Towards Automating Computational Phenotyping: Exploring the Trade-offs of Different Vocabulary Mapping Strategies

Tiffany J. Callahan, MPH[1], Jordan Wyrwa, DO[2], Katy E. Trinkley, PharmD[3], Lawrence E. Hunter, PhD[1], Michael G. Kahn, MD, PhD[4], Tellen D. Bennett, MD, MS[1,4]
[1]Computational Bioscience Program, University of Colorado Denver Anschutz Medical Campus, Aurora, CO; [2]Department of Physical Medicine and Rehabilitation, School of Medicine, University of Colorado Denver Anschutz Medical Campus, Aurora, CO; [3]Department of Clinical Pharmacy, University of Colorado Skaggs School of Pharmacy and Pharmaceutical Sciences, Aurora, CO; [4]Department of Pediatrics, University of Colorado Denver Anschutz Medical Campus, Aurora, CO;

Introduction

The universal adoption of electronic health records (EHRs) presents an unprecedented opportunity to fuel population-scale development of research-grade computational phenotypes (CPs). CPs can enable large-scale biomedical research and ultimately improve therapeutic decision-making and fuel mechanistic insight[1,2]. However, several barriers to the development, validation, and implementation of CPs must be overcome before their potential can be fully realized.

Phenotype knowledgebases like eMERGE′s PheKB are rich repositories of domain expert-derived CPs. Unfortunately, most of the CPs cannot easily be implemented across different EHR systems because they are tailored to specific source vocabularies (SV). Common data models (CDM) provide a practical solution to this problem by enabling the harmonization of multiple SVs to a smaller set of pre-aligned standard terminologies (ST). However, even with robust CDMs like OMOP or i2b2, one could employ different strategies to align the clinical codes (e.g. ICD-9-CM:314.0, LOINC:14288-5) provided in a CP definition to a CDM (e.g. exact string- or manual-mapping, similarity algorithm-derived). Understanding the trade-offs of these different vocabulary mapping strategies is a vital next step towards enabling CDM-driven CP automation.

Recent work by Hripcsak et al.[3] provides one of the first robust examinations of the effects of different vocabulary mapping strategies on patient cohort creation. They translated the diagnosis codes of nine PheKB CPs from ICD-9-CM to SNOMED CT, OMOP′s ST for diagnoses. They demonstrated that for most phenotypes, little information was lost and error rates varied by the mapping approach, when mapping from the SV to the OMOP ST. In some cases, information was gained (i.e. using the SNOMED CT hierarchy enabled the inclusion of additional relevant diagnosis codes). This work had important limitations: (1) mapping only a single SV to an OMOP ST; (2) examining only diagnoses or condition codes; and (3) creating patient cohorts using only the presence of at least one diagnosis code, thus ignoring the clinical logic (also referred to as the phenotype definition). Our objectives were to address these limitations by providing a comprehensive examination of how different vocabulary mapping strategies, using both the clinical code sets and the logical phenotype definitions across all clinical domains (i.e. conditions, medications, labs, procedures, and observations), effects the creation of patient cohorts for both case and control groups.

Methods

Data. We used two independent de-identified datasets: (1) all 11354364 visits, at Childrens Hospital Colorado (CHCO) and (2) 58976 adult intensive care visits from the MIMIC III database. Both datasets were standardized to OMOP V5. The study was approved by the Colorado Multiple Institutional Review Board (15-0445).

Phenotypes. Seven PheKB phenotypes: Attention-deficit/hyperactivity disorder (ADHD), Appendicitis, Crohn′s Disease, Sickle Cell Disease (SCD), Sleep Apnea, Steroid-Induced Osteonecrosis (SIO), and Systemic Lupus Erythematosus (SLE).

Experiments.

We performed the experiments shown in Figure 1. These were designed to elucidate the effects of using (1) different vocabulary mapping strategies (i.e. exact- vs. fuzzy string mapping using varying levels of the ST terminology hierarchies as well additional OMOP resources, like the concept synonym table), (2) different types of clinical data

(i.e. using only condition codes vs. all available clinical domains), and (3) examined their effects when using only the clinical code sets (like Hripcsak et al.[3]) or using both clinical code sets and phenotype definitions. This resulted in 36 mapping strategies x two types of clinical data x two CP approaches for a total of 144 comparisons for each case and control group across the seven CPs. Due to space limitations, we will only report results for cases.

