Electronic health records-driven phenotyping: challenges, recent advances, and perspectives

Jyotishman Pathak, Abel N Kho, Joshua C Denny

With the completion of the Human Genome Project as well as recent advances in genomic science and comparative biological studies, a new era of individualized medicine is evolving where novel biomedical discoveries are leading to more effective prevention, treatment, and diagnosis of disease. Although altered phenotypes are one of the most reliable manifestations of altered gene functions, research in extracting, representing, and analyzing phenotype-genotype relationships is still evolving. This has led to the emergence of a trans-discipline field, called ‘Phenomics’, that aims to capitalize on high-throughput computation and informatics technologies for the systematic study of phenotypes and how they might influence personal genomics. Many comparative phenomics studies in the recent past have demonstrated the power of positively correlating phenotypes with several measures of gene functions. However, despite the advances, research in phenomics is presented with various challenges, including (i) developing approaches for high-throughput extraction and representation of phenotypes, (ii) building techniques for storing, integrating, and querying phenotype data, and (iii) advancing phenotype-driven analysis to derive phenotype-genotype associations. A significant barrier in the discovery of new genetic variants is the requirement to obtain the large sample sizes needed for an effective study (since variants may be rare within a population) leading to time-consuming and onerous sample collection efforts.

Electronic health records (EHRs) can accelerate clinical research and genomic medicine, but are hindered by the limited number of validated processes and tools to enable accurate and rapid phenotype extraction. EHRs are increasing in ubiquity, functionality, and comprehensiveness across the USA, in part due to Meaningful Use standards implemented as part of the Health Information Technology for Economic and Clinical Health (HITECH) Act. One recent advance has been the coupling of DNA biorepositories to EHR data, combined with advances in informatics techniques, such as natural language processing (NLP), to enable genomic discoveries. The Electronic Medical Records and Genomics (eMERGE) network—a network of nine academic medical centers—has demonstrated the effectiveness of EHR-derived phenotyping algorithms for cohort identification to conduct genome- and phenotype-wide association studies. Algorithms to define phenotypes in eMERGE have typically followed an iterative path, and highlight the importance of intermittent chart review to validate phenotype accuracy, typically at more than one site (when a multisite implementation is planned). Once finalized, eMERGE phenotype algorithms (as well as non-eMERGE algorithms) can be viewed at http://phekb.org.

While eMERGE presents an exciting and encouraging demonstration of secondary use of EHR data, evaluating the strengths and limitations for EHRs has important implications for clinical and translational research, including clinical trials, observational cohorts, outcomes research, and comparative effectiveness research. These issues are further amplified within the realm of a learning healthcare system that emphasizes the ability to have real-time access to knowledge, digital capture of care experience, engaged and empowered patients, alignment of incentives to value, and a leadership-instilled culture of learning (figure 1). Hripcsak and Albers highlight several such challenges in leveraging EHRs for research, including data that is often incomplete, inaccurate, highly complex, and biased, and propose a combination of top-down knowledge engineering and bottom-up learning and analysis to address these issues (figure 2). Specifically, understanding the inherent complexity involved in a healthcare process model is critically important towards achieving a scalable and high-throughput phenotyping process.

Several have acknowledged that there is a recognized tension between facilitating data entry by allowing narrative text in EHRs, and utilizing these data for research. A key component for identifying patient cohorts in the EHR is to define inclusion and exclusion criteria that algorithmically select sets of patients based on stored clinical data. This process allows the definition of phenotypes over structured data (eg, demographics, diagnoses, medications, and laboratory measurements) as well as unstructured clinical text (eg, radiology reports, encounter notes, and discharge summaries). In general, this process can be quite complex, involving heuristics encoded as rules or machine learning algorithms. Several NLP techniques have been developed specifically for clinical text, and include solutions to address concept extraction, coreference resolution, word sense disambiguation, and temporal relations, to name just a few. The creation of annotated corpora to help develop and test these algorithms has also been the focus of the biomedical and clinical NLP community.

A notable challenge for general-purpose NLP systems applied to phenotyping is the requirement for high precision across many records. Thus, another important effort in phenotyping is purpose-developed NLP solutions designed to extract specific features over entire records with high accuracy. In addition, researchers have generated focused NLP and machine learning systems for common tasks such as general medical extraction and smoking determination.

