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, MSc[2,*], Xue Li, MSc[2], Harry Campbell, MD[2], Tim Varley, B.Sc.[3], Juan Zhao, 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

Abstract

Many studies of Electronic Health Record (EHR) data utilize custom-developed aggregations of billing codes enabling clinical and genetic research, including phenome-wide association studies (PheWAS). One such grouping is the phecode system, originally developed for PheWAS. Phecodes were built upon the International Classification of Diseases, version 9, Clinical Modification (ICD-9-CM). However, many healthcare systems across the world use ICD-10 and ICD-10-CM, and the United States switched from ICD-9-CM to ICD-10-CM in 2015. Here we present our work on developing and validating the mappings for both ICD-10 and ICD-10-CM to phecodes. We first assessed the coverage of both the ICD-10 and ICD-10-CM phecode maps in two large databases: Vanderbilt University Medical Center (VUMC) using ICD-10-CM and the United Kingdom Biobank (UKBB) using ICD-10 codes. We then evaluated the validity of the map for ICD-10-CM by comparing phecode prevalence between ICD-9-CM and ICD-10-CM derived phecodes at VUMC. Approximately 75% of all instances of ICD-10-CM codes and 80% of ICD-10 codes were successfully mapped to phecodes. To demonstrate the utility of the ICD-10-CM map, we further performed a PheWAS using ICD-9-CM and ICD-10-CM maps. This work provides an initial high-coverage map of ICD-10 and ICD-10-CM to phecodes. These codes are publicly available to aid in EHR-based investigation.

Introduction

Electronic Health Records (EHRs) have become a powerful resource for biomedical research in the last decade. EHR-based studies often leverage International Classification of Diseases (ICD) billing codes.[1] Often, relevant billing codes are grouped customarily to reflect biologically meaningful phenotypes or diseases.[2] A number of such schema exist, including the Agency for Healthcare Research and Quality (AHRQ) Clinical Classification Software (CSS) and the phecode system developed originally to facilitate phenome-wide association studies (PheWAS).[3] Our group first introduced phecodes in 2010 with 733 distinct phenotype codes to enable an early PheWAS. Phecodes currently include 1,866 hierarchical phenotypes, linked to relevant control groups. Since the introduction of phecodes, many published studies have used them successfully to replicate hundreds of known genotype-phenotype associations and discover dozens of new ones, some of which have been validated in subsequent studies.[4–14]

We created the initial version of phecodes using the International Classification of Diseases version 9 (ICD-9) system, which was originally 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 containing 22,406 codes. The WHO later developed the 10th ICD version (ICD-10) in 1990, which has since been used widely outside of the U.S.. Since then, NCHS has replaced ICD-9-CM with ICD-10-CM by modifying ICD-10.[15] Although it is based on ICD-10, ICD-10-CM contains more detailed codes than ICD-10 for the billing purposes.

We originally developed the phecode system using ICD-9-CM codes by combining one or more related ICD-9-CM codes into distinct diseases or traits. For example, the two ICD-9-CM codes 174.9 [Malignant neoplasm of breast], and V10.3 [Personal history of malignant neoplasm of breast], were combined into the same phecode: 174.11 [Malignant neoplasm of female breast]. With the help of clinical experts in disparate domains, such as cardiology and oncology, we have continuously updated the phecode groupings. In addition to identifying individuals with phenotypes, phecodes allow an approach to define relevant control populations for each case through lists of individuals who have similar or potentially overlapping disease states (e.g. excluding type 1 diabetes and secondary diabetes mellitus for an analysis of 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.[16] We recently 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.[17]

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 codes. A few studies have previously mapped ICD-10 codes to phecodes by first converting ICD-10 to ICD-9 codes, and then mapping the converted ICD-9 codes to phecodes.[6,13] 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 the current phecode mappings. We evaluate 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 the UKBB has used ICD-10 for over two decades.[18] We then analyze the ICD-10-CM to phecode map by comparing the phecodes mapped from ICD-9-CM and ICD-10-CM in the VUMC EHR.

Methods

EHR Databases

In this study, we used data from the VUMC EHR and UKBB[18] databases. The VUMC EHR contains records of >2.5 million unique individuals, including richly documented longitudinal clinical data for >1 million patients. VUMC's patient population reflects the racial and ethnic makeup of the surrounding community throughout Tennessee and the Southeastern U.S. Individuals of European ancestry make up the majority of records within this database (85%).

