Public Health and Informatics
J. Mantas et al. (Eds.)

doi:10.3233/SHTI210138

Potential Role of Clinical Trial Eligibility Criteria in Electronic Phenotyping

Zhehuan CHEN [a,1], Hao LIU[a,1], Alex BUTLER[a], Anna OSTROPOLETS[a] and Chunhua WENG [a,2]
[a]*Department of Biomedical Informatics, Columbia University, New York, NY, USA*

Abstract. 2,719 distinctive phenotyping variables from 176 electronic phenotypes were compared with 57,150 distinctive clinical trial eligibility criteria concepts to assess the phenotype knowledge overlap between them. We observed a high percentage (69.5%) of eMERGE phenotype features and a lower percentage (47.6%) of OHDSI phenotype features matched to clinical trial eligibility criteria, possibly due to the relative emphasis on specificity for eMERGE phenotypes and the relative emphasis on sensitivity for OHDSI phenotypes. The study results show the potential of reusing clinical trial eligibility criteria for phenotyping feature selection and moderate benefits of using them for local cohort query implementation.

Keywords. Clinical Trial Eligibility Criteria, Electronic Phenotyping

1. Introduction

Electronic phenotyping involves the identification of a cohort of patients with the condition of interest and the characterization of observable traits of this condition using the electronic health records (EHR) data. For example, a rule-based electronic phenotype algorithm for *Type 2 diabetes mellitus* created by the Electronic Medical Records and Genomics (eMERGE) network [1] specifies rules such as "random glucose > 200 mg/dl, fasting glucose >= 125 mg/dl, or Hemoglobin A1c >= 6.5%" to identify patients. The development of phenotyping algorithms involves feature selection and query implementation in heterogeneous databases using clinical data standards, all being clinical knowledge intensive [2]. Therefore, there is a great need to support phenotype knowledge engineering in order to improve the scalability of electronic phenotyping. Separately, all Randomized Clinical Trials (RCTs) define eligibility rules to specify qualifying study cohorts. Many of these rules define phenotype knowledge and hence are similar to rule-based phenotyping algorithms. For example, an eligibility criterion for a Type 2 diabetes mellitus RCT is "7.0% ≤ HbA1c ≤ 10.0%;". HbA1c is a shared phenotype feature in both EHR phenotypes and clinical trial eligibility criteria for Type 2 Diabetes Mellitus. We previously developed a Clinical Trial Knowledge Base (CTKB, www.ctkb.io). It includes structured clinical trial eligibility criteria extracted from all clinical trials (N=314,056 as of August 2020) registered in ClinicalTrials.gov and standardized using the OMOP Common Data Model (CDM) [3] for optimal knowledge reuse. In addition, as an active participant in both the eMERGE consortium and the

[1] Equal-contribution first authors
[2] Corresponding Author, Department of Biomedical Informatics, Columbia University, 622 West 168th Street, PH-20, Room 407, New York, NY 10032, USA. E-mail: chunhua@columbia.edu.

global open science consortium OHDIS (Observational Health Data Sciences and Informatics), we have validated 33 eMERGE phenotyping algorithms (e.g., *Type 2 Diabetes Mellitus*, *Colorectal Cancer*, *Hypothyroidism*) and implemented 143 OHDSI phenotyping algorithms (e.g., *Neck Pain*, *Hypoglycemia*, *Nausea*). With the availability of a large amount public information of clinical trial summaries, including eligibility criteria text, we ***hypothesize*** that clinical trial eligibility criteria promise to facilitate knowledge reuse for electronic phenotyping given their similarities in specifying phenotype characteristics. Therefore, this study aims to test this hypothesis by contributing an original assessment of the overlapping clinical phenotyping concepts between clinical trial eligibility criteria and the electronic phenotyping algorithms from eMERGE and OHDSI. This study will shed light on the feasibility of identifying phenotype features from RCT eligibility criteria.

