Journal of Biomedical Informatics 96 (2019) 103253

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Facilitating phenotype transfer using a common data model

George Hripcsak[a,b,*], Ning Shang[a], Peggy L. Peissig[c], Luke V. Rasmussen[d], Cong Liu[a], Barbara Benoit[e], Robert J. Carroll[f], David S. Carrell[g], Joshua C. Denny[f,h], Ozan Dikilitas[i], Vivian S. Gainer[e], Kayla Marie Howell[j], Jeffrey G. Klann[e], Iftikhar J. Kullo[i], Todd Lingren[k], Frank D. Mentch[l], Shawn N. Murphy[e], Karthik Natarajan[a,b], Jennifer A. Pacheco[d], Wei-Qi Wei[f], Ken Wiley[m], Chunhua Weng[a]

[a] Department of Biomedical Informatics, Columbia University, New York, NY, United States
[b] Medical Informatics Services, NewYork-Presbyterian Hospital, New York, NY, United States
[c] Center for Precision Medicine Research, Marshfield Clinic Research Institute, Marshfield, WI, United States
[d] Northwestern University Feinberg School of Medicine, Chicago, IL, United States
[e] Research Information Science and Computing, Partners Healthcare, Boston, MA, United States
[f] Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, United States
[g] Kaiser Permanente Washington Health Research Institute, Seattle, WA, United States
[h] Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, United States
[i] Department of Cardiovascular Medicine, Mayo Clinic, Rochester, MN, United States
[j] Vanderbilt Institute for Clinical and Translational Research, Vanderbilt University Medical Center, Nashville, TN, United States
[k] Cincinnati Children's Hospital Medical Center, Cincinnati, OH, United States
[l] Center for Applied Genomics, Children's Hospital of Philadelphia, Philadelphia, PA, United States
[m] National Human Genome Research Institute, NIH, Bethesda, MD, United States

ARTICLE INFO

Keywords:
Common data model
Phenotyping
Electronic health records

ABSTRACT

Background: Implementing clinical phenotypes across a network is labor intensive and potentially error prone. Use of a common data model may facilitate the process.
Methods: Electronic Medical Records and Genomics (eMERGE) sites implemented the Observational Health Data Sciences and Informatics (OHDSI) Observational Medical Outcomes Partnership (OMOP) Common Data Model across their electronic health record (EHR)-linked DNA biobanks. Two previously implemented eMERGE phenotypes were converted to OMOP and implemented across the network.
Results: It was feasible to implement the common data model across sites, with laboratory data producing the greatest challenge due to local encoding. Sites were then able to execute the OMOP phenotype in less than one day, as opposed to weeks of effort to manually implement an eMERGE phenotype in their bespoke research EHR databases. Of the sites that could compare the current OMOP phenotype implementation with the original eMERGE phenotype implementation, specific agreement ranged from 100% to 43%, with disagreements due to the original phenotype, the OMOP phenotype, changes in data, and issues in the databases. Using the OMOP query as a standard comparison revealed differences in the original implementations despite starting from the same definitions, code lists, flowcharts, and pseudocode.
Conclusion: Using a common data model can dramatically speed phenotype implementation at the cost of having to populate that data model, though this will produce a net benefit as the number of phenotype implementations increases. Inconsistencies among the implementations of the original queries point to a potential benefit of using a common data model so that actual phenotype code and logic can be shared, mitigating human error in re-interpretation of a narrative phenotype definition.

* Corresponding author at: Department of Biomedical Informatics, Columbia University Medical Center, 622 W 168th Street, PH20 New York, NY 10032, United States.
E-mail address: hripcsak@columbia.edu (G. Hripcsak).

https://doi.org/10.1016/j.jbi.2019.103253
Received 26 March 2019; Received in revised form 11 July 2019; Accepted 16 July 2019
Available online 17 July 2019
1532-0464/

1. Introduction

The availability of electronic health record data and availability of claims data are driving observational comparative effectiveness research, patient-level prediction, and genome- and phenome-wide association analyses. All of these studies require some mapping of raw clinical data to phenotypes of interest. Many of them employ knowledge engineering or, more recently, machine learning to generate queries that map raw data (diagnoses, medications, laboratory tests, etc.) to phenotype definitions (outcomes like heart failure, exposures like a period of time on a medication). Phenotype definitions and their corresponding data representations are not straightforward and are difficult to share [1–5]. Yet sharing is important, both sharing the exact definition among sites in multisite studies and sharing phenotype knowledge from one study to the next.

The Electronic Medical Records and Genomics (eMERGE) Network began in 2007 with a goal to develop, disseminate, and apply approaches to research that combine biorepositories with electronic health record systems for genomic discovery and genomic medicine implementation research [6,7]. A network of sites was assembled, each of which collected a sample of participants for whom they had health record data and genotype (now sequence) data. Each site generates a hypothesis centered on a phenotype of interest, designs the study protocol, shares it for execution across the network, and gathers the results for interpretation and publication. Phenotypes are usually represented by a narrative definition, a list of codes for clinical concepts, and often pseudocode and a flowchart to illustrate the logic. Each site maps this definition to a query appropriate to their local data warehouse. One of the major bottlenecks in the network's research was found to be the implementation of phenotypes [1], both creating the original phenotype definition and having each network site rewrite the definition to run against a local data repository. Several eMERGE projects have studied these challenges and taken on increasing the efficiency and accuracy of generating and sharing phenotypes. Examples include studying and improving the overall phenotyping process [1,2], evaluating phenotype definition complexity [3], assessing vocabulary mappings [8–10], exploiting repeatable patterns to improve development [11], using a formal specification language [12], pulling in other forms of data to improve accuracy [13–15], use of machine learning [16], and using a phenotyping framework to improve portability [17].