The UKBB is a prospective longitudinal cohort designed to allow detailed investigation of genetic and environmental determinants of diseases of public health importance in UK adults.[18] Between 2006 and 2010, 502,632 men and women aged 40-69 years were recruited from across the UK. Study participants underwent baseline measurements, including questionnaire, face to face interviews, and anthropometric or other measurements (e.g., standing and sitting height, waist and hip circumference, etc.). To allow longitudinal 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, death register records). At the time of this study, the UKBB included only inpatient ICD codes. For disease classification, ICD-10 was used from April 1995 to April 2010, and the ICD-10 4th edition from April 2010 onwards.[19]

We highlight two differences between these two datasets particular relevant to this study. First is the timespan of data collection: VUMC switched to using ICD-10-CM codes in October 2015, resulting in >2.5 years of data (approximately 2015-10-01 to 2017-06-01), while the UKBB has been collecting ICD-10 data for 23 years (1992-03-31 to 2015-03-31). Second is the patient type: the VUMC database includes codes for both inpatients and outpatients, whereas the UKBB codes presented in this paper only reflect inpatient data.

Mapping ICD-10-CM codes to Phecodes

The 2018 ICD-10-CM contains 81,593 unique codes,[20] while ICD-10 contains 12,318 unique codes. There are a large number of ICD-10-CM codes that do not exist in ICD-10, and even the 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]. Thus, we sought to create maps for ICD-10 and ICD-10-CM separately. In addition to implementing ICD-10 or ICD-10-CM, each hospital or facility may create unique codes that are used at that location only, which we refer to as "local codes" because they are not widely accepted or used at outside institutions.

Mapping an ICD-10-CM code to a phecode was performed either directly or indirectly (**Figure 1**). We mapped an ICD-10-CM code to a phecode directly (497 ICD-10-CM codes) 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 (81,083 ICD-10-CM codes) used the existing ICD-9-CM-phecode mappings and involved two approaches. In the first approach, we used the General Equivalence Mapping (GEM)[20] to map ICD-10-CM codes to ICD-9-CM codes. We then mapped the ICD-9-CM code to a phecode using our previously created 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)[21], which were then converted to ICD-9-CM using either official mappings from National Library of Medicine (NLM),[22] Observational Health Data Sciences and Informatics (OHDSI),[23] or International Health Terminology Standards Development Organisation (IHTSDO).[24] 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]. We used only 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 ICD-10-CM code to a phecode using the existing ICD-9-CM to phecode mapping, e.g. L01.03 to phecode 686.2 [impetigo]. If an ICD-10-CM (e.g., I10 [Essential 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 only (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 the mappings. This latter association created a new polyhierarchical nature to phecodes. Phecode Map 2.1 (beta) with ICD-10-CM codes ('icd10cm_phecode_map_110218_beta.csv') can be found as a supplementary file on bioRxiv and will be found on https://phewascatalog.org/.

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; since a GEM does not exist between ICD-9-CM and ICD-10 codes, to handle ICD-10 codes that did not have string matches to phecodes, we either used UMLS or 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 kept the mapping to the child phecode only to maintain the granular meaning of the ICD-10 code. If an ICD-10 code was mapped to multiple distinct phecodes, we kept all the mappings. Phecode Map 2.2 (beta) with ICD-10 codes (icd10_phecode_map_110218_beta.csv') can be found as a supplementary file on bioRxiv and will be found on https://phewascatalog.org/.

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 total of number of codes in the two official maps, the number of those codes mapped to phecodes, the number of official codes used in the two EHRs that are either mapped to phecodes or unmapped to phecodes. We also calculated the number of codes in the two EHR databases that are not in the official maps. From the numbers calculated above, we were able to get the number of official codes that are not used in the EHR databases (**Figure 2**).

Comparison of Phecode Prevalence from Both 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. We selected time windows to occur immediately prior to and after VUMC's transition to ICD-10-CM. To reduce any potential confounders, a buffer time of 6 months 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[25] to the Epic EHR 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. The cohort consisted of 55.1% female and 44.9% male who were 45±25 years old (mean± SD).