Evaluation Metrics. Similar to Hripcsak et al.[3], for each comparison, we calculated false negative (FN) and false positive (FP) error rates as the number of incorrectly included or missed patients, respectively. The patient cohorts created from the SV clinical code set provided by the CP authors were used as the gold standard patient cohorts. A small subset of FP/FN patients from each CP were manually verified by our team of clinical experts.

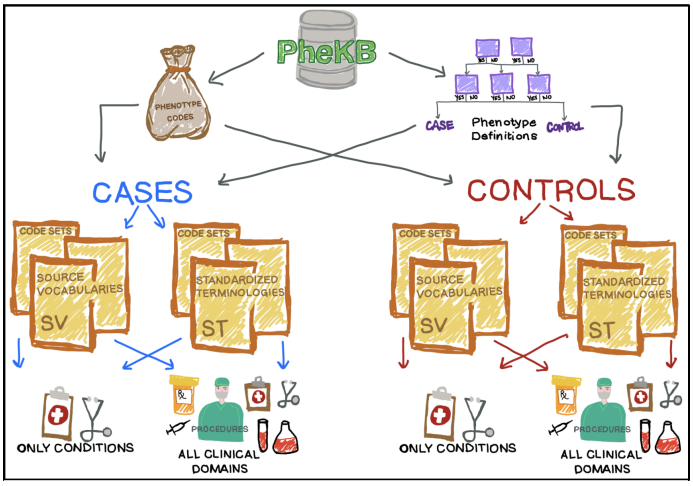

Figure 1. Experimental Design.

Results

The 36 different mapping strategies were first analyzed using only the clinical codes. The created case patient cohort sizes varied widely across the CPs (cohort size created using only conditions/cohort size created using all clinical domains; a single number means the results were the same): ADHD (CHCO: 17639/4472; MIMIC: 131/58), Appendicitis (CHCO: 4178/2948; MIMIC: 30/23), Crohn′s Disease (CHCO: 754/477; MIMIC: 272/189), SCD (CHCO: 333; MIMIC: 13), Sleep Apnea (CHCO: 21631; MIMIC: 2189), SIO (CHCO: 337/108; MIMIC 48/20), and SLE (CHCO: 446/0; MIMIC 178/0). The FP and FN error rates ranged from 0-88% and 0-25%, respectively. In both cases, the highest error rates were observed in the ADHD CP when using a fuzzy-matching mapping strategy that included all of the concept′s synonyms and descendants. Next, we analyzed the mapping strategies using only the clinical codes and the phenotype logic. Similar patterns were observed: ADHD (CHCO: 3624/1706; MIMIC: 17/0), Appendicitis (CHCO: 4178/367; MIMIC: 30/18), Crohn′s Disease (CHCO: 230/199; MIMIC: 1), SCD (CHCO: 1351/1308; MIMIC: 54/9), Sleep Apnea (CHCO: 21631; MIMIC: 2189), SIO (CHCO: 168/61; MIMIC 6/0), and SLE (CHCO: 0; MIMIC: 0). The observed FP and FN error rates ranged from 0-49% (ADHD) and 0-37% (Appendicitis), respectively. Similar to using only clinical codes, the fuzzy-matching mapping strategy that included all of the concept′s synonyms and descendants resulted in the highest error rates. In all analyses, an exact mapping strategy that included the children of each clinical concept resulted in the lowest error.

Conclusion

Our preliminary findings using only clinical codes corroborate prior work by Hripcsak et al.[3]. When including clinical codes and phenotype definitions, we found that utilizing automated vocabulary mapping strategies resulted in lower FP rates, but high FN rates. Work is currently underway which extends these findings by: (1) adding two addition CPs; (2) including new domain expert verified mapping strategies; and (3) performing expert verification of the resulting patient cohorts.

References

1. Richesson RL, Sun J, Pathak J, Kho AN, Denny JC. Clinical phenotyping in selected national networks: demonstrating the need for high-throughput, portable, and computational methods. Artif Intell Med. 2016;71:57-61.
2. Banda JM, Seneviratne M, Hernandez-Boussard T, Shah NH. Advances in electronic phenotyping: from rule-based definitions to machine learning models. Annu Rev Biomed Data Sci. 2018;1:53-68.
3. Hripcsak G, Levine ME, Shang N, Ryan PB. Effect of vocabulary mapping for conditions on phenotype cohorts. J Am Med Inform Assoc. 2018;25:1618-25.