Developing and Evaluating Mappings of ICD-10 and ICD-10-CM Codes to Phecodes

Patrick Wu, BS[1,5,*], Aliya Gifford, PhD[1,*], Xiangrui Meng, PhD[2,*], Xue Li, PhD[2], Harry Campbell, MD[2], Tim Varley, B.Sc.[3], Juan Zhao, PhD[1], Robert Carroll, PhD[1], Lisa Bastarache, MS[1], Joshua C. Denny, MD MS[1,6], Evropi Theodoratou, PhD[2,4], Wei-Qi Wei, MD PhD[1]

[1]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, USA
[2]Centre for Global Health Research, Usher Institute of Population Health Sciences and Informatics, The University of Edinburgh, Edinburgh, UK
[3]Public Health and Intelligence SBU, National Services Scotland
[4]Edinburgh Cancer Research Centre, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, United Kingdom
[5]Medical Scientist Training Program, Vanderbilt University School of Medicine, Nashville, TN, USA
[6]Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, USA

*Authors made equal contribution to the paper

Correspondence to:
Wei-Qi Wei, MD PhD
Email: wei-qi.wei@vumc.org
Department of Biomedical Informatics
Vanderbilt University Medical Center
2525 West End Ave., Suite 1500
Nashville, TN 37203
Tel: (615)343-1956

Keywords: electronic health record; genome-wide association study; phenome-wide association study; phenotyping

Word count: 3,999

ABSTRACT

Objective:

Studies using Electronic Health Record data incorporate custom aggregations of billing codes. One such grouping is the phecode system, which was developed for phenome-wide association studies (PheWAS). Phecodes were built upon the International Classification of Diseases, version 9, Clinical Modification (ICD-9-CM). However, healthcare systems have transitioned to ICD-10/ICD-10-CM. Here we present our work on the development and validation of the mappings for ICD-10/ICD-10-CM to phecodes.

Materials and Methods:

We mapped ICD-10/ICD-10-CM codes to phecodes by matching code descriptions and using ICD-10/ICD-10-CM to SNOMED CT maps, general equivalence maps to ICD-9-CM, and a ICD-9-CM to phecode map. We assessed the coverage of the maps in two databases: Vanderbilt University Medical Center (VUMC) using ICD-10-CM and the UK Biobank (UKBB) using ICD-10. We evaluated the validity of the ICD-10-CM map by comparing phecode prevalence between ICD-9-CM and ICD-10-CM derived phecodes at VUMC and with a PheWAS.

Results:

We mapped >75% of ICD-10-CM and ICD-10 codes to phecodes. Of the unique codes observed in the VUMC (ICD-10-CM) and UKBB (ICD-10) cohorts, >90% were mapped to phecodes. Among the top ten phecodes in the cohorts, essential hypertension and hyperlipidemia overlapped. With regression analysis, we found that the ICD-10-CM map was comparable to the ICD-9-CM to phecode map. An initial PheWAS with a lipoprotein(a) (LPA) genetic variant using ICD-9-CM and ICD-10-CM maps yielded similar genotype-phenotype associations.

Conclusions:

This study introduces initial maps of ICD-10/ICD-10-CM codes to phecodes, which will enable researchers to leverage accumulated ICD-10/ICD-10-CM data for high-throughput clinical and genetic research.

BACKGROUND AND SIGNIFICANCE

Electronic Health Records (EHRs) have become a powerful resource for biomedical research in the last decade, and studies based on EHR data often leverage International Classification of Diseases (ICD) billing codes,[1] which are grouped to reflect biologically meaningful phenotypes or diseases.[2] A number of such schemas exist, including the Agency for Healthcare Research and Quality (AHRQ) Clinical Classification Software (CSS) and the phecode system developed to facilitate phenome-wide association studies (PheWAS).[3,4] Our group introduced the initial version of phecodes in 2010 with 733 distinct phenotype codes to enable an early PheWAS. Since the introduction of phecodes, many studies have used them to replicate hundreds of known genotype-phenotype associations and discover dozens of new ones, some of which have been validated in subsequent investigations.[5–16]

We created the initial version of phecodes using the International Classification of Diseases version 9 (ICD-9) system, which was developed in 1979 by the World Health Organization (WHO) to track mortality and morbidity. To improve its application to clinical billing, the United States National Center for Health Statistics (NCHS) modified ICD-9 codes to create ICD-9-CM, which was last released in 2011. The WHO later developed the 10th ICD version (ICD-10) in 1990,[17] which the NCHS used to develop ICD-10-CM to replace ICD-9-CM. Whereas countries outside of the U.S. have used ICD-10 for many years, the U.S. moved to ICD-10-CM in October 2015. Although it is based on ICD-10, ICD-10-CM contains more granular codes than ICD-10 for billing purposes.

To develop the phecode system, we combined one or more related ICD-9-CM codes into distinct diseases or traits. For example, the three ICD-9-CM codes 311 [Depressive disorder NEC], 296.31 [Major depressive disorder, recurrent episode, mild degree], and 296.21 [Major depressive disorder, single episode, mild degree] are condensed to phecode 296.2 [Depression]. With the help of clinical experts in disparate domains, such as cardiology and oncology, we have updated the phecode groupings. In addition to identifying individuals with phenotypes, phecodes allow the specification of relevant control populations for each case using lists of individuals who have similar or potentially overlapping disease states (e.g. patients with codes representing type 1 diabetes or secondary diabetes mellitus would not serve as controls for type 2 diabetes). The latest version of the phecode system consists of 1,866 hierarchical phenotype codes that map to 13,707 ICD-9-CM diagnostic codes.[18] Recently, we demonstrated that phecodes better align with diseases mentioned in clinical practice and better facilitate genomic studies than ICD-9-CM codes and the AHRQ CCS for ICD-9.[19]

Although the phecode system is effective at replicating and identifying novel genotype-phenotype associations, PheWAS using phecodes have largely been limited to using ICD-9 and ICD-9-CM. A few studies have mapped ICD-10 codes to phecodes by converting ICD-10 to ICD-9 codes, and then mapping the converted ICD-9 codes to phecodes.[7,14] However, these studies limited their mappings to only ICD-10 (not ICD-10-CM) codes and did not evaluate the accuracy of these maps.

In this paper, we developed and evaluated a map of ICD-10 and ICD-10-CM codes to phecodes. To our knowledge, this is one of the first published ICD-10-CM to phecode map. We assessed the coverage performance using data from two large EHR systems: Vanderbilt University Medical Center (VUMC) and the UK Biobank (UKBB). VUMC started using ICD-10-CM for billing in October 2015, and UKBB participants have medical records from health systems that have used ICD-10 for over two decades.[20] We then analyzed the ICD-10-CM to phecode map by comparing the phecode prevalence mapped from ICD-9-CM and ICD-10-CM in the VUMC EHR. Lastly, we performed and compared the results of an initial PheWAS using ICD-9-CM and ICD-10-CM to phecode maps.

METHODS

EHR Databases

In this study, we used data obtained from the VUMC and UKBB[20] databases. The VUMC EHR contains records of >2.5 million (with longitudinal clinical data for >1 million) unique individuals. VUMC's patient population reflects the racial makeup of the surrounding community throughout Tennessee, and U.S. individuals of European ancestry make up the majority of records (85%) within this database.

