Research and Applications

Effect of vocabulary mapping for conditions on phenotype cohorts

George Hripcsak,1,2,3 Matthew E Levine,1,2 Ning Shang,1,2 and Patrick B Ryan1,2,4

1Department of Biomedical Informatics, Columbia University, New York, New York, USA, 2Observational Health Data Sciences and Informatics (OHDSI), New York, New York, USA, 3Medical Informatics Services, NewYork-Presbyterian Hospital, New York, New York, USA, and 4Epidemiology Analytics, Janssen Research and Development, Titusville, New Jersey, USA

Corresponding Author: George Hripcsak, MD, MS, Department of Biomedical Informatics, Columbia University Irving Medical Center, 622 W 168th St, PH20, New York, NY 10032, USA (hripcsak@columbia.edu)

Received 27 April 2018; Revised 13 August 2018; Editorial Decision 22 August 2018; Accepted 3 September 2018

ABSTRACT

Objective: To study the effect on patient cohorts of mapping condition (diagnosis) codes from source billing vocabularies to a clinical vocabulary.

Materials and Methods: Nine International Classification of Diseases, Ninth Revision, Clinical Modification (ICD9-CM) concept sets were extracted from eMERGE network phenotypes, translated to Systematized Nomenclature of Medicine - Clinical Terms concept sets, and applied to patient data that were mapped from source ICD9-CM and ICD10-CM codes to Systematized Nomenclature of Medicine - Clinical Terms codes using Observational Health Data Sciences and Informatics (OHDSI) Observational Medical Outcomes Partnership (OMOP) vocabulary mappings. The original ICD9-CM concept set and a concept set extended to ICD10-CM were used to create patient cohorts that served as gold standards.

Results: Four phenotype concept sets were able to be translated to Systematized Nomenclature of Medicine - Clinical Terms without ambiguities and were able to perform perfectly with respect to the gold standards. The other 5 lost performance when 2 or more ICD9-CM or ICD10-CM codes mapped to the same Systematized Nomenclature of Medicine - Clinical Terms code. The patient cohorts had a total error (false positive and false negative) of up to 0.15% compared to querying ICD9-CM source data and up to 0.26% compared to querying ICD9-CM and ICD10-CM data. Knowledge engineering was required to produce that performance; simple automated methods to generate concept sets had errors up to 10% (one outlier at 250%).

Discussion: The translation of data from source vocabularies to Systematized Nomenclature of Medicine - Clinical Terms (SNOMED CT) resulted in very small error rates that were an order of magnitude smaller than other error sources.

Conclusion: It appears possible to map diagnoses from disparate vocabularies to a single clinical vocabulary and carry out research using a single set of definitions, thus improving efficiency and transportability of research.

Key words: vocabulary, terminology mapping, observational research, phenotyping

INTRODUCTION

Much observational research relies on structured data such as diagnoses, medications, procedures, and laboratory tests. Each area draws its structured codes from some combination of disparate vocabularies and local coding schemes. Diagnoses are among the most used in phenotype definitions in observational research, and in the United States, they include International Classification of Diseases, Ninth Revision, Clinical Modification (ICD9-CM)\(^1\) billing...
While it is possible to define phenotypes made of sets of concepts defined separately from each of the above vocabularies, the process is difficult because of the number of vocabularies and because query authors do not generally have access to databases with all of the vocabularies to train or test on; the result is also hard to maintain. Limiting the number of vocabularies in the phenotype definition limits the generalizability of the phenotype. For example, with only 5% of the world population, the United States can study hypotheses on only more prevalent diagnoses, treatments, and effects. Focusing only on ICD codes, as most U.S. phenotyping activities do, relies on coarse diagnostic codes and suffers from the limitations of ICD’s hierarchical organization.

The Observational Health Data Sciences and Informatics (OHDSI)6,7 initiative has produced and maintains mappings from 80 source vocabularies to a smaller set of “standard” vocabularies that are usually queried. SNOMED CT is OHDSI’s standard vocabulary for diagnoses, which are called “conditions” in the OHDSI data model. SNOMED CT was chosen for its international reach, its clinical as opposed to billing focus, its fine granularity, its extensive hierarchy, and its increasing use in clinical data entry methods such as natural language processing and problem lists. Source data are mapped to standard vocabularies, and both the mapped and source data are stored in OHDSI’s data model, commonly referred to as the Observational Medical Outcomes Partnership (OMOP) Common Data Model,9 named after OHDSI’s predecessor, OMOP.

