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

Tiffany J. Callahan, MPH¹, Jordan Wyrwa, DO², Katy E. Trinkley, PharmD³, Lawrence E. Hunter, PhD¹, Michael G. Kahn, MD, PhD⁴, Tellen D. Bennett, MD, MS¹,⁴
¹Computational Bioscience Program; ²Department of Physical Medicine and Rehabilitation, School of Medicine; ³Department of Clinical Pharmacy; ⁴Department of Pediatrics, University of Colorado Denver Anschutz Medical Campus, Aurora, CO

Introduction

The universal adoption of electronic health records 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 drive mechanistic insight¹,². However, several barriers to the development, validation, and implementation of CPs must be overcome before their potential can be fully realized.

Phenotype knowledge-bases like eMERGE’s PheKB are rich repositories of domain expert-derived CPs. Unfortunately, most of the CPs cannot easily be implemented across different electronic health records systems because they are tailored to specific source vocabularies (SV). Common data models 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 common data models like the Observational Medical Outcomes Partnership (OMOP), one could employ different strategies to align clinical codes (e.g. ICD-9CM, LOINC) provided in a CP definition to a common data model (e.g. exact string- or manual-mapping). Understanding the trade-offs of these different vocabulary mapping strategies is vital for enabling common data model-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 condition (or diagnosis) codes; and (3) creating patient cohorts using only the presence of at least one diagnosis code, thus ignoring the clinical logic. 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 clinical logic across all clinical domains (i.e. conditions, medications, labs, procedures, and observations), impacts the creation of patient cohorts.

Methods

We used two independent de-identified datasets: (1) all 11,354,364 Children's Hospital Colorado pediatric patient visits and (2) 58,976 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. Using these data, five PheKB phenotypes were implemented: Attention-deficit/hyperactivity disorder, appendicitis, Crohn’s disease, sickle cell disease, and systemic lupus erythematosus. We performed the experiments outlined in Figure 1, which 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

Figure 1. Experimental design.
OMOP resources, like the concept synonym table), (2) different data types (i.e. using only condition codes vs. all available clinical domains), and (3) different CP approaches (i.e. using only the clinical code sets, like Hripcsak et al.\(^1\), or using clinical codes and clinical logic) on constructing patient cohorts. This resulted in 36 mapping strategies, 2 types of clinical data, and 2 CP approaches for a total of 144 comparisons for each case and control group across the five CPs. Similar to Hripcsak et al.\(^1\), for each comparison, we calculated false negative (FN) and false positive (FP) error rates as the number of incorrectly missed patients divided by the total number of patients with that condition and the number of incorrectly included patients divided by the total number of patients without that condition, respectively. The patient cohorts created by the CP authors were used as the gold standard patient cohorts, using the same gold standard definition defined by Hripcsak for comparability\(^2\). Due to space limitations, only the results for the Children's Hospital Colorado case cohorts are presented here.

Results

As shown in Table 1, when using only codes, the FP and FN error rates ranged from 0-88% (only conditions) and 0-25% (all data). For both data types, the highest error rates were observed in the ADHD CP when using a fuzzy-matching mapping strategy that included concept synonyms/descendants. When using codes and clinical logic, the observed FP and FN error rates ranged from 0-49% (only conditions) and 0-37% (all data). For both evaluations, the fuzzy-matching mapping strategy using all concept synonyms/descendants resulted in the highest error rates.

**Table 1.** Performance of mapping strategies for the Children's Hospital Colorado case cohort-derived phenotypes.

<table>
<thead>
<tr>
<th>APPROACH</th>
<th>ONLY CODES</th>
<th>CODES + CLINICAL LOGIC</th>
<th>ALL CLINICAL DATA</th>
</tr>
</thead>
<tbody>
<tr>
<td>METRIC(^a)</td>
<td>Actual</td>
<td>Predicted</td>
<td>FP</td>
</tr>
<tr>
<td>ADHD</td>
<td>19980</td>
<td>17639 20464</td>
<td>0 2853</td>
</tr>
<tr>
<td>Appendicitis</td>
<td>4178</td>
<td>2629 4178</td>
<td>0 0</td>
</tr>
<tr>
<td>Crohn's Disease</td>
<td>754</td>
<td>751 754</td>
<td>0 0</td>
</tr>
<tr>
<td>SCD</td>
<td>333</td>
<td>333 338</td>
<td>0 5</td>
</tr>
<tr>
<td>SLE</td>
<td>446</td>
<td>446 447</td>
<td>0 1</td>
</tr>
</tbody>
</table>

\(^a\)Cell values contain the min and max (min-max) of all 144 examined mapping strategies. A single number means the min/max were the same.

SCD: Sickle Cell Disease; SLE: Systemic Lupus Erythematosus.

Conclusions

When using only codes and only condition data, our results overlapped with Hripcsak et al.\(^1\); an exact mapping strategy that included all concept descendants resulted in the lowest error. Our findings extend Hripcsak’s work by: (1) using all clinical data types and (2) defining CPs using codes and clinical logic. We found that using codes and clinical logic with all clinical data resulted in lower FNs and caused some previously eligible CPs to become ineligible (i.e. no patients returned from cohort query). Future work will: (1) evaluate additional CPs; (2) include new domain expert verified mapping strategies; and (3) perform expert verification of resulting patient cohorts.

References


777