Introduction to the Special Section on Clinical Data Mining

Shipeng Yu
Siemens Healthcare
51 Valley Stream Parkway
Malvern, PA 19355
shipeng.yu@siemens.com

Bharat Rao
Siemens Healthcare
51 Valley Stream Parkway
Malvern, PA 19355
bharat.rao@siemens.com

ABSTRACT

Mining clinical data is a fast-evolving field, ranging from mining patient data of a particular type (e.g., images, genomics) to mining the increased amount of mixed-format information (databases, free text, images, labs, etc) in electronic health records (EHR), to selecting, extracting and synthesizing relevant knowledge from large medical corpuses, to the promise of personalized medicine where therapy and prevention are tailored to smaller and smaller patient subpopulations, down to the individual patient. Clinical data mining can be a key asset in driving vast systemic improvements in healthcare, leading to improved patient outcomes and reduced healthcare costs. In this report we briefly survey the latest advancements in this field, and introduce four selected articles that cover both state-of-the-art data mining techniques for clinical data and discuss emerging clinical data mining applications.

1. THE HEALTH CRISIS

Healthcare is facing a crisis. Consider the following trends:

- **Explosion in electronic patient information**. The available information about a single patient and his or her disease is increasing tremendously: images of greater resolution, increasing number of IVD tests, new tumor markers, and soon, whole genome sequences for patients and tumors. Partly motivated by the incentives from the American Recovery and Reinvestment Act (ARRA) in the US, EHR systems are being adopted more widely, leading to the collection of vast stores of patient data.
- **Explosion in medical knowledge**. The amount of medical information (e.g., evidence-based knowledge) and published knowledge is also growing. Studies indicated a 2-fold increase in the number of published journal articles in medicine over the last decade (with 800,000 new articles being published in 2010).
- **The unprecedented number of therapeutic and diagnostic options**. The number of available therapies (drugs) and diagnostic tests is also growing rapidly. A recent trend is the development of therapies with companion tests (often genetic tests) that identify the subset of the population for whom the therapy is likely to be most effective.

Yet, despite all these amazing advances, we do not see corresponding improvement in health outcomes. This is due, primarily, to two reasons. First, there is the unimaginable *complexity of the science of modern healthcare*, as detailed above, which far outstrips the capability of the existing health systems in which clinicians are trained, deployed, and reimbursed. Second, the relentless escalation in the *cost of healthcare* is widely understood to be unsustainable and wasteful. Simply put, clinicians are overwhelmed by the deluge of data, knowledge and available options, and lack the tools to deliver the most effective and cost-efficient care for the patient at hand.

Data mining can (and will) play a key role in tackling both fundamental challenges: first by mining the growing corpuses of patient data, knowledge and therapeutic outcomes, and second by identifying inefficiencies and wasted efforts in the existing healthcare system. The acceleration of two fundamental technologies will provide a foundation for maximizing the power of clinical data mining, and change it from an challenging research exercise into a methodology for transforming healthcare. These technological advances are:

- **Computational power** has become cheap and easily available. It is now possible to mine the vast amount of electronic information captured in patient data, medical knowledge and therapeutic options.
- **Connectivity** will make it possible to do real-time learning from clinical data and deliver the results to clinicians and patients at the point where it can have most impact. Mobile networks, and smart healthcare apps for consumers will also play a growing role in improving healthcare.

Additionally, we will need to continue research in maintaining patient privacy, patient empowerment (social networks can play a big role here), balancing costs with outcomes (cost-sensitive learning), managing patient populations (this has different challenges and trade-offs compared to managing individual patients), and dealing with health reform and associated legislative and regulatory barriers.

Most promising of all, there is a growing realization in the scientific, clinical and legislative community of the imperative for clinical data mining, most strikingly illustrated in the recent report from the Institute of Medicine, *Best care at lower cost: The path to continuously learning health care in America* [1].

2. CONTRIBUTED ARTICLES

In this special section we selected four articles that cover a variety of advanced data mining tools and discuss diverse clinical data mining applications. Interested readers can explore the references in each of the articles to find out more about relevant fields and techniques.

- **Sparse Methods for Biomedical Data** by *Jieping Ye and Jun Liu.* This paper discusses state-of-the-art data mining methods on sparse learning and their applications for many biomedical problems. Sparse learning is motivated from the important observation that although the investigation of massive biomedical data with growing scale, diversity, and complexity has taken a center stage in modern data analysis, the underlying representations of many biomedical data are often sparse. For instance, for many diseases (such as leukemia) only a few genes are relevant to the specific disease even though humans have tens of thousands of genes; therefore, a gene network is sparse since a regulatory pathway involves only a small number of genes. The paper surveys the latest developments in the sparse learning community and presents several success stories for applications as diverse as biomarker selection, biological network construction, and magnetic resonance imaging.