Observational Health Data Sciences and Informatics (OHDSI) [18] is an international open science initiative focused on generating evidence from observational data. It comprises a data model, vocabularies, statistical methods, software, user interfaces, a network of databases, and clinical researchers. Its data model, the Observational Medical Outcomes Partnership (OMOP) Common Data Model [19] retains the name of OHDSI's predecessor. OMOP has a data schema organized around the major types of structured information, such as patient demographics, visit information, conditions (diagnoses), drugs, procedures, measurements (laboratory results and vital signs), and observations (a flexible table to hold other data types). The schema is optimized for very-large-scale observational research. OHDSI maintains mappings among 100 international vocabularies, storing data with codes from its "standard" vocabularies, such as SNOMED CT [20], RxNorm [21], and LOINC [22], as well as preserving the original codes. Using this data model or one from other recent collaborative networks [23–25] should allow for executable phenotype queries to be distributed and run at each site without a need for reimplementation and the delay and variability that such efforts introduce. Despite this obvious solution, we know of no publications that assess the actual effect on phenotyping after implementing a common data model across a large disparate network where each institution started with a different local data model. In this paper, we describe the experiences of the eMERGE network in the evaluation of OMOP as a common phenotyping platform and fill in this gap.

2. Methods

The National Human Genome Research Institute awarded a supplement in 2016 for the ten eMERGE sites to convert data to a common data model. The OHDSI OMOP Common Data Model [19] was chosen, having been used in other large research efforts and having been shown to support large-scale studies [18].

The Phenotype Knowledgebase (PheKB) is an online repository that holds the eMERGE phenotype definitions, including the narrative description, clinical concept lists, flowchart, and pseudocode [26]. As shown in Fig. 1 (left side), in the original (pre-OMOP) implementation of each phenotype, each study site reviewed the central phenotype definition and wrote a local phenotype query to run on their local clinical research database structure, returning the results to the study initiator.

In this study, sites were asked to convert their local clinical research database to the OMOP Common Data Model [19]. Using supplied tools and example extract-transfer-load (ETL) software (bottom of Fig. 1), they moved their structured data for eMERGE subjects into a local physical OMOP instance, including relevant demographics, visits, conditions, drugs, procedures, laboratory tests, and other relevant structured information loosely grouped as observations. We limited the scope to structured data; sharing phenotype definitions with narrative elements is a separate current eMERGE research initiative. Sites had to move the data into a physical OMOP data schema, carrying out OHDSI's vocabulary translations using supplied mappings so that ICD9-CM [27] and ICD10-CM [28] mapped to SNOMED CT, drug codes mapped to RxNorm, and laboratory codes mapped to LOINC. Procedure codes were kept in their original form (e.g., ICD9-CM, ICD10-PCS, CPT), and the original codes for all data types were retained in the schema as source values. The process to do the conversion is as follows. Sites first characterize their local database, surveying what vocabularies they use for each of the above clinical data types. A set of OHDSI tools assist in this task. For coded data, OHDSI supplies direct mappings to the standard vocabularies. For non-coded data, OHDSI supplies a tool to assign codes, but the process is laborious. The data are moved into OMOP tables, and several derivative tables are then generated automatically. Data quality assurance tools assist in reviewing the database.

Sites also installed an open-source software stack called Atlas [29] that includes OHDSI graphical tools for reviewing and curating data and for generating and running phenotypes (called "cohorts" in Atlas). Atlas supports complex phenotype definitions that can exploit multiple ontologies and that can specify intricate temporal relationships and logic. These definitions can be exported from Atlas as JSON code for direct use by other sites using Atlas or in a number of different SQL formats for sites not using Atlas (right side of Fig. 1).

Sites were then asked to run two demonstration phenotypes using OMOP. The first phenotype was type 2 diabetes mellitus (T2DM) [30]; it was chosen because it used a broad variety of structured data types with complex logic but did not rely on narrative data. Furthermore, it was a useful phenotype that would likely be used in future research. The second phenotype was attention deficit and hyperactivity disorder (ADHD) [31], chosen for its greater simplicity and also for its reliance on only structured data. Both phenotypes had already been implemented in the eMERGE network so that sites could potentially compare their original query to OMOP results, and they were the two most frequently accessed eMERGE phenotypes on PheKB [26].

