Artificial Intelligence in Medicine 71 (2016) 57–61

Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

journal homepage: www.elsevier.com/locate/aiim

Clinical phenotyping in selected national networks: demonstrating the need for high-throughput, portable, and computational methods

Rachel L. Richesson[a,*], Jimeng Sun[b], Jyotishman Pathak[c,1], Abel N. Kho[d], Joshua C. Denny[e]

[a] Duke University School of Nursing, 311 Trent Drive, Durham, NC 27710 USA
[b] School of Computational Science and Engineering, Georgia Institute of Technology, 266 Ferst Drive, Atlanta, GA 30313, USA
[c] Department of Health Sciences Research, 200 1st Street SW, Mayo Clinic, Rochester, MN, 55905, USA
[d] Departments of Medicine and Preventive Medicine, Northwestern University, 633 N St. Clair St. 20th floor. Chicago IL 60611, USA
[e] Departments of Biomedical Informatics and Medicine, Vanderbilt University, 2525 West End Ave, Suite 672, Nashville, TN 37203, USA

ARTICLE INFO

Article history:
Received 22 February 2016
Accepted 30 May 2016

Keywords:
Machine learning
Clinical phenotyping
Electronic health records
Networked research
Precision medicine

ABSTRACT

Objective: The combination of phenomic data from electronic health records (EHR) and clinical data repositories with dense biological data has enabled genomic and pharmacogenomic discovery, a first step toward precision medicine. Computational methods for the identification of clinical phenotypes from EHR data will advance our understanding of disease risk and drug response, and support the practice of precision medicine on a national scale.
Methods: Based on our experience within three national research networks, we summarize the broad approaches to clinical phenotyping and highlight the important role of these networks in the progression of high-throughput phenotyping and precision medicine. We provide supporting literature in the form of a non-systematic review.
Results: The practice of clinical phenotyping is evolving to meet the growing demand for scalable, portable, and data driven methods and tools. The resources required for traditional phenotyping algorithms from expert defined rules are significant. In contrast, machine learning approaches that rely on data patterns will require fewer clinical domain experts and resources.
Conclusions: Machine learning approaches that generate phenotype definitions from patient features and clinical profiles will result in truly computational phenotypes, derived from data rather than experts. Research networks and phenotype developers should cooperate to develop methods, collaboration platforms, and data standards that will enable computational phenotyping and truly modernize biomedical research and precision medicine.

1. Introduction

The national Precision Medicine Initiative aims to enroll one million members in a national cohort that will integrate data from biospecimens, sensor and mobile technologies, and healthcare, largely from electronic health record (EHR) data, to advance biomedical discovery and improve health [1]. The realization of this vision will require efficient and effective methods to convert data from EHRs into specific and reliable phenotype characterizations that can be used to predict an individual's risk of disease or response to drug therapy.

Phenotypes are the measurable biological, behavioral and clinical markers of a condition or disease. The process of deriving research-grade phenotypes from clinical data using computer-executable algorithms is called *computational phenotyping* (phenotyping for short) [2]. Phenotyping includes a range of approaches from finding a phenotype using expert-derived rules and those phenotypes emerging from novel computational methods that potentially represent new clinical entities. The widespread adoption of EHRs will increase the reliance on phenotyping for a number of activities, including genomic studies of disease and drug response, clinical predictive modeling, pragmatic clinical trials, and healthcare quality measurement. Current methods face bottlenecks for development, implementation, sharability, and the ability to

* Corresponding author at: Duke University School of Nursing, 2007 Pearson Bldg, 311 Trent Drive, Durham, NC, 27710, USA.
E-mail addresses: rachel.richesson@duke.edu (R.L. Richesson), jsun@cc.gatech.edu (J. Sun), pathak@med.cornell.edu (J. Pathak), a-kho@northwestern.edu (A.N. Kho), josh.denny@Vanderbilt.Edu (J.C. Denny).
[1] Present Address: Division of Health Informatics, Weill Cornell Medical College, 425 East 61st Street, New York City, NY 10065, USA.

http://dx.doi.org/10.1016/j.artmed.2016.05.005
0933-3657/

derive novel, not-foreseen findings. We provide a survey of the approaches to computational phenotyping and challenges experienced by several national research networks with which we are affiliated, and provide supporting literature in the form of a non-systematic review. The aim of this paper is to provide a summary of the approaches and tools that clinical research networks are using to realize the scale of high-throughput computational phenotyping. Based on the common challenges faced by these networks, we suggest cultural change and resources that will be needed to support computational phenotyping on a grand scale and advance data-driven precision medicine research.

