Lessons from CIRCE implementation of eMERGE phenotype definitions into CDM v5 SQL queries

Matthew E. Levine[1,2], Patrick B. Ryan, Ph.D.[1,2,3], George Hripcsak, M.D., M.S.[1,2]
[1]Department of Biomedical Informatics, Columbia University Medical Center, New York, NY, USA;
[2]Observational Health Data Sciences and Informatics , Columbia University Medical Center, New York, NY, USA;
[3]Janssen Research & Development, LLC, Titusville, NJ, USA

Abstract

We have implemented the logic of five eMERGE phenotype definition algorithms from PheKB.org in a clinical study cohort identification tool, CIRCE (Cohort Inclusion and Restriction Criteria Expression), to enable the use of these algorithms on clinical data in the OHDSI (Observational Health Data Sciences and Informatics) network. This work reports the challenges of interpreting and translating the consensus phenotype definitions for research application, and points to important considerations for the representation, presentation, and implementation of electronic phenotype definitions for both human and computer uses. We conclude that EHR phenotyping algorithms should better support both human review and computer execution.

Background

The ability to use EHR data to identify patients with particular characteristics, or phenotypes, is of great importance to the OHDSI community, and the eMERGE (Electronic Medical Records and Genomics) network[1] has developed, tested, and validated over 40 phenotype algorithms (hosted, many publicly, on Phenotype KnowledgeBase at PheKB.org)[2]. We wish to transform the eMERGE phenotyping documents into executable data queries that are compliant with the OMOP CDM. This translation is enabled by tools built by the OHDSI community that provide a human-readable interface for developing and storing data queries[3]. So far, we have implemented five PheKB phenotype definitions (Drug-Induced Liver Injury, Appendicitis, Type 2 Diabetes Mellitus (T2DM), Cataracts, and Hypothyroidism) into standardized computable representations in JSON, which can be compiled into platform-independent SQL code, distributed, and executed across the OHDSI network.

Methods

Human comprehension and interpretation of phenotypes

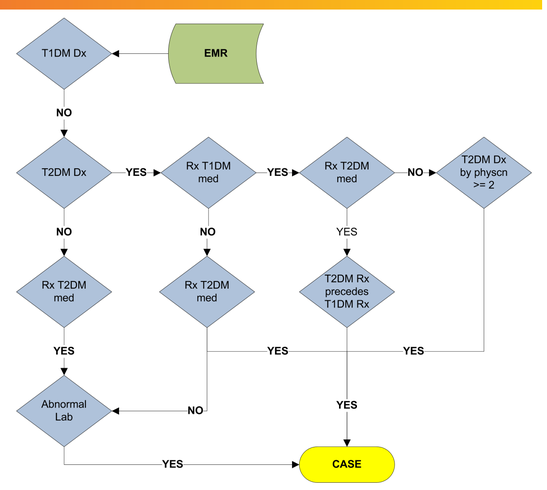
PheKB documentation include pseudo-code, flow-charts, SQL, step-wise directives, and code/ term tables (Fig. 1a). Logic was interpreted as literally as possible.

HERMES: Concept translation

Diagnosis and procedure codes were translated into standard OHDSI vocabularies using HERMES (Health Entity Relationship and Metadata Exploration System), a web-based vocabulary browsing tool for OMOP CDM v5

CIRCE: SQL query generation

HERMES JSON output was imported to CIRCE (Cohort Inclusion and Restriction Criteria Expression) (Fig. 1b), which provides a human-readable interface for query development



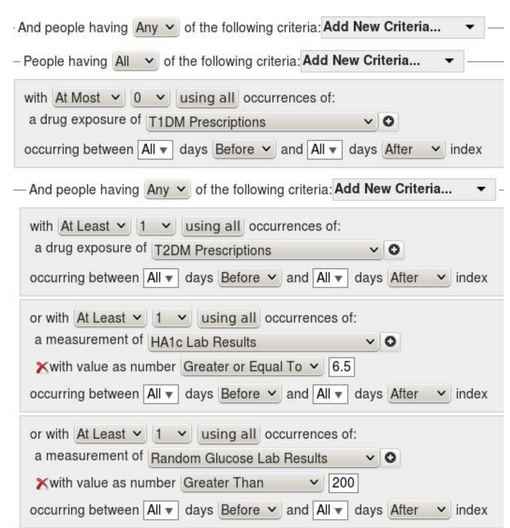


Figure 1. a) Flowchart from T2DM phenotype algorithm b) CIRCE implementation of T2DM phenotype, which is easily stored, shared, and modified online

Results

Challenges of interpretation

Logical interpretations were challenged by the following factors: ill-defined concepts (no codes), linguistic ambiguities, inconsistencies between diagrams and pseudo-code, and overlap of inclusion and exclusion concept sets. By studying the flow-charts, we observed multiple *unintended logical artifacts*. For example, strict interpretation of branches in the T2DM algorithm (Fig. 1a) yields surprising results—adding a T2DM diagnosis code can exclude a case (Fig. 2). We elected to preserve this case in a literal implementation, and will compare its results to a version that removes this provision.

T1 Dx	T2 Dx	T1 Rx	T2 Rx	T2 Rx first	Abnormal Lab	CASE
No	No	Yes	Yes	No	Yes	Yes
No	Yes	Yes	Yes	No	Yes	No

Figure 2. Two paths through the T2DM algorithm are represented by the conditions met in the table—we observed this unique result by studying the diagrams provided by the authors. This demonstrates the value of human-readable versions of algorithms.

Challenges of concept translation

Most relevant ICD-9 codes had standard mappings, and they were typically included along with their descendants in the exported concept set. However, we observed cases in which the standard mapped concept was related to other ICD-9 codes not mentioned in the criteria. In such cases, we evaluated these codes for qualitative similarity to the source concept and to concepts in the exclusion criteria (Fig. 3). Phenotype authors should be consulted for faithful translation, but testing can ensure an effective adaptation.

Standard to non-standard map of 366.8	In Inclusion Set	In Exclusion Set
366 (Cataract)	No	No
366.44 (Cataract associated with other syndromes)	No	No
366.8 (Other cataract)	Yes	No
366.9 (Unspecified cataract)	Yes	No

Figure 3. ICD-9 condition codes 366.8 and 366.9 of the Cataract inclusion concept set both have non-standard to standard maps to the SNOMED concept ID 193570009 (Cataract). ICD-9 mappings of the SNOMED Cataract concept are shown. Although ICD-9 code 366.44 is not in the inclusion set, we elected to use SNOMED Cataract for reasons of similarity.

Conclusions

Five eMERGE phenotype definitions were translated into an OMOP CDM compliant format and stored on CIRCE for use by any institution in the OHDSI community. CIRCE implementations allow for modification and sharing, and we encourage users to store and note changes. In addition, we developed a useful pipeline for reviewing and translating eMERGE phenotype definitions—these processes have elucidated important considerations for the fate and format of such documents. We recommend an increased focus on presenting documents with human readability, developing collaborations between authors and implementers to ensure logical accuracy, and storing fully vetted algorithms in a coded format like that offered by CIRCE to reliably couple human readable information to unambiguous code.

1. Public Phenotypes | PheKB [Internet]. [cited 2015 Oct 12]. Available from: https://phekb.org/phenotypes
2. Newton KM, Peissig PL, Kho AN, Bielinski SJ, Berg RL, Choudhary V, et al. Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network. Journal of the American Medical Informatics Association. 2013 Jun 1;20(e1):e147–e154.
3. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, et al. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. Stud Health Technol Inform. 2015;216:574–8.