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

By

Patrick Wu

Thesis

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

MASTER OF SCIENCE

in

Biomedical Informatics

January 31, 2020

Nashville, Tennessee

Approved:

Wei-Qi Wei, M.D., Ph.D.

QiPing Feng, Ph.D.

Joshua C. Denny, M.D., M.S.

ACKNOWLEDGEMENTS

This work would not have been possible without the generous support of Vanderbilt University Medical Scientist Training Program, School of Medicine, and Department of Biomedical Informatics. This work was also supported by multiple grants from the National Institutes of Health (T32 GM007347, T15 LM007450, P50 GM115305).

I am grateful for all of the people who I have worked with throughout my research career so far over the past ten years. Drs. Tim Huffaker, Beth Lalonde, Zane Bergman, Kristy Blake-Hodek, and Alex Amaro and members of the Huffaker Lab at Cornell University were instrumental in helping me develop a passion for research and continued interest in basic cell biology. Drs. Dorian McGavern, Bernd Zinselmeyer, Jasmin Herz, and members of the McGavern Lab at the National Institutes of Health helped me to appreciate the dynamic nature of the immune system.

For the work presented in this thesis, I would like to acknowledge Drs. QiPing Feng and Joshua Denny for providing extensive professional guidance and teaching me valuable lessons about scientific research. I would like to especially thank my thesis mentor, Dr. Wei-Qi Wei, for his generosity, patience, and guidance in research direction.

TABLE OF CONTENTS

Page

CHAPTERS

LIST OF TABLES

Chapter II

Table Page

Chapter III

Table Page

LIST OF FIGURES

Chapter II

Figure Page

Chapter III

Figure Page

CHAPTER 1

Introduction

The Health Information Technology for Economic and Clinical Health (HITECH) Act passed in 2009 incentivized U.S. hospitals to adopt electronic health record (EHR) systems.[1,2] Adoption of EHRs by U.S. hospitals increased from 72% in 2011 to 96% in 2017.[3] In the last decade, the amount of global EHR data available for biomedical research has grown at an exponential pace and will continue to accumulate in the future.[4,5] The build-up of EHR data coincides with decreases in genotyping costs,[6,7] thereby allowing investigators to connect big longitudinal EHR data with genomic data for use in improving healthcare delivery process, outcome studies, and biomedical research.[8]

Investment in Genomic Medicine

The convergence of big EHR data and decreases in high-throughput genotyping motivated public and private investment in genomic medicine. Examples of large-scale investments that leverage EHRs connected to DNA biobanks include the Electronic Medical Records and Genomics Network (eMERGE)[9,10] the UK Biobank (UKBB).[11] eMERGE was started in 2007 with funding by the National Human Genome Research Institute (NHGRI). The main goal of eMERGE is to investigate methods of combining DNA biorepositories with EHR data for large scale, high-throughput genetic research. As of January 2020, there are 136,078 participants in the network cohort. Studies by the eMERGE network participants have demonstrated the feasibility of creating disease cohorts in the EHR using phenotyping algorithms for performing genome-wide association studies (GWAS). To date, 68 phenotyping algorithms have been developed through eMERGE.[12] With primary support from the Wellcome charity and the

Medical Research Council, the UKBB is a prospective longitudinal study with the aim of improving prevention, diagnosis, and treatment of human disease. Between 2006-2010, 500,000 people were enrolled. The UKBB has genotyping data that are linked to phenotypic information from multiple resources such as questionnaires, EHRs, and physical measurements.[11]

A primary goal of these programs is to improve the understanding of the genetic influences on human diseases through investigations like GWAS.[13] A GWAS typically begins with an investigator selecting a phenotype to study, followed by collection of data required for researchers to label study participants as phenotype cases (those with the phenotype) or controls (those without the phenotype).[14] For example, in one of the first type 2 diabetes (the phenotype) GWAS,[15] cases were individuals who met three requirements: 1) met one of the diagnostic criteria set by the American Diabetic Association (eg, had a fasting plasma glucose of >7.0 mmol/L); 2) had a first degree relative with diabetes; and 3) had a BMI < 30 kg/m². Type 2 diabetes controls were individuals with normal fasting plasma glucose and had a BMI < 27 kg/m². To obtain genotype data from study participants, investigators can use high-throughput methods,[16] like microarrays[17] and whole-genome sequencing.[18] With phenotype and genotype data, the investigator scans the genome to identify single nucleotide polymorphisms (SNPs) associated with the phenotype.

In the late 2000s, though genetic data was relatively inexpensive to obtain due to decreases in cost of genotyping,[6,7] most genetic studies using ad hoc cohorts[15,19,20] were bottlenecked by limited phenotypic data.[21] For example, in 2007, the Wellcome Trust Case Control Consortium published a GWAS that involved >50 research groups in the UK.[20] The study’s main experiment comprised of 500,568 SNPs, but only 7 phenotypes. In general, developing accurate high-throughput portable phenotyping algorithms is monetarily expensive, takes a long time,[22] and is labor-intensive. To develop 13 validated phenotyping algorithms using EHR data involved the trans-institutional collaboration of biomedical informaticians, domain experts, clinicians, geneticists, and others.[23] Thus, in contrast to the many assays available to investigators for interrogating the genome, researchers did not have a similar tool for obtaining accurate information across human diseases.[24,25]

High-throughput Phenotyping in the EHR

Accumulation of EHR data presented opportunities for investigators to decrease the cost of creating disease cohorts. For example, a 2007 study by Wilke et al. used a combination of diagnosis codes, laboratory data, and medication history to identify patients with diabetes.[26] Studies such as Wilke et al. decreased the cost of recruiting study participants, but were limited by the resources spent on informaticians to perform natural language processing (NLP) queries and manual chart reviews by clinicians to obtain a final phenotyping algorithm.[9,22,27–29] Other barriers to high-throughput phenotyping in the EHR included data fragmentation,[30] and data sparsity/irregularity.[31]

To address the difficulties with phenotyping at a massive scale, Denny et al. introduced phecodes in 2010.[32] In the first phenome-wide association study (PheWAS), they developed phecodes as a custom-grouping of International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) codes. In contrast to GWAS, PheWAS starts with one SNP and look for associated phenotypes. They conducted the study at Vanderbilt University Medical Center, where they had access to one of the first de-identified electronic health records (EHRs) that were linked to a DNA biobank.[33] Phecodes offered investigators a single tool to collect phenotype data across human diseases. Further, using only ICD-9-CM diagnosis codes was easier to implement at other sites, given the widespread use of ICD-9-CM codes in billing.

The 2010 PheWAS study replicated four of seven known SNP-disease associations.[32] In 2013, a systematic evaluation of PheWAS by the same research team replicated 66% of the associations in the GWAS catalog that were adequately powered.[34] In 2017, Wei et al. demonstrated that compared to alternative phenotyping methods using diagnostic billing codes, phecodes replicated more known genotype-phenotype associations than ICD-9-CM and clinical classification software (CCS) codes.[35]

Since 2010, the phenotypes represented in phecodes have increased in number and been refined. To increase the statistical power to find genotype-phenotype associations, the initial PheWAS study grouped >10,000 ICD-9-CM codes into 733 phecodes. In 2013, a second version of phecodes was released with 1358 unique phecodes,[34] followed by a third iteration with 1864 phecodes (version 1.1), and a fourth iteration with 1866 phecodes (version 1.2).[36] In 2014, Carroll et al. released an R package[37] that has allowed investigators to easily conduct PheWAS at outside institutions.

Motivation and Research Aims

The studies presented in this thesis were motivated by the absence of a tool to translate ICD-10 and ICD-10-CM codes to phecodes. ICD-10 codes have been used internationally for over 2 decades, and ICD-10-CM codes have been used in the U.S. since 2015.[38] In this thesis, I describe the process used to develop new maps to allow investigators to convert ICD-10 and ICD-10-CM codes to phecodes. These resources will allow investigators to perform high-throughput PheWAS in the EHR containing ICD-10-CM and ICD-10 codes.

This thesis consists of four chapters. The first chapter describes the motivation for my research. The second chapter primarily focuses on the creation of the ICD-10/ICD-10-CM codes to phecode maps. The third chapter focuses on the evaluation of the ICD-10-CM to phecode map. I summarize this work In the last chapter with a discussion of the limitations and future directions.

CHAPTER 2

Mapping ICD-10 and ICD-10-CM Codes to Phecodes: Workflow Development and Initial Evaluation

This manuscript was published in JMIR Medical Informatics as follows:

Wu P, Gifford A, Meng X, Li X, Campbell H, Varley T, Zhao J, Carroll R, Bastarache L, Denny JC, Theodoratou E, Wei WQ

Mapping ICD-10 and ICD-10-CM Codes to Phecodes: Workflow Development and Initial Evaluation

JMIR Med Inform 2019;7(4):e14325

URL: https://medinform.jmir.org/2019/4/e14325

DOI: 10.2196/14325

PMID: 31553307

PMCID: 6911227

Abstract

Background

The phecode system was built upon the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) for phenome-wide association studies (PheWAS) in the electronic health record (EHR).

Objectives

We present our work on the development and evaluation of maps from ICD-10 and ICD-10-CM codes to phecodes.

Methods

We mapped ICD-10 and ICD-10-CM codes to phecodes using a number of methods and resources, such as concept relationships and explicit mappings from the Centers for Medicare & Medicaid Services, the Unified Medical Language System, Observational Health Data Sciences and Informatics, Systematized Nomenclature of Medicine - Clinical Terms, and the National Library of Medicine. 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 assessed the fidelity of the ICD-10-CM map in comparison to the gold-standard ICD-9-CM phecode map by investigating phenotype reproducibility and conducting a PheWAS.

