



# Smart Health and Wellbeing

Hsinchun Chen, *University of Arizona*

**T**oday's healthcare has become cost-prohibitive for many, and it suffers substantially from medical errors and waste.<sup>1</sup> An often cited reference is the 1998 Institute of Medicine report, which estimated that preventable medical errors lead to as many as 98,000 deaths per year in the US.<sup>2</sup> Gerard Anderson and Patricia Markovich also reported that the US spends \$1.7 trillion annually (16 percent of GDP) on healthcare, yet produces significantly lower health outcomes than many other developed countries.<sup>3</sup>

Many new government initiatives to improve this situation have been started recently. For example, President Obama's \$787 billion federal stimulus package was announced in 2009. In it, the Health Information Technology for Economic and Clinical Health (HITECH) Act for healthcare information technology (IT) stipulates that healthcare entities in the US need to use IT to fix ingrained problems, with \$50 billion allocated to this effort over the next five years and \$19 billion in 2011 alone. Similarly, China's government announced a plan in January 2009 to spend more than \$120 billion on the first phase of a 10-year overhaul of its healthcare system.<sup>4</sup>

## Research Initiatives

In academia there has also been significant recent interest in adopting and advancing IT for effective healthcare. The 2009 US National Research Council report on "Computational Technology for Effective Health Care" suggests an overarching research grand challenge of developing "patient-centered cognitive support."<sup>1</sup> It also points out several representative research challenges for the IT community, including virtual patient modeling, healthcare automation, healthcare data sharing and collaboration, and healthcare data management at scale. The National Academy of Engineering's 2008 report<sup>5</sup> featured advancing health informatics as

one of the 14 major challenges of engineering. A major research goal is to develop trusted healthcare systems that offer relevant decision support to clinicians and patients and offer "just in time, just for me" advice at the point of care. Other suggested healthcare IT advances include, for example, better (wearable) devices for patient monitoring and effective public health surveillance.

In 2010, the US National Science Foundation (NSF) announced an innovative cross-division program called Smart Health and Wellbeing (see [www.nsf.gov/funding/pgm\\_summ.jsp?pims\\_id=503556](http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=503556)). The then assistant director of NSF's Computer and Information Science and Engineering (CISE) Jeanette Wing explained, "We are looking for your great ideas for how advances in computer and information science and engineering can transform the nature and conduct of healthcare and wellness as we know it today."

The goal of the Smart Health and Wellbeing program is to seek improvements in safe, effective, efficient, equitable, and patient-centered health and wellness services through innovations in computers, information science, and engineering. Doing so requires leveraging the scientific methods and knowledge bases of a broad range of computing and communication research perspectives. Projects submitted to this program should be motivated by specific challenges in health and wellbeing. The Smart Health and Wellbeing program aims to facilitate large-scale discoveries that yield a long-term, transformative impact in how we treat illness and maintain our health.

In light of such overwhelming interest from governments and academia, there are great opportunities for researchers and practitioners alike to invest effort in conducting innovative and high-impact healthcare IT research. This *IEEE Intelligent Systems* Trends and Controversies (T&C) Department hopes to raise awareness and highlight

selected recent research to move us toward such goals.

### In This Issue

This T&C Department includes three articles on Smart Health and Well-being from distinguished experts in computer science, information systems, and medicine.

In “Can Computer Science Save Healthcare?” Howard Wactlar, Misha Pavel, and Will Barkis discuss ways that computer science research could help resolve some aspects of the healthcare crises in America. They propose a program of research and development along four technology thrusts to enable this vision:

- creating an interoperable, digital infrastructure of universal health data and knowledge;
- utilizing diverse data to provide automated and augmented insight, discovery, and evidence-based health and wellness decision support;
- a cyber-based empowering of patients and healthy individuals that enables them to play a substantial role in their own health and treatment; and
- monitoring and assisting individuals with intelligent systems (including sensors, devices, and robotics) to maintain function and independence.

In the second article, David A. Hanauer, Kai Zheng, Naren Ramakrishnan, and Benjamin J. Keller’s “Opportunities and Challenges in Association and Episode Discovery from Electronic Health Records” summarizes their research in associations in medical diagnoses using large-scale, longitudinal electronic health records (EHRs). They also consider temporal relations between events to better elucidate patterns of disease progression. They argue that

applying data mining approaches with visualization could open up opportunities for healthcare discoveries that were previously impractical.