For ICD9-CM and ICD10-CM conditions, OHDSI’s primary source of mappings is a combination of the National Library of Medicine’s Unified Medical Language System Metathesaurus Mapping Project9 and a mapping from the United Kingdom National Health Service Terminology Service. OHDSI contracts a knowledge engineering vendor (Odysseus Data Services, Cambridge, MA) to import these mappings, expand and correct them as needed, and, as appropriate, suggest additions and corrections back. Mappings for other vocabularies may have other sources or may be generated by the vendor. All mappings are freely available (OHDSI.org). Typical mappings are ICD9-CM 3-digit non-billing code 410 “Acute myocardial infarction” to SNOMED CT 57054005 “Acute myocardial infarction,” and ICD9-CM 5-digit billing code 410.00 “Acute myocardial infarction of anterolateral wall, episode of care unspecified” to SNOMED CT 70211005 “Acute myocardial infarction of anterolateral wall.” As will be seen below, mapping codes in concept sets is different from mapping codes in patient records, and several approaches are possible. A secondary goal was therefore to assess the performance of different approaches to mapping concept sets. We assessed both the effect on the codes in the concept sets and the effect on patient cohorts generated by applying those concept sets to our clinical database. See Figure 1 for an overview of the study.

Source of phenotypes
We used 9 phenotypes11–17 from the eMERGE18 initiative (Table 1). This initiative was chosen because the phenotype definitions were validated, because the phenotypes were explained in each case, thus allowing us to assess intent, and because the sets were made available on the Internet. The phenotypes were chosen based on having a predominant concept set (as opposed to, say, relying primarily on laboratory values). Our study addressed only the concept sets, not the logic that surrounds them, because our goal was specifically to study the mappings. For example, a phenotype definition may require multiple diagnosis concept sets, impose temporal constraints, combine diagnosis evidence with evidence from medications and other areas, or exploit narrative data.

Concept sets
We used several versions of concept sets to query the data (Table 2). The original ICD9-CM concept set served as a baseline, and it was run on the unmapped ICD9-CM patient data. The rest of the concept sets comprised SNOMED CT codes and were run on the mapped data. A hand-engineered SNOMED CT concept set was intended to mimic the behavior of the original ICD9-CM concept set. A second hand-engineered concept set was optimized to extend the original query author’s intent to ICD10-CM codes (and SNOMED CT codes, but we had no such data for testing). These two concept sets might not be identical because adding a SNOMED CT concept that pulls in a needed ICD10-CM code might pull in unwanted ICD9-CM codes or because a SNOMED CT concept that is needed to pull in an ICD9-CM code might also pull in an unwanted ICD10-CM code that has many more patients.
Several concept sets were generated automatically by applying OHDSI data mappings to the ICD9-CM codes in the original concept sets to generate SNOMED CT concept sets. The first version, “no descendants,” takes only the SNOMED CT concepts that are directly mapped from ICD9-CM codes in the original concept set. This works for the ICD9-CM patient data, but can miss much of the ICD10-CM data because ICD10-CM has greater granularity than ICD9-CM. If one includes descendants of the SNOMED CT concepts, one can often pull in these needed ICD10-CM codes. Therefore a second choice is to include all descendants of the SNOMED CT concepts. Because that often also pulls in many unwanted ICD9-CM and ICD10-CM codes, we also studied a pair of intermediate algorithms. They pull in the descendants of a SNOMED CT concept only if none of that concept’s children or descendants (depending on which algorithm) is also in the list. The intuition in these algorithms is that if a query author is selecting some children or descendants and not others, then there may be a reason that those other children or descendants are excluded.

We then assessed concept sets by determining whether the SNOMED CT codes included in a given set would retrieve data that were mapped from the desired ICD9-CM codes and implied ICD10-CM codes (see “Gold standards,” below). The concept sets are available in the Supplementary Materials and online (https://github.com/mattlevine22/emerge2ohdsi_information_loss_CURATED.git).