For the evaluation, we asked sites to report the dates when each phenotype was downloaded, when work started, when work was completed, and the effort to complete the query. We also asked about experiences with running the original phenotype. Concordance between the cohorts created by the original implementation of the phenotypes and the OMOP implementation were recorded. We asked sites what modifications were required to run the OMOP query locally and what were the implementation barriers: coding issues (e.g., codes

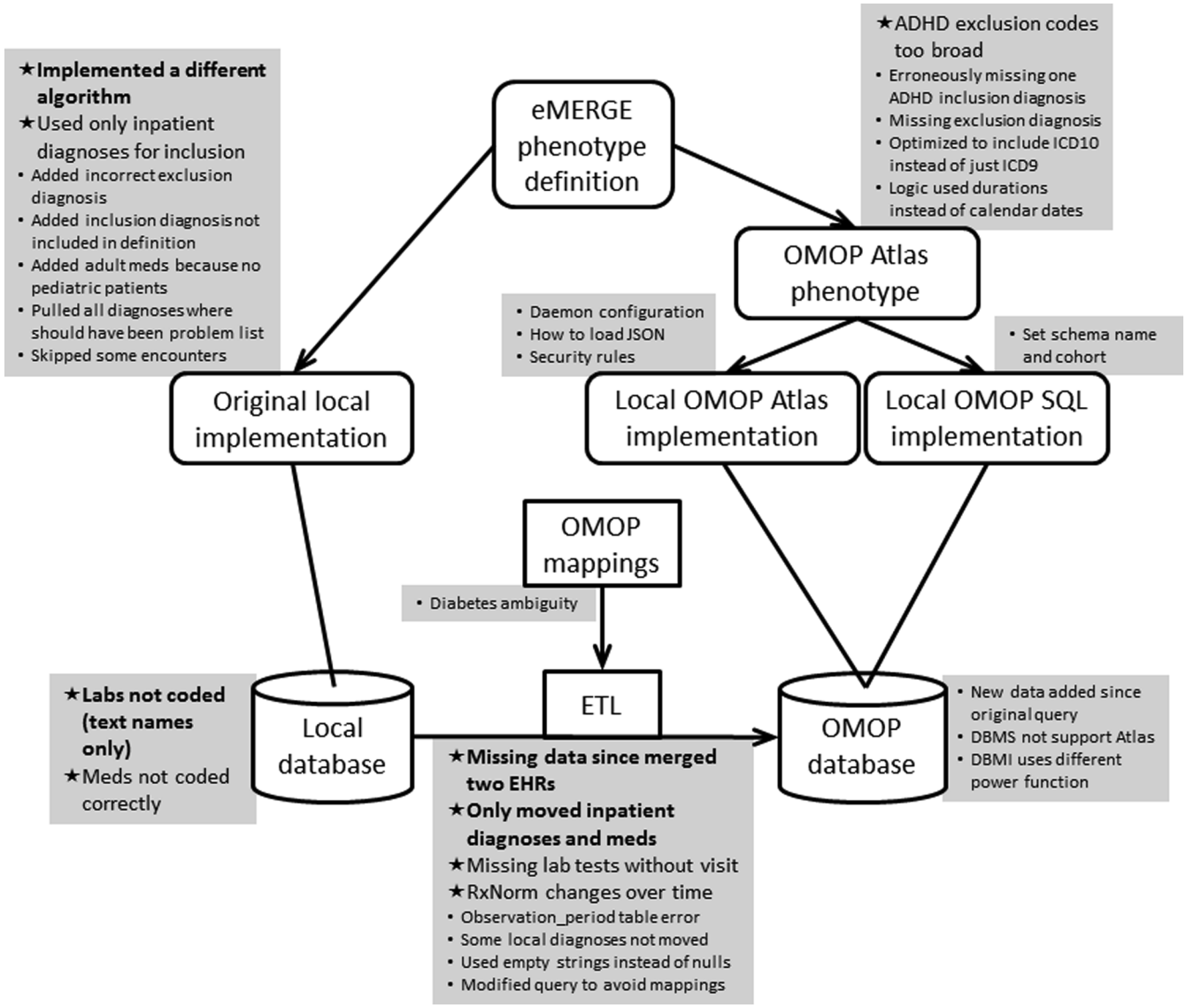

Fig. 1. Phenotyping flowchart and issues reported. Phenotypes were based on the published eMERGE phenotype definition, which included a narrative definition, high-level concept code lists, a flowchart, and pseudocode, and it was taken as the gold standard definition and therefore had no issues. The original eMERGE implementation had each local site write software to query their local database based on the eMERGE definition. In this study, each site converted its data to the OHDSI OMOP database using extract-transfer-load (ETL) software and OMOP vocabulary mappings. The eMERGE definition was encoded as a single OMOP phenotype using the OHDSI Atlas tool. A site could use a local copy of Atlas running on their OMOP database to run the phenotype or use SQL that was generated automatically by Atlas for five database management systems. Where possible, the OMOP result of the phenotype query was compared to the original result. Sites reported issues that they encountered, which are shown in grey squares adjacent to the most relevant step. Issues that caused a significant difference between the original and OMOP query are marked with a star in bold. Those that caused a moderate difference are marked with a star in non-bold. Issues that caused little or no change are marked with a smaller round bullet in a smaller font.

missing from the supplied query), data issues (e.g., site does not collect a required data type), query issues (e.g., query had syntactic errors), database management system issues (e.g., database management system generated errors), software stack (e.g., software stack errors), organizational issues (e.g., insufficient staffing), and all other issues. For each issue, sites were asked to identify what went wrong and what could be done to fix it.

For each phenotype, we report time to implement and complete the query; prevalence according to the OMOP query; concordance between the original query and the OMOP query as number of cases returned by only one, only the other, both, and neither; and we summarize the reported issues. We calculated positive specific agreement (PSA) and negative specific agreement (NSA) [32]. In addition, issues were categorized roughly on their impact on the concordance between the queries, with major issues causing a 2% or more difference in the sample, minor issues causing a difference less than 0.2%, and moderate issues in between. Non-quantitative reports were judged on their likely level of impact (e.g., if something is reported as the "main" issue on a phenotype with a large difference between queries, it would be categorized as major).