The UKBB is a prospective longitudinal cohort study designed to investigate the genetic and environmental determinants of diseases in UK adults.[20] Between 2006-2010, the study recruited 502,632 men and women aged 40-69 years. The study recorded participants' baseline measurements, including questionnaires, interviews, and anthropometric measurements. To allow follow-up of disease incidence and mortality, participants consented to allow their data to be linked to their medical records (including hospital episode statistics, general practice records, cancer registry records, and death register records). For disease classification, ICD-10 was used from April 1995 to April 2010, and the ICD-10 4th edition from April 2010 onwards.[21]

We highlight two characteristics that differ between these datasets. First is the timespan of data collection: VUMC switched to ICD-10-CM codes in October 2015, resulting in >2.5 years of data (~2015-10-01 to 2017-06-01), while the UKBB has ICD-10 data for a longer period of time (~1995-04-01 to 2015-03-31). Second is the patient type: the VUMC database includes codes for inpatient and outpatient encounters; the UKBB codes in this paper are only inpatient codes.

Mapping ICD-10-CM Codes to Phecodes

The ICD-10-CM contains 71,704 unique codes,[22] while ICD-10 contains 12,318 unique codes. We created separate maps for each system, as there are a large number of ICD-10-CM codes that do not exist in ICD-10, and even codes sharing the same number may have slightly different meanings. For example, ICD-10-CM A18.6 [Tuberculosis of (inner) (middle) ear] and ICD-10 A18.6 [Tuberculosis of ear].

An ICD-10-CM code was mapped either directly or indirectly to a phecode (**Figure 1**). We mapped an ICD-10-CM code to a phecode directly (549 mappings) if the two descriptions matched each other regardless of capitalization, e.g. ICD-10-CM H52.4 [Presbyopia] and phecode 367.4 [presbyopia]. Indirect mapping (67,430 mappings) used the existing ICD-9-CM-phecode map and involved two approaches (note: due to different mapping instances that result in the same ICD-10-CM to phecode mapping, the total number of ICD-10-CM to phecode mapping instances is larger than the number of mapped ICD-10-CM codes (67,979 vs. 63,951)). In the first approach, we used the General Equivalence Mapping (GEM)[22] to map ICD-10-CM codes to ICD-9-CM codes. We then mapped the ICD-9-CM code to a phecode using the ICD-9-CM to phecode map. The mapping of ICD-10-CM Z85.3 [Personal history of malignant neoplasm of breast] to phecode 174.11 [Malignant neoplasm of female breast] illustrates this indirect approach: ICD-10-CM Z85.3 → ICD-9-CM V10.3

[Personal history of malignant neoplasm of breast] → phecode 174.11. In the second approach to indirect mapping, ICD-10-CM codes were first mapped to Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT),[23] which were then converted to ICD-9-CM using either official mappings from National Library of Medicine (NLM),[24] Observational Health Data Sciences and Informatics (OHDSI),[25] or International Health Terminology Standards Development Organisation (IHTSDO).[26] In this route, we first considered an ICD-10-CM and a SNOMED CT code identical if both codes shared the same Concept Unique Identifier (CUI) in the current Unified Medical Language System (UMLS), e.g. ICD-10-CM L01.03 [Bullous impetigo] and SNOMED CT 399183005 [Impetigo bullosa]. To map SNOMED CT to ICD-9-CM, we used one-to-one and many-to-one (where SNOMED CT code is more specific) mappings, e.g. SNOMED CT 399183005 [Impetigo bullosa (disorder)] to ICD-9-CM 684 [impetigo]. We then mapped the mapped ICD-9-CM code to a phecode using the existing ICD-9-CM map, e.g. ICD-9-CM 684 to phecode 686.2 [impetigo]. If an ICD-10-CM (e.g., I10 [Essential (primary) hypertension]) could be mapped to both a child phecode (e.g., 401.1 [Essential hypertension]) and its parent code (e.g., 401 [Hypertension]), we removed the mapping to the parent code and kept the mapping to the child phecode (i.e. 401.1) to maintain the granular meanings of the ICD-10-CM codes (since the phecodes themselves are arranged hierarchically). If an ICD-10-CM (e.g., D57.812 [Other sickle-cell disorders with splenic sequestration]) could be mapped to multiple distinct phecodes (e.g., 282.5 [Sickle cell anemia] and 289.5 [Diseases of spleen]), we kept all mappings. This latter association created a polyhierarchical nature to phecodes that did not exist in previous versions.

Phecode v1.2 ICD-10-CM code map (beta) can be found at https://phewascatalog.org/phecodes_icd10cm.

Mapping ICD-10 Codes to Phecodes

ICD-10 codes were mapped to phecodes in a similar manner to ICD-10-CM codes, with one exception. Given that a GEM does not exist between ICD-9-CM and ICD-10 codes, to map ICD-10 codes that did not have string matches to phecodes, we used SNOMED CT relationships to map an ICD-10 code to an ICD-9-CM code, which was then mapped to its corresponding phecode (**Figure 1**). When an ICD-10 code was mapped to both a child phecode and its parent phecode, we only kept the mapping to the child phecode to maintain the granular meaning of the ICD-10 code. If an ICD-10 code was mapped to multiple distinct phecodes, we kept all mappings.

Phecode v1.2 ICD-10 code map (beta) can be found at https://phewascatalog.org/phecodes_icd10

Evaluation of Phecode Coverage of ICD-10 and ICD-10-CM in UKBB and VUMC

To evaluate the phecode coverage of ICD-10 and ICD-10-CM in UKBB and VUMC, we calculated the number of codes in the two official maps, the number of those codes that mapped to phecodes, and the number of official codes used in the two EHRs that were mapped or unmapped to phecodes (**Figure 2**).

Comparison of Phecode Prevalence from ICD-9-CM and ICD-10-CM

One aim of this work was to determine whether the prevalence of phecodes mapped from ICD-10-CM is comparable to the phecodes mapped from ICD-9-CM. To accomplish this, we selected a cohort of patients in the VUMC EHR who have both ICD-9-CM and ICD-10-CM codes within two 18-month windows. Note that these codes also include "local" codes at VUMC, which are not in the official classification systems or used at outside institutions. We selected time windows to occur prior to and after VUMC's transition to ICD-10-CM. To reduce any potential confounders, a 6 month buffer was left surrounding the transition from ICD-9-CM to ICD-10-CM. Further, the ICD-10-CM 18-month window ended before VUMC switched from its locally developed EHR[27] to the Epic system. This created two 18-month windows ranging from 2014-01-01 to 2015-06-30 for ICD-9-CM codes, and 2016-01-01 to 2017-06-30 for ICD-10-CM codes (**Figure 3**). Limiting the cohort of patients to those with at least one ICD-9-CM in the first time window and at least one ICD-10-CM in the second time window resulted in a cohort of 357,728 individuals. Aside from these requirements, we did not place other limitations on the cohort, which consisted of 55.1% female and 44.9% male who were mean (standard deviation, SD) 45 (25) years old.

After defining the cohort, we extracted all distinct ICD-9-CM and ICD-10-CM codes from these two 18-month windows for each patient, along with their age at the occurrence of each code. We used Wilcoxon signed-rank test to compare the distributions of ICD-9-CM and ICD-10-CM codes. We mapped these codes to phecodes using Phecode v1.2 ICD-9-CM map [28,29] and the Phecode v1.2 ICD-10-CM code map. We split the codes into two groups to investigate the codes that were (1) mapped or (2) unmapped to phecodes.

To analyze the codes that did not map to phecodes, we first removed codes that we did not consider to be candidates for phecode mapping. These included removing all the ICD-9-CM codes starting with "V" or "E" and all the ICD-10-CM codes starting with "X", "Y", or "Z". We did not map these codes to phecodes because they deal largely not with active biological processes, but rather with healthcare processes, codes denoting "past medical history of", and external factors. Additionally, we removed mappings from procedure and local ICD-9-CM and ICD-10-CM codes.