### Table 1. Original ICD9-CM definition of concept sets used in phenotypes

| Algorithm                        | Original ICD9-CM concept set |
|----------------------------------|------------------------------|
| Heart failure (HF)¹¹             | 428.*                        |
| Heart failure as exclusion       | 428.*                        |
| diagnosis (HF2)¹¹                | 250.x1, 250.x3               |
| Type-1 diabetes mellitus (T1DM)¹²| 250.x0, 250.x2               |
| Type-2 diabetes mellitus (T2DM)¹²|                             |
| Appendicitis (Appy)¹³           | 540.*                        |
| Attention deficit hyperactivity  | 314, 314.0, 314.1, 314.2,    |
| disorder (ADHD)¹⁴               | 314.8, 314.9                 |
| Cataract (Catar)¹⁵               | 366.10, 366.12, 366.13, 366.14, |
|                                 | 366.15, 366.16, 366.17, 366.18, |
|                                 | 366.19, 366.21,             |
|                                 | 366.30, 366.41, 366.45, 366.48, |
|                                 | 366.9                       |
| Crohn’s disease (Crohn)¹⁶       | 555, 555.0, 555.1, 555.2, 555.9 |
| Rheumatoid arthritis (RA)¹⁷     | 714, 714.0, 714.1, 714.2     |

¹Within a code list, “*” means one or more digits or a period; “x” means one digit.

¹Only rheumatoid arthritis also had ICD10-CM codes in its original definition, namely, M05* and M06*, and these were used in the second gold standard.

Patient cohorts

We applied the concept sets to the New York Presbyterian/Columbia University Irving Medical Center patient database. The database has over 5 million patients. It has ICD9-CM codes for about 30 years and ICD10-CM codes since October 2015. While our concept sets would also retrieve data originally stored as SNOMED CT codes, our OHDSI database did not actually include SNOMED CT codes as source concepts (we use them for natural language processing, and that has not been pulled into the database yet). We defined cohorts of patients as those who had at least one of the codes from the concept set ever in their records. This OHDSI study was approved by our institutional review board.
Gold standards
We used 2 gold standards to assess patient cohorts. The first was the simple application of the original ICD9-CM concept sets to the original ICD9-CM patient data. This gold standard reflects the basic fidelity of the translated query. The second gold standard was created by the authors to extend the query beyond ICD9-CM codes to ICD10-CM and SNOMED CT codes based on the original intent described by the phenotype definition authors. It was created by casting a broad net using mappings and search terms on the source vocabularies, enumerating every code in the hierarchies under those terms, and—code by code in all 3 vocabularies—deciding whether it matched the phenotype authors’ intent. In 2 cases, heart failure as an exclusion diagnosis and cataract, ICD9-CM codes were also added because it appeared that the query authors had missed the codes. In general, we assumed the query authors were correct unless there were other included codes that made it clear that new codes were intended and that the new codes would improve the phenotype. These gold standard concept sets were applied to the source codes (not the mapped data) in the database to create gold standard patient cohorts. For the evaluation, we counted the number of patients inappropriately included in (false positive or FP) or missing from (false negative or FN) the patient cohort generated by the new SNOMED CT concept sets compared to the patient cohorts generated from the gold standard.

RESULTS
Patient cohorts
Here we show how mapping source data affected patient cohorts and, further, how different approaches to mapping concept sets affected that performance (examples will be provided in the diagnostics section). Table 3 shows how the different concept set mapping methods performed for the specific task of mimicking what the original ICD9-CM concept sets returned for only ICD9-CM patients.