- **Supervised Patient Similarity Measure of Heterogeneous Patient Records** by *Jimeng Sun, Fei Wang, Jianying Hu, and Shahram Edabollahi.* Determining if two patients are "similar" is a long-standing challenge in medicine; being able to do this well could support a variety of therapy selection and prevention approaches. This paper describes a suite of metrics or similarity learning approaches that tackle the problem of whether two patients are clinically similar or not, based on their key clinical indicators. The approach has wide applications in the context of patient cohort identification for comparative effectiveness studies and clinical decision support. The paper also discusses how to incorporate explicit feedback (i.e., which ones are similar and which ones are dissimilar) into the learning system, and how to integrate the individual distance metrics from each physician into a globally consistent unified metric. The authors also present a clinical decision support prototype system powered by these patient-similarity methods, and evaluate their proposed methods on patient EHR data against several baselines.

- **Mining Anatomical, Physiological and Pathological Information from Medical Images** by *Xiang Sean Zhou, et al.* Despite the promise of genomics, medical imaging is likely to remain as the primary diagnostic and therapy-monitoring tool for physicians for many important diseases for a long time. The paper surveys data mining techniques for information extraction from medical images, more specifically to mine the anatomical, physiological and pathological information. This is an important and fast growing field for imaging-based clinical data analysis. The paper discusses three aspects related to information mining in this domain: the target user groups, the information to mine, and technologies to enable mining. The authors also describe representative methods and algorithms that are effective for solving these mining problems. Many of these algorithms are state-of-the-art machine learning algorithms that are also shown to be applicable to other clinical domains.

- **Data Mining Methodologies for Pharmacovigilance** by *Mei Liu, Michael E. Matheny and Hua Xu.* This paper surveys data mining tools for pharmacovigilance, which is a research field that is concerned with detection, assessment, understanding and prevention of adverse effects related to drugs. Adverse drug events (ADEs) are a huge problem in the US healthcare system – around 770,000 people are injured or die in hospitals annually due to ADEs, which are estimated to cost between $2-5B every year. ADEs are also very important for the pharmaceutical companies that develop these drugs. The ultimate goal is to reduce adverse drug reactions (ADRs), which are defined as "any undesirable effect of a drug beyond its anticipated therapeutic effects occurring during clinical use". This paper describes both pre-clinical screening and post-market surveillance tools to identify ADRs, and also discusses future perspectives and challenges in this field.

3. CONCLUDING REMARKS

Two recent articles in the popular press describe clinical data mining applications at two ends of the spectrum, which have the potential for tremendous impact. The first example (the Time magazine, Aug 27) comes from Uganda, where the health system is under severe strain, particularly when it comes to dealing with treatable diseases like malaria (which accounts for a quarter of all deaths of children under age 5 in Uganda). Malaria can be treated with drugs, but often Ugandans line up in local clinics which have run out of drugs. The problem is not one of scarcity, but of distribution, with drugs in over-supply at one clinic but in short supply in another. Earlier this information was only available by paper, but UNICEF and WHO have tasked health workers to begin using text messages to centrally record information about shortages and epidemic outbreaks (mobile phones with text capabilities are plentiful in Uganda). This data is collected, collated into a dashboard, and used to move drugs quickly between clinics. Further, this dashboard is only as accurate as the information entered, and clinics may not be forthcoming about gaps in their supply. So, a secondary source of data is used – crowdsourcing. Consumers send information anonymously about service problems and outbreaks into the system which is analyzed to identify unreported gaps. Finally, UNICEF has created a social network group of 140,000 people who communicate entirely by mobile texts, to send and receive information about health, and can alert people to necessary health services (e.g., tell mothers about free vaccinations in their area). Although the data analytics being performed are rudimentary at best, the impact is large, and the potential to multiply the impact with better techniques is immense. Most impressively, all this is being done at the monthly cost of $14 per district.

The second example comes from the personal genomics company, 23andMe, which reports that it is seeking approval from the US Food and Drug Administration for its Personal Genome Service which enables individuals to explore their own DNA and provides 200 health and trait reports as well as genetic ancestry information. FDA approval would be a huge step for the nascent genetic testing industry, perhaps presaging a future previously limited to the minds of science fiction authors and scriptwriters where genetic testing is widely used to identify multiple aspects of health risk, treatment effectiveness and disease susceptibility.

The two examples discussed here are illustrative of the diversity of global health issues. In much of the developing world, the key problem is that of "access" – namely, access to what would be considered basic care in more developed parts of the world. In the developed nations (e.g., the G10), the problems are of escalating healthcare costs, health science complexity, and the reduction of unnecessary, ineffective and harmful medicine. In both these scenarios, the science and practice of mining health data will play an increasingly central role in tackling global health challenges.

4. ACKNOWLEDGEMENTS

We would like to thank all the authors who contributed to this special section.

5. REFERENCES

[1] Institute of Medicine. *Best Care at Lower Cost: The Path to Continuously Learning Health Care in America*. The National Academies Press, Washington, DC, 2012.