After removing codes that were not mapped or should not be mapped to phecodes, we used logistic regression to analyze the 10,988 ICD-9-CM and 25,857 ICD-10-CM mapped codes, to determine whether phecode prevalence from ICD-9-CM and ICD-10-CM were comparable. For the analysis, we compiled a table that included two rows for each patient-phecode pair, one each for phecodes mapped from ICD-9-CM and ICD-10-CM (**Supplementary Figure 1**). Given the phecodes mapped from ICD-9-CM and ICD-10-CM, a a "case" for a phecode was a patient having at least one ICD code mapped to that phecode, and a "control" for the same phecode was a patient who did not have a code mapped to that phecode. Each patient-phecode was labeled with the ICD code source (10-CM vs. 9-CM) of that phecode. We regressed the outcome (case/control status) for each phecode against the predictors (ICD source), age (at end of each observation period), and sex). For each phecode, this produced a p-value and odds ratio (OR) indicating whether the phecode was

more likely to come from ICD-9-CM or ICD-10-CM, with OR > 1 indicating the phecode more likely comes from ICD-9-CM.

To evaluate our hypothesis that incorrectly mapped and/or mistakenly unmapped ICD-10-CM codes drove the differences observed in the analysis of the ICD-10-CM to phecode map, we selected phecodes with >200 cases that crossed Bonferroni correction. We reviewed ten of these phecodes with the highest and lowest ORs, for a total of 20 reviewed phecodes.

Comparative PheWAS analysis of lipoprotein(a) (LPA) single-nucleotide polymorphism (SNP)

To evaluate the accuracy of the ICD-10-CM to phecode map, we performed two PheWAS on a LPA genetic variant (rs10455872) using mapped phecodes from ICD-10-CM and ICD-9-CM, respectively. We chose rs10455872 as the predictor because it is associated with increased risks of developing hyperlipidemia and cardiovascular diseases.[30,31]

We used data from BioVU, the de-identified DNA biobank at VUMC to conduct the PheWAS.[32] We identified 13,900 adults (56.9 % female, 59 (15) years old in 2014), who had rs10455872 genotyped, and at least one ICD-9-CM and ICD-10-CM code in their respective time windows. We observed 86.7% AA, 12.8% AG, and 0.5% GG for rs10455872. We used the 1,632 phecodes that overlapped in the time windows for PheWAS using R PheWAS package,[18] adjusting for age, sex, and race.

RESULTS

Phecode coverage of ICD-10-CM and ICD-10 in VUMC and UKBB

Of the official ICD-10-CM codes,[22] 63,951 (89.2%) mapped to at least one phecode, with 7,881 (9.6%) mapping to >1 phecode. Of all possible ICD-10 codes, 9,354 (75.9%) mapped to at least one phecode, and 300 (3.2%) mapped to more than one phecode. For example, ICD-10-CM E13.36 [Other specified diabetes mellitus with diabetic cataract] maps to three phecodes: 366 [Cataract], 250.7 [Diabetic retinopathy], and 250.23 [Type 2 diabetes with ophthalmic manifestations]; and ICD-10-CM S12.000K [Unspecified displaced fracture of first cervical vertebra, subsequent encounter for fracture with nonunion] maps to both 733.8 [Malunion and nonunion of fracture] and 805 [Fracture of vertebral column without mention of spinal cord injury]. Whereas, ICD-10 code B21.1 [HIV disease resulting in Burkitt's lymphoma] maps to two phecodes: 071.1 [HIV infection, symptomatic] and 202.2 [Non-Hodgkin's lymphoma].

Among the 36,858 ICD-10-CM codes used at VUMC since they were adopted on October 2015, 34,793 (94.4%) codes can be mapped to phecodes. Of the 6,245 ICD-10 codes used in UKBB, 5,823 (93.2%) codes can be mapped to phecodes (**Table 1, Figure 2**). An additional 29,158 ICD-10-CM and 3,531 ICD-10 codes are mapped to phecodes, although they have not been used in the VUMC or UKBB systems. Considering all the instances of ICD-10-CM and ICD-10 codes used at each site, we generated a total count of unique codes grouped by patient and date, and a subset of these codes that are mapped to phecodes (**Table 1**). Among the total number of ICD-10-CM and ICD-10 codes used, 89.7% and 83.7%, were mapped to phecodes.

Table 1. ICD-10-CM and ICD-10 codes data summary.

	ICD-10-CM (No.)	**ICD-10** (No.)
Official classification systems		
Unique codes	71,704	12,318
Unique codes mapped	63,951 (89.2%)	9,354 (75.9%)
Official codes used in cohorts		
Unique codes	36,858	6,245
Unique codes mapped	34,793 (94.4%)	5,823 (93.2%)
Total patients (with ICD-10-CM or ICD-10 codes)	651,649	391,181

Total instances of all ICD codes	19,682,697	5,114,363
Instances mapped to phecodes	17,658,470 (89.7%)	4,279,544 (83.7%)

*ICD-10-CM data for patients derived from VUMC includes inpatient and outpatient data; ICD-10 data derived from UK Biobank.

We also show the top ten most commonly used ICD-10-CM (at VUMC) and ICD-10 (at UKBB) codes that were successfully mapped, ranked by the total number of unique patients for each code (**Tables 2 and 3**). In both EHR databases, phecode 401.1 [Essential hypertension] has the largest number of unique patients. Further, in the two databases, phecodes 401.1 and 411.4 [Coronary atherosclerosis] appear in the top ten most common phecodes.

Table 2. The top ten most commonly used ICD-10-CM codes, that map to phecodes, at VUMC

VUMC ICD-10-CM	ICD-10-CM Description	Phecode	Phecode Description	Total Unique People
I10	Essential (primary) hypertension	401.1	Essential hypertension	67,208
E78.5	Hyperlipidemia, unspecified	272.1	Hyperlipidemia	26,980
E78.2	Xanthoma tuberosum	272.13	Mixed hyperlipidemia	26,807
R05	Cough	512.8	Cough	25,680
K21.9	Gastro-esophageal reflux disease without esophagitis	530.11	GERD	23,645
J06.9	Upper respiratory infection NOS	465	Acute upper respiratory infections of multiple or unspecified sites	23,401

E11.9	Type 2 diabetes mellitus without complications	250.2	Type 2 diabetes	22,118
I25.10	Athscl heart disease of native coronary artery w/o ang pctrs	411.4	Coronary atherosclerosis	21,576
E03.9	Myxedema NOS	244.4	Hypothyroidism NOS	18,229
M54.5	Lumbago NOS	760	Back pain	17,810

*does not contain ICD-10-CM codes that start with X, Y, or Z

Table 3. The top ten most commonly used ICD-10 codes, that map to phecodes, in the UKBB

UKBB ICD-10	ICD-10 Description	Phecode	Phecode Description	Total Unique People
I10	Essential (primary) hypertension	401.1	Essential hypertension	96,867
R07.4	Chest pain, unspecified	418	Nonspecific chest pain	53,752
K44.9	Diaphragmatic hernia without obstruction or gangrene	550.2	Diaphragmatic hernia	44,162
K57.3	Diverticular disease of large intestine without perforation or abscess	562.1	Diverticulosis	42,706
E78.0	Pure hypercholesterolaemia	272.11	Hypercholesterolemia	41,640
I25.1	Atherosclerotic heart disease	411.4	Coronary atherosclerosis	40,191
R10.4	Other and unspecified abdominal pain	785	Abdominal pain	38,163

R31	Unspecified haematuria	593	Hematuria	34,934
J45.9	Asthma, unspecified	495	Asthma	34,720
Z86.4	Personal history of psychoactive substance abuse	306	Other mental disorder	33,546

Comparing phecodes from ICD-9-CM and ICD-10-CM

During the defined 18-month time windows, a cohort 357,728 patients had both ICD-9-CM and ICD-10-CM codes (**Figure 3**). The cohort had a median (interquartile range, IQR) of 8.0 (11.0) ICD-9-CM codes/individual and 8.0 (11.0) ICD-10-CM codes/individual (Wilcoxon signed-rank test p-value = 0.77). This resulted in 10,988 of 14,593 (75.3%) ICD-9-CM codes, and 25,837 of 34,261 (75.4%) ICD-10-CM codes used in our patient population that successfully mapped to 1,753 phecodes.