Table 2. Methods to generate concept sets from ICD9-CM concept set

| Method | Description |
|--------|-------------|
| Original (no mapping) | Original concept set. |
| ICD9 set | Original ICD9-CM concept set generated by the phenotype author. This set is always run against the patients’ original ICD9-CM terms to show what would have happened before either data or concept sets were mapped. |
| Knowledge engineered (automatically map data; manually translate concept sets) | These SNOMED CT concept sets were created by hand. They are run against data in the form of SNOMED CT terms that were generated by mapping data from ICD9-CM and ICD10-CM to SNOMED CT using the OHDSI vocabulary mappings. |
| SNOMED mimic | SNOMED CT concept set designed to mimic the original ICD9-CM concept set as much as possible, ignoring data from other vocabularies. |
| SNOMED optimize | SNOMED CT concept set designed to carry out phenotype author’s intent to ICD9-CM, ICD10-CM, and SNOMED CT. |
| Automatically generated (automatically map data and concept sets) | These SNOMED CT concept sets were generated automatically from the original ICD9-CM set using OHDSI vocabulary mappings. Like knowledge engineered, they are run against data in the form of SNOMED CT terms that were generated by mapping data from ICD9-CM and ICD10-CM to SNOMED CT using the OHDSI vocabulary mappings. |
| SNOMED no desc | SNOMED CT concept set generated by using OHDSI vocabulary mappings to map from ICD9-CM terms to SNOMED CT, not using the SNOMED hierarchy. |
| SNOMED all desc | Like “SNOMED no desc,” but includes all terms in the SNOMED CT hierarchy that are descendants of the mapped terms. |
| SNOMED desc x child | Like “SNOMED no desc,” but includes descendants for mapped terms only if none of the term’s children are also in the concept set. Can be seen as limited descendants. |
| SNOMED desc x desc | Like “SNOMED no desc,” but includes descendants for mapped terms only if none of the term’s descendants is also in the concept set. Can be seen as more limited descendants. |

Code mapping diagnostics
The following analysis is based on the “SNOMED optimize” concept set, which represents the current authors’ best effort at generating a concept set.

Error-free translations
Some phenotype concept sets such as appendicitis could be translated without inappropriately including or losing codes and therefore
without FP or FN patients. The optimized query used a single code, SNOMED CT 85189001 “Acute appendicitis,” and all its descend-ants. Similarly, Crohn’s Disease was encoded without FP or FN using two codes, SNOMED CT 34000006 “Crohn’s disease” and 1085911000119103 “Complication due to Crohn’s disease,” and all their descendants. Heart failure as an exclusion diagnosis was straightforward without FP or FN, using 3 codes, SNOMED CT 84114007 “Heart failure,” 371037005 “Systolic dysfunction,” 3545003 “Diastolic dysfunction,” and all their descendants. Heart failure as an inclusion diagnosis—which emphasizes specificity over sensitivity—was less straightforward, including 29 SNOMED CT terms, some with and some without descendants, but could be mapped to the intended terms without FP or FN. (This latter phenotype achieved higher specificity despite more terms than heart failure as an exclusion diagnosis because each of the terms was more specific.)

Multiple source codes (ICD) to one standard code (SNOMED CT) The primary difficulty we found related to ambiguity when 2 or more ICD9-CM or ICD10-CM source codes mapped to the same SNOMED CT standard code. The difficulty arises when a phenotype concept set includes 1 of the source codes but excludes another. There is then no way in the mapped data to get them all right; some must be erroneously included or excluded. Take attention deficit hyperactivity disorder as an example. The original definition includes ICD9-CM 314.0 “Attention deficit disorder of childhood” but excludes its child, 314.00 “Attention deficit disorder without mention of hyperactivity.” Both of these terms map to SNOMED CT 192127007 “Child attention deficit disorder.” Because 314.0 is not reimbursable and should have fewer cases, it was deemed more important to exclude 314.00, which is a reimbursable code, which also meant excluding SNOMED code 192127007.

Type-1 diabetes mellitus had 8 standard codes with multiple source codes, 1 of which did cause consequential ambiguities (“consequential” here means it caused FP or FN in the patient cohorts): SNOMED CT 420662003 “Coma associated with diabetes mellitus” was mapped from ICD9-CM 250.30 “Diabetes with other coma, type II or unspecified type, not stated as uncontrolled,” 250.31 “Diabetes with other coma, type I [juvenile type], not stated as uncontrolled,” and 2 others (thus this set of type 1 and type 2 patients could not be separated after mapping). Type-2 diabetes mellitus had 8 standard codes with multiple source codes, and 2 of these caused consequential ambiguities. Cataract also had several ambiguities: ICD9-CM 366 “Cataract” was included because it was mapped to SNOMED 193570009 “Cataract,” which was needed to pull in other source codes; 1 other extra code was included, and 7 codes were excluded because of ambiguities.

