AMIA Jt Summits Transl Sci Proc, 2018. TBI: p. 261-262.

Towards Patient-Driven Phenotyping and Similarity for Precision Medicine

Tiffany J. Callahan, MPH[1], Olivier Bodenreider, MD, PhD[2], Michael G. Kahn, MD, PhD[3]

[1]Computational Bioscience Program, University of Colorado Denver Anschutz Medical Campus, Aurora, CO; [2]Lister Hill National Center for Biomedical Communications, National Library of Medicine, National Institutes of Health, Bethesda, MD; [3]Department of Pediatrics, University of Colorado Denver Anschutz Medical Campus, Aurora CO

Abstract

Clinical phenotyping provides important insight into the manifestation and outcome of rare and complex diseases. Traditional phenotyping techniques often require multiple iterations of refinement with a domain expert, lack interoperability, and have limited reproducibility. In comparison, patient similarity-based techniques derive personalized patient risk models that are highly accurate, even when applied to sparse data or poorly characterized diseases/outcomes. We introduce a novel, semi-supervised data-driven method for applying clinical similarity to pediatric phenotyping.

Introduction

Clinical phenotyping is a technique designed to provide clinicians with important insight into the development, progression, and outcome of complex diseases for a population of patients. Many clinical phenotyping techniques have been developed[1] including rule-based,[2,3] text-mining or natural language processing based,[4,5] and statistical learning-based (i.e., machine and deep learning).[6,7] While there is a large body of research supporting their success, these techniques require multiple iterations of refinement with a domain expert, often lack generalizability and interoperability, and have limited reproducibility. Compared to traditional phenotyping approaches, patient similarity-based techniques aim to derive personalized patient-level risk models. When compared to other methods, patient similarity-based approaches have shown to be more accurate,[8] even when applied to sparse data or poorly characterized diseases/phenotypes.[9] An example of expert-driven supervised patient-similarity-based methods is Longhurst and Shah's "Green Button".[10] While this approach, and others like it, are scalable and accurate, they still suffer from poor handling of missing data, lack both internal and external validation, and maintain reliance on domain expertise.[11] The current project aims to address the limitations of existing phenotyping and patient similarity-based methods by developing a novel, semi-supervised data-driven method to measure patient-level clinical similarity

Methods

The composite clinical similarity algorithm was specifically developed to leverage the Observational Medical Outcomes Partnership (OMOP) common data model (CDM)[12] in order to take advantage of pre-normalized data standardized to a specific set of clinical terminologies. Our algorithm combines existing pairwise[13] and groupwise[14] semantic similarity measures, in a novel way, to identify groups of clinically similar patients. First, pairwise similarity scores are calculated for each clinical attribute by incorporating hierarchical relations from standard clinical terminologies (e.g., LOINC and SNOMED CT). Pairwise similarity scores are also calculated for demographic attributes, accounting for binary (e.g., gender), categorical (e.g., race), and continuous (e.g., age) variables. Groupwise similarity measures are then used to compare sets of clinical and demographic attributes among patients. The final composite clinical similarity score, where scores range from 0.0 (completely dissimilar) to 1.0 (perfect similarity), between two patients is calculated as a weighted average of the individual demographic and clinical groupwise similarities. While individual attribute weights can be learned, or user-generated, no differential weighting was applied in this experiment.

A proof-of-concept demonstration of the composite clinical similarity algorithm was performed using de-identified Children's Hospital of Colorado (CHCO) data. CHCO data conforms to the structure defined by the PEDSnet data network, which is an adaptation of the OMOP CDM version 5.0.[12,15] From the condition occurrence, drug exposure, measurement, observation, and procedures tables, we retrieved demographic and clinical data and constructed two distinct groups of 10 patients having the highest counts of cystic fibrosis (CF; SNOMED CT 190905008) and Huntington's Chorea (HC; SNOMED CT 58756001) encounter-diagnoses. To ensure an unbiased assessment of the method, all SNOMED CT codes for CF and HC used to define the two groups were excluded. Agglomerative hierarchical clustering with complete linkage and Euclidean distance were used to generate clusters of similar patients

in the expectation that the two groups of patients would separate into distinct clusters. This project was approved by the Colorado Multiple Institutional Review Board (15-0445).