Results

We mapped >75% of ICD-10 and ICD-10-CM codes to phecodes. Of the unique codes observed in the UKBB (ICD-10) and VUMC (ICD-10-CM) cohorts, >90% were mapped to phecodes. We observed 70-75% reproducibility for chronic diseases and <10% for an acute disease for phenotypes sourced from the ICD-10-CM phecode map. Using the ICD-9-CM and ICD-10-CM maps, we conducted a PheWAS with a lipoprotein(a) (*LPA*) genetic variant, rs10455872, which replicated two known genotype-phenotype associations with similar effect sizes: coronary atherosclerosis (ICD-9-CM: *P*=1.96E-15, odds ratio (OR) = 1.60, 95% confidence interval (CI): 1.43-1.80 vs. ICD-10-CM: *P*=8.63E-16, OR = 1.60, 95% CI: 1.43-1.80) and chronic ischemic heart disease (ICD-9-CM: *P*=4.18E-10, OR = 1.56, 95% CI: 1.35-1.79 vs. ICD-10-CM: *P*=5.21E-05, OR = 1.47, 95% CI: 1.22-1.77).

Conclusions

This study introduces the “beta” versions of ICD-10 and ICD-10-CM to phecode maps that enable researchers to leverage accumulated ICD-10 and ICD-10-CM data for PheWAS in the EHR. The maps are available from https://phewascatalog.org and incorporated in the PheWAS R package, https://github.com/PheWAS/PheWAS.

Keywords

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

Introduction

Background

Electronic health records (EHRs) have become a powerful resource for biomedical research in the last decade, and many studies based on EHR data have used International Classification of Diseases (ICD) codes.[22] When linked to DNA biobanks, healthcare information in EHRs is a tool to discover genetic associations using billing codes in phenotyping algorithms. The phenome-wide association study (PheWAS) paradigm was introduced in 2010 as an approach that scans across a range of phenotypes, similar to genome-wide association studies. Studies using PheWAS have replicated hundreds of known genotype-phenotype associations and discovered dozens of new ones.[39–49] The initial version of phecodes consisted of 733 custom groups of ICD Ninth Revision, Clinical Modification (ICD-9-CM) diagnosis codes. The most recent iteration of phecodes consists of 1,866 hierarchical phenotype codes that map to 15,558 ICD-9-CM codes.[36,37] However, many health systems and international groups use ICD-10 or ICD-10-CM codes,[38] therefore necessitating a new phecode map.

Transition from ICD-9 to ICD-10

In 1979, the World Health Organization (WHO) developed ICD-9 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, whose end-of-life date was scheduled around the year 2000, but was delayed until October 2015.[38] In 1990, the WHO developed ICD-10,[50] which the NCHS used to create ICD-10-CM to replace ICD-9-CM.

Moving from ICD-9-CM to ICD-10-CM led to major structural changes in the coding system. First, the structure moved from a broadly numeric-based system in ICD-9-CM (e.g. 474.11 for “Hypertrophy of tonsils alone”) to an alphanumeric system in ICD-10-CM (e.g. J35.1 for the same condition). Second, ICD-10-CM contains much more granular information than

ICD-9-CM, as seen with the approximately tenfold increase in the number of diabetes-related codes in ICD-10-CM. ICD-10-CM also differs from ICD-9-CM in terms of semantics and organization.[38,51]

Compared to ICD-10, ICD-10-CM has more codes and granularity. While the 2018AA Unified Medical Language System (UMLS)[52] contains 94,201 unique ICD-10-CM codes, it has 12,027 unique ICD-10 codes after exclusion of range codes (e.g. ICD-10-CM A00-A09). Further, there are ICD-10 codes that do not exist in ICD-10-CM, and vice versa, such as ICD-10 A16.9 "Respiratory tuberculosis unspecified, without mention of bacteriological or histological confirmation", which has no ICD-10-CM equivalent.

Prior Work

To develop the original phecode system, one or more related ICD-9-CM codes were combined into distinct diseases or traits. For example, three depression-related ICD-9-CM codes 311, 296.31, and 296.2 are condensed to phecode 296.2 "Depression". With the help of clinical experts in disparate domains, such as cardiology and oncology, we have iteratively updated the phecode groupings.[34]

The phecode scheme is unique because it has built-in exclusion criteria to prevent contamination by cases in the control cohort. This is an important feature, as case contamination of control groups decreases the statistical power to find genotype-phenotype associations.[35] For each disease phenotype, we defined exclusion criteria by using our clinical knowledge and by consulting physician specialists.

An example for how users can use phecode exclusion criteria is illustrated by a type 2 diabetes study in the EHR. To define cases of type 2 diabetes, users include patients with ICD codes that map to phecode 250.2 "Type 2 diabetes". To create the control cohort, they include patients without phenotypes in the "DIABETES" group, which is comprised of phecodes in the range of 249-250.99. This prevents contamination of the control group by patients with diseases such as “Type 1 diabetes” (phecode 250.1) and “Secondary diabetes mellitus” (phecode 249). Excluded patients also include those with signs and symptoms commonly associated with type 2 diabetes, such as “Abnormal glucose” (phecode 250.4), which may indicate someone who has not yet been diagnosed with diabetes.

Though the phecode system is effective at replicating and identifying novel genotype-phenotype associations, PheWAS have largely been limited to using ICD-9-CM codes. A few studies have mapped ICD-10 codes to phecodes by converting ICD-10 to ICD-9-CM, and then mapping the converted ICD-9-CM codes to phecodes.[40,47] However, these studies limited their mappings to ICD-10 (non-CM) codes, did not provide a map to translate ICD-10-CM codes to phecodes, and did not evaluate the accuracy of these maps.

Goal of this Study

In this study, we developed and evaluated maps of ICD-10 and ICD-10-CM codes to phecodes. The primary aims of this study were to create an initial “beta” map to perform PheWAS using ICD-10 and ICD-10-CM codes and to focus the analyses on PheWAS-relevant codes. Our goal was to demonstrate that researchers should expect similar results from the ICD-10-CM phecode map compared to the gold-standard ICD-9-CM map. To accomplish this goal, we investigated phecode coverage, phenotype reproducibility, and the results from a PheWAS.

Methods

Databases

In this study, we used data obtained from the Vanderbilt University Medical Center (VUMC) and UK Biobank (UKBB) databases. The VUMC EHR contains clinical information derived from the medical records of >3 million unique individuals. The UKBB is a prospective longitudinal cohort study designed to investigate the genetic and environmental determinants of diseases in UK adults. Between 2006-2010, the study recruited >500,000 men and women aged 40-69 years. Participants consented to allow their data to be linked to their medical records. EHR records of UKBB were obtained under an approved data request application (ID:10775).

At the time of this study, VUMC had >2.5 years of ICD-10-CM data (~2015-10-01 to 2017-06-01), while the UKBB had >2 decades of ICD-10 data[53] (~1995-04-01 to 2015-03-31). VUMC includes codes for inpatient and outpatient encounters, whereas UKBB codes in this study are only inpatient codes.

Mapping ICD-10-CM and ICD-10 Codes to Phecodes

We extracted ICD-10-CM codes from the 2018AA release of the UMLS,[52] and used a number of automated methods to translate ICD-10-CM diagnosis codes to phecodes (Figure 1). We mapped 515 ICD-10-CM codes directly to phecodes by matching code descriptions regardless of capitalization, e.g. ICD-10-CM H52.4 “Presbyopia” to phecode 367.4 “Presbyopia”. We mapped 82,287 ICD-10-CM codes indirectly to phecodes using the existing ICD-9-CM phecode map.[36] To convert ICD-10-CM codes indirectly to phecodes, we used General Equivalence Mappings (GEMS) provided by the Centers for Medicare & Medicaid Services, that maps ICD-10-CM to ICD-9-CM and vice versa.[54] We included both equivalent and non-equivalent GEMS mappings (i.e. where the "approximate" flag was either "0" or "1"). As an example of this indirect approach, to map ICD-10-CM E11.9 “Type 2 diabetes mellitus without complications” to phecode 250.2 “Type 2 diabetes”: ICD-10-CM E11.9 to ICD-9-CM 250.0 “Diabetes mellitus without mention of complication” to phecode 250.2.

Since the GEMS do not provide mappings for all ICD-10-CM codes,[51] we complemented this approach with UMLS semantic mapping,[55] Observational Health Data Sciences and Informatics (OHDSI) concept relationships,[56,57] and National Library of Medicine (NLM) maps.[58] In this approach to indirect mapping, we first mapped ICD-10-CM codes to Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) through UMLS

Concept (CUI) equivalents, which were then converted to ICD-9-CM through either UMLS CUI equivalents,[52,55] OHDSI,[57] or NLM maps.[58] For example, ICD-10-CM L01.00 “Impetigo, unspecified” to CUI C0021099 to SNOMED CT 48277006 to OHDSI Concept ID 140480 to OHDSI Concept ID 44832600 to ICD-9-CM 684 to phecode 686.2 “Impetigo”.

There were two general instances when an ICD-10-CM code mapped to more than one phecode. First, some ICD-10-CM codes mapped to a parent phecode and one if its child phecodes that was lower in the hierarchy. To maintain the granular meanings of ICD-10-CM codes, we only kept the mappings to child phecodes, a decision that we could make due to the hierarchical structure of phecodes. For example, ICD-10-CM I10 “Essential (primary) hypertension” was mapped to phecodes 401 “Hypertension” and 401.1 “Essential hypertension”, but we only kept the mapping to phecode 401.1. Second, we kept all the mappings for ICD-10-CM codes that were translated to phecodes that were not in the same family. This can be seen in the mapping of ICD-10-CM D57.812 “Other sickle-cell disorders with splenic sequestration” to phecodes 282.5 “Sickle cell anemia” and 289.5 “Diseases of spleen”. This latter association created a polyhierarchical nature to phecodes that did not previously exist. To map ICD-10 (non-CM) codes to phecodes, we used ICD-10 codes also from the 2018AA UMLS.[52] ICD-10 codes were mapped to phecodes in a similar manner to ICD-10-CM, but since a GEMS to translate ICD-10 to ICD-9-CM was not available, we used only string matching and previously manually-reviewed resources from the UMLS,[55] NLM,[58] and OHDSI.[56,57]