Rheumatoid arthritis was similarly affected: ICD9-CM 714 “Rheumatoid arthritis and other inflammatory polyarthropathies” was included but pulled in some other inappropriate codes (they

Table 3. Performance on ICD9-CM source data mapped to SNOMED CT (FP false positive, FN false negative)

| Pheno  | #Cases | Original | Knowledge engineered | Automated concept set creation |
|--------|--------|----------|----------------------|-------------------------------|
|        |        | ICD9 set | SNOMED mimic | SNOMED optimize | SNOMED no desc | SNOMED all desc | SNOMED desc x child | SNOMED desc x desc |
|        |        |          | FP | FN | FP | FN | FP | FN | FP | FN | FP | FN | FP | FN |
| HF     | 75312  | 0        | 0  | 0  | 0  | 0  | 1262| 0  | 1054| 0  | 1054| 0  | 1054| 0  | 1054|
| HF2    | 75312  | 0        | 0  | 0  | 0  | 0  | 1262| 0  | 1054| 0  | 1054| 0  | 1054| 0  | 1054|
| T1DM   | 27861  | 0        | 0  | 0  | 0  | 0  | 108 | 943| 0  | 943| 0  | 943| 0  | 108|
| T2DM   | 125342 | 0        | 3  | 30 | 30 | 30 | 1348| 1318| 0  | 104| 34 | 0  | 104| 34 |
| Appy   | 9887   | 0        | 0  | 0  | 0  | 0  | 108 | 943| 0  | 943| 0  | 943| 0  | 108|
| ADHD   | 14399  | 0        | 0  | 0  | 0  | 0  | 1362| 1362| 1362| 1362|
| Catar  | 50879  | 0        | 0  | 50 | 0  | 74 | 0  | 2491| 0  | 80 | 0  | 80 | 0  | 80|
| Crohn  | 4679   | 0        | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  |
| RA     | 9655   | 0        | 0  | 0  | 0  | 0  | 0  | 25103| 0 | 0  | 0  | 0  | 0  | 0  |

Table 4. Performance on ICD9-CM and ICD10-CM source data mapped to SNOMED CT (FP false positive, FN false negative)

| Pheno  | #Cases | Original | Knowledge engineered | Automated concept set creation |
|--------|--------|----------|----------------------|-------------------------------|
|        |        | ICD9 set | SNOMED mimic | SNOMED optimize | SNOMED no desc | SNOMED all desc | SNOMED desc x child | SNOMED desc x desc |
|        |        |          | FP | FN | FP | FN | FP | FN | FP | FN | FP | FN | FP | FN |
| HF     | 75626  | 0        | 314| 0  | 0  | 0  | 0  | 0  | 1116| 0  | 1116| 0  | 1116| 0  | 1116|
| HF2    | 75958  | 0        | 1646| 0  | 1332| 0  | 0  | 1116| 0  | 108| 0  | 1116| 0  | 1116|
| T1DM   | 27935  | 0        | 74  | 0  | 23  | 0  | 0  | 0  | 0  | 0  | 0  | 216| 0  | 0  | 216|
| T2DM   | 126828 | 0        | 1486| 3  | 1412| 3  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  |
| Appy   | 9920   | 0        | 33  | 0  | 0  | 0  | 1359| 0  | 1359| 0  | 1359| 0  | 1359| 0  | 1359|
| ADHD   | 14547  | 0        | 148 | 0  | 39  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  |
| Catar  | 50953  | 0        | 194 | 39 | 26  | 39 | 0  | 39 | 26  | 0  | 39 | 26 | 0  | 39 | 26 |
| Crohn  | 4679   | 0        | 0  | 0  | 0  | 0  | 25150| 0  | 0  | 25150| 0  | 0  | 25150| 0  | 0  |
| RA     | 9793   | 0        | 138 | 0  | 25  | 0  | 25 | 0  | 0  | 25 | 0  | 0  | 25 | 0  | 0  |

This column is used as the gold standard and is run on unmapped source data and therefore must have perfect performance.
turned out to have no consequence), and ICD10-CM M06.4 “Inflammatory polyarthropathy” was not included because it pulled in too many inappropriate codes (it did turn out to cause some FN). Based on the definitions, we do not believe either one should have been in the original concept set, but the original authors included them, and we acquiesced. A number of codes related to specific joints were ambiguous: ICD10-CM M08.011 “Unspecified juvenile rheumatoid arthritis, right shoulder” had to be inappropriately included because it mapped to a more general SNOMED CT term for the joint, such as SNOMED CT 201766009 “Rheumatoid arthritis of shoulder”; it was not consequential.

