Towards Patient-Driven Phenotyping and Similarity for Precision Medicine

Tiffany J. Callahan, MPH1, Olivier Bodenreider, MD, PhD2, Michael G. Kahn, MD, PhD3

1Computational Bioscience Program, University of Colorado Denver Anschutz Medical Campus, Aurora, CO; 2Lister Hill National Center for Biomedical Communications, National Library of Medicine, National Institutes of Health, Bethesda, MD; 3Department 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 present preliminary results from a novel, unsupervised data-driven method for applying patient similarity to pediatric phenotyping.

Introduction

Clinical phenotyping, or the classification of patients with or without an outcome or disease, is a technique designed to provide clinicians with important insight into the development, progression, and outcome of complex diseases for a population of patients. While there is a large body of research supporting the successful implementation of existing phenotyping methods, these techniques often require multiple iterations of refinement, lack generalizability and interoperability, and have limited reproducibility.1–3 Compared to traditional phenotyping approaches, patient similarity-based techniques aim to derive personalized risk models for each patient. When compared to other methods, patient similarity-based approaches have shown to be more accurate,4 even when applied to sparse data or poorly characterized diseases/phenotypes.5 An example of clinician driven supervised patient-similarity-based methods is Longhurst and Shah’s “Green Button”.6 While this approach, and others like it, are scalable and accurate, they still suffer from poor handling of missing data, lack robust internal and external validation, and maintain reliance on domain expertise.7 The current project aims to address some of the limitations of traditional phenotyping and existing expert-driven patient similarity-based methods by developing a novel, unsupervised data-driven algorithm for patient-level phenotyping and similarity.

Methods

A composite patient similarity algorithm was designed specifically for use with the Observational Medical Outcomes Partnership (OMOP) common data model (CDM).8 By developing our method specifically for this CDM, we can take advantage of pre-normalized data standardized to a specific set of clinical terminologies and can provide a tool that can be readily adopted by members of the worldwide OMOP community. Our composite patient similarity algorithm leverages existing pairwise9 and groupwise10 semantic similarity measures. Pairwise similarity scores were calculated for demographic attributes, accounting for binary (e.g., gender), categorical (e.g., race), and continuous (e.g., age) variables. Pairwise similarity scores were calculated for clinical attributes by incorporating hierarchical relations from standard clinical terminologies (e.g., LOINC, RxNorm, and SNOMED CT). Furthermore, we used groupwise similarity measures to compare sets of codes among patients. The final composite 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 patient similarity algorithm was performed using de-identified Children’s Hospital of Colorado (CHCO). CHCO data conforms to the structure defined by the PEDSnet, which is an adaptation of the OMOP CDM version 5.0.5,11 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. Results were described and interpreted using dendrograms and heat maps. 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 semantically similar patients with scores ranging from 0.36 to 1.0 (Figure 1, https://tinyurl.com/y6uzlx6u). The red (n=3) and blue (n=2) clusters only contained HC patients. On average, HC red cluster patients were younger (17 vs. 26 years) than blue cluster HC patients. Red cluster patients were distinguished by more frequent Parkinson’s disease (16.39%), dystonia (12.61%), and failure to thrive (7.56%) encounter-diagnoses. There were no occurrences of these diagnosis codes among any of the blue cluster HC patients. Further, medical nutrition therapy, which occurred in only two encounters in blue cluster HC patients, was the only frequently co-occurring red and blue cluster HC patient encounter-procedure. The white cluster (n=9) only contained CF patients. Headache (7.50%), anxiety disorder (6.80%), and asthma (5.81%) were the most frequent encounter-diagnoses. Pressurized or nonpressurized inhalation treatment for acute airway obstruction (18.05%), manipulation of chest wall to facilitate lung function (9.99%), and demonstration/evaluation of patient utilization of an aerosol generator (7.68%) were the most frequent encounter-procedures. The final magenta cluster (n=6) contained 5 HC patients and 1 CF patient. These patients were most frequently diagnosed with post inflammatory pulmonary fibrosis (6.51%), hypoxemia (5.41%), and congenital iodine deficiency (4.31%). Their most frequent encounter-procedures were noninvasive ear/pulse oximetry for oxygen saturation (8.68%), pressurized or nonpressurized inhalation treatment for acute airway obstruction (7.47%), and collection of venous blood by venipuncture (5.45%). A detailed description and comparison of the patient’s clinical attributes, by cluster, is provided in Figures 1-2 (https://tinyurl.com/y6uzlx6u).

Discussion

We are currently developing a novel unsupervised data-driven method to measure patient 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 diagnoses. Future work is underway to address current limitations including: conducting a more comprehensive evaluation, accounting for changes in clinical variables over time, and learning of variable weights.

References