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. This resulted in mean ± standard deviation of 21.14 ± 48.06 ICD-9-CM codes/individual and 22.18 ± 51.67 ICD-10-CM codes/individual (**Figure 3**). We mapped these codes to phecodes using phecode map version 1.2 for ICD-9-CM codes [26,27] and the newly generated ICD-10-CM phecode map. We then 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 which 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 external causes, 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 the codes that were not mapped or should not be mapped to phecodes, we used logistic regression to analyze the remaining ICD-9-CM (10,988) and ICD-10-CM (25,857) codes that we successfully mapped, to determine whether phecode prevalence from ICD-9-CM and ICD-10-CM were comparable. To perform the logistic regression analysis, a table was compiled that included every patient twice: once for the ICD-9-CM data and a second time for ICD-10-CM data. Given all the phecodes which resulted from mapping both ICD-9-CM and ICD-10-CM codes, a patient was labeled a "case" for that phecode if they had at least one code which mapped to that phecode, and the code source which mapped to that phecode was noted as coming from ICD-9-CM or ICD-10-CM. Additionally, the maximum age was calculated for each patient from their last ICD-9-CM and ICD-10-CM code date considered in this dataset. The patient's gender and maximum age were combined with the case/control data to perform the logistic regression. We regressed Case/control status for each phecode against the ICD source (10-CM vs. 9-CM), age, and gender (**Figure 4**). For each phecode, this produced a p-value and odds ratio (OR) indicating whether the phecode was more likely to come from an ICD-9-CM or ICD-10-CM code, with an OR > 1 indicating the phecode more likely comes from ICD-9-CM.

Comparative PheWAS analysis of LPA SNP, rs10455872

To evaluate the accuracy of the ICD-10-CM to phecode map, we performed two PheWAS analyses on a LPA genetic variant (rs10455872) using mapped phecodes from ICD-10-CM and ICD-9-CM, respectively. We chose lipoprotein(a) (LPA) SNP (rs10455872) as the predictor because it is known to be associated with increased risks of developing hyperlipidemia and cardiovascular diseases (CVD).[28,29]

We used data from BioVU, the de-identified DNA biobank at VUMC to conduct the PheWAS.[30] We identified 13,900 adults (56.9 % female, 59 +/– 15 years old in 2014), who had LPA variant genotyped, and at least one ICD-9-CM and ICD-10-CM code in their respective time windows (**Figure 3**). We observed 86.7% AA, 12.8% AG, and 0.5% GG. We used the 1,632 phecodes that overlapped in the two time windows for PheWAS. We performed the study using the R PheWAS package,[31] adjusting for age, sex, and race.

Results

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

In total, of all possible ICD-10-CM codes, 81,580 codes successfully map to at least one phecode, with 7,846 (9.6%) codes mapping to more than one phecode. Of all possible ICD-10 codes, 9,354 mapped to at least one phecode, and 300 (3.4%) codes to more than one phecode. For example, the 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]. Whereas, ICD-10-CM E13.36 [Other specified diabetes mellitus with diabetic cataract] maps to two phecodes: 366 [Cataract] 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]

Among the 43,282 ICD-10-CM codes used at VUMC since they were adopted on October 2015, 31,282 (72.3%) codes can be mapped to phecodes. Comparatively, of the 8,144 ICD-10 codes used in UKBB, 5,823 (71.5%) codes can be mapped to phecodes (**Figure 2**). An additional 34,434 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. A total of 9,617 and 1,899 codes billed at VUMC and UKBB did not exist in the current ICD-10-CM and ICD-10 versions, respectively. Therefore, we were unable to map these codes to phecodes.

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 (**Figure 2**). We observed that 84.6% of all the instances of ICD-10-CM codes used at VUMC, and 71.5% of all instances of ICD-10 codes in the UKBB are mapped to phecodes.

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 1 and 2**).

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

A total of 357,728 patients had both ICD-9-CM and ICD-10-CM codes during the defined 18-month time windows. They had 14,593 distinct ICD-9-CM and 34,261 distinct ICD-10-CM codes. We found that the codes that did not map to phecodes were local, procedural, or supplementary classification codes. After these codes were removed, 45 ICD-9-CM and 915 ICD-10-CM codes remained unmapped. 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 successfully mapped to 1,842 phecodes.

Removing the ICD-10-CM codes which should not be mapped left 2,190 ICD-10-CM codes which did not map to a phecode. Of these, 1,520 were X, Y, and Z codes, leaving 670 remaining codes. All of these unmapped ICD-10-CM codes in this study’s cohort have less than 200 unique individuals (<0.1% of the total cohort), and the majority of the codes with a >10 unique individuals do not represent diseases that are caused by biological processes. 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 to phecode map, we compared the map to that of the most recent ICD-9-CM phecode map using logistic regression. Specifically, we aimed to see whether there were any significant differences in the prevalence of phecodes in the two 18-month periods. The distribution of the resulting odds ratios (OR) is shown in **Figure 4**.

We hypothesized a few possible causes underlying the differences observed in the logistic regression analysis of the ICD-10-CM to phecode map. Specifically, we aimed to determine whether incorrectly mapped ICD-10-CM codes and/or ICD-10-CM codes that were mistakenly not mapped primarily drove these differences. To accomplish this task, we first selected phecodes that crossed Bonferroni correction with >200 cases. Second, from this list, we manually reviewed ten phecodes with the highest and lowest ORs (for a total of 20 phecodes selected for review) (**Table 3**).