3. Results

All ten sites reported successfully implementing the OMOP data schema, populating the schema with eMERGE data, and converting the coded data to the OHDSI standard terminologies, completing the task in 4 to 12 months (elapsed time, not total effort). Two sites, however, reported that the conversion of laboratory and procedure data was ongoing. Five sites reported installing the OHDSI stack successfully, one installed a partial stack, one could not install it because the site's

Table 1
Phenotypes by site and agreement with original phenotypes.

Phenotype	Site	Time (days)	Prevalence	Overlap between original and OMOP phenotypes (number)				Positive specific agreement	Negative specific agreement	Issues*
				Overlap	Original only	OMOP only	Neither			
T2DM	1	< 1	0.008	38	0	4	5465	0.950	1.000	OMOP query included "Diabetic oculopathy" even though it includes both type 1 and type 2 diabetes, leading to 3 FPs; original eMERGE query implementation differed from logic in original eMERGE description on how to handle problem list; noted need for Observation table
T2DM	2	< 1	0.224	1179	95	30	4086	0.950	0.985	Original eMERGE query included "Diabetic macular edema" for type 1 diabetes even though it includes type 2; OMOP query missing "Diabetic oculopathy due to type I diabetes mellitus" for type 1 diabetes; database missing some lab codes; changes in RxNorm mappings over time
T2DM	3	23	0.087	242	381	250	4804	0.434	0.938	Missing and incorrect LOINC codes in database; reported that error was mainly in original eMERGE query
T2DM	4	< 1	0.003							Did not have full labs in database
T2DM	5	< 1	0.038	735	1165	18	396	0.554	0.401	Erroneous and missing source data due to EHR merger
T2DM	6	< 1	0.108							(No additional issues)
T2DM	7	144	0.191	3139	819	1588	19,143	0.723	0.941	Drugs in database were not coded so rewrote medication component; difficulty debugging without OHDSI stack; did NOT compare to the original eMERGE query but to a site-specific query
T2DM	8	< 1	0.022							OMOP vocabulary files were loaded as empty string instead of null
T2DM	9	< 1								(No additional issues)
ADHD	1	< 1	0.001	7	0	0	5500	1.000	1.000	(No additional issues)
ADHD	2	< 1	0.004	23	11	1	5355	0.793	0.999	Some coding issues around descendants in OMOP query, e.g., exclusion concept "Organic mental disorder" had concept "Postconcussion syndrome" as a descendant but should not be excluded.
ADHD	3	< 1	0.003							(No additional issues)
ADHD	4	< 1	0.123	1761	507	48	12,282	0.864	0.978	Drugs for alternate visit types not migrated to OMOP database (388 of 507 mismatches); original query added ADHD "predominantly inattentive type" not in definition (187 of 507 mismatches); OMOP query did not distinguish in-person encounters but original eMERGE query used only in-person encounter inclusion codes and all exclusion codes; missing some exclusions in OMOP query compared to definition; original eMERGE query did not include ICD10-CM exclusion diagnoses; OMOP query used year count instead of exact days for age requirement
ADHD	5									Did not have permission for retrospective mental health
ADHD	6	84	0.017	65	15	19	4861	0.793	0.997	Additional medications were added to the original eMERGE query because the site had only adults; original query added "Attention deficit hyperactivity disorder, predominantly inattentive type"; two years of data were added to the database for the OMOP query
ADHD	7	< 1	0.008							(No additional issues)
ADHD	8	14	0.002							Rewrote OMOP query to use ICD codes instead of SNOMED CT
ADHD	9	< 1								Lab codes were manually mapped to LOINC

* Most sites additionally commented on the need to map the schema names in the OMOP query to local schema names and to create an empty cohort definition.

database management system was not supported, one site could not get security clearance for its installation, and the rest are still reported as in progress.

Nine sites carried out the phenotyping exercise (Table 1). Seven of nine sites for T2DM and six of eight for ADHD completed work on the first day they started, and the others took 14 to 144 days of elapsed time. The longer times were either because the task was put aside for other work or because an error in the database was found and data were reloaded. The time to complete the original eMERGE version of the phenotypes queries was not recorded by the sites. Based on recent network phenotypes, however, elapsed time is reported as two to three weeks for phenotypes without natural language processing, with some taking shorter or longer depending on the complexity of the algorithm and availability of the structured data required.

Prevalence of the conditions based on the OMOP query varied widely, from 0.3% to 22.4% for T2DM and 0.1% to 12.3% for ADHD. This is in part due to the vast differences in how eMERGE cohorts were assembled. For example, some sites have only pediatric patients and some have none. Some cohorts were assembled from patients with specific diseases, such as renal failure.