Results

Patients were predominately white (90%) and female (60%) with a median age of 19. Hierarchical clustering resulted in four groups of clinically similar patients with scores ranging from 0.36 to 1.0. Clusters 1 (n=3) and 2 (n=2) only contained HC patients. On average, Cluster 1 HC patients were younger (17 vs. 26 years) had more frequent Parkinson's disease (16.4%) and dystonia (12.6%) encounter-diagnoses than Cluster 2 patients. Cluster 3 (n=9) only contained CF patients. Headache (7.5%) and anxiety disorder (6.8%) were the most frequent encounter-diagnoses. Pressurized or nonpressurized inhalation treatment for acute airway obstruction (18.1%) and manipulation of chest wall to facilitate lung function (10.0%) were the most frequent encounter-procedures. The final cluster (n=6) contained 5 HC patients and 1 CF patient. These patients were most frequently diagnosed with post inflammatory pulmonary fibrosis (6.5%) and hypoxemia (5.4%). Their most frequent encounter-procedures were noninvasive ear/pulse oximetry for oxygen saturation (8.7%) and pressurized or nonpressurized inhalation treatment for acute airway obstruction (7.5%).

Discussion

We are currently developing a novel semi-supervised data-driven method to measure patient-level clinical similarity and provided an initial proof-of-concept using a sample of pediatric patients. Preliminary results highlight the ability of our approach to successfully identify clinically distinguishable groups and sub-groups of similar patients, in the absence of the patient's primary encounter-diagnoses. Future work is underway to address current limitations including: conducting a more robust and comprehensive evaluation, accounting for changes in clinical variables over time, and learning of variable weights.

References

1. Shivade C, Raghavan P, Fosler-Lussier E, et al. A Review of Approaches to Identifying Patient Phenotype Cohorts using Electronic Health Records. *J Am Med Inform Assoc*. 2014;21:221–30.
2. Agarwal V, Podchiyska T, Banda JM, et al. Learning Statistical Models of Phenotypes using Noisy Labeled Training Data. *J Am Med Inform Assoc*. 2016;23:1166–73.
3. Li D, Simon G, Chute CG, Pathak J. Using Association Rule Mining for Phenotype Extraction from Electronic Health Records. In: *AMIA Jt Summits Transl Sci Proc*. 2013;142–6.
4. Jensen K, Soguero-Ruiz C, Mikalsen KO, et al. Analysis of Free Text in Electronic Health Records for Identification of Cancer Patient Trajectories. *Sci Rep*. 2017;7:46226.
5. Collier N, Groza T, Smedley D, et al. PhenoMiner: from Text to a Database of Phenotypes Associated with OMIM diseases. *J Biol Databases Curation*. 2015;27.
6. Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep Learning for Healthcare: Review, Opportunities and Challenges. *Brief Bioinform*. 2017.
7. Chiu P-H, Hripcsak G. EHR-Based Phenotyping: Bulk Learning and Evaluation. *J Biomed Inform*. 2017;70:35–51.
8. Ng K, Sun J, Hu J, Wang F. Personalized Predictive Modeling and Risk Factor Identification using Patient Similarity. *AMIA Summits Transl Sci Pro*c. 2015;132–6.
9. Beaulieu-Jones BK, Greene CS, et al. Semi-Supervised Learning of the Electronic Health Record for Phenotype Stratification. *J Biomed Inform*. 2016;64:168–78.
10. Longhurst CA, Harrington RA, Shah NH. A "Green Button" for using Aggregate Patient Data at the Point of Care. *Health Aff*. 2014;33:1229–35.
11. Sharafoddini A, Dubin JA, Lee J. Patient Similarity in Prediction Models Based on Health Data: A Scoping Review. *JMIR Med Inform*. 2017;5:e7.
12. Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a Common Data Model for Active Safety Surveillance Research. *J Am Med Inform Assoc*. 2012;19:54–60.
13. Batet M, Sánchez D, Valls A. An Ontology-Based Measure to Compute Semantic Similarity in Biomedicine. *J Biomed Inform*. 2011;44:118–25.
14. Azuaje F, Wang H, Bodenreider O. Ontology-Driven Similarity Approaches to Supporting Gene Functional Assessment. In: *ISMB SIG on Bio-ontologies*. 2005:9–10.
15. Forrest CB, Margolis PA, Bailey LC, et al. PEDSnet: A National Pediatric Learning Health System. *J Am Med Inform Assoc*. 2014;21:602–6.