The following are examples of ICD-10-CM codes that were incorrectly mapped. Among the top 10 phecodes ranked by OR, 250.6 [polyneuropathy in diabetes] had an OR = 9.28 (favoring ICD-9-CM) with 3,312 ICD-9-CM and 391 ICD-10-CM cases. We found that the mapping of ICD-10-CM E11.42 [Type 2 diabetes mellitus with diabetic polyneuropathy] drove this imbalance. 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, then the number of ICD-10-CM cases for phecode 250.6 increases to 3,879 thereby decreasing the difference of case numbers from +2,921 in favor of ICD-9-CM to +567 in favor of ICD-10-CM.

In the same group of phecodes that we analyzed, there were also some imbalances that we expected. 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. Phecode 395.3 [Nonrheumatic tricuspid valve disorders] with OR = 0.11 is another example of an expected imbalance, which is driven by ICD-10-CM I36.1 [Nonrheumatic tricuspid (valve) regurgitation]. Given the older population in the ICD-10-CM observation period, an increase in left-sided congestive heart failure could lead to elevated risk of developing secondary tricuspid regurgitation.[32]

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 LPA SNP, rs10455872; one PheWAS was conducted using ICD-9-CM to phecode map and another was conducted using the ICD-10-CM to phecode map. They both showed significant positive correlations (p-value < 0.05) between the minor allele of rs10455872 and coronary atherosclerosis and chronic ischemic heart disease, replicating the previous findings (**Figure 5**).

Discussion

This work demonstrates the results of mapping ICD-10-CM and ICD-10 codes to phecodes and evaluation in two distinct databases. These results show that the vast majority of billed instances are mappable to phecodes using either ICD-9-CM, ICD-10-CM, or ICD-10. Further, though many codes appear have variations in billing between ICD-9-CM and ICD-10-CM in our study cohort, the vast majority do not vary greatly in their frequency of mapping to phecodes. As the use of ICD-10-CM and ICD-10 codes increases, so does the need for convenient and reliable methods of aggregating the codes for clinical and genetic research. Since 2010, many studies have demonstrated the value of aggregating ICD-9-CM codes such as is done with phecodes. This resource will allow researchers in the biomedical sciences who use EHR data to leverage the increasing amount of clinical data represented by ICD-10 and ICD-10-CM for phecode analysis alongside their ICD-9 codes.

As seen in **Figure 2**, there are codes that are included in ICD-10-CM or ICD-10 (25,733 and 2,542 codes, respectively), but have not been observed in either EHR database, as well as codes that have been used, but are not included in the current versions of ICD-10-CM or ICD-10 (black section: 9,617 and 1,899 codes, respectively). Neither of these code groupings are mapped to phecodes, as they are likely composed of encounter, procedural codes, or “local codes” developed and used solely at that institution. Beyond these unmapped codes, the mapping results demonstrate a few areas in the phecode aggregate system that could be updated, such as personal history and single overlooked codes.

Further, as expected, the most frequent ICD-10-CM codes include hypertension, type 2 diabetes, and coronary atherosclerosis, as these are common conditions and are consistent with our routine and clinical observations at Vanderbilt.

Examining the codes that are unable to be mapped is similarly enlightening. The majority of VUMC ICD-10-CM codes which are unable to be mapped are encounter or procedural codes (i.e. codes beginning with Z, such as Z00.00 [Encounter for gynecological examination (general) (routine) without abnormal findings] or supplementary codes (i.e. codes beginning with Y). These codes should not be mapped to any phecodes because they are not a description of a specific phenotype or disease.

Analyzing the UKBB ICD-10 codes that do not map to phecodes revealed the missing element of family history from the current phecode version. In addition to the codes that we anticipated would not map to phecodes, the majority of the most frequent inpatient ICD-10 codes involve personal or family history. This reveals a potential shortcoming of the current phecode system, and demonstrates an area of the phecodes that could be expanded.

This work also sheds light on codes that should be further reviewed and included in an existing phecode map or used to create a new phecode. For example, Z37.0 [Single live birth] occurs in both the VUMC and UKBB data at a high rate, but does not currently map to a phecode, as events and procedures typically have not been mapped into phecodes.

An initial PheWAS using the ICD-9-CM and ICD-10-CM to phecode maps and found similar significant associations with phenotypes using either code system. The results of this PheWAS comparison further demonstrates the general accuracy of ICD-10-CM to phecode map when compared to the current gold-standard phecode map.