Lastly, Yu-Kai Lin, Randall A. Brown, Hung Jen Yang, Shu-Hsing Li, Hsin-Min Lu, and Hsinchun Chen’s article “Data Mining Large-Scale Electronic Health Records for Clinical Support” presents their recent research on symptom-disease-treatment (SDT) association rule mining using comprehensive EHRs containing approximately 2.1 million records from a major Taiwan hospital. Based on selected International Classification of Diseases version 9 (ICD-9) codes, they were able to identify clinically relevant and accurate SDT associations from patient records in seven distinct diseases, ranging from various cancers to chronic and infectious diseases. Their recent analysis also involves scenario-based SDT association mining on different patient age groups and of either gender to determine if diagnoses and treatments are appropriate or relevant to patients’ demographic backgrounds.

Each of these articles presents unique perspectives, advanced computational methods, and selected results and examples. They demonstrate the breadth of computing areas that can help address the challenges in using IT for effective healthcare. As we move forward, researchers will need to combine and/or go beyond traditional areas of computing, information science, and engineering to meet specific challenges in healthcare. The hope is that future studies and discoveries will transform how we care for patients, treat illness, gather patient data, and maintain health.

### Acknowledgments

This research is supported in part by US National Science Foundation grants CNS-070933, CBET-0730908, and IIS-0428241.

### References

1. Nat’l Research Council, *Computational Technology for Effective Health Care: Immediate Steps and Strategic Directions*, Nat’l Academy Press, 2009.
2. Inst. of Medicine, *To Err Is Human: Building a Safer Health System*, Nat’l Academy Press, 2000.
3. G. Anderson and P. Markovich, *Multinational Comparisons of Health Systems Data*, The Commonwealth Fund, 2009.
4. G. Fairclough, “In China, Rx for Ailing Health System,” *Wall Street J.*, 15 Oct. 2009, p. A1; <http://online.wsj.com/article/SB125556557369186287.html>.
5. Nat’l Academy of Eng., *Grand Challenges for Engineering*, 2008; [www.engineeringchallenges.org/cms/challenges.aspx](http://www.engineeringchallenges.org/cms/challenges.aspx).

**Hsinchun Chen** is the director of the Artificial Intelligence Lab at the University of Arizona. Contact him at [hchen@eller.arizona.edu](mailto:hchen@eller.arizona.edu).

### Can Computer Science Save Healthcare?

**Howard Wactlar**, *Carnegie Mellon University*

**Misha Pavel**, *Oregon Health and Science University*

**Will Barkis**, *American Association for the Advancement of Science*

Healthcare in America is in crisis. We spend \$1.7 trillion on healthcare, yet have significantly lower health outcomes than many other developed countries, including life expectancy.<sup>1</sup> Today’s problems will be exacerbated by tomorrow’s age demographics, chronic diseases, and lack of resources. This will account for an increase from 16 to 25 percent of GDP between 2009 and 2014 and significant care-giver shortfalls.

This poor performance is in part due to the fragmented, uncoordinated, and inefficient uses of resources and information, disengagement of patients, and misaligned economic incentives. A confluence of factors—social, political, technological, and scientific—lead us to conclude that now is the time to catalyze a cross-fertilization between computer science, healthcare, and health/wellness systems through means well beyond the meaningful use of electronic health records (EHRs) legislated as part of the Health Information Technology for Economic and Clinical Health (HITECH) Act. Increasing demands of improved quality of life, increased efficiency, and reduced costs are driving radical transformations in the US healthcare system and creating unprecedented opportunities for disruptive change through technological advances.

Digitizing health data and existing processes will likely help, but as in many prior application domains, the initial gains achieved by digitizing data alone are marginal.<sup>2</sup> Healthcare must evolve in more fundamental ways. We look to “game-changers” such as moving from healthcare to health, reactive care to proactive and preventive care, clinic-centric to patient-centered practice, training- and experience-based interventions to globally aggregated evidence, and episodic response to continuous well-being monitoring and maintenance.

Fundamental research challenges must be addressed by the broad computer, networking, and information science communities to achieve such transformative goals. We propose a program of research and development along four technology thrusts to enable this vision. First, we must create an interoperable, digital infrastructure of universal health data and knowledge. Second, we need to

utilize diverse data to provide automated and augmented insight, discovery, and evidence-based health and wellness decision support. Third, a cyber-based empowering of patients and healthy individuals is necessary to enable them to play a substantial role in their own health and treatment. Lastly, we must find ways to monitor and assist individuals with intelligent systems (including sensors, devices, and robotics) to maintain function and independence.

