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

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

Abstract
We have implemented the logic of five eMERGE phenotype definition algorithms from PheKB.org in CIRCE (Cohort Inclusion and Restriction Criteria Expression) to enable the use of these algorithms on clinical data in the OHDSI network. Here, we enumerate the types of challenges faced when interpreting the consensus phenotype definitions for research application, and point to important considerations for the development, presentation, and distribution of electronic phenotype definitions.

Introduction
The Phenotype KnowledgeBase (PheKB.org) provides expert validated algorithms from the eMERGE (Electronic Medical Records and Genomics) network for identifying characteristics of patients in electronic health records (EHR), and describes these consensus definitions through a variety of online documents, which include a combination of SQL code, pseudo-code, flow-charts, step-wise directives, and tables (Excel, Word, and PDF) of relevant codes and terms. So far, we have transformed five phenotype definitions (Type 2 Diabetes, Hypothyroidism, Cataracts, Appendicitis, and ADHD) into actionable SQL code that can be applied across the OHDSI network.

Methods
The transformation from an eMERGE phenotype definition to callable OMOP CDM v5 SQL code requires three key steps: 1) Appropriate interpretation of the published algorithms, 2) Reliable translation of non-standard data formats, such as ICD-9, into the standardized concepts of the OMOP CDM, and 3) Generation of the SQL queries using the logic described in the phenotype definition.

We first reviewed the documentation made available by PheKB that articulates definitions of the phenotype algorithms, and a logical interpretation was written in pseudo-code. We then performed medical concept translation and aggregation with HERMES (Health Entity Relationship and Metadata Exploration System), a web-based vocabulary browsing tool for OMOP CDM v5 that allows exploration of concept relationships and provides a JSON output of rich concept inclusion/exclusion sets. CIRCE (Cohort Inclusion and Restriction Criteria Expression) was then used to integrate our pseudo-code interpretation of the original phenotype with the standard concept sets exported from HERMES to produce a sophisticated OMOP CDM v5 SQL query that reflects the eMERGE phenotype definition. The experience of working with these definitions and implementing this pipeline provided important insights for the continuation of this work.

Results
It was found that the eMERGE phenotype definitions provide sufficient information to translate the inclusion/exclusion concept sets into sensible standardized concept sets with HERMES. Data formats from the definitions were limited to ICD-9, CPT, lab tests, and drug names, all of which have standard concept mappings in HERMES. It was found that RxNorm drug ingredient concepts and their descendants most efficiently reflected the drugs listed in the documents, whether they were enumerated by brand name or ingredient. Lab tests were consistently well-represented by LOINC codes. The primary challenge during this stage was faithfully redefining sets of ICD-9 codes in standard data formats. The majority of ICD-9 codes in the phenotypes' inclusion/exclusion sets had standard mappings in HERMES, and the mappings and all their descendants were typically included in the new exported concept set. However, we often discovered related concepts that, although not descendants of the primary mapped concepts, were similar to the original ICD-9 code of interest. In these cases, it is the choice of the user to either ignore or include these additional concept IDs. We assumed that an unmapped, yet similar, concept that lacks an ICD-9 mapping was not available for consideration by the phenotypers (and included it for its similarity), but assumed that a standard concept with a separate ICD-9 mapping was likely to be intentionally omitted by the phenotypers, since its concept was available to the developers in ICD-9.

We elected to ignore provisions in the algorithms that addressed pathology reports and natural language processing (NLP) on clinical notes, given that these data are not routinely available. None of the five reviewed phenotypes required these data for a case. Two of the definitions only addressed them if the data was available in the record, and one of the definitions (hypothyroidism) provided ICD-9 billing codes as an alternative to text-based search of exclusion criteria.

Interpretation of the overall logical structure of the published algorithms presented a few key concerns regarding the precision of the definitions. In the case of the definitions of Type 2 Diabetes Mellitus, Appendicitis, and Cataracts, the algorithms were written as sequential checks through a path of criteria. We expect that the ordering of these checks restricts the cohorts more than was intended by the developers. An important example is in the Type 2 Diabetes Mellitus (T2DM) algorithm (Figure 1), in which adding a Type 2 Diabetes diagnosis code excludes a case.