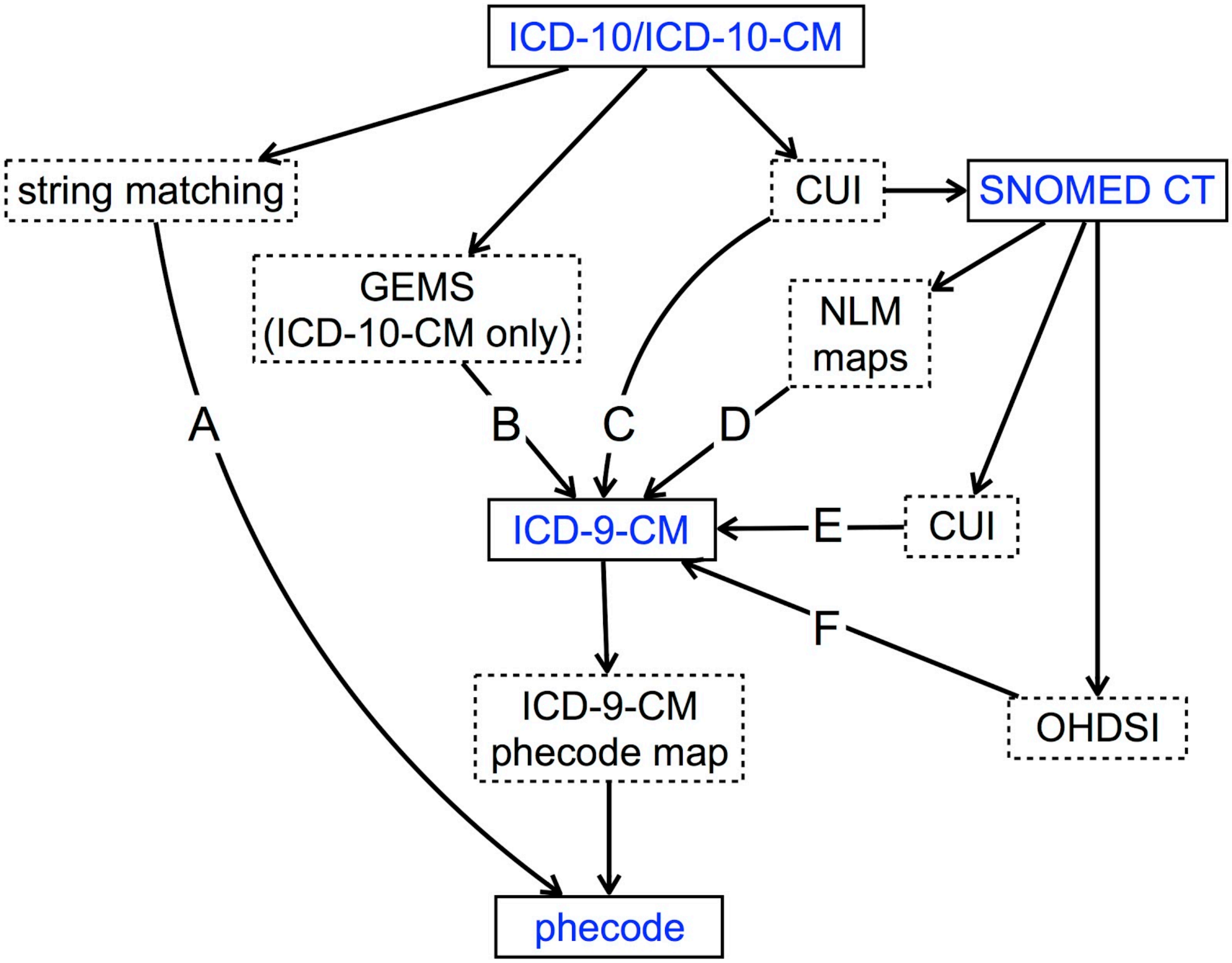

Figure 1. Mapping strategy for ICD-10 (non-CM) and ICD-10-CM diagnosis codes to phecodes. We mapped ICD-10-CM codes directly by matching code descriptions (path A) or indirectly to phecodes, using a number of manually-validated mapping resources (paths B, C, D, E, and F). In path D, we used NLM's SNOMED CT to ICD-9-CM one-to-one and many-to-one maps.[58] To map ICD-9-CM codes to phecodes, we applied Phecode Map 1.2 with ICD-9 Codes (ICD-9-CM phecode map).[36] Boxes with solid lines indicate clinical terminologies, and those with dashed lines describe the resources and mapping methods used. ICD-10-CM: International Classification of Diseases, Tenth Revision, Clinical Modification. SNOMED CT: Systematized Nomenclature of Medicine Clinical Terms. GEMS: General Equivalence Mappings. UMLS: Unified Medical Language System. OHDSI: Observational Health Data Sciences and Informatics. CUI: Concept Unique Identifier. NLM: National Library of Medicine.

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 source codes in UKBB and VUMC, respectively, we calculated the number of source codes in the 2018AA UMLS, number of source codes mapped to phecodes, and number of mapped and unmapped source codes that were used in the two EHRs (Figure 2). To identify potential limitations of our automated mapping approach, two authors with clinical training (P.W., W.Q.W.) manually reviewed all the unmapped ICD-10 and ICD-10-CM codes that were used at UKBB and VUMC, respectively.

Comparison of Phenotypes Generated from the ICD-10-CM Phecode Map

We aimed to provide evidence that the ICD-10-CM phecode map resulted in phenotypes similar to those sourced from the ICD-9-CM phecode map. First, we selected 357,728 patients in the VUMC EHR who had ≥1 ICD-9-CM and ≥1 ICD-10-CM codes in two 18-month windows. We selected windows to occur prior to and after VUMC's transition to ICD-10-CM. To reduce potential confounders, we left a six-month buffer after ICD-9-CM was replaced with ICD-10-CM. Further, the ICD-10-CM observation window ended before VUMC switched from its locally developed EHR[59] to the Epic system. This created two windows ranging from 2014-01-01 to 2015-06-30 for ICD-9-CM, and 2016-01-01 to 2017-06-30 for ICD-10-CM (Figure 3). The final cohort consisted of 55.10% female with mean (standard deviation, SD) 45 (25) years old. From the two observation periods, we extracted all ICD-9-CM and ICD-10-CM codes for each patient. We then mapped these codes to phecodes using the ICD-9-CM phecode[36] and ICD-10-CM phecode maps.

We used the patient cohort to test our hypothesis that the ICD-10-CM phecode map created phenotype definitions that were comparable to those generated using the gold-standard ICD-9-CM phecode map. For this analysis, we used four common chronic diseases (Hypertension, Hyperlipidemia, Type 1 Diabetes, and Type 2 Diabetes) and chose one acute disease (Intestinal infection) as a negative control. We expected that a large majority of the chronic disease patients and small minority of the acute disease patients from the ICD-9-CM era would reproduce the same phenotypes during the ICD-10-CM era. We defined the phenotype cases as follows: Hypertension with phecodes 401.* ("*" means one or more digits or a period);

Hyperlipidemia, phecodes 272.*; Type 1 diabetes, phecodes 250.1*; Type 2 diabetes, phecodes 250.2*; Intestinal infection, phecodes 008.*.

For each phenotype, we reported the number of ICD-9-CM cases and the number of those individuals who were also ICD-10-CM cases. To identify the possible reasons for individuals who were not identified as phenotype cases in the ICD-10-CM period, two authors with clinical training (P.W., W.Q.W.) manually reviewed the EHRs of ten randomly selected patients from each chronic disease group, except Type 1 diabetes, for a total of thirty patients.

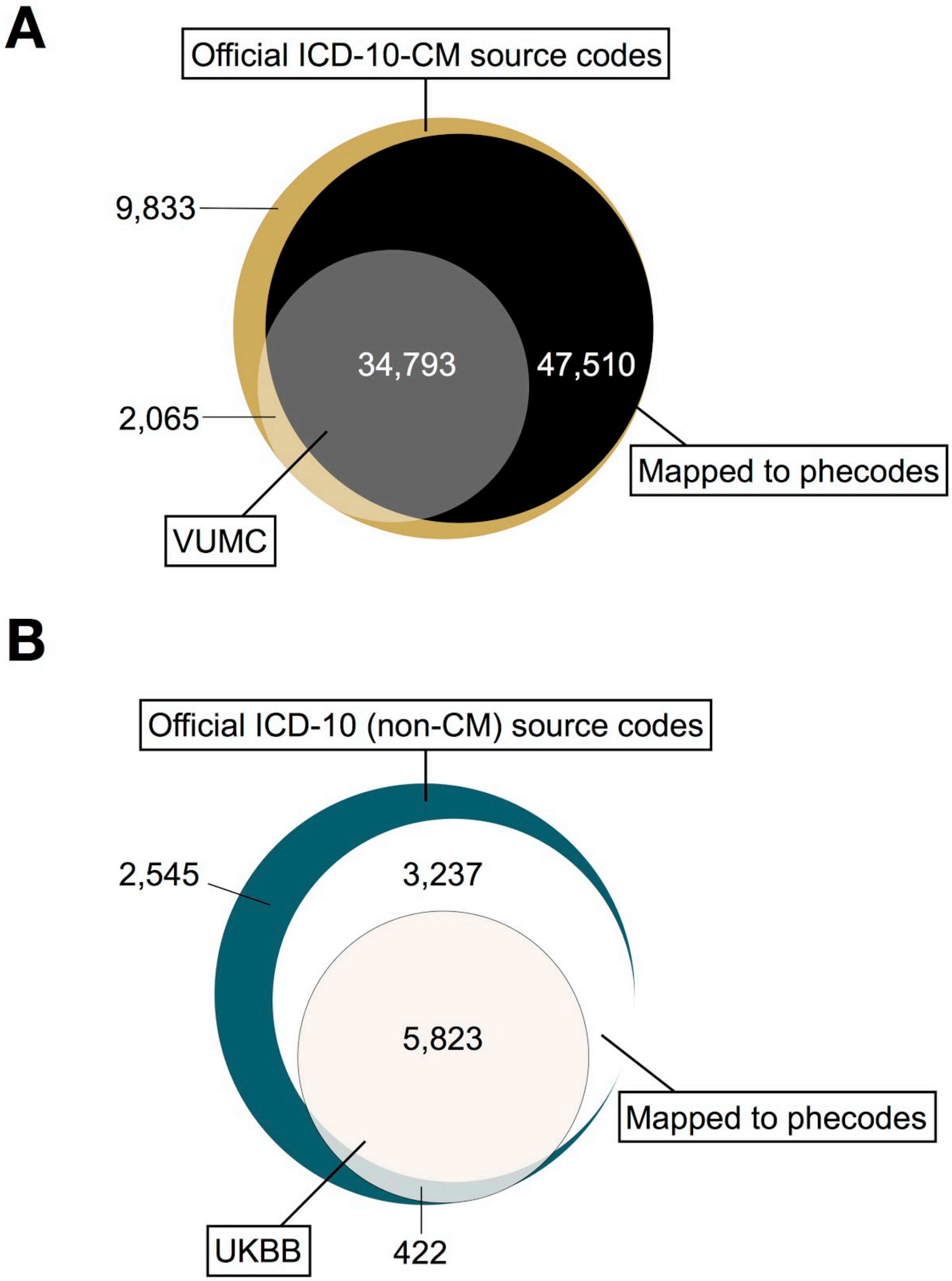

Figure 2. Counts of distinct ICD-10-CM source codes at VUMC and ICD-10 (non-CM) source codes in UKBB. (A) Number of unique ICD-10-CM codes in each category. For example, there were 34,793 unique codes (grey section) that were in the official ICD-10-CM system, observed in the VUMC dataset, and mapped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