Only 50% of the potential comparisons between OMOP and the original query were carried out, with the following reasons for failure to compare: the site joined the network after that original phenotype was implemented (that is, the site never implemented the phenotype on their local data model, and the work to do it now on their local data model was too great), the site did not execute the original phenotype because of a low expected case count, the original phenotype results were lost, a recent change in privacy policy prevented running the OMOP phenotype, or the reason was not stated. Of the sites that could compare the current OMOP phenotype implementation with the original eMERGE phenotype implementation, agreement varied from 100% to 43%. For T2DM, site 3 had more cases returned by one query or the other than returned in common; it reported that the primary error appeared to be in the original eMERGE query, not the OMOP query. Site 5 reported a new major data source issue unrelated to the query. Site 7 reported that the original query was not the eMERGE query but a site-specific version that had been used early in the network's history; both algorithms had been evaluated and had high positive predictive values. For ADHD, the three sites with imperfect agreement (79% to 86% in sites 2, 4, 6) reported issues related to the inclusion or exclusion of codes, differences in inclusion of visit types, and coding of medication data in the database. While each site reported somewhat different specific issues, this is not surprising because under the original implementation, each site developed its own specific query based on an eMERGE description, concept lists, flowchart, and pseudocode. For example, the original ADHD pseudocode supplied drug names, and each site could potentially map them to different local codes. This led to different implementations and potentially different errors.

The issues are summarized in Fig. 1, shown grouped together next to the most relevant step in the study execution. Most issue reports were not quantitative, so categorization as major or minor is approximate. Nevertheless, the trends were clear: Getting the system running and fixing the phenotype logic were generally minor issues. The major issues were related to the data—including data conversion—and the concept encodings, and we observed the original queries had at least as many issues as the OMOP queries. The fidelity of the OMOP concept mappings between vocabularies was only a minor issue, but the selection of the right concepts in the target vocabulary was important. Many of the differences related to decisions not to move a site's entire eMERGE database to the OMOP instance; in some cases this was due to lack of coding in the local database (usually laboratory test names).

4. Discussion

The main findings of this common data model phenotyping exercise were related to efficiency and consistency. Efficiency was high, with most sites implementing each phenotype query within a day. This is not surprising given that each site put in the effort ahead of time to adopt a common data model, allowing quick implementation of the query. This contrasts with several weeks usually needed to implement a phenotype at each site when the data models differ. The savings are offset by the work needed to populate the common data model beforehand. This latter effort generally took months of elapsed time. As a rule of thumb in OHDSI, a site without previous OMOP experience will have an analyst take two months to convert a large database to OMOP, and then the site iterates to improve the conversion as errors are found or as new data are supplied. This study supported approximately 40% of one year of an analyst to populate the data model and run and evaluate the test queries, supporting the OHDSI estimate. Given this, a likely breakeven point for time saved on queries versus time spent on implementing the model is 10 to 20 phenotypes (based on two weeks for a phenotype and four months for the database, with a large buffer). As research moves to larger-scale studies with many phenotypes, this return on investment may be reached quickly. There are other issues, such as the costs to support another database as well as potential benefits such as consistency in networks like eMERGE and ability to participate in additional networks.

Consistency varied by site and phenotype, as measured by agreement between the original phenotype implementation and the OMOP phenotype implementation. The cause, however, was not simply the adoption of the common data model. The original phenotype was run in the past for many sites, and newly accumulated data on the same patients could change their statuses (that some sites could not easily rerun their original phenotype queries shows the brittleness of local non-standard implementations). When sites noted a difference between the shared OMOP query and their own implementation of the original eMERGE query, each site reported a different way in which the OMOP version differed from their original version. That is, despite being given the same original phenotype definition, high-level concept code list, flowchart, and pseudocode, each site produced a somewhat different original implementation for their own local data model. While we did not test this directly, it is possible that adopting a common data model would produce more consistent phenotype query implementations because the same query code would be executed at each site.

Some differences were due to instances of missing or included codes in the OMOP or the original phenotype query. The good news with a common data model is that as such errors are discovered, the query can be updated and rerun across the network without significant effort. Furthermore, the quick turn-around was beneficial in implementation and debugging. Because the data model, a public version of the Atlas tools, and sometimes a local version of the Atlas tools were shared, when a site had a question about a query, the query authoring site and implementing site could converse in real time, reviewing the query on the Internet and alternately running different versions in their respective databases in real time.

Other differences occurred because the data in the common data model were incomplete, miscoded, or not coded (Table 1, Fig. 1). The OMOP data schema with a separate table for patients, visits, diagnoses, procedures, drugs, lab tests, etc. is relatively straightforward. Medication data are converted to RxNorm codes, which most sites had some experience in implementing. Diagnosis codes are usually available as ICD9-CM or ICD10-CM codes, and OMOP maps them to SNOMED CT. The mappings are supplied, so the process is straightforward and usually automated as part of the extract-transform-load process used to populate the OMOP tables. Such mapping can lead to errors because they are not always 1-to-1. Nevertheless, a previous study of eMERGE phenotypes showed that high accuracy was possible in the nine phenotypes studied, with a maximum error rate of 0.26% in hand-written queries like ours [9], and we found only a minor issue in this study (Fig. 1). Laboratory data turned out to be the greatest challenge. Most sites had not encoded their laboratory data as LOINC codes and instead used local codes or text strings, often accumulated over many years. Converting thousands of laboratory tests to LOINC is a substantial

undertaking, and many sites chose to map laboratory data as needed for each phenotype.