Removing the ICD-10-CM codes which should not be mapped (i.e. local or supplementary classification codes) left 2,065 ICD-10-CM codes that did not map to a phecode. Of these, 1,395 were X, Y, and Z codes, leaving 670 remaining codes. All of the unmapped ICD-10-CM codes in this study's cohort have <200 unique individuals (<0.1% of the total cohort), and the majority of the codes with >10 unique individuals do not represent biologically-relevant diseases. For example, 59% (287/485) of the unmapped ICD-10-CM codes represented external causes of morbidity, such as assault and injuries due to motor vehicle accidents.

To evaluate the accuracy of the ICD-10-CM map, we compared it to the most recent ICD-9-CM phecode map using logistic regression (**Figure 4**). We aimed to see whether there were any significant differences in the prevalence of phecodes in the two 18-month periods. The median (IQR) of the resulting odds ratios (OR) was 0.96 (0.80-1.16).

To determine whether incorrectly mapped or mistakenly unmapped ICD-10-CM codes drove these differences, we reviewed ten phecodes with the highest and lowest ORs (**Table 4**); results from this analysis supported our initial hypothesis. For example, phecode 250.6 [polyneuropathy in diabetes] with OR = 9.28 had 3,312 ICD-9-CM and 391 ICD-10-CM cases. We found that the difference in cases was due to the mapping of ICD-10-CM E11.42 [Type 2 diabetes mellitus with diabetic polyneuropathy] in this map. While E11.42 is accurately mapped to phecode 250.24 [Type 2 diabetes with neurological manifestations], which has 3,488 unique cases, this code can also be mapped to phecode 250.6. With this additional mapping, the difference of cases changes from +2,921 in favor of ICD-9-CM to +567 in favor of ICD-10-CM.

Table 4. The ten phecodes with the highest (A) and lowest (B) odds ratios.

A. Phecodes more commonly derived from ICD-9-CMs

Phecode	Phecode Description	ICD-9-CM cases	ICD-10-CM cases	p-value	Odds Ratio
509.3	Pulmonary insufficiency or respiratory failure following trauma and surgery	1,441	220	< 0.001	6.66
716.9	Arthropathy NOS	4,027	705	< 0.001	6.29
963.1	Antineoplastic and immunosuppressive drugs causing adverse effects	1,288	240	< 0.001	5.48
772	Symptoms of the muscles	1,123	213	< 0.001	5.29
291.8	Alteration of consciousness	2,261	450	< 0.001	5.16
509.2	Respiratory insufficiency	2,362	483	< 0.001	5.04
625	Pain and other symptoms associated with female genital organs	4,051	844	< 0.001	4.83
425.2	Secondary/extrinsic cardiomyopathies	2,623	603	< 0.001	4.74
217.1	Nevus, non-neoplastic	1,078	258	< 0.001	4.39
841	Sprains and strains of back and neck	3,286	763	< 0.001	4.36

B. Phecodes more commonly derived from ICD-10-CMs

Phecode	Phecode Description	ICD-9-CM cases	ICD-10-CM cases	p-value	Odds Ratio
306	Other mental disorder	616	10,479	< 0.001	0.06
611	Abnormal findings on mammogram or breast exam	248	2,977	< 0.001	0.09

729	Other disorders of soft tissues	407	4,319	< 0.001	0.09
1019	Other ill-defined and unknown causes of morbidity and mortality	457	4,268	< 0.001	0.10
395.3	Nonrheumatic tricuspid valve disorders	427	4,308	< 0.001	0.11
196	Radiotherapy	1,192	8,946	< 0.001	0.13
357	Inflammatory and toxic neuropathy	802	5,119	< 0.001	0.16
965	Poisoning by analgesics, antipyretics, and antirheumatics	596	3,050	< 0.001	0.20
478	Throat pain	1,979	8,501	< 0.001	0.22
842	Other sprains and strains	401	1,733	< 0.001	0.23

In the same group of phecodes, there were imbalances that we expected due to biological mechanisms. For example, phecode 625 [Pain and other symptoms associated with female genital organs] with OR = 5.05 (favoring ICD-9-CM) had 4,282 ICD-9-CM and 844 ICD-10-CM cases with only 152 overlapping cases. We expected this small number of overlapping cases due to the acute pathophysiology of diseases that are associated with this phecode, i.e., an individual who is affected by one of these diseases is not very likely to be affected again.

Comparative PheWAS analysis of LPA SNP, rs10455872

To further evaluate the ICD-10-CM to phecode map, we performed and compared the results of PheWAS for rs10455872. One PheWAS was conducted using ICD-9-CM map and another was conducted using the ICD-10-CM map. Both analyses replicated previous findings, showing significant (p-value < 0.05) positive correlations between the minor allele of rs10455872 and coronary atherosclerosis (ICD-9-CM: $p < 0.001$, OR = 1.60; ICD-10-CM: $p < 0.001$, OR = 1.60) and with chronic ischemic heart disease (ICD-9-CM: $p < 0.001$, OR = 1.56; ICD-10-CM: $p < 0.001$, OR = 1.47) (**Figure 5**).

DISCUSSION

This work demonstrates the results of mapping ICD-10-CM/ICD-10 codes to phecodes and evaluation in two databases. These results show that the majority of billed instances can be mapped to phecodes using either ICD-9-CM, ICD-10-CM, or ICD-10. Further, though many codes appear to have variations in billing between ICD-9-CM and ICD-10-CM in our study cohort, the majority do not vary greatly in their odds of mapping to phecodes (**Figure 4**). As use of ICD-10-CM/ICD-10 codes increases, so does the need for convenient and reliable methods of aggregating codes for biomedical research. Since the introduction of phecodes, many studies have demonstrated the value of aggregating ICD-9-CM codes such as is done with phecodes. This resource will allow biomedical researchers using EHR data to leverage clinical data represented by ICD-10 and ICD-10-CM codes for their studies.

As expected, the most frequent ICD-10-CM codes include hypertension, type 2 diabetes, and coronary atherosclerosis, as these are common conditions and consistent with our clinical observations at VUMC. Among the top ten codes used at UKBB and VUMC, phecodes 401.1 [Essential hypertension] and 411.4 [Coronary atherosclerosis] are the only overlapping phecodes (**Tables 2 and 3**). When broadening the analysis to parent phecodes, there is an additional overlap of phecodes that represent hyperlipidemia (272.1 [Hyperlipidemia] and 272.13 [Mixed hyperlipidemia] at VUMC vs. 272.11 [Hypercholesterolemia] at UKBB). The small number of overlapping phecodes between the two systems is most likely multifactorial. Factors such as patient population heterogeneity, data structure (inpatient and outpatient codes at VUMC vs. inpatient codes only for UKBB), and differences between ICD-10 and ICD-10-CM systems (e.g. variations in descriptions for similar phenotypes).