One source code (ICD) to multiple standard codes (SNOMED CT)
In some cases, 1 source code mapped to multiple standard codes. This usually occurs because the source code is a compound concept that exists only as separate codes in the standard vocabulary. This is generally easily addressed in the mapped concept set by including a conjunction of both terms. In type-1 diabetes mellitus, ICD9-CM 250.03 “Diabetes mellitus without mention of complication, type I [juvenile type], uncontrolled” mapped to both SNOMED CT 46639007 “Type 1 diabetes mellitus” and 444073006 “Type 1 diabetes mellitus uncontrolled,” which was likely just an oversight in the mapping process. No conjunction was necessary because the first subsumes the second. Type-2 diabetes mellitus had a similar circumstance with ICD9-CM 250.02 “Diabetes mellitus without mention of complication, type II or unspecified type, uncontrolled,” which also had no consequence.

Missing OMOP codes
In some cases, source codes had no corresponding OHDSI code and therefore could not be mapped to a standard code. This generally reflected a lag between the creation of new ICD10-CM codes and their incorporation into OHDSI. Type-1 diabetes mellitus had 57 such codes at the time of the evaluation, such as ICD10-CM E10.3211 “Type 1 diabetes mellitus with mild nonproliferative diabetic retinopathy with macular edema, right eye.” We verified that the codes were new enough that they had not been used in patient care yet in our institution. Type-2 diabetes mellitus also had 57 missing codes, also with no consequence.

Information gain
We found that the mapping process also produced some benefits. For example, in heart failure as an exclusion diagnosis, the original ICD9-CM query eliminated heart failure only under ICD9-CM 428 “Heart failure.” The SNOMED CT hierarchy pulled in relevant exclusion diagnoses that were not under 428, but under other codes such as 398, 402, 404, and 415. Because the goal was for the phenotype to be specific, these codes were deemed to be relevant to an exclusion diagnosis and included in the intent gold standard.

DISCUSSION
Main finding: source data mapping produces minimal error
We found that the vocabulary mapping process produced little error in creating cohorts. For 4 of 9 phenotypes, the concept set mapping was straightforward without ambiguity. For the other 5, some number of ambiguities arose, although the number was always small compared to the number of concepts involved. For the patient cohorts, the differences were very small at a few patients per thousand (0.26%) or less. Compare that rate to the rate of erroneous diagnosis codes at 14% or the rate of entering notes on the wrong patient at 0.5%. The consequential ambiguities were always in the form of 2 or more source codes (ICD9-CM or ICD10-CM) mapping to 1 SNOMED CT code. Most of those ambiguities involved ICD9-CM codes, so we guess that over time as more billing data are encoded in ICD10-CM, the error rates will drop further.

Concept set mapping
The OHDSI vocabulary mappings were designed to map data from source vocabularies to their standard vocabularies. They were not designed to map concept sets. We found that none of our automated algorithms to map concept sets performed well, with error rates up to 10% (with 1 larger outlier). The alternative is manually translating the concept sets using knowledge engineering. In our study, the process of creating the gold standard was merged with the process of translating the concept set, and we estimate times from 1 hour to 2 days to create an optimized concept set depending on complexity.

While we found that knowledge engineering was necessary to optimize the query, we also found that the process could improve the concept sets. ICD9-CM has a strict hierarchy, so that a term can have only 1 parent; for example, an infection of an anatomical structure must be stored with infections or with the relevant structure but not with both. SNOMED CT is a multiple hierarchy and can place concepts under several parents. We found, especially for heart failure as an exclusion diagnosis, that the SNOMED CT hierarchy could identify codes that would have been missed by simply looking at the ICD-CM hierarchy.

Use of SNOMED CT
Our study demonstrates feasibility of using SNOMED CT as the basis of concept sets for accurate phenotype definitions. The use of SNOMED CT brings several benefits. International studies can use a single coding scheme for conditions and distribute studies broadly. Attempting to write every phenotype definition to accommodate ICD9-CM, ICD10-CM, ICD10, SNOMED CT, Read codes, MedDRA, etc., separately is not feasible, especially because no phenotype author will have access to patient databases with all the diagnosis codes to test the accuracy of the phenotype. As electronic health records advance, enabling clinicians to enter clinically relevant data more easily, we should see greater availability of problem lists and clinical documentation that are encoded in SNOMED CT either directly or through natural language processing. We believe that shifting the entire nation toward clinically oriented vocabularies may do more to improve research than carefully querying poorly coded billing data.