The issues that caused the most inconsistency appeared to be related to data in various forms—data sources, data coding, data inclusion, and concept selection—as opposed to problems in the query logic or software. This corroborates our phenotyping research from a quarter century ago, which found that sharing medical logic was mainly constrained by data issues [33]. Many of the differences were due to conscious decisions to either alter the original query or to limit what data got placed in the OMOP database. Some may have been seen as optimizations or as time savings for the database conversion (e.g., insofar as the conversion was seen as a mere exercise for the study).

As noted above, the eMERGE Network has been studying EHR-based phenotyping since its inception [1–3,8–17]. The study closest to ours was done by Pacheco et al. [17] using Phenotype Execution and Modeling Architecture (PhEMA) in the eMERGE network. The basic design was an authoring tool with connectors to each local database. They ran the benign prostatic hypertrophy phenotype over seven eMERGE network sites and one other, comparing the results to the original implementation. Of five sites that had the information, three had specific agreements over 90% and the other two were 68% and 70%, and the newer version had higher positive predictive values than the originals on average. This confirms our findings that agreement can vary but that the fault may lie in the original. This approach complements ours by leaving data in its original form but building connectors. Many of the issues observed in that study related to the implementation and execution of the connectors, and long-term maintenance of the connectors is likely to add to the challenge. The OMOP approach places the data in a common format, which essentially shifts the work to the population of the database instead of the connectors, and maintenance is shifted to the extract-transfer-load code; therefore, both are subject to similar issues. Advantages of the OMOP approach are that it is a standard used in other research efforts so the workload can be shared among disparate projects and that some of the performance challenges seen in the PhEMA project are avoided with OMOP's physical database. The two approaches can be combined, using the PhEMA tools on top of the OMOP database, and in fact the Columbia site used that approach; its 100% overlap reflected the use of the OMOP database for both queries.

Among the desiderata identified for computable representations of phenotype algorithms [2], the move to OHDSI OMOP followed nine out of ten: structuring the clinical data for querying, use of a common model where shared representations are possible (e.g., not yet supporting narrative data), supporting readable and computable representations (through the Atlas tool), use of set and relational operators, representing criteria with structured rules, supporting temporal relations between events, using standard terminologies, providing application programming interfaces for external software algorithms (through the OHDSI stack), and maintaining backward compatibility (e.g., by supporting ICD9-CM). The one not covered was the definition of representations for text searching and natural language processing. While narrative data were not covered by the study, they have been demonstrated to improve performance in phenotyping [13–15]. Nevertheless, much phenotyping work still uses structured data; a recent review of the PheKB eMERGE phenotype knowledge base revealed that 50 of 92 phenotypes used no narrative data [26]. Even the phenotypes that use narrative data can benefit from more efficient and consistent sharing of their structured components. Furthermore, the lessons here may be applicable to future narrative work. For example, natural language processing systems often produce SNOMED CT codes or UMLS codes that are convertible to SNOMED CT. The eMERGE network is currently studying the transportability of natural language processing and will use the OMOP format for its output.

In addition to the limitation of not using narrative data, we only assessed two phenotypes due to the funding and time constraints of the study. While this number limits generalizability to other phenotypes, we believe that the issues that arose are generally applicable, and we note that roughly the same issues arose for both phenotypes despite their difference in complexity and content area.

5. Conclusions

In summary, all sites successfully converted their data to a common data model. Once that work was put in, implementing phenotype queries across the network was more efficient; we estimate a breakeven point to be 10 to 20 phenotypes. Agreement between the original and OMOP phenotype query results varied sometimes because of the original query, sometimes because of the OMOP query, and sometimes because of the data. The experiment revealed inconsistencies among the implementations of the original queries despite working from the same definitions, code lists, flowcharts, and pseudocode, pointing to a possible benefit of a common data model.

IRB

IRB approvals were obtained for this study as part of the eMERGE initiative

Declaration of Competing Interest

None reported.

Acknowledgments

Funding

This phase of the eMERGE Network was initiated and funded by the NHGRI through the following grants: U01HG008657 (Group Health Cooperative/University of Washington); U01HG008685 (Brigham and Women's Hospital); U01HG008672 (Vanderbilt University Medical Center); U01HG008666 (Cincinnati Children's Hospital Medical Center); U01HG006379 (Mayo Clinic); U01HG008679 (Geisinger Clinic); U01HG008680 (Columbia University Health Sciences); U01HG008684 (Children's Hospital of Philadelphia); U01HG008673 (Northwestern University); U01HG008701 (Vanderbilt University Medical Center serving as the Coordinating Center); U01HG008676 (Partners Healthcare/Broad Institute); U01HG008664 (Baylor College of Medicine); and U54MD007593 (Meharry Medical College).

In addition, this work was funded by R01LM006910, Discovering and applying knowledge in clinical databases; R01HG009174, Developing i2b2 into a Health Innovation Platform for Clinical Decision Support in the Genomics Era, OT2OD026553, The New England Precision Medicine Consortium of the All of Us Research Program. Vanderbilt University Medical Center's BioVU is supported by numerous sources: institutional funding, private agencies, and federal grants, including the NIH funded Shared Instrumentation Grant S10RR025141; and CTSA grants UL1TR002243, UL1TR000445, and UL1RR024975.