As seen in **Figure 2**, there 2,542 ICD-10 and 5,688 ICD-10-CM codes not observed in the respective databases and that are not mapped. Examining these unmapped codes illustrates potential areas of refinement in the phecode system. The majority of ICD-10-CM codes that were not mapped are encounter or procedural codes (e.g. codes beginning with Z, such as Z00.00 [Encounter for gynecological examination (general) (routine) without abnormal findings] or supplementary codes, such codes beginning with Y. These codes that are not a description of a phenotype or disease, should not be mapped to phecodes.

Analyzing the unmapped UKBB ICD-10 codes reveals that family history related concepts are missing from the phecode system. In addition to the anticipated unmapped codes, most of the most frequent unmapped codes involve personal or family history. This reveals a limitation of the current phecode system, and demonstrates an area of the phecodes that could be expanded.

Using either ICD-9-CM or ICD-10-CM map, PheWAS found similar significant associations and effect sizes with coronary atherosclerosis and chronic ischemic heart disease, using either code system (**Figure 5**). This analysis further shows the general accuracy of the ICD-10-CM map when compared to the current gold-standard ICD-9-CM to phecode map

The results presented here have limitations. While the data from the UKBB spans 23 years and is derived from a population recruited from the UK, the VUMC ICD-10-CM data encompass just over one year

with billing practices likely still evolving. The VUMC data also represent a single academic medical center, thereby making it difficult to compare the most frequent phecodes between databases.

Additionally, this work did not aim to thoroughly evaluate the accuracy of the mapping, as the resources we used (e.g. GEM) are assumed to be correct. Currently, if an ICD-10-CM or ICD-10 code maps to two or more unlinked phecodes, we currently keep all of the mappings. In subsequent studies, it will be important to further scrutinize these mappings to ensure accuracy through manual review. Future work involves expanding and updating the mappings to phecodes, addressing the currently unmapped codes, and extensive manual validation of mapping accuracy.

CONCLUSION

In this paper, we introduced our work on mapping ICD-10-CM/ICD-10 codes to phecodes. We provide initial maps with high coverage of EHR data in two large databases. In a PheWAS, the ICD-10-CM map performs similarly to the most recent ICD-9-CM to phecode map. These mappings will enable researchers to leverage accumulated ICD-10-CM and ICD-10 data for biomedical and clinical research in a high-throughput manner.

COMPETING INTERESTS

The authors have no competing interests to declare.

FUNDING

The project was supported by NIH grant R01 LM 010685 (A.G., J.C.D., L.B., R.C.,), R01 HL133786 (W.Q.), T32 GM007347 (P.W.), T15 LM007450 (P.W.), P50 GM115305 (P.W.), and AHA Scientist Development Grant 16SDG27490014 (J.Z.). The dataset used in the analyses described were obtained from Vanderbilt University Medical Centers BioVU, which is supported by institutional funding and by the Vanderbilt CTSA grant ULTR000445 from NCATS/NIH. This research was also conducted using the UK Biobank Resource under Application Number 10775. The work conducted in Edinburgh was supported by funding for the infrastructure and staffing of the Edinburgh CRUK Cancer Research Centre. E.T. is supported by a CRUK Career Development Fellowship (C31250/ A22804). X.M. and X.L. are supported by China Scholarship Council studentships.

CONTRIBUTORS

[Study Design]: P.W., A.G., J.C.D., W.Q.W.

[Analysis]: P.W., A.G., X.M., X.L., H.C., E.T, T.V., J.Z., J.C.D., W.Q.W.

[Literature Search]: P.W., A.G.

[Data retrieval]: A.G., X.M., X.L., E.T.

[Data interpretation]: P.W., A.G., R.C., L.B., J.C.D., E.T., W.Q.W

[Initial document draft]: P.W., A.G., J.C.D., W.Q.W

[Figure design and creation]: P.W., A.G., J.C.D., W.Q.W

[Table creation]: P.W., A.G., J.C.D., W.Q.W

All authors revised the document and gave final approval for publication

REFERENCES

1 Kirby JC, Speltz P, Rasmussen LV, *et al.* PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability. *J Am Med Inform Assoc* 2016;**23**:1046–52.

2 Wei W-Q, Denny JC. Extracting research-quality phenotypes from electronic health records to support precision medicine. *Genome Med* 2015;**7**:41.

3 Denny JC, Ritchie MD, Basford MA, *et al.* PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics* 2010;**26**:1205–10.

4 Denny JC, Bastarache L, Roden DM. Phenome-Wide Association Studies as a Tool to Advance Precision Medicine. *Annu Rev Genomics Hum Genet* 2016;**17**:353–73.

5 Cortes A, Dendrou CA, Motyer A, *et al.* Bayesian analysis of genetic association across tree-structured routine healthcare data in the UK Biobank. *Nat Genet* 2017;**49**:1311–8.

6 Gamazon ER, Segrè AV, van de Bunt M, *et al.* Using an atlas of gene regulation across 44 human tissues to inform complex disease- and trait-associated variation. *Nat Genet* 2018;**50**:956–67.

7 Nielsen JB, Thorolfsdottir RB, Fritsche LG, *et al.* Biobank-driven genomic discovery yields new insight into atrial fibrillation biology. *Nat Genet* 2018;**50**:1234–9.

8 Simonti CN, Vernot B, Bastarache L, *et al.* The phenotypic legacy of admixture between modern humans and Neandertals. *Science* 2016;**351**:737–41.

9 Diogo D, Bastarache L, Liao KP, *et al.* TYK2 protein-coding variants protect against rheumatoid arthritis and autoimmunity, with no evidence of major pleiotropic effects on non-autoimmune complex traits. *PLoS One* 2015;**10**:e0122271.

10 Rastegar-Mojarad M, Ye Z, Kolesar JM, *et al.* Opportunities for drug repositioning from phenome-wide association studies. *Nat Biotechnol* 2015;**33**:342–5.

11 Millard LAC, Davies NM, Timpson NJ, *et al.* MR-PheWAS: hypothesis prioritization among potential causal effects of body mass index on many outcomes, using Mendelian randomization. *Sci Rep* 2015;**5**:16645.

12 Ehm MG, Aponte JL, Chiano MN, *et al.* Phenome-wide association study using research participants' self-reported data provides insight into the Th17 and IL-17 pathway. *PLoS One* 2017;**12**:e0186405.

13 Liu J, Ye Z, Mayer JG, *et al.* Phenome-wide association study maps new diseases to the human major histocompatibility complex region. *J Med Genet* 2016;**53**:681–9.

14 Neuraz A, Chouchana L, Malamut G, *et al.* Phenome-wide association studies on a quantitative trait: application to TPMT enzyme activity and thiopurine therapy in pharmacogenomics. *PLoS Comput Biol* 2013;**9**:e1003405.

15 Doshi-Velez F, Ge Y, Kohane I. Comorbidity clusters in autism spectrum disorders: an electronic health record time-series analysis. *Pediatrics* 2014;**133**:e54–63.

16 Li X, Meng X, Spiliopoulou A, *et al.* MR-PheWAS: exploring the causal effect of SUA level on multiple disease outcomes by using genetic instruments in UK Biobank. *Ann Rheum Dis* 2018;**77**:1039–47.

17 Topaz M, Shafran-Topaz L, Bowles KH. ICD-9 to ICD-10: evolution, revolution, and current debates in the United States. *Perspect Health Inf Manag* 2013;**10**:1d.