2. Methods

We used **53** validated eMERGE phenotyping algorithms from http://phekb.org and **190** phenotyping algorithms from OHDSI (www.ohdsi.org). Each eMERGE phenotype uses up to 13 types of variables, including *Diagnosis*, *Family History*, *Problem List*, *Medication*, *Procedure*, *Demographic*, *Observation*, *Phenotype*, *Lab Test*, *Note*, *Visit* and *Enrollment*, with the numbers of variables ranging from 1 (*Dementia*) to 79 (*Autoimmune*). These variables were mapped to the OMOP CDM standard concepts. To mitigate ambiguity (when a variable can be mapped to multiple concepts), we created the mapping from eMERGE variable types to OMOP concept domains in Table 1. For example, eMERGE variable kidney transplant of type *Procedure* can be mapped to a Procedure Concept *Transplant of kidney (ID: 4322471)* or a Condition concept "*Transplanted kidney present (ID: 42539502)*") in the OMOP CDM. For our mapping, "kidney transplant" can only be mapped to the *Procedure* domain instead of the *Condition* domain in the context of this study. Each OHDSI phenotype contains a set of OMOP CDM Condition concepts, which were manually selected initially and then iteratively reviewed and refined by adding descendant concepts, parent concepts, and lexically similar concepts for 2-6 rounds [4]. The numbers of OMOP concepts for each phenotype range from 1 (*Takayasu's disease*) to 4117 (*Malignant neoplastic disease*).

Table 1. Mapping from eMERGE phenotyping variable types to the OMOP CDM domains.

***From* Variable Type**	***To* OMOP CDM Domain**
Diagnosis	Condition
Family History	Condition
Lab Test	Measurement
Note	Condition
Observation	Observation
Phenotype	Condition
Procedure	Procedure
Medication	Drug
Problem List	Condition

The methodology framework is shown in Figure 1. Our methods include concept standardization at the phenotype level (step 1) and at the phenotype variable level (step 2). First, the "Condition" field for all RCT studies and the phenotype names were mapped to MeSH terms using a public tool called Athena (https://athena.ohdsi.org/). We retrained 33 eMERGE phenotypes and 143 OHDSI phenotypes for further analysis after removing unmapped phenotype names. Next, all variables in phenotypes and eligibility criteria

were mapped to the OMOP CDM standard concepts. Third, we filtered out descendent concepts and only retained the high-level concepts used the SNOMED-CT [5] hierarchy. For example, if "*Respiratory failure*" and "*Acute-on-chronic respiratory failure*" are both phenotype concepts while "*Acute-on-chronic respiratory failure*" is the child concept of "*Respiratory failure*", we only retained "*Respiratory failure*". The total count of distinctive eligibility criteria concepts extracted from CTKB was 57,150. For each variable from a phenotype algorithm, we verified if this variable exists in clinical trial eligibility criteria for trial on the same phenotype. A hit was identified if and only if a match is found. The hit rate was defined as the percentage of phenotype variables that match with eligibility criteria concepts. We calculated the hit rates for both eMERGE and OHDSI phenotypes. For eMERGE phenotypes, we explored the effects of eMERGE variable types and OMOP domains on hit rates. For OHDSI phenotypes, we investigated the associations between the number of high-level concepts and hit rates.

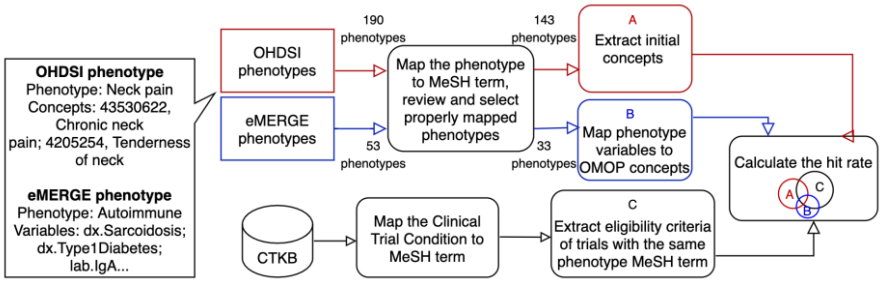

Figure 1. The methodology framework. Blue arrows represent the dataflow of eMERGE phenotypes while red represents the data flow of the OHDSI phenotypes.