The results presented here have limitations. First, while the data from the UKBB spans 23 years and is derived from a population recruited from across the UK, the VUMC ICD-10-CM data encompass just over one year and billing practices are likely still evolving. The VUMC data also represent a single academic medical center.

Additionally, this work did not aim thoroughly evaluate the accuracy of the mapping, as the resource for our mappings (e.g. GEM) are assumed to be correct. If an ICD-10-CM or ICD-10 code maps to two or more unlinked phecodes, we currently keep all of the mappings. It is important for future work 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 manual validation of mapping accuracy.

Conclusion

In this paper, we introduced our work on mapping ICD-10-CM and ICD-10 codes to current phecodes. We provide an initial map with high coverage of EHR data in two large databases. These mappings will enable researchers to leverage accumulated ICD-10-CM and ICD-10 data for clinical research in a high-throughput manner.

Acknowledgements

The project was supported by NIH grant R01 LM 010685, R01 HL133786 and AHA Scientist Development Grant 16SDG27490014. 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. Evropi Theodoratou is supported by a CRUK Career Development Fellowship (C31250/ A22804). Xiangrui Meng and Xue Li are supported by China Scholarship Council studentships. Patrick Wu is supported by T32 GM007347 from the National Institute of General Medical Studies for the Vanderbilt Medical-Scientist Training Program, T15 LM007450 from the National Library of Medicine for the Vanderbilt Biomedical Informatics Training Program, and P50 GM115305 from the National Institute of General Medical Studies for PGRN.

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 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.

5 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.

6 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.

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

8 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.

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

10 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.

11 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.

12 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.

13 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.

14 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.

15 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.

16 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.

17 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.

18 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.

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

20 2018 ICD-10 CM and GEMs. 2017.https://www.cms.gov/Medicare/Coding/ICD10/2018-ICD-10-CM-and-GEMs.html

21 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).

22 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).

23 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).

24 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).

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

26 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).

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

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

29 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.

30 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.

31 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.

32 Lancellotti P, Magne J. Tricuspid valve regurgitation in patients with heart failure: does it matter? *Eur Heart J* 2013;**34**:799–801.

FIGURE LEGENDS

Figure 1. Mapping strategy for ICD-10 and ICD-10-CM to phecode. Mapping methods between various code formats. The gray dashed-outline box indicates previously published work.[3,16] 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 in UKBB and ICD-10-CM at VUMC. Categories include codes that are only in the EHR, mapped to phecode, and/or are part of the official ICD-10/ICD-10-CM systems.

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, with a mean ± standard deviation of 21.14 ± 48.06 ICD-9-CM codes/individual and 22.18 ± 51.67 ICD-10-CM codes/individual.

Figure 4. Distribution of odds ratios (OR) resulting from logistic regression analysis of phecode maps. To assess the broad accuracy of the ICD-10-CM to phecode map, we used logistic regression analysis to calculate the likelihood of an individual being assigned a phecode. Phecodes with < 1 case in the ICD-9-CM or ICD-10-CM periods were removed before plotting the distribution or ORs. 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.

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

TABLE LEGENDS

Table 1. The top ten most commonly used ICD-10-CM codes at VUMC, that map to phecodes

Table 2. The top ten most commonly used ICD-10 codes in the UKBB that successfully map to phecodes.

Table 3. The ten phecodes with the highest (A) and lowest (B) odds ratios.