18 Carroll RJ, Bastarache L, Denny JC. R PheWAS: data analysis and plotting tools for phenome-wide

association studies in the R environment. *Bioinformatics* 2014;**30**:2375–6.

19 Wei W-Q, Bastarache LA, Carroll RJ, *et al.* Evaluating phecodes, clinical classification software, and ICD-9-CM codes for phenome-wide association studies in the electronic health record. *PLoS One* 07 2017;**12**:1–16.

20 Sudlow C, Gallacher J, Allen N, *et al.* UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLoS Med* 2015;**12**:e1001779.

21 Hospital Episode Statistics data in Showcase. 12/2013. http://biobank.ctsu.ox.ac.uk/showcase/docs/HospitalEpisodeStatistics.pdf

22 2018 ICD-10 CM and GEMs. 2017.https://www.cms.gov/Medicare/Coding/ICD10/2018-ICD-10-CM-and-GEMs.html

23 SNOMED CT to ICD-10-CM Map. Published Online First: 29 February 2012.https://www.nlm.nih.gov/research/umls/mapping_projects/snomedct_to_icd10cm.html (accessed 12 Oct 2018).

24 ICD-9-CM Diagnostic Codes to SNOMED CT Map. Published Online First: 14 May 2012.https://www.nlm.nih.gov/research/umls/mapping_projects/icd9cm_to_snomedct.html (accessed 12 Oct 2018).

25 documentation:vocabulary:icd9cm [Observational Health Data Sciences and Informatics]. http://www.ohdsi.org/web/wiki/doku.php?id=documentation:vocabulary:icd9cm (accessed 12 Oct 2018).

26 SNOMED CT United States Edition. Published Online First: 2 March 2016.https://www.nlm.nih.gov/healthit/snomedct/us_edition.html (accessed 12 Oct 2018).

27 Giuse DA. Supporting communication in an integrated patient record system. *AMIA Annu Symp Proc* 2003;:1065.

28 Vanderbilt University School of Medicine. ICD-9 to PheWAS code map, version 1.2 | Center for Precision Medicine. https://www.vumc.org/cpm/center-precision-medicine-blog/icd-9-phewas-code-map-version-12 (accessed 12 Oct 2018).

29 PheWAS - Phenome Wide Association Studies. https://phewascatalog.org/phecodes (accessed 12 Oct 2018).

30 Nordestgaard BG, Chapman MJ, Ray K, *et al.* Lipoprotein(a) as a cardiovascular risk factor: current status. *Eur Heart J* 2010;**31**:2844–53.

31 Wei W-Q, Li X, Feng Q, *et al.* LPA Variants Are Associated With Residual Cardiovascular Risk in Patients Receiving Statins. *Circulation* 2018;**138**:1839–49.

32 Roden DM, Pulley JM, Basford MA, *et al.* Development of a large-scale de-identified DNA biobank to enable personalized medicine. *Clin Pharmacol Ther* 2008;**84**:362–9.

33 Denny JC, Bastarache L, Ritchie MD, *et al.* Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat Biotechnol* 2013;**31**:1102–10.

FIGURE LEGENDS

Figure 1. Mapping strategy for ICD-10 and ICD-10-CM to phecodes. Mapping methods between various code formats. The gray dashed-outline box indicates previously published work.[3,33] Acronyms used: SNOMED CT: Systematized Nomenclature of Medicine Clinical Terms. UMLS: Unified Medical Language System. GEM: General Equivalence Mappings. OHDSI: Observational Health Data Sciences and Informatics. NLM: National Library of Medicine. M:1: many to one.

Figure 2. Counts of distinct ICD-10-CM at VUMC and ICD-10 in UKBB. (A) Numbers represent number of unique ICD-10-CM codes in each category. For example, there are 34,793 unique codes (grey section) that are in the official ICD-10-CM system, observed in the Vanderbilt University Medical Center (VUMC) dataset, and mapped to phecodes. (B) Numbers represent number of unique ICD-10 codes in each category. For example, there are 5,823 unique codes (off-white section) that are in the official ICD-10 system, observed in the UK Biobank (UKBB) dataset, and mapped to phecodes .

Figure 3. Timeline of the two 18-month periods from which the ICD-9-CM and ICD-10-CM codes were analyzed at VUMC. The cohort of 357,728 patients had both ICD-9-CM and ICD-10-CM codes in the respective 18-month windows. Numbers signify median (interquartile range, IQR).

Figure 4. Distribution of odds ratios (OR) resulting from logistic regression analysis of phecode maps. To assess the general accuracy of the ICD-10-CM to phecode map, we compared it to the ICD-9-CM map with logistic regression analysis to calculate the likelihood of an individual being assigned a phecode. For the binary outcome variables, a patient was labeled a “Case” for that phecode if they had at least one code which mapped to that phecode. We regressed Case/Control status for each phecode against the ICD source (10-CM vs. 9-CM), age, and sex. Prior to plotting the distribution of the ORs, phecodes with < 15 cases in the ICD-9-CM or ICD-10-CM periods, OR < 0.1, and OR > 10, were removed before plotting the distribution of the ORs. This resulted in 1,144 phecodes from a starting set of 1,843. An OR > 1 indicates the phecode more likely comes from an ICD-9-CM code, whereas OR < 1 indicates the phecode more likely comes from an ICD-10-CM code. Red vertical lines indicate the IQR (left line: 25th percentile = 0.80; median (not shown) = 0.96; right line: 75th percentile = 1.16).

Figure 5. Comparative PheWAS of lipoprotein(a) (LPA) genetic variant, rs10455872. Coronary atherosclerosis and other chronic ischemic heart disease were top hits associated with rs10455872 in a phenome-wide association study (PheWAS) analysis conducted using ICD-9-CM (top) and ICD-10-CM (bottom) to phecode maps. Analysis was adjusted for age, sex, and race.