References

[1] Newton KM, Peissig PL, Kho AN, Bielinski SJ, Berg RL, Choudhary V, Basford M, Chute CG, Kullo IJ, Li R, Pacheco JA, Rasmussen LV, Spangler L, Denny JC. J Am Med Inform Assoc. 2013 Jun;20(e1):e147-54. Validation of electronic medical record–based phenotyping algorithms: results and lessons learned from the eMERGE network. J Am Med Inform Assoc 2013;20(e1):e147–54.

[2] H. Mo, W.K. Thompson, L.V. Rasmussen, J.A. Pacheco, G. Jiang, R. Kiefer, Q. Zhu, J. Xu, E. Montague, D.S. Carrell, T. Lingren, F.D. Mentch, Y. Ni, F.H. Wehbe, P.L. Peissig, G. Tromp, E.B. Larson, C.G. Chute, J. Pathak, J.C. Denny, P. Speltz, A.N. Kho, G.P. Jarvik, C.A. Bejan, M.S. Williams, K. Borthwick, T.E. Kitchner, D.M. Roden, P.A. Harris, Desiderata for computable representations of electronic health records-driven phenotype algorithms, J. Am. Med. Inform. Assoc. 22 (6) (2015 Nov) 1220–1230, https://doi.org/10.1093/jamia/ocv112.

[3] M. Conway, R.L. Berg, D. Carrell, J.C. Denny, A.N. Kho, I.J. Kullo, J.G. Linneman, J.A. Pacheco, P. Peissig, L. Rasmussen, N. Weston, C.G. Chute, J. Pathak, Analyzing the heterogeneity and complexity of electronic health record oriented phenotyping algorithms, AMIA Ann. Symp. Proc. 2011 (2011) 274–283.