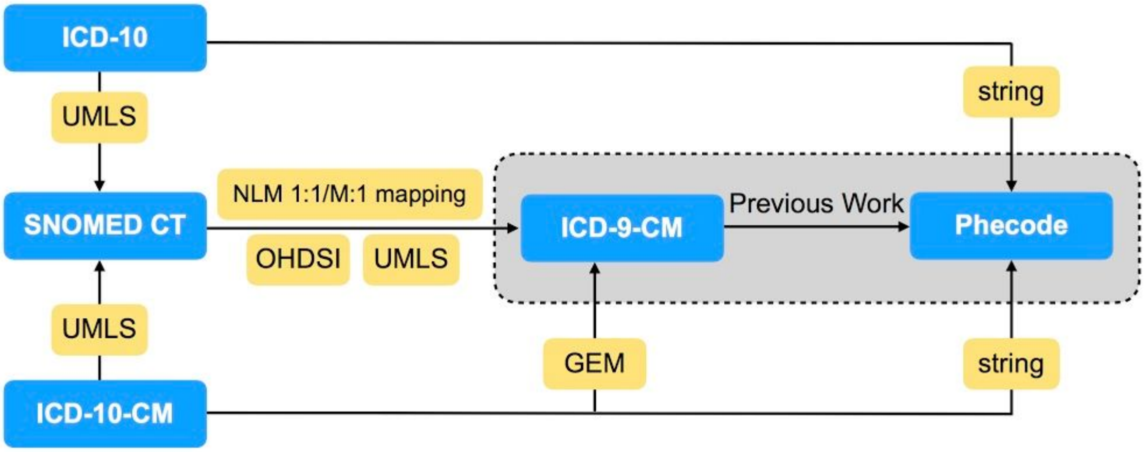

ICD-10
UMLS
SNOMED CT
UMLS
ICD-10-CM
NLM 1:1/M:1 mapping
OHDSI
UMLS
ICD-9-CM
Previous Work
Phecode
string
GEM
string

Figure 1

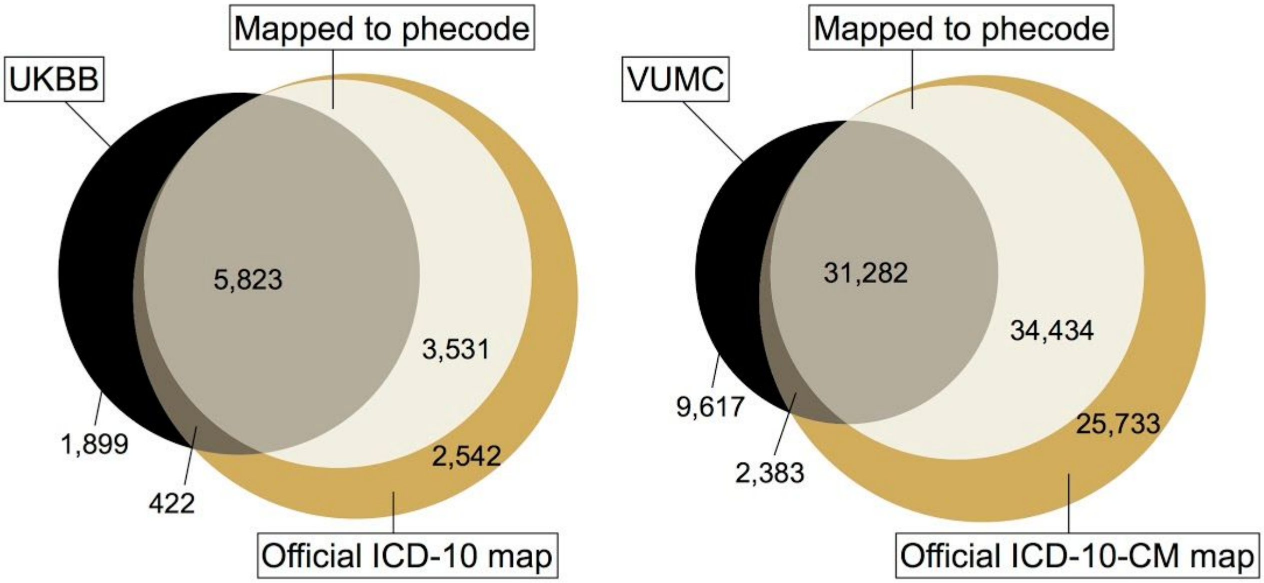

Mapped to phecode
UKBB
5,823
3,531
1,899
2,542
422
Official ICD-10 map
Mapped to phecode
VUMC
31,282
34,434
9,617
25,733
2,383
Official ICD-10-CM map