Related work
Our work corroborates previous studies about vocabulary mapping. In the study closest to ours, Reich et al. looked at ICD9-CM, SNOMED CT, and MedDRA diagnosis concept sets in OMOP, and they found the same kinds of ambiguities in mapped concept sets due to many-to-one source-to-standard mappings. They then carried out drug-outcome studies using those concept sets, pre- and post-mapping, and they found that the ambiguities caused minimal changes in the study results. Our study uses independently validated concept set definitions, uses newer versions of the OHDSI mappings, bridges ICD9-CM and ICD10-CM, and includes an optimized
knowledge engineered version of the mapped concept sets that are intended to minimize the differences before and after mapping.

In related work, Defalco et al. showed incomplete OMOP mappings among 3 drug classification schemes but did not look at the consequences of those differences. A number of studies look at coverage. For example, Cartagena et al. measured the incomplete overlap between SNOMED CT problem lists and ICD10-CM codes using National Library of Medicine SNOMED-to-ICD mappings in the Unified Medical Language System, on which the OHDSI mappings are largely based. Fung et al. look at ways to automate mappings from ICD9-CM to ICD10-CM exploiting the Centers for Medicare and Medicaid Services General Equivalent Maps (GEMs). With respect to the potential switch from billing to more clinical vocabularies, Bodenreider looked at using SNOMED CT to enter clinical concepts related to drug reactions and the possible subsequent mapping to MedDRA for research and reporting, and Elkin et al. showed that the use of a more clinically oriented terminology such as SNOMED-RT outperformed the ICD9-CM billing terminology for encoding and querying clinical text diagnoses.

Alternatives

One could carry out an analogous study going from ICD9-CM to ICD10-CM instead of SNOMED CT using Centers for Medicare and Medicaid Services mappings. We do not currently have our database encoded that way, but we expect similarly good performance, especially given that ICD9 is the precursor to ICD10. Some ambiguities do occur. For example, the attention deficit hyperactivity disorder ICD9-CM concept set includes 314.01 but excludes 314.00, but both map to the same ICD10-CM term. We still advocate for the SNOMED CT mapping for the broader reach and clinical focus.

While our study measures inaccuracies produced by data mappings and concept set mappings, it does not imply that these inaccuracies are properties of the OHDSI data model. OHDSI retains the source data so that queries can always go back to the original data if desired, and no records are lost even if mappings do not exist yet.

Limitations

This study has several limitations. Only a small number of phenotypes were studied. The size was limited by the amount of work necessary to create the gold standard based on the original authors’ intent, which represented the bulk of the work. Once that gold standard was created, the rest of the knowledge engineering followed logically. We chose phenotypes from the eMERGE set based on the presence of a concept set that steered the majority of the phenotype definition. No phenotype definitions were rejected based on performance, good or bad. A second limitation is that we did not measure the clinical accuracy of the assignment of patients in the cohort but relied on their billing codes in both gold standards. Based on our very low error rate (0.26%), however, we can reuse the original authors’ clinically derived error rates (around 5%) to infer continued good performance. A third limitation is that we cannot prove that the errors in our patient cohorts are not the patients who would have been most important in the study. Our very low error rate points against a large effect, the codes that caused errors did not seem to be especially clinically unique, and the Reich et al. study corroborated little effect. Finally, we optimized our codes for our database, which had mostly ICD9-CM codes. The optimal concept set for a database of mostly ICD10-CM codes might be different.

CONCLUSION

Mapping data from small source ICD billing codes to SNOMED CT codes produced only a very small effect on the generated patient cohorts, and one can infer that the corresponding phenotype definitions would maintain their accuracies. Only the concept sets that were hand-engineered achieved that performance; simple automated translation of concept sets did not work as well. The implication is that it should be feasible to define phenotypes with a single diagnostic vocabulary.

FUNDING

This work was funded by grants from the National Institutes of Health R01 LM006910 “Discovering and applying knowledge in clinical databases” and U01 HG008680 “Columbia GENIE (GENomic Integration with Ehr).”

Conflict of interest statement. None.