1. no T1DM dx + no T2DM dx + T1DM rx + T2DM rx + T2DM rx follows T1 rx + Abnormal Lab -> case
2. no T1DM dx + T2DM dx + T1DM rx + T2DM rx + T2DM rx follows T1 rx + Abnormal Lab -> no case

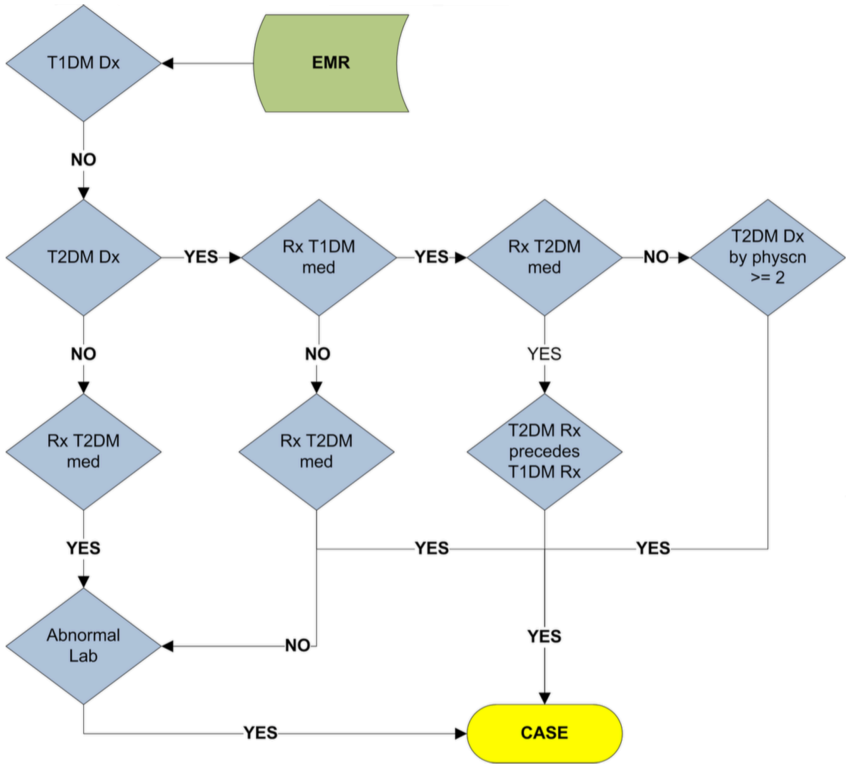

Figure 1. Algorithm for identifying Type 2 Diabetes Mellitus (T2DM) in the EHR

We also found multiple inconsistencies within the algorithms. The appendicitis documentation included a flowchart that did not match the logic of its corresponding pseudo-code, and invoked an undefined concept ("History of Appendicitis"), which we chose to omit. In addition, the Cataracts definition listed an ICD-9 code in its inclusion set that was a descendant of one of its listed exclusion codes. We also encountered linguistic ambiguity—for instance, during our review of the cataracts definition cataracts, it was not immediately clear whether diagnoses prior to 1960 should simply be ignored, or if having a diagnosis prior to 1960 was a criterion for exclusion.

Conclusion

In light of the significant challenges we faced in implementing the consensus phenotype definitions from PheKB.org, we believe it is important to reconsider the format of such documents. We recommend explicit definition of the criteria sets needed for satisfying a phenotype in place of ordered criteria checks, which can become unnecessarily cumbersome when dealing with many steps. Moreover, we believe that these algorithms need a *transferable precision*, in which all instructions, conditions, and concept sets are unquestionably clear to the reader. This can be accomplished in pseudo-code or carefully worded English; however, it may be beneficial for the experts who developed these phenotypes to work directly with the team responsible for translation into SQL in order to solidify these definitions in an immutable and actionable form. In order to further validate published phenotypes, we recommend an additional focus on verifying the accuracy with which flow-charts reflect the authors' original intentions—with improved diagrams, researchers can verify their final output from CIRCE by comparing the original flow-charts with the manual process created in CIRCE, and even check against auto-generated flow diagrams from the CIRCE SQL code.