Figure 2

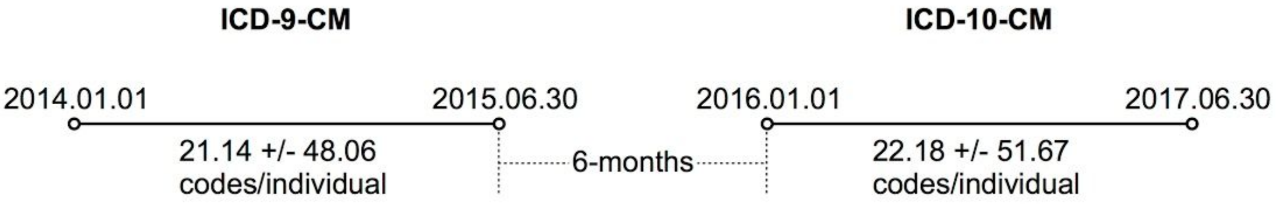
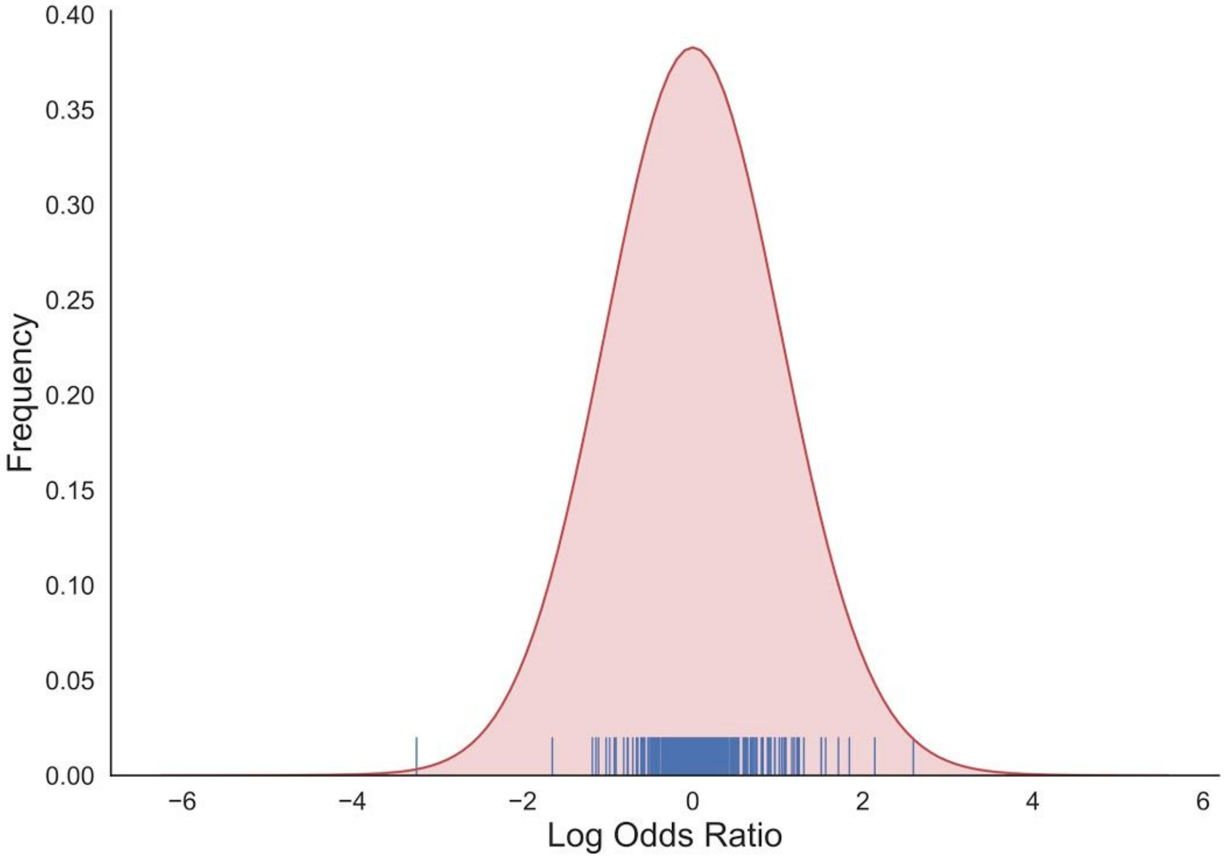

ICD-9-CM
ICD-10-CM
2014.01.01
2015.06.30
2016.01.01
2017.06.30
21.14 +/- 48.06
codes/individual
6-months
22.18 +/- 51.67
codes/individual

Figure 3



Logistic Regression Model

case/control status ~ [icd source] + [sex] + [age]

Interpretation of Odds Ratios (OR)