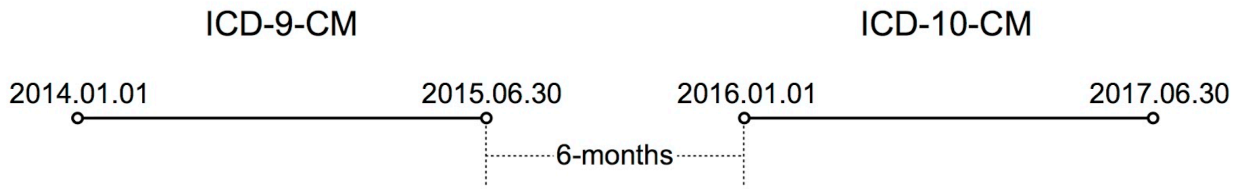

Figure 3. Timeline of the two 18-month periods from which ICD-9-CM and ICD-10-CM codes from VUMC were analyzed. The cohort of 357,728 patients had at least one ICD-9-CM and one ICD-10-CM code in the respective 18-month windows.

Comparative PheWAS Analysis of Lipoprotein(a) (LPA) Single-nucleotide polymorphism (SNP)

To evaluate the accuracy of the ICD-10-CM phecode map, we performed two PheWAS on an *LPA* genetic variant (rs10455872) using mapped phecodes from ICD-9-CM and ICD-10-CM. The *LPA* SNP is associated with increased risks of developing hyperlipidemia and cardiovascular diseases.[60–62]

We used data from BioVU, the de-identified DNA biobank at VUMC to conduct the PheWAS.[33] We identified 13,900 adults (56.90 % female with mean (standard deviation, SD) 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. For rs10455872, we observed 86.7% AA, 12.8% AG, and 0.5% GG. We used 1,632 phecodes that overlapped in the time windows for PheWAS using the R PheWAS package[37] with binary logistic regression, adjusting for age, sex, and race.

Results

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

Of all possible ICD-10-CM codes,[52] 82,303 (87.37%) mapped to at least one phecode, with 7,881 (8.37%) mapping to >1 phecode. For example, ICD-10-CM I25.708 “Atherosclerosis of coronary artery bypass graft(s), unspecified, with other forms of angina pectoris” mapped to phecodes 411.3 “Angina pectoris” and 411.4 “Coronary atherosclerosis”. Of all possible ICD-10 codes, 9,060 (75.33%) mapped to at least one phecode, and 289 (2.40%) mapped to >1 phecode. For example, ICD-10 code B21.1 “HIV disease resulting in Burkitt lymphoma” maps to phecodes 071.1 “HIV infection, symptomatic” and 202.2 “Non-Hodgkins lymphoma”.

Among the 36,858 ICD-10-CM codes used at VUMC, 34,793 (94.40%) codes were mapped to phecodes. In the UKBB, 5,823 (93.24%) of the ICD-10 codes mapped to phecodes (Table 1, Figure 2). 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 those codes that mapped to phecodes (Table 1). Among the total number of codes used, 89.72% of ICD-10-CM and 83.68% of ICD-10 codes were mapped to phecodes.

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

	ICD-10-CM (No.) (VUMC)	**ICD-10 (No.) (UKBB)**
Official classification systems		
Unique codes	94,201	12,027
Unique codes mapped	82,303 (87.37%)	9,060 (75.33%)
Official codes used in cohorts		
Unique codes	36,858	6,245
Unique codes mapped	34,793 (94.40%)	5,823 (93.24%)
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 (89.72%)	4,279,544 (83.68%)

Analysis of Unmapped ICD-10 and ICD-10-CM Codes

Majority of the unmapped ICD-10 codes used in the UKBB dataset represented medical concepts related to personal (i.e. past medical history) or family history of disease. For ICD-10-

CM, removing codes used at VUMC that we expected to be unmapped (i.e. local or supplementary classification codes) left 2,065 ICD-10-CM codes that did not map to a phecode. After excluding X, Y, and Z codes (1,395 codes), 670 codes remained, majority of which represented either “external causes of morbidity” or “factors influencing health status and contact with health services”. All of the remaining unmapped ICD-10-CM codes in this cohort had <200 unique individuals (i.e. <.1% of the cohort), and majority of the ICD-10-CM codes with >10 unique individuals were phenotypes that are most likely due to non-genetic factors. For example, 287 (59.2%) of the unmapped ICD-10-CM codes represented external causes of morbidity, such as assault and injuries due to motor vehicle accidents.

Reproducibility Analysis of the ICD-10-CM Phecode map

In the defined 18-month time windows, a cohort 357,728 patients had both ICD-9-CM and ICD-10-CM codes (Figure 3). For the chronic diseases, 70-75% of individuals with the relevant phecodes in the ICD-9-CM observation period also had the same phecodes of interest during the ICD-10-CM period. On the contrary, for the reproducibility analysis with an acute disease, we observed that <10% of individuals who had phecodes 008.* (Intestinal infection) in the ICD-9-CM period also had the same phecodes in the ICD-10-CM period (Table 2).

Table 2. ICD-10-CM phecode map reproducibility analysis.

Phenotype	Phecodes[a]	No. ICD-9-CM cases	No. Individuals (%), (ICD-10-CM case | ICD-9-CM case)[b]
Hypertension	401.*	65,216	49,468 (75.85%)
Hyperlipidemia	272.*	51,187	36,187 (70.70%)
Type 1 diabetes	250.1*	5,782	4,412 (76.31%)
Type 2 diabetes	250.2*	25,077	19,066 (76.03%)
Intestinal infection	008.*	3,410	273 (8.01%)

[a] In the phecode column, "*" means ≥1 digits or a period, e.g. phecode 401.* = phecodes 401, 401.1, 401.3, 401.22, 401.21, 401.2.
[b] In the last column, "ICD-10-CM case | ICD-9-CM case" indicates patients who were cases for the phenotype of interest during the ICD-9-CM period who were also ICD-10-CM cases.

To identify the reasons that may explain why some patients were not identified as cases for the phenotype of interest during the ICD-10-CM period, we manually reviewed their medical records. Thirty patients were selected for review, ten each from the Hypertension, Hyperlipidemia, and Type 2 diabetes cohorts (Table 3). We found that none of the patients had a relevant ICD-10-CM code for the phenotype being studied in the 18-month observation period. Reasons for patients not being ICD-10-CM cases include: patients were labeled with the relevant ICD-10-CM code(s) outside of the short ICD-10-CM observation window (8 patients), patients had <2 visits at VUMC during the ICD-10-CM period and/or were only seen by physician specialists (10 patients; e.g. patient with hypertension was only seen by their neurologist during the ICD-10-CM period), and patients were inconsistently diagnosed (2 people; e.g. patient with Type 1 diabetes given Type 2 diabetes ICD-9-CM code). No cases were missed due to errors in the ICD-10-CM phecode map.

Table 3. ICD-10-CM reproducibility analysis, manual chart review results.

Phenotype	Group	Number of People
Hypertension	absence of ICD-10-CM only	2
Hyperlipidemia	absence of ICD-10-CM only	4
Type 2 diabetes	absence of ICD-10-CM only	4
Hypertension	short observation window	1
Hyperlipidemia	short observation window	4
Type 2 diabetes	short observation window	3
Hypertension	limited number of visits/specialist	7
Hyperlipidemia	limited number of visits/specialist	2
Type 2 diabetes	limited number of visits/specialist	1
Hypertension	inconsistent diagnosis	0
Hyperlipidemia	inconsistent diagnosis	0
Type 2 diabetes	inconsistent diagnosis	2

Comparative PheWAS Analysis of LPA SNP, rs10455872

To further evaluate the ICD-10-CM phecode map, we performed and compared the results of PheWAS analyses for rs10455872. One PheWAS was conducted using the ICD-9-CM map and another was conducted using the ICD-10-CM map. Both analyses replicated previous findings with similar effect sizes: coronary atherosclerosis (ICD-9-CM: *P*=1.96E-15, odds ratio (OR) = 1.60, 95% confidence interval (CI): 1.43-1.80 vs. ICD-10-CM: *P*=8.63E-16, OR = 1.60, 95% CI: 1.43-1.80) and chronic ischemic heart disease (ICD-9-CM: *P*=4.18E-10, OR = 1.56, 95% CI: 1.35-1.79 vs. ICD-10-CM: *P*=5.21E-05, OR = 1.47, 95% CI: 1.22-1.77) (Figure 4).