### **Digital Infrastructure Systems and Data**

The digital health information infrastructure starts with the vision that an individuals’ lifelong health information can follow them from all modes of production to every point of health and healthcare action. The data might include physiology, behavior, symptoms, treatments, and outcomes as well as genomic, proteomic, metabolomic, and a host of other “-omic” data. The data must be accessible to both the individual and the practitioner, albeit with customized and appropriate interfaces, applications, and access controls. Most importantly, the information architecture must enable data integration and reconciliation across all individuals and the repositories of all systems. These multimodal data—for example, unstructured and structured text, images, audio, video, and streaming—are essential for medical history representation, clinical decision support, biomedical research, epidemiology, population health, and other vital activities. In this context, information is a broad concept, ranging from EHRs, sensor and instrumentation data, trends analysis of diseases, and semantic rules in knowledge management systems. At its highest level, we need to ensure the continuous collection, integration, fusion, and dissemination

of all personal and public health data and knowledge from diverse sources in multiple formats and media, while respecting privacy and proprietary rights.

There are numerous challenges in the development of such an infrastructure. One of the main issues concerning knowledge representation is that the content of EHRs is highly variable among different providers and data sources. This variability includes the diverse terminology and definitions existing across providers, data provenance, data uncertainty, and so forth. As recognized by the President’s Council of Science and Technology Advisors,<sup>3</sup> a central issue is the development of a universal exchange language that would mediate translation of concepts, provide metatagging, and represent data provenance. This would support probabilistic reasoning and associated algorithms for the harmonization and fusion of diverse data types necessary for inference and data mining. Continuous extension of this approach requires rich semantic knowledge representation, such as the techniques used in the Semantic Web.

Any architecture supporting integration and exchange among these systems must be characterized by decentralized data, development, and operational control and authority; accessibility to both professionals and individuals; and highly trustworthy infrastructure ensuring privacy, authenticity, security, and reliability while providing individually prescribed access control. A recent Institute of Medicine (IOM) study<sup>4</sup> suggests using principles of an ultra-large-scale (ULS) systems approach, “a virtual system ... in which a few key elements, such as interchange representation, may be standardized, but whose many participants have diverse and even conflicting goals, so adaptability is key.”<sup>5</sup>

Another essential characteristic is open architecture with simple and transparent, but protective protocols allowing easy implementation of third-party applications, biomedical apps, analogous perhaps to the iPhone architecture, to be layered onto an EHR system and to function seamlessly with the other components. Thus, small and large companies alike would have incentives to develop new and improved applications and services that could be rapidly applied in the real world.

### **Reasoning under Uncertainty**

With growing numbers of EHRs and patient-driven personal health records (PHRs), results of clinical trials, and the biomedical research literature, combined with an operational infrastructure, healthcare is entering the realm of Big Data. In addition to optimizing point-of-care clinical decisions, data mining, machine learning, and natural language processing, data visualization will enable secondary use of health data for evidence-based medicine and comparative effectiveness research. Ultimately, we will realize automated and assistive discovery from the massive longitudinal care and individual “-omic” data accruing. Novel and diverse data will be captured by ubiquitous mobile sensors about the patients’ activities, behaviors, and physiologies and the environment around them.

To take advantage of these data, the health and wellness system needs to continuously adapt and learn. The resulting system, conceptualized by IOM as a learning health system, will shape the decisions of individuals, providers, and policymakers. Rich, real-time evidence will account for variations in patient needs and circumstances and will support the

next generation of insights. Integrating continuous home-based behavioral and physiological monitoring data will enable early detection and intervention of progressive conditions (such as multiple sclerosis, depression, or Alzheimer’s). Ultimately, we foresee the ability to tailor therapy to the molecular/genetic mechanisms in combination with environmental circumstances tempering the efficacy of therapies. These analytic algorithms within the learning health system will enable early detection of epidemics and discoveries of subtle relationships between environmental, social, and clinical aspects.

These goals of the learning health system present significant challenges to computer science and related technologies. Optimizing clinical decisions in real time will require algorithms for transformation and fusion of heterogeneous, distributed data with diverse provenance at multiple temporal scales and multiple levels ranging from molecular medicine to organs, systems, behavioral, and even social data. Fundamental to the knowledge representation will be computational models of compromised and healthy processes, including their dynamics. Quantitative, computational relationships between the underlying physical mechanisms and their symptoms are essential for therapeutic interventions tailored to the patient’s unique biology. These models would also enable precise definitions and assessment of behavioral markers required for personalized healthcare.

