become depressed when their cat bites them? Do individuals with depression simply own more cats? Or are some of these just meaningless associations that came to our attention by chance? After all, in a dataset with hundreds of thousands of associations statistically significant at the 0.001 level, we might still expect hundreds of these associations to be "significant" by chance alone.

Indeed, when we initially explored the results of our dataset, we found some unusual associations that we ruled out based on further exploration. An example included an association between autism and intestinal candidiasis. Due to the ongoing controversy surrounding the etiology and treatment of autism, we looked into this relationship further and found that a single physician in our health system had made all the entries in which these two diagnoses appeared together. Thus, rather than new knowledge being derived from the "wisdom of the crowds," we were likely uncovering the bias introduced by a single clinician.

One of the biggest challenges that remain is the lack of a reference standard for what is known versus unknown. When hundreds of thousands of associations are discovered, how can we practically sort through them to identify those that are novel, if there does not exist an automated way to filter out the well-known ones? At this point, we still need to manually review a subset of results using our clinical judgment. If we filter by the most significant associations, we are more likely to uncover findings that are well known, but as we reduce the level of significance the number of possibilities quickly becomes unmanageable, even though these less significant associations are not likely to be described in the literature. Data visualization can help us rapidly scan the results efficiently, and additional approaches for visualizing the datasets would be welcomed.

Additionally, constraining the data to include in the analysis will likewise constrain the range of potential discoveries. For example, our analyses have not included medications, precluding discovery of associations between drugs and adverse events—including the recently reported association between the diabetes drug Avandia (rosiglitazone) and subsequent heart attacks and strokes.[5]

Finally, the variability of the many code sets in use, with their alphabet soup of abbreviations (see Table 1), might make it difficult to merge data from different sources and preserve meaning across them. For example, in 2013, all practices in the US must switch from ICD-9 to ICD-10, which Europe has been using for years. But for those planning to combine data from both ICD code sets, the mapping will not always be straightforward, and this might present its own set of challenges tangential to the underlying analysis.

Together, the application of computationally intensive data mining approaches with visualization of the results have opened up possibilities for discoveries that were previously impractical, if not impossible, and demonstrates the power of what can be done at the intersection of clinical informatics, healthcare, and computer science.

Clinical practices that have recently implemented EHRs are only now starting to capture electronic data. Ten or 20 years from now, however, we will have captured a tremendous amount of longitudinal data that will likely be used for new computational and visualization approaches that have yet to be developed.

References

1. S.T. Rosenbloom et al., "Data from Clinical Notes: A Perspective on the Tension Between Structure and Flexible Documentation," *J. Am. Medical Informatics Assoc.*, vol. 18, no. 2, 2011, pp. 181–186.
2. D.A. Hanauer, D.R. Rhodes, and A.M. Chinnaiyan, "Exploring Clinical Associations Using '-omics' Based Enrichment Analyses," *PLoS One*, 2009;4(4):e5203.

Table 1. A sampling of code sets used in clinical medicine.

Code set	Full name	Main uses	Number of codes/concepts*	Copyright/ownership
CPT-4	Current Procedural Terminology, 4th edition	Procedural billing and coding	10,000	American Medical Association
ICD-9 and ICD-10	International Statistical Classification of Diseases and Related Health Problems, versions 9 and 10	Diagnoses billing and coding	21,000 (ICD-9) and 155,000 (ICD-10)	World Health Organization
ICD-O-3	International Classification of Diseases for Oncology, 3rd edition	Oncology billing and coding	9,500	World Health Organization and College of American Pathologists
IMO	Intelligent Medical Objects	Clinical concept matching for medical records	190,000+	Intelligent Medical Objects
LOINC	Logical Observation Identifiers, Names and Codes	Laboratory testing and results	58,000	Regenstrief Institute
MEDCIN	MEDCIN	Clinical documentation in medical records	250,000	Medicomp Systems
MeSH	Medical Subject Headings	Medical literature concepts	25,000	National Library of Medicine
NDC	National Drug Code	Drugs names and ingredients	110,000	US Food and Drug Administration
NDF-RT	National Drug File, Reference Terminology	Drugs, ingredients, physiologic effects, and so forth	130,000	US Department of Veteran Affairs
SNOMED-CT	Systematized Nomenclature of Medicine, Clinical Terms	Research, clinical data organization	> 1,000,000	Originally the College of American Pathologists, now the International Health Terminology Standards Development Organisation

* These are estimates; actual numbers frequently change.

3. D. Patnaik et al., "Experiences with Mining Temporal Event Sequences from Electronic Medical Records: Initial Successes and Some Challenges," *Proc. 17th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining* (KDD 11), 2011.
4. US Food and Drug Administration, "FDA Review Indicates Possible Association between Breast Implants and a Rare Cancer," 26 Jan. 2011; www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/ucm241090.htm.
5. US Food and Drug Administration, "FDA Significantly Restricts Access to the Diabetes Drug Avandia," 23 Sept. 2010; www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/ucm226975.htm.

David A. Hanauer is a clinical assistant professor in the Department of Pediatrics at the University of Michigan Medical School. Contact him at hanauer@umich.edu.

Kai Zheng is an assistant professor in the University of Michigan School of Public Health Department of Health Management and Policy and School of Information. Contact him at kzheng@umich.edu.

Naren Ramakrishnan is a professor in the Department of Computer Science at the Virginia Tech. Contact him at naren@cs.vt.edu.

Benjamin J. Keller is an associate professor in the Department of Computer Science at Eastern Michigan University. Contact him at bkeller@emich.edu.

Data Mining Large-Scale Electronic Health Records for Clinical Support

Yu-Kai Lin and Randall A. Brown, *University of Arizona*
Hung Jen Yang, *Min Sheng General Hospital, Taiwan*
Shu-Hsing Li and Hsin-Min Lu, *National Taiwan University, Taiwan*
Hsinchun Chen, *University of Arizona*

A clinical consultation process usually involves two major steps: diagnostic reasoning and treatment planning. First, a physician tries to identify a patient's health problems based on the presented signs or symptoms using his or her own medical expertise. Next, based on the best conclusion about the patient's conditions, the physician plans the most suitable treatments for the patient.

Providing clinical support is challenging because tens of thousands of symptoms, diseases, and treatments (SDT) constitute an extremely high dimensional search and decision space for physicians. Although the traditional divisions of different medical specialties help reduce the complexity, the degrees of knowledge depth and breadth in each area