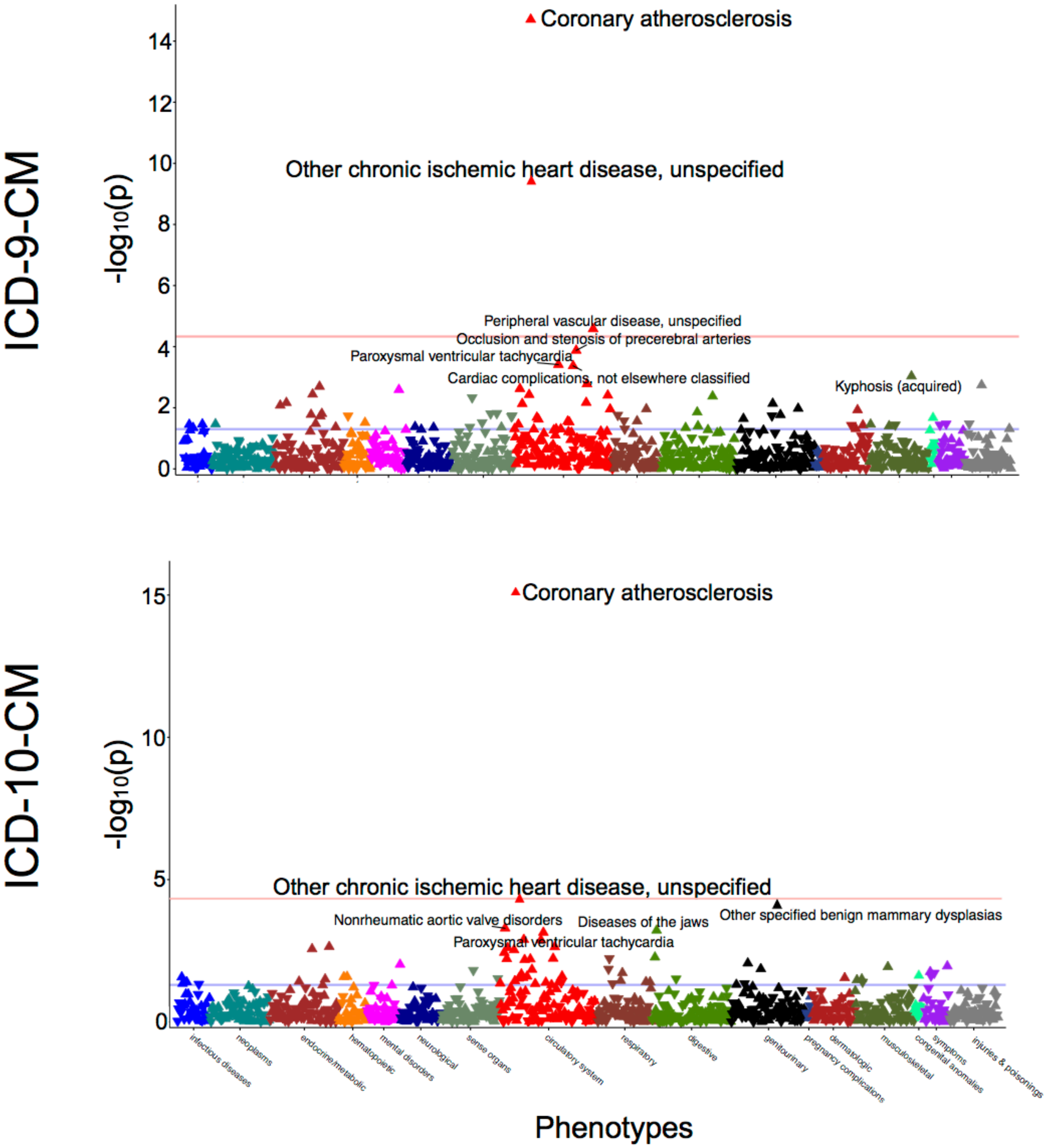

Figure 4. Comparative PheWAS of lipoprotein(a) (*LPA*) genetic variant, rs10455872. “Coronary atherosclerosis” (phecode 411.4) and “Other chronic ischemic heart disease” (phecode 411.8) were top hits associated with rs10455872 in a PheWAS analysis conducted using ICD-9-CM (top) and ICD-10-CM (bottom) phecode maps. Analyses were adjusted for age, sex, and race.

Discussion

Main Findings: Maps of ICD-10 and ICD-10-CM Codes to Phecodes have High Coverage and Yield Similar Results as the ICD-9-CM Phecode Map.

In this study, we described the process of mapping ICD-10 and ICD-10-CM codes to phecodes, and evaluated the results of the new maps in two databases. These results show that the majority of the ICD-10 and ICD-10-CM codes used in EHRs were mapped to phecodes. Our analyses suggest that researchers can expect that phenotypes sourced using the ICD-10-CM phecode map will be similar to those sourced from the gold-standard ICD-9-CM phecode map. As the use of ICD-10 and ICD-10-CM codes increases, so does the need for convenient and reliable methods of aggregating codes to represent clinically meaningful phenotypes.

Since the introduction of phecodes, many studies have demonstrated the value of aggregating ICD-9-CM codes for genetic association studies. These maps will allow biomedical researchers to leverage clinical data represented by ICD-10 and ICD-10-CM codes for their large-scale PheWAS in the EHR. They will also allow researchers to combine phenotypes as phecodes mapped from ICD-9 and ICD-10 based coding systems, thereby increasing the size of their patient cohorts and statistical power of their studies. The maps are available from the PheWAS Resources page[36] and are incorporated in the PheWAS R package, version 0.99.5-2.[37,63]

ICD-10 and ICD-10-CM Codes not Mapped to Phecodes

Analysis of the unmapped ICD-10 codes demonstrates a possible area of expansion for phecodes. The ICD-10 phecode map did not include medical concepts representing personal history or family history of disease.

We observed that a majority of the unmapped ICD-10-CM codes represented concepts that we did not expect to have phecode equivalents. Majority of the codes were from ICD-10-CM chapters 20 “External causes of morbidity” and 21 “Factors influencing health status and contact with health services”. Codes from chapter 19 “Injury, poisoning, and certain other consequences of external causes" also made up a large proportion of unmapped codes, such as ICD-10-CM T38.3X6A "Underdosing of insulin and oral hypoglycemic [antidiabetic] drugs, initial encounter". We did not expect ICD-10-CM T38.3X6A to map to a phecode, as it is an encounter code that is not relevant to PheWAS. Three-digit codes that are not frequently used for reimbursement purposes, such as ICD-10-CM I67 "Other cerebrovascular diseases", also made up a large number of unmapped codes. A few potential clinically meaningful phenotypes, such as ICD-10-CM O04.6 "Delayed or excessive hemorrhage following [induced] termination of pregnancy", were unmapped and represent areas of potential expansion for phecodes.

ICD-10-CM Phecode Map Phenotype Reproducibility Analysis

In general, our analysis suggests that in the majority of the cases in which phenotypes are not reproduced in the ICD-10-CM observation period are not due to errors in the ICD-10-CM phecode map. This study's reproducibility analysis (Table 2) demonstrates that the vast majority of patients (70-75%) with phecodes of four chronic diseases sourced from ICD-9-CM codes were also phenotype cases in the ICD-10-CM era. In comparison, when the same experiment is repeated for an acute disease (Intestinal infection), a minority (<10%) of patients had the same phenotype in the ICD-10-CM period.

Using the ICD-9-CM and ICD-10-CM maps, PheWAS found significant genetic associations with similar effect sizes for coronary atherosclerosis and chronic ischemic heart disease (Figure 4). Results of this analysis provide additional support for the accuracy of the ICD-10-CM map when compared to the gold-standard ICD-9-CM phecode map.

PheWAS Using ICD-10 Phecode Map

Two published studies have used the ICD-10 phecode map to identify genotype-phenotype associations using UKBB data. Zhou et al. used the map to demonstrate a method that adjusts for case-control imbalances in a large genome-wide PheWAS.[64] Li et al. used the same map to estimate the causal effects of elevated serum uric acid across the phenome.[65]

Utilization of Phecodes Outside of PheWAS

In addition to being employed for PheWAS, phecodes have been used to answer a range of questions in biomedicine. Phecodes have been used to identify features in radiographic images that are associated with disease phenotypes,[66] and used in machine learning models to improve cardiovascular disease prediction.[67] In a recent study to understand public opinion about diseases, Huang et al. identified articles about diseases and mapped them to phecodes.[68] Motivated by the difficulties in automatically translating diagnosis codes in the EHR, Shi et al. used phecodes to map ICD-9-CM diagnosis codes from one health system to another.[69] Phecodes have also been applied to identify conditions for aggregation in “phenotype risk scores”, much as SNPs are aggregated as a genetic risk score, to identify Mendelian diseases and determine pathogenicity of genetic variants.[70]

Related Work

The Clinical Classification Software (CCS) is another maintained system for aggregating ICD codes into clinically meaningful phenotypes. CCS was originally developed by the Agency for Healthcare Research and Quality (AHRQ) to cluster ICD-9-CM diagnosis and procedure codes to a smaller number of clinically meaningful categories.[71] CCS has been used for many purposes, such as to measure outcomes[72] and to predict future health care usage.[73] In a previous study, we showed that phecodes better aligned with diseases mentioned in clinical practice and that are relevant to genomic studies, than CCS for ICD-9-CM (CCS9) codes.[35] We found that phecodes outperform CCS9 codes, in part because CCS9 was not as granular as phecodes. Since CCS for ICD-10-CM (CCS10) is of similar granularity as CCS9 (283 vs. 285

disease groups),[71] we believe that the phecode map would likely still better represent clinically meaningful phenotypes in genetic research.

Limitations

This study has limitations. First, only 84.14% (1570/1866) of phecodes are mapped to at least one ICD-10 code. This may be due in part to the automated strategy that we used to map ICD-10 to ICD-9-CM. Second, the VUMC data are from a single site, thereby making it difficult to generalize the results of our accuracy studies (e.g. phenotype reproducibility analysis and *LPA* SNP PheWAS) to patient cohorts in other EHRs. Third, we have not yet manually reviewed all of the mappings in these “beta” phecode maps, and our assumptions that the manually-reviewed resources (e.g. NLM and OHDSI) are highly accurate could have affected the accuracy of the new phecode maps. For example, in the 2009 ICD-10-CM to ICD-9-CM GEMS, >90% of the mappings were “approximate” (i.e. non-equivalent).[38] For this study’s purposes, we aimed to maximize phecode coverage of ICD source codes, and thus included both equivalent and non-equivalent 2018 GEMS translations, which could have decreased mapping performance.