3. Results

3.1. Hit Rate Analysis with eMERGE Phenotypes

The 33 eMERGE phenotype algorithms contain 351 variables, with 199 (56.7%) mapped to the OMOP CDM *Condition* domain, 77 (21.9%) to *Measurement*, 38 (10.8%) to *Drug*, 26 (7.4%) to *Procedure* and 11 (3.1%) to *Observation*, and 176 (50.1%) to *Diagnosis*, 77 (21.9%) to *Lab Test*, 98 (27%) belong to other types.

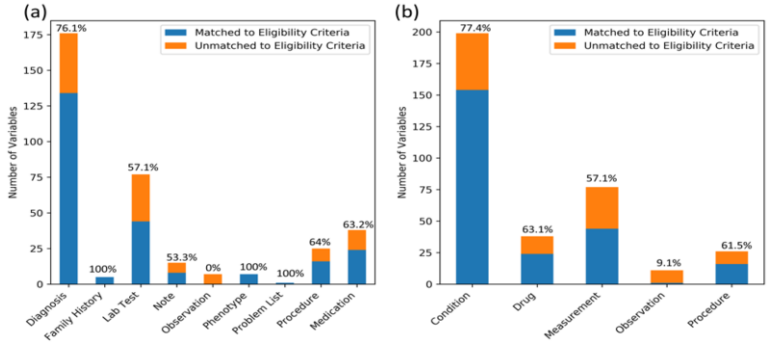

Figure 2. Hit rate between eMERGE phenotyping variables and CTKB eligibility criteria by (a) variable type and (b) domain. The percentage on each bar shows the hit rate.

The average hit rate of the 33 phenotypes was 69.5% with a median of 75%, a

maximum of 100% and a minimum of 33.3% (for *Anxiety*). As shown in Figure 2, blue shows the number of variables matched to eligibility criteria, and yellow shows the number of unmatched variables. Among those different variable types, *Diagnosis* (76.1%), *Family History* (100%), *Phenotype* (100%) and *Problem List* (100%) have the highest hit rates; *Lab Test* (57.1%), *Note* (53.3%), *Procedure* (64%) and *Medication* (63.2%) have moderate hit rates; *Observation* (0%) has the lowest hit rates. As for the hit rate by different domains, the *Condition* domain has the highest hit rate with the value of 77.4%, whereas the hit rate of the *Observation* domain is only 9.1%.

3.2. Hit Rate Analysis with OHDSI Phenotypes

The 143 OHDSI phenotypes contain 10,872 concepts, including 2,368 high-level (top class in the terminology) concepts. The phenotype *Renal impairment* has most high-level concepts (125), whereas 16 phenotypes each have only one high-level concept. The average hit rate of the 143 OHDSI phenotypes was 47.6%, with 20 phenotypes having the hit rate of 100% and 4 having the hit rate of 0%. As shown in Figure 3, the average hit rates for phenotypes with high-level concept counts between 1 and 10, 11 and 20, 21 and 30, 31 and 40, and larger than 40 are 56.5%, 45.9%, 32.3%, 34.9%, and 26.9%.

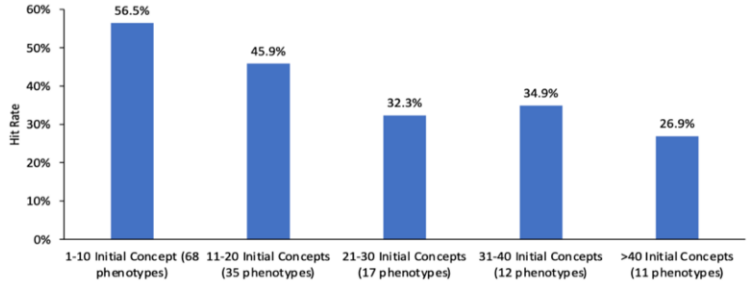

Figure 3. Average hit rates of OHDSI phenotypes with different initial concept number range.