2. National networks and phenotyping activity

A number of national research and public health surveillance networks have leveraged data from EHRs for defining conditions and risk. The Electronic Medical Records & Genomics (eMERGE) Network, formed in 2007 and arguably the pioneer of computational phenotyping, has investigated more than 40 phenotypes for genomic studies using algorithms that combine billing codes, medication data, laboratory and test results, and natural language processing of clinical notes [3,4]. Sites from the Pharmacogenomics Research Network (PGRN) have used EHR data to identify genetic predictors of drug-response phenotypes across multiple sites [5–7]. The Mini-Sentinel surveillance initiative, funded by the U.S. Food and Drug Administration, uses phenotype algorithms to define conditions from administrative data from 18 national health plans to identify adverse drug outcomes [8–15]. In addition, provider networks use computational phenotyping to identify patients with particular conditions for health services or population-level research. These include the Health Care Systems Research Network, formerly known as the HMO Research Network, and the Observational Medical Outcomes Partnership (OMOP) [16], now part of the Observational Health Data Sciences and Informatics (OHDSI) collaborative [17].

A number of disease-specific research networks and multi-site registries have developed and validated EHR-based phenotype definitions for specific conditions [18,19]. In the National Institutes of Health's Health Care Systems Research Collaboratory, a number of multi-site pragmatic clinical trial demonstration projects are using computable phenotypes for cohort identification, development of interventions, and study outcomes [20–22] More recently, the Patient Centered Outcomes Research Institute funded the National Patient-Centered Clinical Research Network (PCORnet) to conduct comparative effectiveness studies across 13 Clinical Data Research Networks and 21 Patient Powered Research Networks [23]. Partnering institutions are expected to support up to 200 queries in the next few years, signifying the imminent need for high-throughput and reproducible phenotyping methods.

Although the aforementioned research networks have unique objectives and constraints, they share common challenges related to the use of clinical data for research, including heterogeneous EHR systems, a lack of standardized data, concerns about data completeness and inherent biases, and variation in medical diagnosis, procedures, treatments, and data documentation across providers, organizations, and regions. In response, several networks have published methodological guides for data quality assurance [24–29].

3. Evolution of phenotyping methods

Research networks by their very nature require scalable approaches that can be implemented quickly with reproducible performance characteristics in multiple settings and information systems. There are several broad classes of methods to computational phenotyping that are continuously improving.

The use of *expert-defined rules* is most widely adopted method for phenotyping, and this approach was used for the early phenotypes developed from the eMERGE network, such as type 2 diabetes [30] and cataracts [31]. This approach begins with the manual development of an algorithm – often using Boolean logic, scoring thresholds, or a decision tree – based on domain expertise. The logic is then iteratively enhanced through validation and chart review on EHR data. Advantages of this approach are that it yields human-interpretable algorithms, which can be portable to other sites [32], and the number of charts needed to review to train/validate an algorithm can be lower. However, the effort and time for developing the algorithms can be significant, requiring clinical and informatics knowledge, and this approach cannot be used to identify phenotypes not first envisioned by a researcher.

Machine learning methods rely on data patterns to develop the phenotype definitions, and can reduce the effort required from clinical domain experts. *Supervised learning* aims to construct classifiers to differentiate cases (positive for the phenotype) and controls (negative for the phenotype). The high level steps involve (1) characterizing patients as feature vectors, (2) determining the class label (case vs. control) for each patient, (3) building and optimizing the classifier. Typically the number of charts reviewed is higher than required for rule-based algorithms, a time-consuming task requiring domain experts. Chen et al. explored active learning as a more efficient labeling process, demonstrating reduction in the number of cases needed [33]. However, machine learning classification models can be difficult to interpret, require significant training data, and may not transfer well to other sites, as a model may learn features that are unique to an institution (e.g., physician name, local note type, or clinical unit). Yu et al. extracted clinical features from publicly-available knowledge sources to develop more "interpretable" machine learning algorithms that performed as well as or better than expert-derived algorithms [34].

Unsupervised learning provides approaches to cluster EHR data into patient groups corresponding to phenotypes or subtypes. Unsupervised learning does not require expert labels, which tremendously reduces the time needed for manual chart review. However, the validation of the resulting phenotypic groups is challenging, as no clear ground truth on those groups are given. While these methods require very large volumes of training data, they do not carry costs of manually labeling individuals as cases or controls. Various tensor factorization methods have been developed for unsupervised phenotyping [35–37]. Deep learning is another approach which has successfully identified patterns in clinical data representing distinct phenotypes [38].

Because important relevant clinical data is included in narrative clinical notes rather than structured data elements or standardized coding systems, natural language processing methods can be used to extract phenotypes from clinical notes [39,40] and to process data for more advanced machine learning techniques. Phenotype definitions including general purpose natural language processing (NLP) tools [41–43] have accelerated the widespread use of NLP, which is an important component of some complex phenotypes [44].

4. Toward a future of higher throughput phenotyping

The planned Precision Medicine Initiative study will require higher-throughput, more easily shared computational approaches than have been demonstrated to date. Scalable precision medicine will require clinical phenotypes that can be rapidly developed, executed in high volume, and easily adapted to new sites with high algorithm reliability (Fig. 1).

The vision of rapid, portable phenotyping implies that multiple providers and applications can reuse computational methods and definitional logic, enhanced by accessible repositories for phe-