Fourth, our automated approach to map >80,000 ICD-10-CM and >9,000 ICD-10 codes to phecodes with minimal human-engineering could have decreased the accuracy of the final maps. Hripcsak et al.[74] recently evaluated the effects of translating ICD-9-CM codes to SNOMED CT codes on the creation of patient cohorts. In general, they found that mapping source billing codes to a standard clinical vocabulary (e.g. ICD-9-CM to SNOMED CT) did not greatly affect cohort selection. Their findings suggested that optimized domain knowledge-engineered mappings outperformed simple automated translations between clinical vocabularies. Using four phenotype concept sets, they showed that automated mappings resulted in errors of up to 10% and that domain-knowledge engineered mappings to have errors of <.5%. Other studies have also found that mapping performance is generally better with smaller value sets.[51] To create a more comprehensive and accurate map between ICD-9-CM and ICD-10-CM, future mapping studies could consider using an iterative forward and backward mapping approach using GEMS.[51]

Future Directions

Currently, if an ICD-10 or ICD-10-CM code maps to ≥2 codes unlinked phecodes, we keep all of the mappings. In subsequent studies, it will be important to further scrutinize these mappings to ensure accuracy through manual review. As new ICD-10-CM codes are released, we plan to assess their relevance to clinical practice and genetic research, and decide whether we should translate them to phecodes. We intend to address the unmapped source codes (e.g. ICD-10-CM E78.41 “Elevated Lipoprotein(a)”) by potentially expanding the phecode system, and to systematically evaluate the mappings with input from users.

Conclusions

In this paper, we introduced our work on mapping ICD-10 and ICD-10-CM codes to phecodes. We provide initial “beta” maps with high coverage of EHR data in two large databases. Results from this study suggested that the ICD-10-CM phecode map created phenotypes similar to those generated by the ICD-9-CM phecode map. These mappings will enable researchers to leverage accumulated ICD-10 and ICD-10-CM data in the EHR for large PheWAS.

Abbreviations

PheWAS: phenome-wide association studies

EHR: electronic health record

ICD: International Classification of Diseases

AHRQ: Agency for Healthcare Research and Quality

CM: Clinical Modification

WHO: World Health Organization

NCHS: National Center for Health Statistics

UMLS: Unified Medical Language System

GEMS: General Equivalence Mappings

SNOMED CT: Systematized Nomenclature of Medicine Clinical Terms

CUI: Concept Unique Identifier

OHDSI: Observational Health Data Sciences and Informatics

NLM: National Library of Medicine

VUMC: Vanderbilt University Medical Center

UKBB: UK Biobank

SD: standard deviation

OR: odds ratio

LPA: lipoprotein(a)

SNP: single nucleotide polymorphism

CCS: Clinical Classification Software

CHAPTER 3

Impact of Vocabulary Mapping on High-throughput PheWAS

Introduction

In the past decade, many genetic studies have transitioned from traditional purpose-built cohorts to using data collected in the electronic health record (EHR). For example, national programs such as the UK Biobank (UKBB),[53] and Million Veterans Program (MVP)[75] are longitudinal cohort studies that have collected health-related information, including genetic and EHR phenotype data, to better understand human health and disease. Recently, NIH started a similar project called the *All of Us* Research Program (AoU; formerly known as the Precision Medicine Initiative Cohort Program) with the goal of creating an ethnically diverse cohort of 1 million or more individuals.[76]

One of the main goals of these national programs is to discover genetic correlates of disease to enable genome medicine.[76] Genetic information from participants (ie, genotyping) is collected using modern high-throughput tools, like sequencing and microarrays.[11] Though there are some challenges in genotype calling,[77,78] it is relatively easy to map the raw signals from these platforms to discrete locations and classify base identities, as DNA nucleotides are generally located in the same physical location on a chromosome. For example, rs10455872, lipoprotein(a) single-nucleotide polymorphism (SNP) is mapped to position 160589086 on chromosome 6 in Genome Reference Consortium Human Build 38. Thus, mapping genetic information from disparate sources to a reference build is relatively straightforward.

On the contrary, there are many ways to define a phenotype, which results in many difficulties in creating a patient cohort for many diseases using just one tool. For example, creating a cohort to study hypertension using EHR data can be done through multiple methods.

One can identify cases of hypertension as individuals with systolic pressures over 120 mmHg or diastolic pressures over 80 mmHg. One can also define cases as patients with hypertension ICD code(s) or who have ever been exposed to medications used to lower blood pressure. Research has shown the best method uses a combination of approaches.[79]

One commonly used high-throughput phenotyping method is to map ICD codes to phecodes. Phecodes were created to largely represent phenotypes that were affected by variations in the genome. When phecodes were first introduced, there was only a map to translate ICD-9-CM billing codes, but a beta version to map ICD-10 and ICD-10-CM codes to phecodes was recently released.[80] Since the UKBB contains many ICD-10 codes and the U.S. health system recently transitioned to the ICD-10-CM system, leading to more and more phenotype data contained in ICD-10-CM codes, these maps allow researchers to translate ICD-10/ICD-10-CM data to phecodes.

A few studies have used the ICD-10 phecode map to conduct PheWAS, primarily using UKBB data.[64,65] In the paper where we described the development and initial demonstration of the ICD-10/ICD-10-CM phecode maps, we applied the ICD-10-CM phecode map to replicate known associations between a lipoprotein(a) genetic variant (rs10455872) and increased likelihood of developing coronary heart disease. However, a more comprehensive evaluation of the ICD-10-CM phecode map has not been evaluated for use in identifying genotype-phenotype associations.

In this study, we evaluate the performance of the ICD-10-CM phecode map using PheWAS with 5 genetic variants that were studied in the paper demonstrating the potential of PheWAS using ICD-9-CM diagnostic codes.

Methods

Study setting

We used data from the Synthetic Derivative, Vanderbilt University Medical Center’s (VUMC's) de-identified EHR, which has clinical data from >3 million unique patients[33]. ICD-9-CM codes in this study were from inpatient and outpatient visits between October 1987-November 2017. ICD-10-CM codes in this study were from inpatient and outpatient visits between April 2014-November 2017.

Cohort selection

In this study, we used a case-control study design. All individuals had ≥1 ICD-9-CM and ≥1 ICD-10-CM codes. Cases were defined as individuals who had relevant ICD-9-CM/ICD-10-CM code on at least two distinct days. Controls were individuals with no ICD-9-CM/ICD-10-CM code in the corresponding range. For each phenotype, patients were excluded from the analysis if they had a phecode that was included in the exclusion criteria.

Phenome-wide association analysis (PheWAS)

Using the `Phecode v1.2 ICD-10-CM code map beta 1`, we conducted PheWAS for 5 SNPs that were included in the original PheWAS paper[32]: rs3135388, rs17234657, rs2200733, rs1333049, and rs6457620. For each SNP, we calculated the association with 1,856 unique phenotypes. We conducted the PheWAS using the R PheWAS package[37] with binary logistic regression assuming an additive genetic model, adjusting for age at the start of 2019, sex, and race.

GWAS catalog reproducibility analysis

We determined the proportion of known genotype-phenotype associations on GWAS catalog[81] that were reproduced. This evaluation only considered phenotypes on GWAS catalog that overlapped with PheWAS phenotypes. Only associations with odds ratios (ORs) in the expected direction and p-value <0.05 were considered reproduced.

PheWAS discovery analysis

For genotype-phenotype associations not found in the GWAS catalog as of December 19, 2019, and had a p-value less than the Bonferroni corrected threshold ($0.05/1817=2.8x10^{-5}$), we considered as potentially novel associations.

Evaluation of ICD-10-CM phecode map PheWAS performance

To evaluate the ICD-10-CM phecode map, we compared the PheWAS ORs for phecodes mapped ICD-9-CM, ICD-10-CM, and Combined (ICD-9-CM+ICD-10-CM). In addition to comparing ORs for all tests, we also compared only the genotype-associations associations with p<0.05 in both PheWAS analyses being compared. The output of the comparisons were Pearson correlation coefficient (R) and p-value. We used a Bonferroni corrected P-value (p<0.05/6 = 0.008) to determine significance for the Pearson measurements.

Statistical analysis

Analyses in python (version 3.6.7), were conducted using these software packages: pandas (version 0.23.4),[82] numpy (version 1.16.0),[83] statsmodels (version 0.9.0),[84] scipy (version 1.2.0),[85] and jupyter notebook (version 5.7.4)[86]. Analyses in R (version 3.5.1) were conducted using these software packages: PheWAS (version 0.99.5-3)[63], readr (version 1.3.1), ggplot2 (version 3.2.1), and tidyverse (version 1.2.1).

Results

Genotype selection and population characteristics

Individuals were genotyped using the Infinium® Expanded Multi-Ethnic Genotyping Array (MEGA$_{EX}$). Starting from a cohort of 96,921 unique individuals in the MEGA genotyped cohort, we studied 56,929 multi-ethnic individuals who had ≥1 ICD-9-CM and ≥1 ICD-10-CM code, in this PheWAS. Demographics of the study cohort are in Table 1**.** The average age was 49.9 years, 58.3% were female, and 76.1% were of European descent. Subjects had a mean follow-up of 10.4 (SD=5.9) years.

Table 1. Population characteristics

Description	**Value, or mean (SD)**
Patients (n)	56,929
Age (years)	49.9 (22.3)
Female (%)	58.3
European (%)	76.1
Total ICD-9-CM codes/person	187 (249)
Unique ICD-9-CM codes/person	60 (53)
Total ICD-10-CM codes/person	111 (195)
Unique ICD-10-CM codes/person	26 (32)
EHR follow-up (years)	10.4 (5.9)

-Age is age at start of 2019

GWAS catalog reproducibility analysis

For the 5 SNPs and 1,856 phenotypes analyzed in this study, a total of 6 genotype-phenotype associations were listed in the GWAS catalog at the time of analysis. Using the ICD-9-CM phecode map, we replicated 5/6 (83.33%) known phenotype associations from the GWAS catalog for the SNPs tested (Table 2). Using the ICD-10-CM phecode map, we replicated the same 5/6 phenotype associations (Table 2). We obtained the same results for the combined (ICD-9-CM+ICD-10-CM) PheWAS (Table 2).