Several efforts have demonstrated the portability of phenotype algorithms across sites and EHR platforms, although implementation still requires significant manual effort. Efforts such as populHieN9 and SHRINE70 are attempting to provide common platforms to enable federated querying, although currently using simpler methodologies than invoked through networks such as eMERGE. The National Quality Forum’s Quality Data Model may also represent one such standard. As in any other process, use of EHR phenotypes across institutions and across applications will not be complete without standardization. While some standards that are specific for research applications are continually being refined, existing standards that were designed for clinical applications also have high relevance in EHR phenotyping.
SUMMARY OF PAPERS IN THE SPECIAL ISSUE
The focus of this special issue of the journal is to provide a forum for presenting methodologies, tools, and algorithms to enable high-throughput phenotype extraction from EHR data. The journal has traditionally published articles related to EHR use for clinical decision support associated with patient safety and screening, as well as therapy management, but the extension of EHR use for research is relatively recent, reflecting the growth of clinical research informatics.

Judging by the response, interest in EHR-based phenotyping is strong: of over 60 papers submitted for consideration in this issue, 20 articles were accepted for publication. These articles were evaluated through a rigorous process involving three guest associate editors (JP, JD, AK) in addition to the editor and one additional associate editor who dealt with any submissions a guest associate editor could not handle due to conflicts of interest.

Phenotyping algorithm development, implementation, and ecology emerged as major themes within the accepted manuscripts, along with applications in clinical trials and clinical decision support. The first article by Overby et al presents a collaborative approach for developing an EHR-driven phenotyping algorithm for drug-induced liver injury. This work, done within the eMERGE consortium, highlights the challenges in algorithm development and portability across multiple EHR systems, and emphasizes the need for robust validation methods complementing local, institution-specific algorithm implementation processes. The article by Tian et al presents a similar approach for identifying patients with chronic pain at a multi-site community...
health center. Here the authors demonstrate that by combining multiple different classes of data—diagnosis, medications, and pain scores—one can achieve much higher performance in terms of sensitivity and specificity. Finally, the article by Ludvigsson et al.101 also argues that a combination of structured data with information extracted from unstructured text via NLP—in this case, for detecting patients with celiac disease—achieves a much higher performance.

With respect to applying more advanced machine learning, text mining, and statistical approaches for phenotype extraction, Chen et al.96 demonstrated that active learning was useful in the identification of patient cohorts for rheumatoid arthritis, colorectal cancer, and venous thromboembolism. Not only did active learning methods outperform passive learning techniques, but the authors concluded that machine learning and feature engineering principles can be combined to develop efficient and generalizable phenotyping algorithms on a larger scale. The article by Klann et al.97 explored Bayesian structured learning methods for population phenotyping to prioritize and tailor pediatric preventative care services. This work demonstrates how population phenotyping models can be built automatically, using prior data, without any human intervention for prioritizing pediatric screening questions and reminders in a patient-tailored manner. Similarly, the article by Deleger et al. illustrates the application of conditional random fields to risk stratify abdominal pain patients based on pediatric appendicitis scores. The information driven approach demonstrated performance comparable to physician chart reviews, and represents a promising new approach for future computerized decision support applications. On the application of statistical based approaches, Lyalina et al.102 applied automated text processing pipelines to annotate clinical notes with Unified Medical Language System concepts, and used dimensionality reduction to study patient-level phenotypic variations for neuro-psychiatric disorders. The authors argue that such methods enable large-scale cohort building for clinical and genomic studies. Similarly, Gundlapalli et al. demonstrated the high-throughput annotation process for extracting psychosocial concepts from approximately 1 billion documents from the medical facilities of the Department of Veterans Affairs. This work further illustrates the need to leverage high-yield documents and clinical notes, as opposed to the entire unstructured text corpus within an EHR system. Finally, in a similar effort, the article by Davis et al. demonstrates the applicability of using NLP methods and scalable annotations to identify patients with multiple sclerosis and the key clinical traits for the disease course.