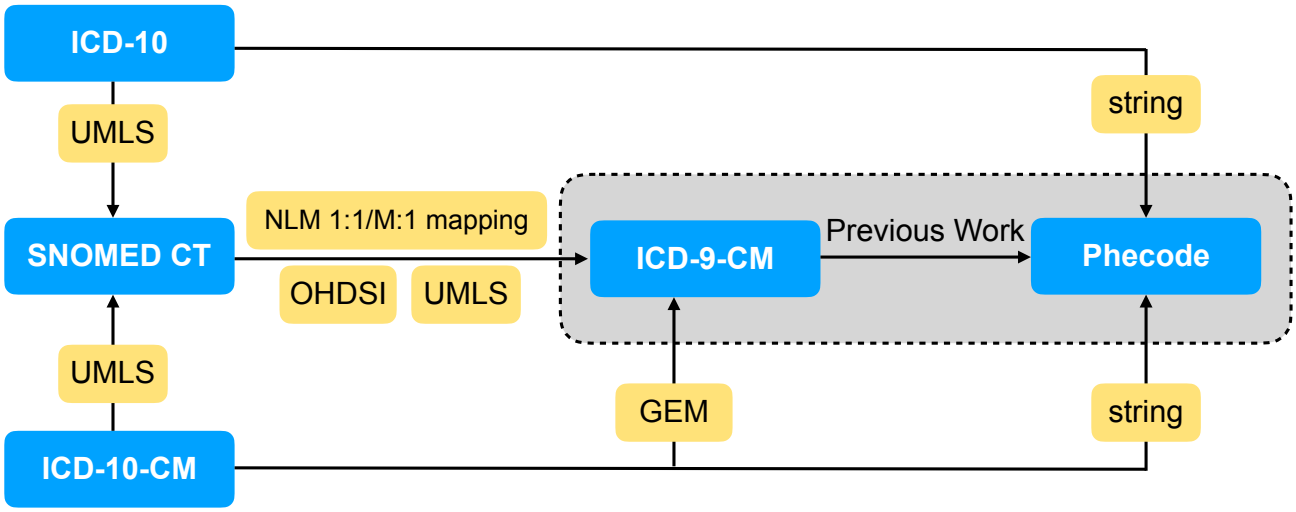


Figure 1

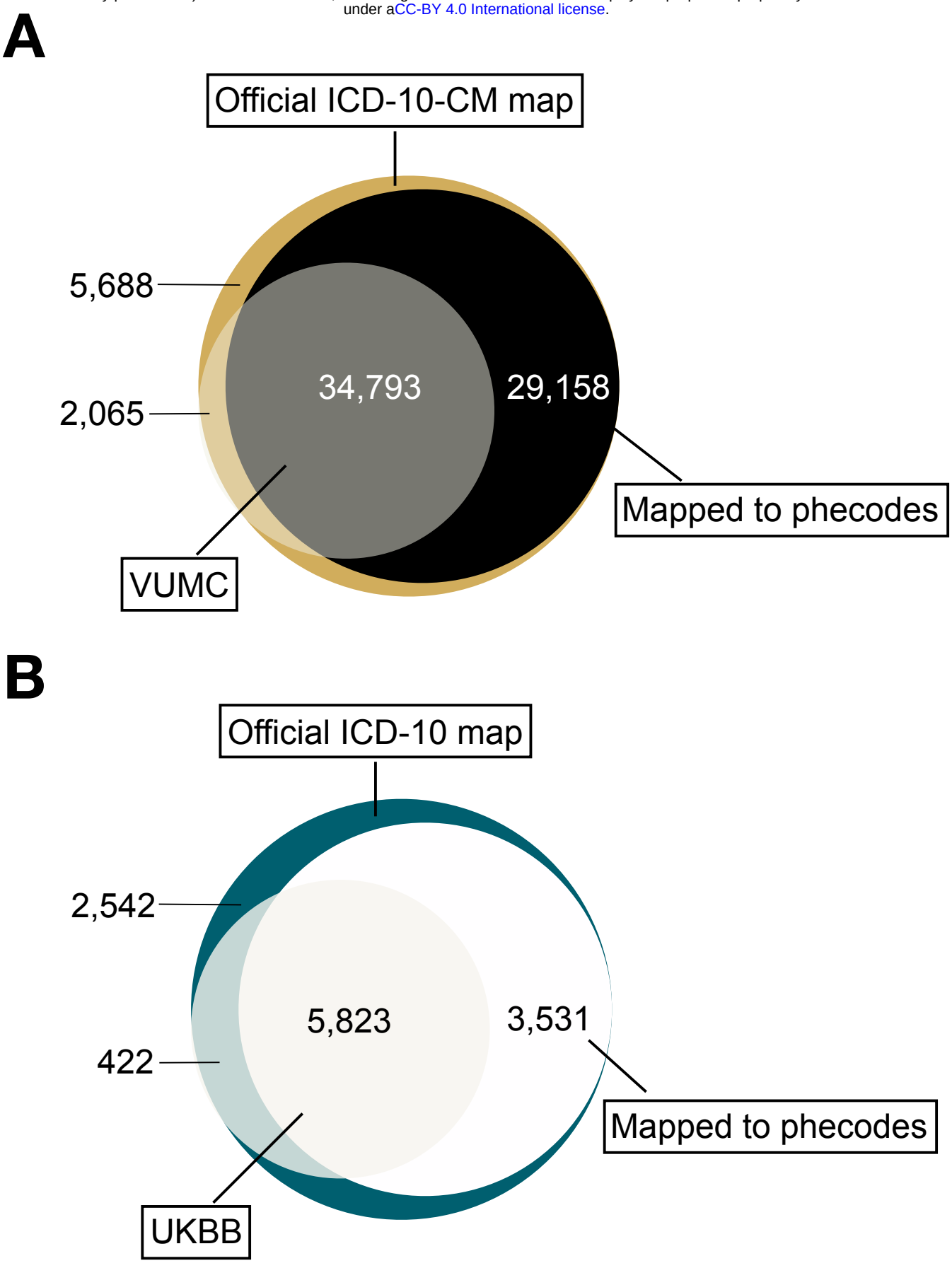


Figure 2

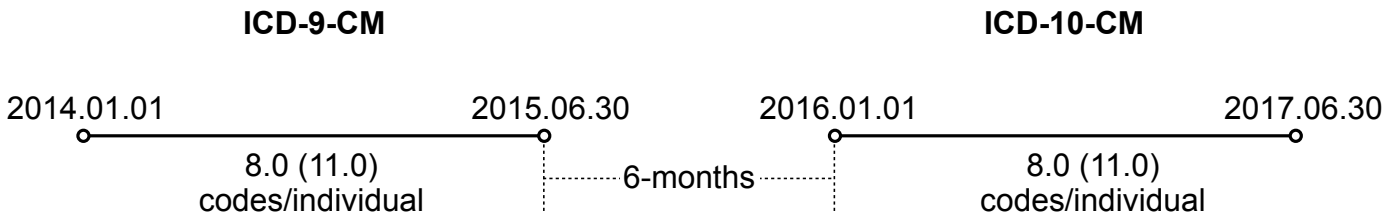


Figure 3

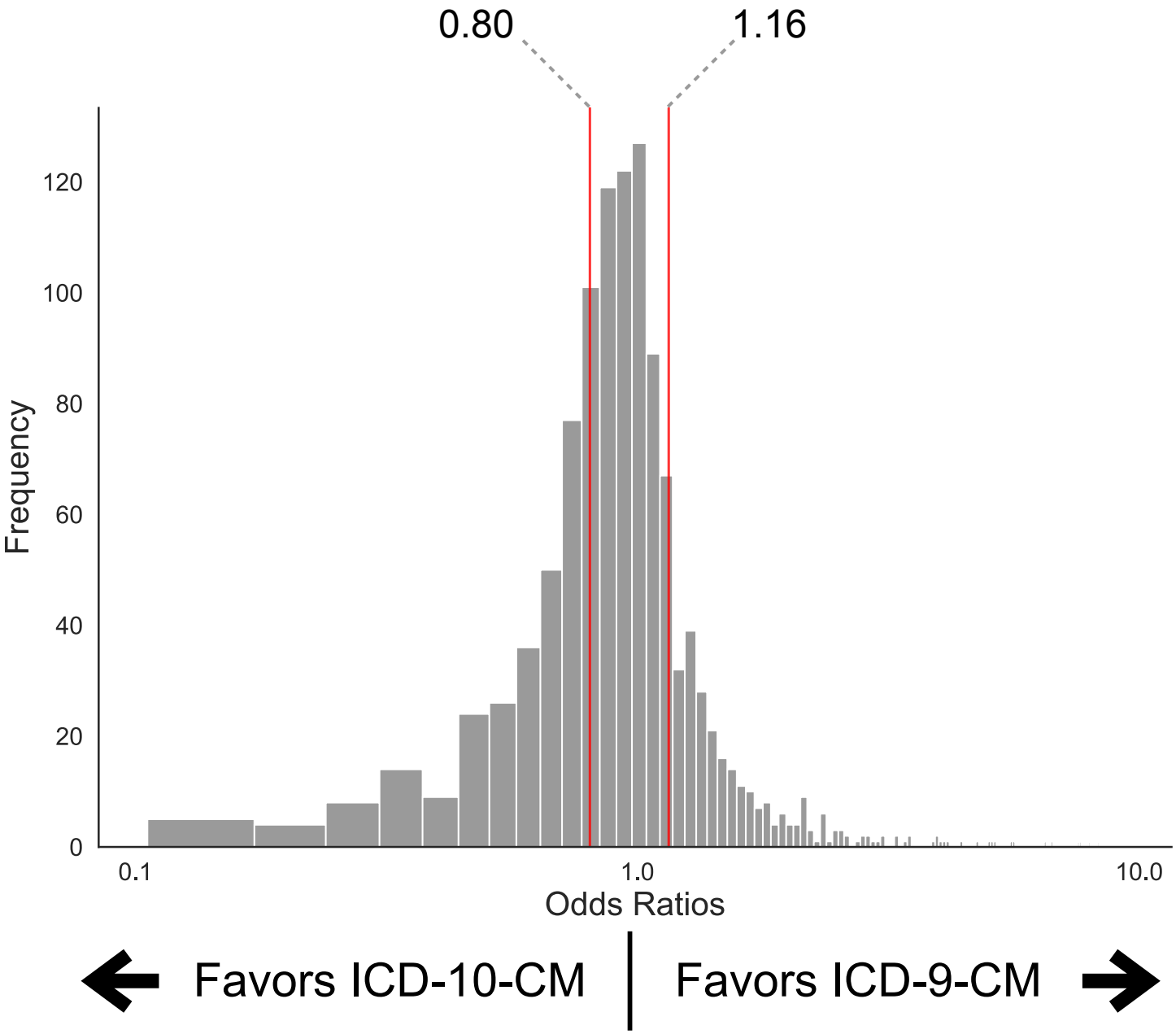


Figure 4

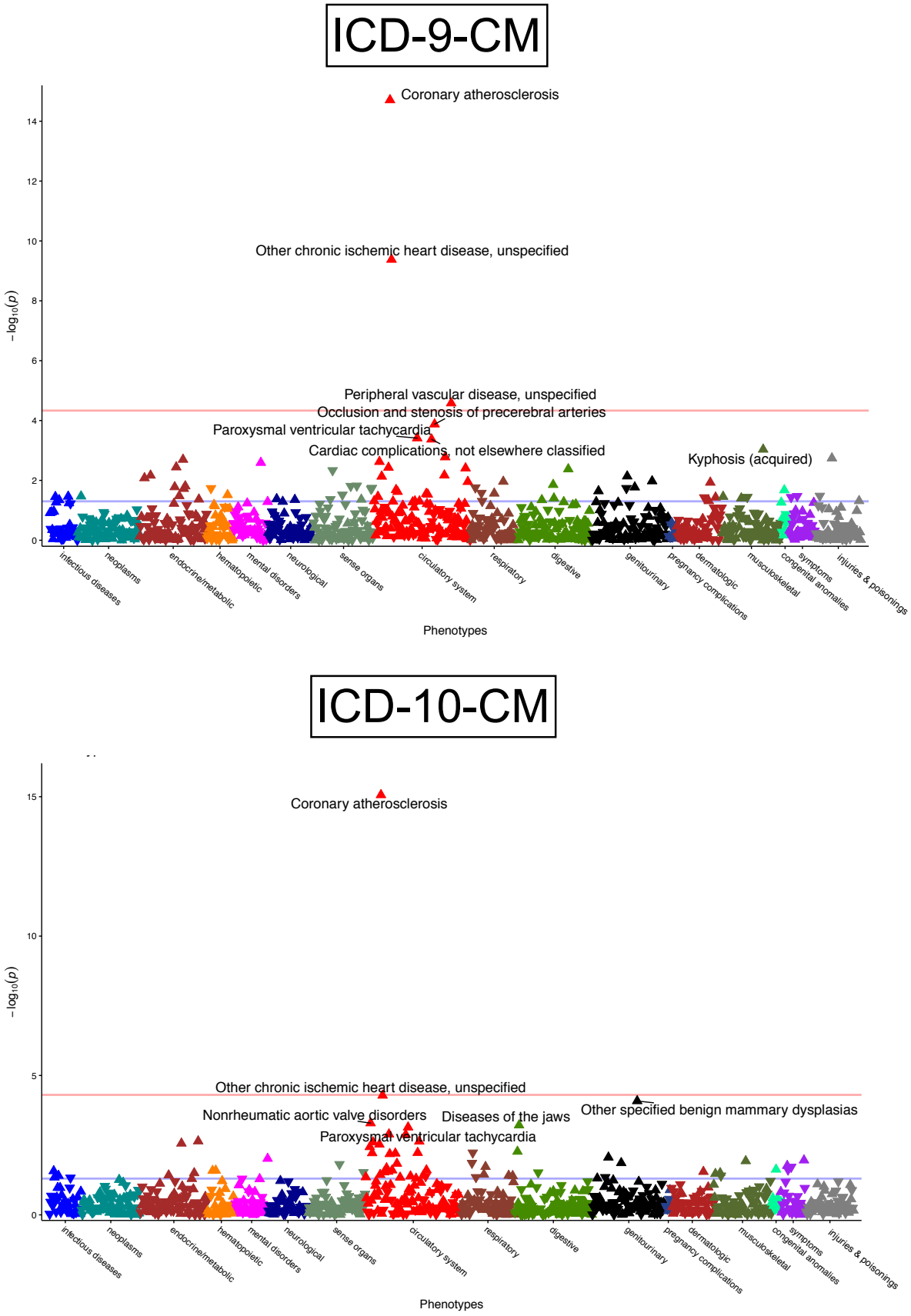
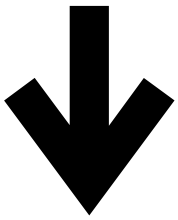


Figure 5

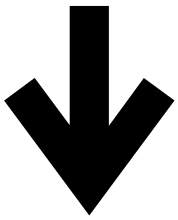
Supplementary Figures

Supplementary Figure 1: Example workflow used for calculating phecode prevalence from ICD-9-CM and ICD-10-CM. Example workflow used for calculating phecode prevalence from ICD-9-CM and ICD-10-CM by logistic regression analysis: [case_control] ~ [icd_source] + [sex] + [age]. [case_control] = 1 if patient had at least one code mapped to that phecode. [icd_source] = 1 if phecode was mapped from ICD-9-CM and 0 if phecode was mapped from ICD-10-CM. For example, in the first row, the patient is a case for phecode 296.2 and the phecode was mapped from an ICD-9-CM code ([case_control] =1, [icd_source]=1). In the second row, the same patient from the first row also is a case for phecode = 296.2, but the phecode was mapped from an ICD-10-CM code ([case_control]=1, [icd_source]=0). In the third row, a different patient is a case with the phecode from ICD-9-CM ([case_control]=1, [icd_source]=1), but does not have a ICD-10-CM code that maps to phecode 296.2, and so is a control for the phecode ([case_control]=0, [icd_source]=0). [age] is the patient's age at the end of each observation period.

id	phecode	count
pt1_icd9	296.2	1
pt1_icd10	296.2	1
pt2_icd9	296.2	1
pt2_icd10	296.2	0



id	phecode	case_control	icd_source	age	sex
pt1_icd9	296.2	1	1	14	F
pt1_icd10	296.2	1	0	16	F
pt2_icd9	296.2	1	1	25	M
pt2_icd10	296.2	0	0	27	M



[case_control] ~ [icd_source] + [sex] + [age]

Supplementary Fig 1