Table 2. GWAS catalog reproducibility analysis for ICD-9-CM, ICD-10-CM, and Combined (ICD-9-CM+ICD-10-CM) PheWAS

SNP	Gene	PheWAS phenotype	Previous OR	PheWAS P-value	PheWAS OR
rs3135388	HLA-DRA	Multiple sclerosis	2.75	ICD-9-CM: 1.54E-71 ICD-10-CM: 4.97E-71 Combined: 1.03E-71	ICD-9-CM: 2.49 ICD-10-CM: 2.59 Combined: 2.48
rs17234657	RNU1-150P, AC093277.1	Regional enteritis	1.54	ICD-9-CM: 9.76E-11 ICD-10-CM: 8.16E-09 Combined: 1.31E-09	ICD-9-CM: 1.36 ICD-10-CM: 1.31 Combined: 1.33
rs2200733	PITX2, LINC01438	Atrial fibrillation and flutter	1.72	ICD-9-CM: 9.53E-32 ICD-10-CM: 9.4E-24 Combined: 1.46E-29	ICD-9-CM: 1.54 ICD-10-CM: 1.45 Combined: 1.48
rs2200733	PITX2, LINC01438	Occlusion of cerebral arteries, with cerebral infarction	1.26	ICD-9-CM: 0.23 ICD-10-CM: 0.17 Combined: 0.22	ICD-9-CM: 1.08 ICD-10-CM: 0.75 Combined: 1.07
rs1333049	CDKN2B-AS1	Coronary atherosclerosis	1.24	ICD-9-CM: 3.81E-15 ICD-10-CM: 2.61E-16 Combined: 7.46E-15	ICD-9-CM: 1.19 ICD-10-CM: 1.21 Combined: 1.18
rs6457620[a]	HLA-DQB1, MTCO3P1	Rheumatoid arthritis	2.55	ICD-9-CM: 1.44E-24 ICD-10-CM: 6.01E-20 Combined: 6.64E-23	ICD-9-CM: 1.52 ICD-10-CM: 1.49 Combined: 1.45

[a] rs6457620 minor allele = G

Evaluation of ICD-10-CM phecode map PheWAS performance

For each SNP, we made a total 6 of comparisons of PheWAS ORs: ICD-9-CM vs. ICD-10-CM, ICD-9-CM vs. Combined, and ICD-10-CM vs. Combined (Table 3, Figure 7). The 6 analyses were comprised of comparing ORs of all tests performed (3 comparisons) and only comparing ORs where the association p-value <0.05 in both PheWAS (3 comparisons). Among the same PheWAS analysis, the Pearson correlation coefficients were higher when only comparing ORs with association p-values <0.05. For example, for the ICD-9-CM vs. ICD-10-CM PheWAS comparison, the Pearson correlation coefficient was mean (SD) 0.38 (0.06) when all ORs were compared and 0.95 (0.03) when only comparing ORs with p<0.05. The correlations between the all Combined PheWAS ORs (ICD-9-CM vs. Combined = 0.85 (0.01); ICD-10-CM vs. Combined = 0.73 (0.03)) were also larger than that of ICD-9-CM vs. ICD-10-CM. All Pearson correlations had p<0.008.

Table 3. Evaluation of ICD-10-CM phecode map PheWAS performance

SNP	**9 vs.10**	**9 vs.10, sig**[a]	**9 vs. Combined**	**9 vs. Combined, sig**[a]	**10 vs. Combined**	**10 vs. Combined, sig**[a]
rs3135388	0.43	0.98	0.86	0.99	0.75	0.92
rs17234657	0.40	0.91	0.86	0.98	0.75	0.94
rs2200733	0.28	0.97	0.83	0.97	0.70	0.94
rs6457620	0.43	0.97	0.86	0.98	0.75	0.97
rs1333049	0.35	0.94	0.86	0.99	0.70	0.96

-Values are Pearson correlation (R)
-All Pearson correlations have p<0.008
[a]Pearson correlation (R) for PheWAS associations with p<0.05

PheWAS Discovery Analysis

We observed that 4/5 SNPs tested had genotype-phenotype associations that were not listed in the GWAS catalog (Bonferroni corrected p-value, Table 4). The exception was rs2200733, for which all significant phenotype associations were present in the GWAS catalog. Many phenotype associations were found for rs3135388, like thyroid-related diseases, such as phecode 244 “Hypothyroidism” (ICD-9-CM OR = 0.84). Other potential novel phenotype associations include negative correlations of rs3135388 with phecode 250.1 “Type 1 diabetes” (Combined OR=0.5), 250.2 “Type 2 diabetes” (ICD-9-CM OR = 0.83), and 272.1 “Hyperlipidemia” (ICD-9-CM OR = 0.88) (Figure 1, Table 4). For rs17234657, potential novel phenotype associations include phecode 280.1 “Iron deficiency anemias, unspecified or not due to blood loss” (ICD-9-CM OR = 1.19) and phecode 379.5 “Disorders of iris and ciliary body” (ICD-9-CM OR = 2.00) (Figure 2, Table 4). For rs1333049, potential novel phenotypic associations include phecode 282.5 “Sickle cell anemia” (ICD-10-CM OR = 0.59) and phecode 440.22 “Atherosclerosis of native arteries of the extremities with intermittent claudication” (ICD-10-CM OR = 1.63) (Figure 4, Table 4). For rs6457620, potential novel associations include phecode 335 “Multiple sclerosis” (ICD-10-CM OR = 0.61), phecode 557.1 “Celiac disease” (Combined OR = 0.56), and phecode 575.1 “Cholangitis” (ICD-9-CM OR = 0.65) (Figure 5, Table 4).

Table 4. PheWAS Associations with Significant p-values post-Bonferroni Correction

SNP	PheWAS Phenotype	Phecode	Map: # cases, OR, p-value
rs3135388	Neuroendocrine tumors	209	ICD-10-CM: 45, 2.66, 2.43E-05
rs3135388	Hypothyroidism	244	ICD-9-CM: 5817, 0.84, 2.85E-08 ICD-10-CM: 4937, 0.85, 8.39E-07 combined: 6689, 0.85, 5.07E-08
rs3135388	Hypothyroidism NOS	244.4	ICD-9-CM: 5347, 0.80, 5.13E-11 ICD-10-CM: 4447, 0.82, 5.90E-08 combined: 6198, 0.82, 1.67E-10
rs3135388	Thyroiditis	245	ICD-10-CM: 517, 0.61, 9.47E-06 combined: 938, 0.68, 1.25E-06
rs3135388	Chronic thyroiditis	245.2	ICD-10-CM: 473, 0.58, 3.93E-06 combined: 645, 0.62, 1.02E-06
rs3135388	Chronic lymphocytic thyroiditis	245.21	ICD-10-CM: 464, 0.58, 5.57E-06 combined: 636, 0.62, 1.43E-06
rs3135388	Diabetes mellitus	250	ICD-9-CM: 7656, 0.79, 1.07E-15 ICD-10-CM: 7000, 0.81, 7.27E-12 combined: 8674, 0.81, 2.29E-13
rs3135388	Type 1 diabetes	250.1	ICD-9-CM: 1797, 0.47, 9.40E-28 ICD-10-CM: 1075, 0.31, 6.03E-28 combined: 1924, 0.50, 1.18E-26
rs3135388	Type 1 diabetes with ketoacidosis	250.11	ICD-9-CM: 236, 0.07, 5.62E-08 combined: 289, 0.08, 7.58E-10
rs3135388	Type 1 diabetes with renal manifestations	250.12	ICD-9-CM: 239, 0.28, 1.06E-07 ICD-10-CM: 192, 0.36, 2.26E-05 combined: 300, 0.32, 1.74E-08
rs3135388	Type 1 diabetes with ophthalmic manifestations	250.13	ICD-9-CM: 197, 0.19, 8.47E-08 ICD-10-CM: 138, 0.16, 7.44E-06 combined: 252, 0.19, 1.60E-09
rs3135388	Type 1 diabetes with neurological manifestations	250.14	ICD-9-CM: 368, 0.42, 4.23 ICD-10-CM: 214, 0.30, 6.42E-07 combined: 445, 0.43, 3.10E-09
rs3135388	Type 2 diabetes	250.2	ICD-9-CM: 7260, 0.83, 5.40E-10 combined: 8292, 0.84, 3.60E-09
rs3135388	Insulin pump user	250.3	ICD-9-CM: 1845, 0.65, 1.13E-12 ICD-10-CM: 1735, 0.65, 5.61E-12 combined: 2491, 0.66, 1.94E-15

Table 4. Continued

SNP	PheWAS Phenotype	Phecode	Map: # cases, OR, p-value
rs3135388	Hypoglycemia	251.1	ICD-9-CM: 419, 0.56, 1.77E-05
rs3135388	Disorders of lipoid metabolism	272	ICD-9-CM: 14679, 0.88, 2.20E-07 combined: 16506, 0.88, 2.83E-07
rs3135388	Hyperlipidemia	272.1	ICD-9-CM: 14638, 0.88, 2.41E-07 combined: 16491, 0.88, 2.86E-07
rs3135388	Mixed hyperlipidemia	272.13	ICD-9-CM: 7702, 0.85, 3.64E-07 combined: 8998, 0.86, 7.66E-07
rs3135388	Rheumatoid arthritis and other inflammatory polyarthropathies	714	ICD-9-CM: 1951, 0.79, 1.86E-05 combined: 2271, 0.78, 1.17E-06
rs17234657	Iron deficiency anemias, unspecified or not due to blood loss	280.1	ICD-9-CM: 2375, 1.19, 2.17E-05
rs17234657	Disorders of iris and ciliary body	379.5	ICD-9-CM: 124, 2.00, 2.94E-06
rs1333049	Sickle cell anemia	282.5	ICD-9-CM: 261, 0.64, 1.19E-05 ICD-10-CM: 254, 0.59, 1E-06 combined: 321, 0.64, 1.61E-06
rs1333049	Atherosclerosis of native arteries of the extremities with intermittent claudication	440.22	ICD-10-CM: 152, 1.63, 2.86E-05
rs6457620[a]	Hypothyroidism NOS	244.4	ICD-9-CM: 5345, 0.91, 1.94E-05
rs6457620[a]	Multiple sclerosis	335	ICD-9-CM: 1059, 1.59, 1.01E-24 ICD-10-CM: 935, 1.65, 4.14E-25 combined: 1081, 1.59, 3.28E-25
rs6457620[a]	Other demyelinating diseases of central nervous system	341	ICD-9-CM: 503, 1.40, 1.15e-07 combined: 521, 1.40E-08
rs6457620[a]	Abnormal movement	350	ICD-9-CM: 3262, 1.15, 5.56E-08 combined: 4177, 1.11, 3.75E-06
rs6457620[a]	Lack of coordination	350.3	ICD-9-CM: 1169, 1.21, 4.27E-06 ICD-10-CM: 770, 1.30, 3.01E-07 combined: 1279, 1.22, 5.56E-07