### **Empowering Patients**

When symptoms of illness or injury occur, the patients or their families perform initial triage, deciding whether to seek formal care. Thus, patients are and should be healthcare decision participants, and they need

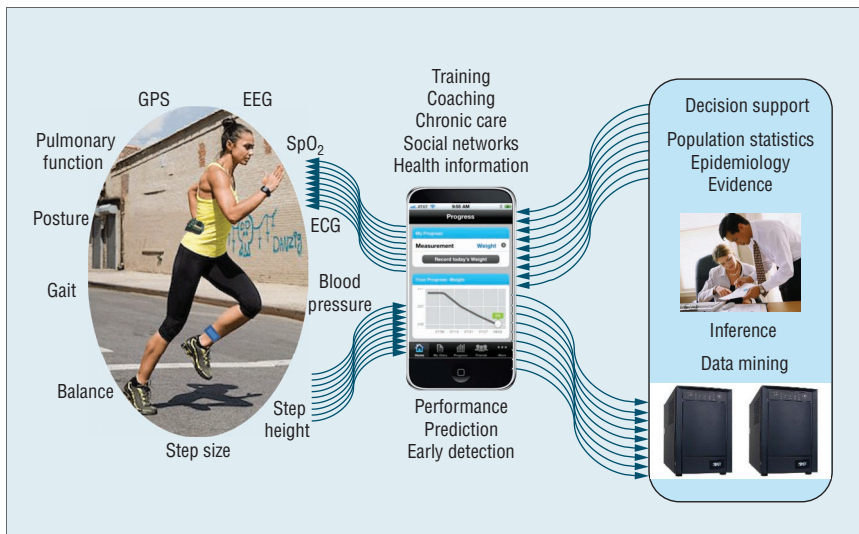
to be empowered to do so through their entire lives. Patients who actively participate in their healthcare have better outcomes and better perceived quality of life than those who do not.<sup>6</sup> They are more likely to adhere to their therapeutic regimens, engage in healthy behaviors, and thereby improve outcomes. Moreover, involving patients in decision making reduces the costs of care.

The Kaiser Permanente Hawaii health system implemented an experimental, integrated health IT system supporting both EHRs and electronic communication between cooperating physician teams and patients. With more than 225,000 members, the Kaiser system demonstrated a 26.2 percent reduction in annual office visits between 2004 and 2007, significantly increasing efficiency while maintaining quality and increasing satisfaction.<sup>7</sup>

Empowerment of patients and healthy individuals consists of several key components:

- *knowledge-to-action*—sufficient literacy, combining understanding with the ability to influence decisions by communicating subjective values, selecting interventions, and assessing outcomes;
- *patient self-efficacy*—confidence that patients can perform a given behavior, including decision making and a belief in their ability to change the situation;
- *availability of ubiquitous, just-in-time support*—with mechanisms that incorporate known principles of health-behavior change, based on the user models with representations of motivations and barriers to change;
- *availability of social networks* that are appropriate in terms of composition and educational level; and





**Figure 1. Life cycle of health data from body sensors to local assessment, aggregation with personal health records and comparable population-wide data, analysis, and feedback to the individual.**

- *telemedicine and assistive technologies*—availability of devices, systems, and the corresponding support for home care and rehabilitation.

Technical challenges to the computer science community include auto-tailored interfaces appropriate to user vocabulary, knowledge, and sophistication levels; assessment of quality of information; models of cognitive and emotional states of individuals; health-domain-specific social networking; and coaching platforms. A critical extension of this incorporates emerging mobile health technologies that enable relevant information and data access and the ability to intervene anytime, anywhere. These technological advances combined with virtual reality techniques and computational user models will enable optimization of coaching platforms and behavior-changing processes, thereby supporting individuals' attempts to modify their behaviors and achieve their health goals.

Another challenge involves developing technology to support effective social networks while enabling individuals to control the privacy of

their information. Finally, the goal of deploying easy-to-use, fail-safe networked home diagnostic and auto-adjustable rehabilitation instruments will provide individuals with support, decrease costs, and increase responsibility for their own health and wellness.

### **Intelligent Systems: Sensors, Devices, and Robotics**

The broad class of computationally intelligent systems that sense and respond to events and conditions in their environment by applying coordinated forces and movements and/or initiating communications and dialogue are game-changing for health. State-of-the-art technology systems are beginning to assist complex surgical procedures (such as Intuitive Surgical daVinci), monitor personal activity and physiological state (such as BodyMedia FIT), rehabilitate stroke victims at home (such as the University of California, Irvine, Java Therapy), safely propel wheelchairs over curbs and up stairs (such as Independence Technology iBot), and augment diminished body movement and strength with exoskeletons (such as Cyberdyne HAL).

We have barely started to tap the potential for devices and instrumented environments assisting mobility, manipulation, perception, and cognition. The functioning and performance of prosthetics and implants are yet to be dramatically enhanced by embedded intelligence and in-body networks. The categories of orthotics in our future will stretch from aids for walking to those for planning and reasoning in the performance of the instrumental activities of independent daily living.