A key aspect for portability of phenotyping algorithms is standardization. The article by Richesson et al.105 highlights this need by comparing seven different algorithms to identify diabetes mellitus. These algorithms yielded different cohorts when applied to the same population within the Duke University Health System. Similar results were observed in the article by Fan et al. wherein an algorithm for identifying patients with peripheral arterial disease performed differently depending on the medical specialties involved. Not only do such differences lead to different interpretations of results and data, but multiple phenotype definitions can potentially affect their application within a healthcare organization. The article by Pathak et al. attempts to address this challenge by leveraging standardized information modeling and Meaningful Use standards for representation of phenotyping algorithms. This work has led to the development of a publicly accessible library for standardized phenotype definitions. In addition to the standardized definitions, there is also the need for uniform infrastructure to enable cohort identification. The article by Fernandez-Breis et al. proposes the use of standardized EHR models, archetypes, and ontologies for colorectal cancer screening. The authors argue that emerging semantic web technologies can facilitate the much needed interoperability among EHR data and systems. Finally, the article by Bache et al. proposes an architecture for identifying patient cohorts by explicit query modeling and support for temporal reasoning. The authors illustrate that such an approach, while initially burdensome to establish (eg, compared to direct SQL-based queries), is more scalable and adaptable to heterogeneous data sources.

Understanding the inherent complexities associated with the healthcare process is vital for achieving scalable and high-throughput phenotyping. Along these lines, the article by Hripcsak and Albers studies correlation between EHR variables and healthcare process events. The authors illustrate that variable groups represent not only clinical and physiological properties, but also characteristics related to the way the information is gathered and recorded during the healthcare process. Similarly, the article by Boland et al. introduces the concept of a ‘verotype’ (the Latin word ‘verus’ means ‘true’ or ‘actual’) to represent the ‘true’ population of similar patients for treatment purposes through the integration of genotype, phenotype, and disease subtype (eg, specific glucose value pattern in patients with diabetes) information. Both these works have implications for how phenotype extraction methods are used for real-world applications. The article by Richesson et al.104 highlights these challenges within the context of clinical trials for the NIH Health Care Systems Collaboratory initiative. Of equal importance is the ability not only to implement and execute phenotyping algorithms, but also to understand the associated healthcare process events to achieve meaningful phenotype extraction. This issue is specifically highlighted in the article by London et al.106 whereby the authors use a research data mart—powered by i2b2—for identifying patients eligible for clinical trials. In a similar effort, Warner et al.107 describe a temporal phenotype analysis to create a visual analysis of phenomic associations and healthcare process events. This work presents a new methodology for visual analytics and testable hypothesis generation from EHR data, which reveals patterns of context-specific complications with clinical implications.

Finally, the article by Rosenbloom et al. discusses the ethical and legal implications of opt-out biobanks, such as Vanderbilt’s DNA biobank, BioVU. Opt-out models have the advantage of rapid and inexpensive data collection; however, the characteristics of patients who opt-out versus those included in the biobank is not currently known, nor are the reasons why patients opt out. The importance of such research is highlighted by a recent Advance Notice of Proposed Rule-Making announced by the Department of Health and Human Services.110

**NEXT STEPS AND THE FUTURE**

The articles presented in this issue provide a glimpse into the opportunities that lie ahead. The rapid proliferation of EHRs, and the ability to connect and integrate data across multiple EHR systems through the robust application of data standards, will create ever richer clinical data for high-throughput phenotyping algorithms. These same algorithms can and are being adapted for direct clinical applications, such as automation of clinical quality measures111 and of clinical documentation. Across large populations, EHR data have potential for novel
discovery of associations between disease and genetic, environmental, or process measures. Increasingly, electronic data will be available from what have been considered non-clinical sources, such as patient behavior/activity (eg, Fitbit) or social networks, and these can be combined with EHR-derived data to create more comprehensive ecological views of patients. These opportunities will naturally uncover issues and challenges around integration, analysis, interpretation, and sharing of ‘big data.’ It is hoped that this issue of the journal will serve as a useful reference and guide over the next few years. The technologies presented here will mature and evolve towards scalable and high-throughput integrative phenotyping that is needed to facilitate research, patient care, and healthcare management.

Funding This work was supported in part by National Institutes of Health grant R01-GM105688, eMERGE Network grants (U01-HG006379, U01-HG003388, U01-HG003678), R01-LM010685, and a Pharmacogenomics Research Network grant (U19-GM013888).

Competing interests None.

Provenance and peer review Commissioned; internally peer reviewed.

To cite Pathak J, Kho AN, Denny JC. J Am Med Inform Assoc 2013;20:e206–e211.

Accepted 18 October 2013

J Am Med Inform Assoc 2013;20:e206–e211. doi:10.1136/amiajnl-2013-002428

REFERENCES