CONTRIBUTORS

All authors made substantial contributions to the conception and design of the work; drafted the work or revised it critically for important intellectual content; had final approval of the version to be published; and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

SUPPLEMENTARY MATERIAL

Supplementary material is available at Journal of the American Medical Informatics Association online.

REFERENCES

1. 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 April 24, 2018.
2. 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 April 24, 2018.
3. SNOMED CT. http://www.snomed.org/snomed-ct Accessed April 24, 2018.
4. MedDRA. Medical dictionary for regulatory activities. http://www.meddra.org/Accesed April 24, 2018.
5. RCD (Read Codes) – Synopsys. https://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/RCD Accessed April 24, 2018.
6. Hirpscsak G, Duke JD, Shah NH, et al. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. MEDINFO’15; August 19–23, 2015; São Paulo, Brazil.
7. Hirpscsak G, Ryan PB, Duke JD, et al. Characterizing treatment pathways at scale using the OHDSI network. Proc Natl Acad Sci USA 2016; 113 (27): 7329–36.
8. Overyhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a common data model for active safety surveillance research. J Am Med Inform Assoc 2012; 19 (1): 54–60.
9. Unified Medical Language System (UMLS) Metathesaurus - Mapping Projects. https://www.nlm.nih.gov/research/umlsknowledge_sources/metathesaurus/mapping_projects/index.html Accessed June 15, 2018.
10. Reich C, Ryan PB, Stang PE, Rocca M. Evaluation of alternative standardized terminologies for medical conditions within a network of observational healthcare databases. J Biomed Inform 2012; 45 (4): 689–96.
11. Newton KM, Peissig PL, Kho AN, et al. 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.
12. The eMERGE Network. Type 2 diabetes mellitus. https://phekb.org/phenotype/type-2-diabetes-mellitus Accessed April 24, 2018.
13. The eMERGE Network. Appendicitis. https://phekb.org/phenotype/appendicitis Accessed April 24, 2018.
14. The eMERGE Network. ADHD phenotype algorithm. https://phekb.org/phenotype/adhd-phenotype-algorithm Accessed April 24, 2018.
15. The eMERGE Network. Cataracts. https://phekb.org/phenotype/cataracts Accessed April 24, 2018.
16. The eMERGE Network. Crohn’s disease—demonstration project. https://phekb.org/phenotype/crohns-disease-demonstration-project Accessed April 24, 2018.
17. The eMERGE Network. Rheumatoid arthritis (RA). https://phekb.org/phenotype/rheumatoid-arthritis-ra Accessed April 24, 2018.
18. eMERGE Network. Heart failure (HF) with differentiation between preserved and reduced ejection fraction. https://phekb.org/phenotype/heart-failure-hf-differentiation-between-preserved-and-reduced-ejection-fraction Accessed April 24, 2018.
19. Hogan WR, Wagner MM. Accuracy of data in computer-based patient records. J Am Med Inform Assoc 1997; 4 (5): 342–55.
20. Wilcox AB, Chen YH, Hrpscak G. Minimizing electronic health record patient-note mismatches. J Am Med Inform Assoc 2011; 18 (4): 511–4.
21. Defalco FJ, Ryan PB, Soledad Cepeda M. Applying standardized drug terminologies to observational healthcare databases: a case study on opioid exposure. Health Serv Outcomes Res Methodol 2013; 13 (1): 58–67.
22. Cartagena FP, Schaeffer M, Rifai D, Doroshenko V, Goldberg HS. Leveraging the NLM map from SNOMED CT to ICD-10-CM to facilitate adoption of ICD-10-CM. J Am Med Inform Assoc 2015; 22 (3): 659–70.
23. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res 2004; 32 (90001): 267D–D270.
24. Fung KW, Richesson R, Smerek M, et al. Preparing for the ICD-10-CM transition: automated methods for translating ICD codes in clinical phenotype definitions. eGEMs 2016; 4 (1): 4–1211.
25. Bodenreider O. Using SNOMED CT in combination with MedDRA for reporting signal detection and adverse drug reactions reporting. AMIA Annu Symp Proc 2009; 2009: 45–9.
26. Elkin PL, Ruggieri AP, Brown SH, et al. A randomized controlled trial of the accuracy of clinical record retrieval using SNOMED-RT as compared with ICD9-CM. Proc AMIA Symp 2001; 159–63.