The same underlying sensor and communication technologies can continuously gather human activity and behavioral data, monitor medication adherence, reveal early indicators of disease and depression, and thus enable proactive prediction, early intervention, and reduced severity of conditions. Figure 1 illustrates such a data transformation life cycle for data cumulatively captured by wearable body sensors. By also incorporating social robots and home assistants, there is the potential for behavioral therapy, life-style modification, and disease prevention. The time spent in hospitals, institutions, and even traveling for physician visits can be dramatically reduced with commensurate decrease in expense and increase in quality of life.

Enormous challenges remain ranging from the physical realization of sensors and precisely controlled actuators to power management, complexity of data analysis, data fusion, pattern recognition, and more; many of these require near real-time performance. First and foremost are the many issues of easy, intuitive, safe, and resilient human-robot interaction: the need to plan and navigate in rapidly changing home or external environments; the need for soft robot structures that cannot injure people or their environments; task-appropriate

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.<sup>1</sup>

### **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.<sup>2</sup> 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,<sup>3</sup> 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.

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 osteoarthritis and a joint symptom followed by

a femoral neck fracture. Both pelvic osteoarthritis 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

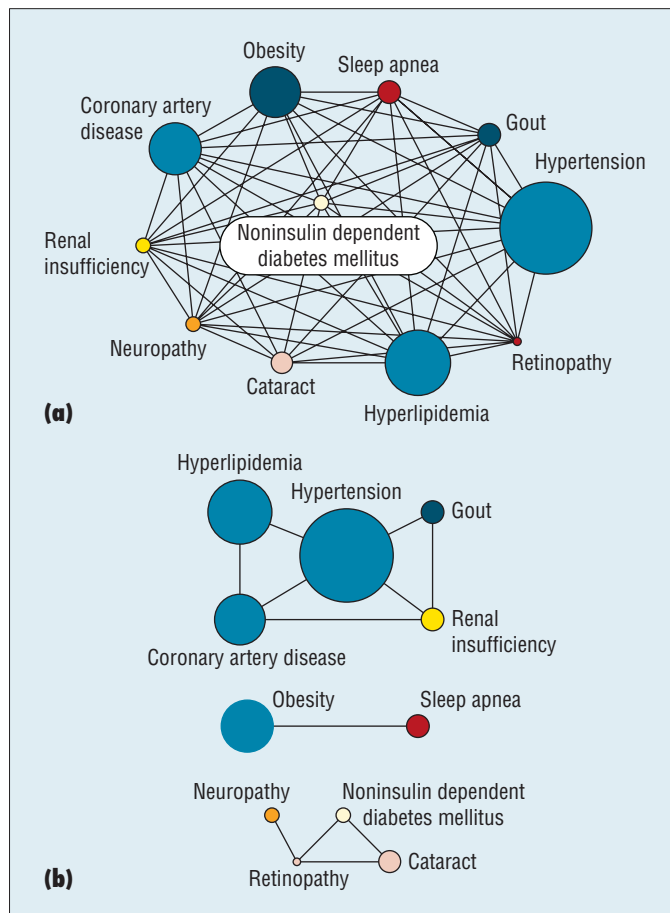
be reported, including one recently announced by the US Food and Drug Administration related to the possible association between breast implants and the development of a rare cancer known as anaplastic large cell lymphoma (ALCL).<sup>4</sup>

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.

prompt practitioners for necessary investigations.

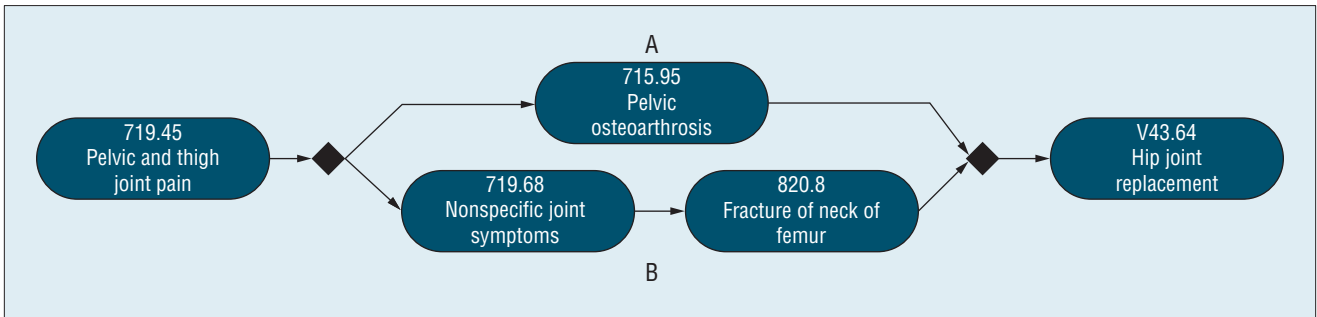
## Challenges

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



**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.**





**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.<sup>5</sup>

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](http://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](http://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](mailto: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](mailto: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](mailto: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](mailto: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

are still high. Elements in the clinical decision space are often interrelated. A physician might need to examine comorbidity before prescribing optimal drugs to a patient. Different drugs could potentially interact with each other and cause undesirable effects. When taken together, the dimensionality and interrelationships in clinical practices severely increase the difficulty of diagnostic reasoning and treatment planning.