[4] G. Hripcsak, D.J. Albers, High-fidelity phenotyping: richness and freedom from bias, J. Am. Med. Inform. Assoc. (2017), https://doi.org/10.1093/jamia/ocx110.
[5] G. Hripcsak, D.J. Albers, Next-generation phenotyping of electronic health records, J. Am. Med. Inform. Assoc. 20 (2013) 117–121, https://doi.org/10.1136/amiajnl-2012-001145.
[6] C.A. McCarty, R.L. Chisholm, C.G. Chute, I.J. Kullo, G.P. Jarvik, E.B. Larson, R. Li, D.R. Masys, M.D. Ritchie, D.M. Roden, J.P. Struewing, W.A. Wolf, eMERGE team. The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies, BMC Med. Genom. 26 (4) (2011) 13.
[7] O. Gottesman, H. Kuivaniemi, G. Tromp, W.A. Faucett, R. Li, T.A. Manolio, S.C. Sanderson, J. Kannry, R. Zinberg, M.A. Basford, M. Brilliant, D.J. Carey, R.L. Chisholm, C.G. Chute, J.J. Connolly, D. Crosslin, J.C. Denny, C.J. Gallego, J.L. Haines, H. Hakonarson, J. Harley, G.P. Jarvik, I. Kohane, I.J. Kullo, E.B. Larson, C. McCarty, M.D. Ritchie, D.M. Roden, M.E. Smith, E.P. Böttinger, M.S. Williams, eMERGE Network. The electronic medical records and genomics (eMERGE) network: past, present, and future, Genet. Med. 15 (10) (2013) 761–771.
[8] J. Pathak, J. Wang, S. Kashyap, M. Basford, R. Li, D.R. Masys, C.G. Chute, Mapping clinical phenotype data elements to standardized metadata repositories and controlled terminologies: the eMERGE network experience, J. Am. Med. Inform. Assoc. 18 (4) (2011) 376–386.
[9] G. Hripcsak, M.E. Levine, N. Shang, P.B. Ryan, Effect of vocabulary mapping for conditions on phenotype cohorts, J. Am. Med. Inform. Assoc. 25 (12) (2018) 1618–1625.
[10] J. Pathak, H. Pan, J. Wang, S. Kashyap, P.A. Schad, C.M. Hamilton, D.R. Masys, C.G. Chute, Evaluating phenotypic data elements for genetics and epidemiological research: experiences from the eMERGE and PhenX network projects, AMIA Jt Summits Transl. Sci. Proc. 2011 (2011) 41–45.
[11] L.V. Rasmussen, W.K. Thompson, J.A. Pacheco, A.N. Kho, D.S. Carrell, J. Pathak, P.L. Peissig, G. Tromp, J.C. Denny, J.B. Starren, Design patterns for the development of electronic health record-driven phenotype extraction algorithms, J. Biomed. Inform. 51 (2014) 280–286.
[12] W.K. Thompson, L.V. Rasmussen, J.A. Pacheco, P.L. Peissig, J.C. Denny, A.N. Kho, A. Miller, J. Pathak, An evaluation of the NQF quality data model for representing electronic health record driven phenotyping algorithms, AMIA Ann. Symp. Proc. 2012 (2012) 911–920.
[13] P.L. Peissig, L.V. Rasmussen, R.L. Berg, J.G. Linneman, C.A. McCarty, C. Waudby, L. Chen, J.C. Denny, R.A. Wilke, J. Pathak, D. Carrell, A.N. Kho, J.B. Starren, Importance of multi-modal approaches to effectively identify cataract cases from electronic health records, J. Am. Med. Inform. Assoc. 19 (2) (2012) 225–234.
[14] V.M. Castro, D. Dligach, S. Finan, S. Yu, A. Can, M. Abd-El-Barr, V. Gainer, N.A. Shadick, S. Murphy, T. Cai, G. Savova, S.T. Weiss, R. Du, Large-scale identification of patients with cerebral aneurysms using natural language processing, Neurology 88 (2) (2017) 164–168.
[15] K.P. Liao, A.N. Ananthakrishnan, V. Kumar, Z. Xia, A. Cagan, V.S. Gainer, S. Goryachev, P. Chen, G.K. Savova, D. Agniel, S. Churchill, J. Lee, S.N. Murphy, R.M. Plenge, P. Szolovits, I. Kohane, S.Y. Shaw, E.W. Karlson, T. Cai, Methods to develop an electronic medical record phenotype algorithm to compare the risk of coronary artery disease across 3 chronic disease cohorts, PLoS One 10 (8) (2015) e0136651.
[16] S. Yu, Y. Ma, J. Gronsbell, T. Cai, A.N. Ananthakrishnan, V.S. Gainer, S.E. Churchill, P. Szolovits, S.N. Murphy, I.S. Kohane, K.P. Liao, T. Cai, Enabling phenotypic big data with PheNorm, J. Am. Med. Inform. Assoc. 25 (1) (2018) 54–60.
[17] J.A. Pacheco, L.V. Rasmussen, R.C. Kiefer, T.R. Campion, P. Speltz, R.J. Carroll, S.C. Stallings, H. Mo, M. Ahuja, G. Jiang, E.R. LaRose, P.L. Peissig, N. Shang, B. Benoit, V.S. Gainer, K. Borthwick, K.L. Jackson, A. Sharma, A.Y. Wu, A.N. Kho, D.M. Roden, J. Pathak, J.C. Denny, W.K. Thompson, A case study evaluating the portability of an executable computable phenotype algorithm across multiple institutions and electronic health record environments, J. Am. Med. Inform. Assoc. 25 (11) (2018) 1540–1546.
[18] G. Hripcsak, P.B. Ryan, J.D. Duke, N.H. Shah, R.W. Park, V. Huser, M.A. Suchard, M.J. Schuemie, F.J. DeFalco, A. Perotte, J.M. Banda, C.G. Reich, L.M. Schilling, M.E. Matheny, D. Meeker, N. Pratt, D. Madigan, Characterizing treatment pathways at scale using the OHDSI network, Proc. Natl. Acad. Sci USA 113 (2016) 7329–7336.
[19] J.M. Overhage, P.B. Ryan, C.G. Reich, A.G. Hartzema, P.E. Stang, Validation of a common data model for active safety surveillance research, J. Am. Med. Inform. Assoc. 19 (1) (2012) 54–60.
[20] SNOMED CT. http://www.snomed.org/snomed-ct (accessed 2019 March 14).
[21] U.S. National Library of Medicine. Unified Medical Language System (UMLS) RxNorm. https://www.nlm.nih.gov/research/umls/rxnorm/ (accessed 2019 March 14).
[22] Regenstrief Institute. LOINC. https://loinc.org/ (accessed 2019 March 14).
[23] E.H. Wagner, S.M. Greene, G. Hart, et al., Building a research consortium of large health systems: the Cancer research network, J. Natl. Cancer Inst. Monogr. 35 (2005) 3–11.
[24] R.L. Fleurence, L.H. Curtis, R.M. Califf, R. Platt, J.V. Selby, J.S. Brown, Launching PCORnet, a national patient-centered clinical research network, J. Am. Med. Inform. Assoc. 21 (4) (2014) 578–582.
[25] United States Food and Drug Administration. Sentinel Distributed Database and Common Data Model. 2017; https://www.sentinelinitiative.org/sentinel/data/distributed-database-common-data-model. Accessed 2017 Jul 21..
[26] J.C. Kirby, P. Speltz, L.V. Rasmussen, M. Basford, O. Gottesman, P.L. Peissig, J.A. Pacheco, G. Tromp, J. Pathak, D.S. Carrell, S.B. Ellis, T. Lingren, W.K. Thompson, G. Savova, J. Haines, D.M. Roden, P.A. Harris, J.C. Denny, PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability, J. Am. Med. Inform. Assoc. 23 (2016) 1046–1052.
[27] International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM), National Center for Health Statistics. http://www.cdc.gov/nchs/icd/icd9cm.htm (accessed 2019 March 14).
[28] International Classification of Diseases, Tenth Revision, Clinical Modification (ICD-10-CM), National Center for Health Statistics. http://www.cdc.gov/nchs/icd/icd10cm.htm (accessed 2019 March 14).
[29] ODHSI. ATLAS. https://github.com/OHDSI/Atlas (assessed 2019 March 14).
[30] PheKB. Type 2 Diabetes Mellitus. https://phekb.org/phenotype/type-2-diabetes-mellitus (accessed 2019 March 7).
[31] PheKB. ADHD phenotype algorithm. https://phekb.org/phenotype/adhd-phenotype-algorithm (accessed 2019 March 7).
[32] G. Hripcsak, A.S. Rothschild, Agreement, the F-measure, and reliability in information retrieval, J. Am. Med. Inform. Assoc. 12 (2005) 296–298.
[33] T.A. Pryor, G. Hripcsak, Sharing MLM's: an experiment between Columbia-Presbyterian and LDS hospital, Proc. Ann. Symp. Comput. Appl. Med. Care (1993) 399–403.