4. Discussion

On average 69.5% percent of eMERGE phenotype variables are found in clinical trial eligibility criteria for the corresponding phenotype, implying that criteria are useful for phenotype knowledge engineering. We noticed a low hit rate on *Observation* variables, which is attributed to the catch-all nature and open-ended definition of *Observation* (https://www.ohdsi.org/web/wiki/doku.php?id=documentation:cdm:observation) so that some concepts are defined in multiple domains: e.g., phenotyping variable "*smoking status*" could be mapped to eligibility criteria concepts such as "*smoker*", "*non-smoker*", "*smokeless*", and "s*moke*", which are distributed in the *Condition* or the *Observation* domains. Many *Measurement* variables were not matched to eligibility criteria because they tend to be highly specific and span multiple domains, such as "*Fecal occult blood negative*" (*Condition* "*Fecal occult blood negative*" or *Measurement* "*occult blood*" with the attribute "*negative*"). The other reason lies in the complexity of the OHDSI OMOP CDM. As OMOP CDM integrates multiple disparate vocabularies into one, mapping a measurement correctly can be challenging without additional context.

On average 47.6% of OHDSI phenotype variables are found in eligibility criteria, much lower than that of eMERGE phenotype (69.5%). This discrepancy can be attributed

to two possible causes. First, eMERGE phenotypes emphasize specificity more while OHDSI phenotypes emphasize sensitivity more given its requirement to support cohort queries across multiple distributed databases. Second, eMERGE phenotype variables represent important clinical features used for phenotyping while OHDSI phenotype concepts tend to be all-inclusive to reduce overheads in database queries. This finding underscores the utility of reusing clinical trial eligibility criteria for phenotyping feature extraction. On the contrary, the low coverage of OHDSI concepts in eligibility criteria suggests limited gains from reusing clinical trials in the implementation phase of phenotyping. The OHDSI phenotypes contain a large portion (78.2%) of low-class specific concepts derived from database implementations and hence are usually absent from eligibility criteria. For example, for the phenotype *Renal impairment* (hit rate: 7.2%), it is coded with compound concepts (e.g., mixing with pathological or anatomical descriptors) like "*Hypertensive heart and chronic kidney disease with heart failure and stage 1 through stage 4 chronic kidney disease, or unspecified chronic kidney disease*" was rarely used in eligibility criteria for recruiting patients with renal impairment. Also, the average hit rate is lower with relatively large number of initial concepts (56.5% with 1-10 initial concepts vs. 26.9% with >40 initial concepts).

5. Conclusions

Our study shows that clinical trial eligibility criteria can serve as a valuable and reusable source of phenotype knowledge, particularly for guiding phenotyping feature selection. Our results also show the moderate benefits of extracting phenotype concepts relevant for local algorithm implementation, largely due to the absence of specific coded concepts in clinical trial protocols. A potential solution is to improve concept standardization for clinical trial eligibility criteria using widely adopted standards such as OMOP CDM. Future work should improve formal knowledge representation for eligibility criteria using widely adopted clinical data standards, such as the OMOP CDM.

Acknowledgments

This study was sponsored by National Library of Medicine grant 5R01LM009886-11 and National Center for Advancing Clinical and Translational Science grant UL1TR001873 and 1OT2TR003434-01.

References

[1] McCarty CA, Chisholm RL, Chute CG, Kullo IJ, Jarvik GP, Larson EB, et al. The eMERGE Network: A consortium of biorepositories linked to electronic medical records data for conducting genomic studies. BMC Med Genomics. 2011;4(1):1–11.
[2] Shang N, Liu C, Rasmussen L V., Ta CN, Caroll RJ, Benoit B, et al. Making work visible for electronic phenotype implementation: Lessons learned from the eMERGE network. J Biomed Inform. 2019 Nov 1.
[3] OMOP Common Data Model – OHDSI, Available at: https://www.ohdsi.org/data-standardization/the-common-data-model/.
[4] Ostropolets A. Concept Prevalence Design diagnostics. In 2020 OHDSI Global Symposium;
[5] SNOMED CT, Available at: https://www.nlm.nih.gov/healthit/snomedct/index.html.