This challenge can lead to inferior care and even serious impact on patient wellbeing. In fact, a large population survey study reported that the top causes of adverse events, after operative mistakes, are related to drug, medical procedure, incorrect, or delayed diagnosis and therapy.<sup>1</sup> Hence, providing a technique that supports and supplements the diagnostic healthcare process is an urgent and desperate need.

### Association Rule Mining

In essence, diagnostic reasoning and treatment planning are tasks that involve linking SDT together. Following this vein of thinking, association rule mining (ARM), a well-established data mining method, is a valid design choice to support these clinical processes. The ARM technique is widely used in the market basket problem, which helps determine which products are likely to co-occur in a shopper's basket. The problem was first formulated to efficiently generate associations.<sup>2</sup>

Association rules are often presented in the following form: {item set 1} → {item set 2}. Each item set contains one or multiple items. The item set on the left-hand side is a given condition, whereas the one on the right is a prediction. The quality of rules is often evaluated by support, the effect size of the rule, and confidence, the strength of the association.

In light of the growing number of electronic health records (EHRs) in medical facilities in the past decade, there has been increased interest in applying ARM in the medical context. Specifically, ARM has recently been applied to disease prediction, problem verification, comorbidity analysis, disease clustering, automatic order generation, treatment suggestion, and many other medical scenarios. For example, drawing from the structured EHRs in their institute, Adam Wright and his colleagues located clinically accurate associations among medications, laboratory results, and diseases, aiming to identify and mitigate gaps in the patient problem list.<sup>3</sup>

In this research, we are motivated to apply data mining techniques to SDT associations using large-scale EHRs, not only because of its promising development demonstrated in previous research, but more importantly to provide a tool that can support physicians' diagnostic reasoning and treatment planning in a highly complex and time-pressed environment.<sup>4</sup>

### Experimental Settings and Results

We obtained comprehensive EHRs from a major 600-bed hospital with six campuses located in northern Taiwan. The dataset contains EHRs of 894,061 registered patients in which there are approximately eight years of outpatient records (around 2,100,000 records from January 2002 to July 2010) and six years of inpatient records (about 140,000 records from November 2003 to July 2010). The data can be grouped broadly into four categories: registration (patient background), diagnosis (symptom and disease), treatment (procedure and order), and transaction (payment). The EHRs contain

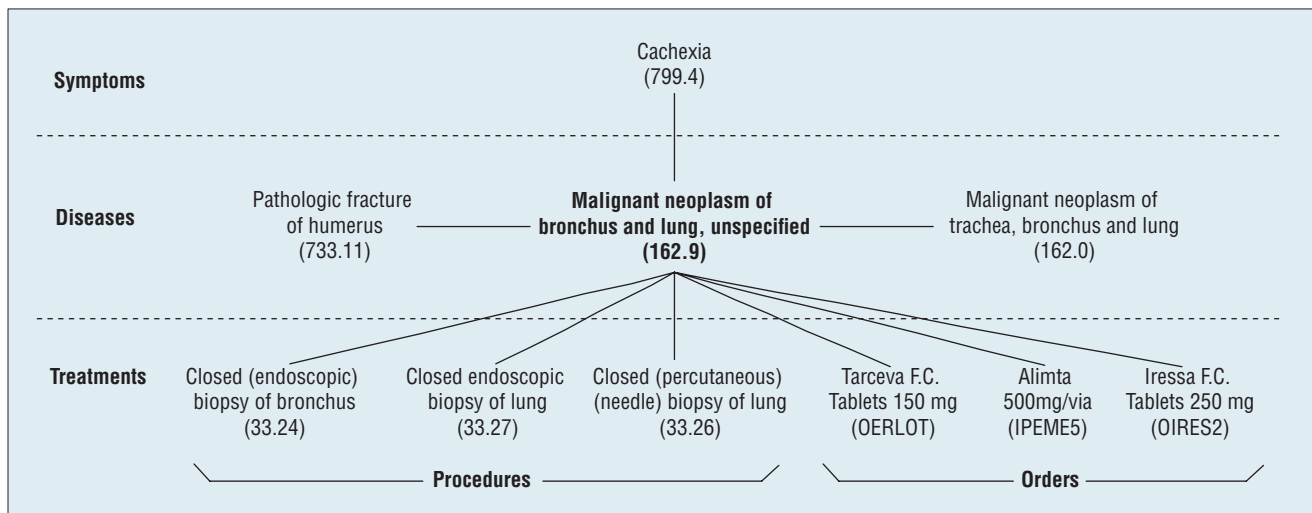
a significant amount of structured data; symptom, disease, and procedure are denoted following the coding standard of the International Classification of Diseases version 9 (ICD-9), whereas orders are coded with the internal scheme that links to pricing information aligned with Taiwan's National Health Insurance (one of the best national health insurance plans in the world).