OR > 1: phecode more likely to come from ICD-9-CM

OR < 1: phecode more likely to come from ICD-10-CM

Figure 4

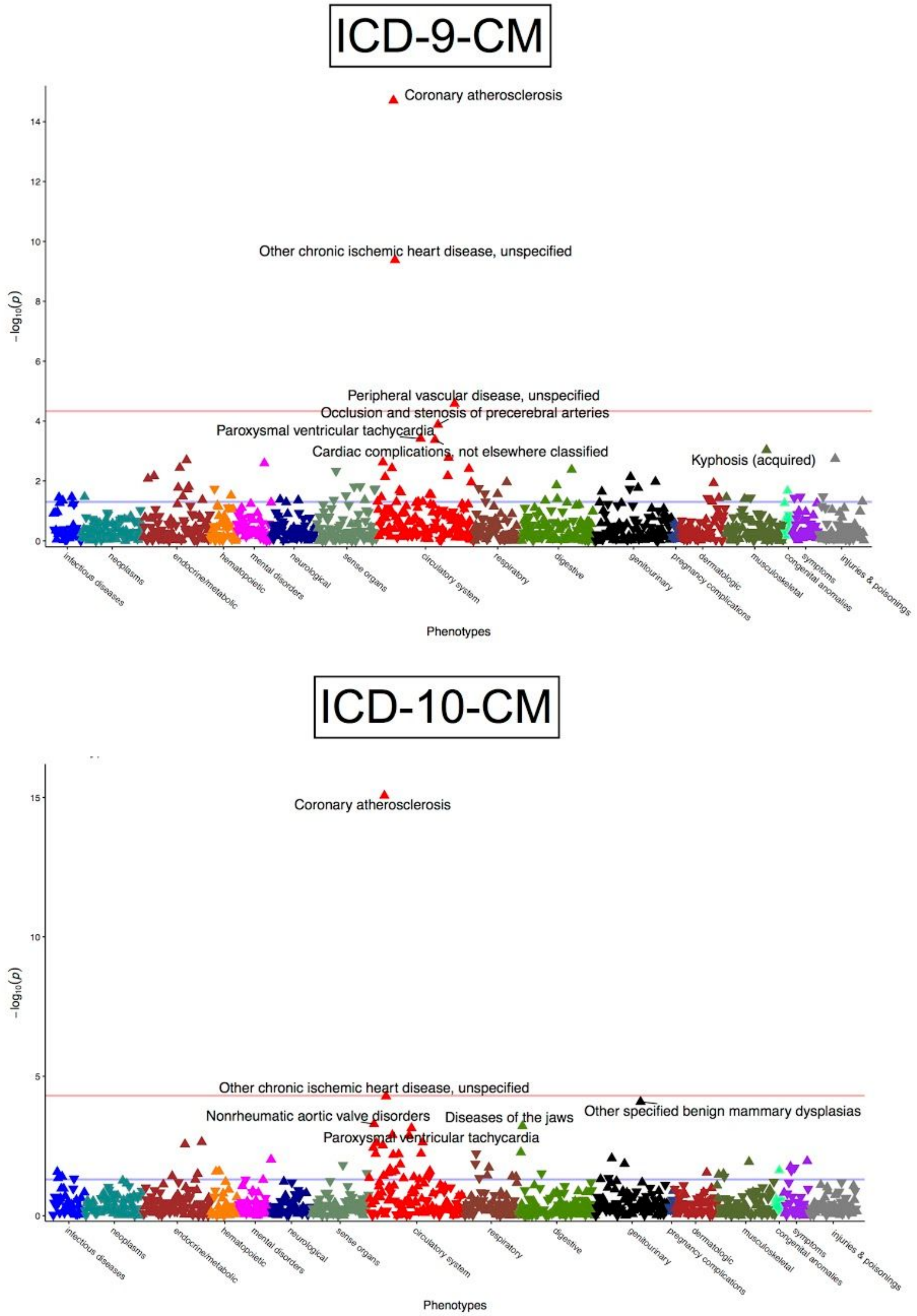

ICD-9-CM
Coronary atherosclerosis
Other chronic ischemic heart disease, unspecified
Peripheral vascular disease, unspecified
Occlusion and stenosis of precerebral arteries
Paroxysmal ventricular tachycardia
Cardiac complications, not elsewhere classified
Kyphosis (acquired)
$-\log_{10}(p)$
0
2
4
6
8
10
12
14
infectious diseases
neoplasms
endocrine/metabolic
hematopoietic
mental disorders
neurological
sense organs
circulatory system
respiratory
digestive
genitourinary
pregnancy complications
dermatologic
musculoskeletal
congenital anomalies
symptoms
injuries & poisonings
Phenotypes
ICD-10-CM
Coronary atherosclerosis
Other chronic ischemic heart disease, unspecified
Other specified benign mammary dysplasias
Nonrheumatic aortic valve disorders
Diseases of the jaws
Paroxysmal ventricular tachycardia
$-\log_{10}(p)$
0
5
10
15
infectious diseases
neoplasms
endocrine/metabolic
hematopoietic
mental disorders
neurological
sense organs
circulatory system
respiratory
digestive
genitourinary
pregnancy complications
dermatologic
musculoskeletal
congenital anomalies
symptoms
injuries & poisonings
Phenotypes

Figure 5

Table 1

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 2

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

Table 3

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	427	4,308	< 0.001	0.11

	valve disorders				
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