We followed a relevant previous study in ARM and set our minimum support at five and a minimum confidence at 0.1.<sup>3</sup> To simplify the verification of association rules, we limited each item set to one item. The verification process proceeded as follows. First, we consulted WebMD.com for initial verification and examined whether the discovered associations were indeed described in WebMD. If not, we then resorted to Web search engines to determine if the associations are discussed in other reliable sources, such as academic medical databases (PubMed), medical journals, or authentic medical organization webpages.

The SDT association mining can be applied to any symptom, disease, or treatment with or without a specific target in mind. Our preliminary experiments were conducted on seven distinct diseases, ranging from cancers to chronic diseases to infectious diseases.

Given its large patient population and severe life threat, here we use lung cancer (with ICD-9 code 162.9) to demonstrate our results. The associations are built upon 975 lung cancer inpatient visits between 2003 and 2010. Figure 4 shows that SDT elements that are relevant to lung cancer.

Our technique demonstrates a high degree of accuracy. For example, cachexia is a symptom of fatigue and weakness frequently concurrent



**Figure 4. Network representation of symptoms, diseases, and treatments (SDT) association mining on lung cancer patient EHRs. The line width is roughly proportional to its confidence value. The codes in parentheses are either International Classification of Diseases version 9 (ICD-9) codes or order codes.**

with serious or chronic diseases, including lung cancer. On the other hand, a pathologic fracture of the humerus is associated with lung cancer when the cancer metastasizes to bone. With regard to treatments, the procedures in Figure 1 are laboratory examinations in lung cancer assessments, whereas the orders are chemotherapy drugs.

Table 2 provides a detailed analysis and verification on the top 10 orders. Iressa, Alimta, and Tarceva are chemotherapy treatments of advanced non-small cell lung cancer, and Taxotere and Etoposide are drugs to control the growth of cancer cells.

Perhaps, a more interesting and promising way to represent SDT associations is to confine the associations within certain subgroups or scenarios, which is considered closer to actual clinical practice. We have conducted the scenario-based SDT association mining on different patient age groups and genders to determine if diagnoses and treatments are relevant to patients' demographic background. We also compared SDT associations from different physicians, aiming to find how physicians' specialties and preferences might affect prescriptions. (Due to space

**Table 2. Verification of top lung cancer treatment orders.**

	Order code	Order name	Conf.	Verification	
				WebMD	Others*
1	OIRES2	Gefitinib F.C. (Iressa)	0.90	x	
2	IPEME5	Pemetrexed disodium heptahydrate (Alimta)	0.85	x	
3	OERLOT	Erlotinib (Tarceva)	0.82	x	
4	OGFI2	Gefitinib F.C. (Iressa)**	0.78	x	
5	IETOP1	Etoposide injection (Etoposide-Teva)	0.74	x	
6	IDOCE8	Docetaxel (Taxotere)	0.60	x	
7	70213	Radical Lymphadenectomy	0.56		x
8	SMEGEOS	Megestrol acetate suspension (Megest)	0.43		x
9	37038	Intravenous chemotherapy <=1 hours	0.39		x
10	33103	CT guided biopsy	0.37	x	

\* Other sources here include PubMed and the *New England Journal of Medicine*.

\*\* This order name is repeated due to different drug pricing for the national health insurance in Taiwan.

limitations, we are not able to include these results here.) Overall, the scenario-based SDT association mining demonstrates the usefulness of identifying intriguing treatment patterns within and between subgroups.

**I**n the clinical context, the importance of providing intelligent systems that fit and support physicians'

diagnostic reasoning and treatment planning cannot be overemphasized. Our experimental results on the SDT association mining technique show great potential toward this goal. Specifically, we presented the applicability of ARM to physicians' diagnostic thinking, discussed SDT associations in a disease case, and verified the accuracy of the results.



Our future studies involve several major directions. First, we will leverage the results from the SDT association mining to identify outlier individuals and abnormal patterns, which could be due to either exceptional medical decisions or clinical mistakes. In addition, we will explore knowledge patterns evolved over time as new problems, diagnoses, and treatments continue to change for different patient populations and diseases. We are also in the process of incorporating laboratory results and medical images to augment or adjust the identified associations for various diseases. Advanced text mining and natural language processing techniques will be explored to extract and represent textual narratives frequently expressed in clinical notes and lab reports. Lastly, the association rules will be extended to the outcomes of these medical interventions, instead of stopping at what treatments the patients receive. ■

### Acknowledgments

This work was supported in part by the US National Science Foundation through grant IIS-0428241. We thank Chung-Ting (William) Shing, Chung-Huan Hsieh, and Polin Li for their research support.

### References

1. E.J. Thomas et al., "Incidence and Types of Adverse Events and Negligent Care in Utah and Colorado," *Medical Care*, vol. 38, no. 3, 2000, pp. 261–271.
2. R. Agrawal, T. Imieliński, and A. Swami, "Mining Association Rules Between Sets of Items in Large Databases," *ACM SIGMOD Record*, vol. 22, no. 2, 1993, pp. 207–216.
3. A. Wright, E.S. Chen, and F.L. Maloney, "An Automated Technique for Identifying Associations Between Medications, Laboratory Results and Problems," *J. Biomedical Informatics*, vol. 43, no. 6, 2010, pp. 891–901.
4. G.D. Schiff and D.W. Bates, "Can Electronic Clinical Documentation Help Prevent Diagnostic Errors?" *New England J. Medicine*, vol. 362, no. 12, 2010, p. 1066.

**Yu-Kai Lin** is a PhD student in the Department of Management Information Systems and a research associate in the Artificial Intelligence Lab at the University of Arizona. Contact him at [yklin@email.arizona.edu](mailto:yklin@email.arizona.edu).

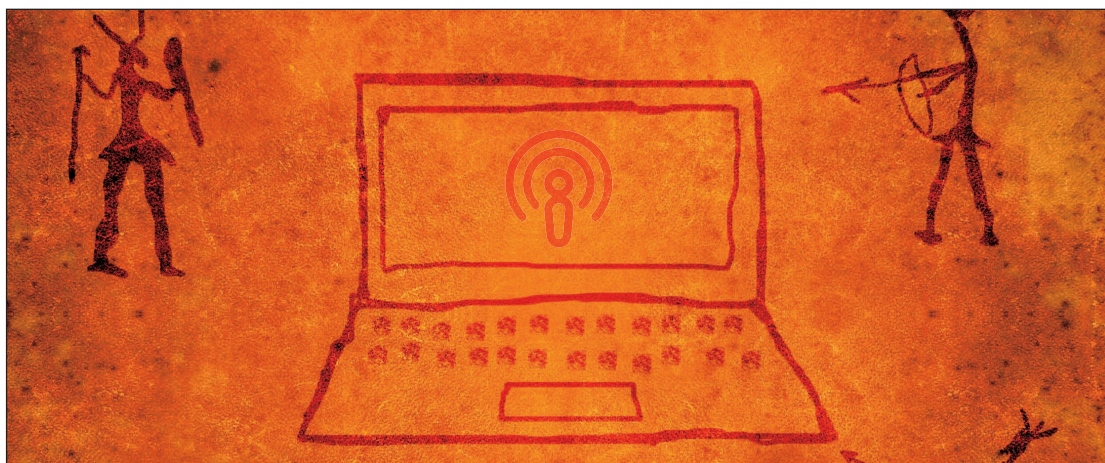
**Randall A. Brown** is an internal medicine physician in the University Medical Center at the University of Arizona.

**Hung Jen Yang** is the president and chief executive officer at the Min Sheng General Hospital, Taiwan.

**Shu-Hsing Li** is the dean of the College of Management and a professor in the Department of Accounting at the National Taiwan University.

**Hsin-Min Lu** is an assistant professor in the Department of Information Management at the National Taiwan University.

**Hsinchun Chen** is the director of the Artificial Intelligence Lab at the University of Arizona. Contact him at [hchen@eller.arizona.edu](mailto:hchen@eller.arizona.edu).



# COMPUTING LIVES

[www.computer.org/annals/computing-lives](http://www.computer.org/annals/computing-lives)

The "Computing Lives" podcast series of selected articles from the *IEEE Annals of the History of Computing* cover the breadth of computer history. This series features scholarly accounts by leading computer scientists and historians, as well as firsthand stories by computer pioneers.