Table 4. Continued

SNP	PheWAS Phenotype	Phecode	Map: # cases, OR, p-value
rs6457620[a]	Celiac disease	557.1	ICD-9-CM: 268, 1.80, 8.36E-11 ICD-10-CM: 208, 1.93, 2.50E-10 combined: 304, 1.77, 2.01E-11
rs6457620[a]	Cholangitis	575.1	ICD-9-CM: 197, 1.54, 2.82E-05

-Table excludes those associations in GWAS catalog or those related to the phenotype, i.e. same phecode family or is a treatment for the disease.
[a]rs6457620 minor allele = C

Discussion

In this study, we showed that using the ICD-10-CM phecode map, we replicated a majority of the known phenotype associations for 5 SNPs in the GWAS catalog. Further, we demonstrated that PheWAS phenotypes mapped from ICD-10-CM codes were similar to those mapped from ICD-9-CM codes. For most of the SNPs that we evaluated, we observed novel phenotype associations that were not listed in the GWAS catalog.

In the GWAS catalog replication study, we did not replicate the association between rs2200733 and EFO_0000712 “stroke”. The original study,[87] found an association with ischemic stroke (OR = 1.26, p=2.18x10^{-10}) in patients from a registry of Icelandic patients with 1,661 cases and 10,851 controls. In this registry, stroke was confirmed clinically by neurologists and supported by imaging. This positive correlation between rs2200733 and ischemic stroke was replicated in two external datasets. In contrast, this study did not find a statistically significant association between rs2200733 and stroke, either using phenotypes mapped from ICD-9-CM or ICD-10-CM codes. Further, many of the tests had enough cases to trust the association statistic, like phecode 433.21 “Occlusion of cerebral arteries” with 1,075 cases identified using ICD-9-CM (OR = 0.93, p=0.29), which one could confidently infer as having an opposite effect compared to the original study and in subsequent studies (OR = 1.39, p=6.5x10^{-32}).[88]

There was a statistically significant linear positive correlation between the PheWAS ORs for phenotypes mapped from ICD-9-CM and ICD-10-CM codes. Further, when comparing associations that had $p<0.05$, the positive correlation strengthened. The stronger correlation between ORs in the associations with $p<0.05$ was also observed in the other comparative analyses (ie, ICD-9-CM vs. Combined and ICD-10-CM vs. Combined). The correlation between PheWAS ORs between phenotypes mapped from both ICD-9-CM and ICD-10-CM codes was larger than ICD-9-CM vs. ICD-10-CM alone. For example, rs3135388 ICD-9-CM vs. ICD-10-CM Pearson correlation (R) = 0.43 vs rs3135388 ICD-9-CM vs. combined Pearson correlation (R) = 0.86 (Table 4, Figure 1). We expected this result as the combined analysis has contributions from both ICD-9-CM and ICD-10-CM codes.

We examined genotype-phenotype associations that were not listed in the GWAS catalog. We found that many of these associations were found in other databases. For example, the negative correlations between (rs3135388 and Type 1 diabetes) and (rs6457620 and Celiac disease) have been observed in a preliminary GWAS using UKBB data.[89] Thus, it may be worth expanding the associations included in the GWAS catalog, as it is commonly used as a reference for human genetic studies.[81]

The associations between rs3135388 and thyroid-related phenotypes are particularly interesting, as these associations, to our knowledge, has only been observed in the UKBB dataset.[89] The causal genotype-phenotype association driving the other associations could originate from a lymphocytic thyroiditis, resulting in destruction of the thyroid gland and hypothyroidism in the affected patient. This conjecture is supported by the mapped gene for rs3135388 in HLA-DRA that encodes MHC Class II receptor. Further, the negative correlation with Type 1 diabetes (ICD-10-CM OR = 0.31, p-value = 6.03×10^{-28}) for rs3135388 supports the negative correlation with phecode 245.21 “Chronic lymphocytic thyroiditis” (Combined PheWAS OR=0.62, p-value = 1.43×10^{-6}) (i.e. Hashimoto’s thyroiditis). On the other hand, rs3135388 has a positive association with phecode 335 “Multiple sclerosis” (Combined OR=2.48, p-value=1.03x10^{-71}) which is unexpected as both diseases are considered autoimmune disorders.

The main limitation of this study is that we only evaluated the performance of the ICD-10-CM phecode map for five SNPs and did not do so for all the SNPs available on the MEGA platform. In a follow-up analysis with more SNPs, it will be interesting to see whether the associations captured by ICD-10-CM are still largely similar to that of PheWAS performed using phenotypes mapped from ICD-9-CM codes. Another limitation of this study is that the regression equations were adjusted using patient race as reported in the EHR and not with principal components (PCs). The lack of PC adjustment in the regression equation could explain the unexpected association between Sickle cell anemia (phecode 282.5) and rs1333049, as many studies have demonstrated the correlation between admixture and Sickle cell disease.[90]

Conclusion

In conclusion, the results of this study demonstrate that researchers can be confident in the phenotypes generated from the beta ICD-10-CM v1.2 phecode map for their PheWAS analyses in the EHR.

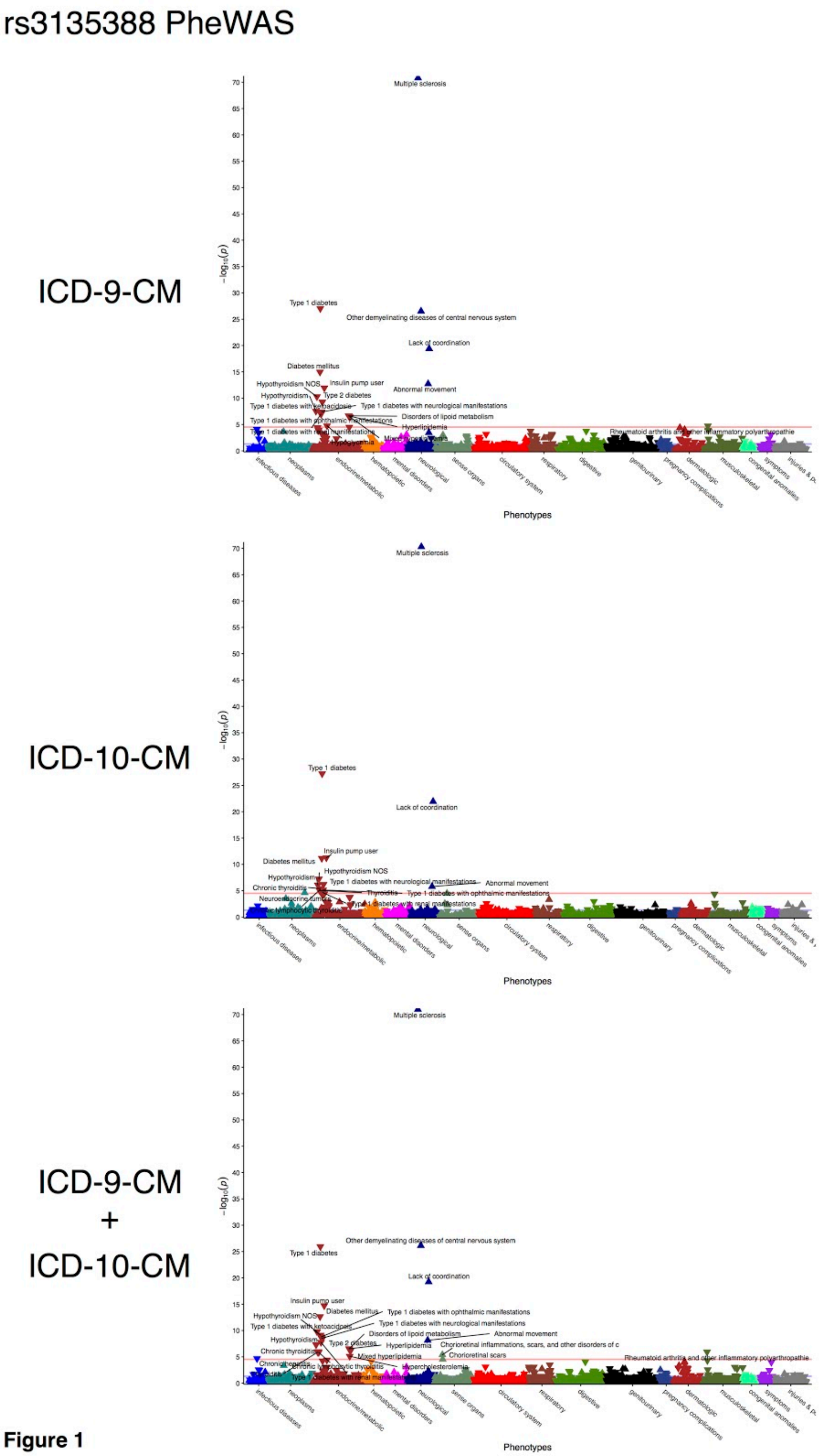

Figure 1

Figure 1. PheWAS manhattan plots for rs3135388.

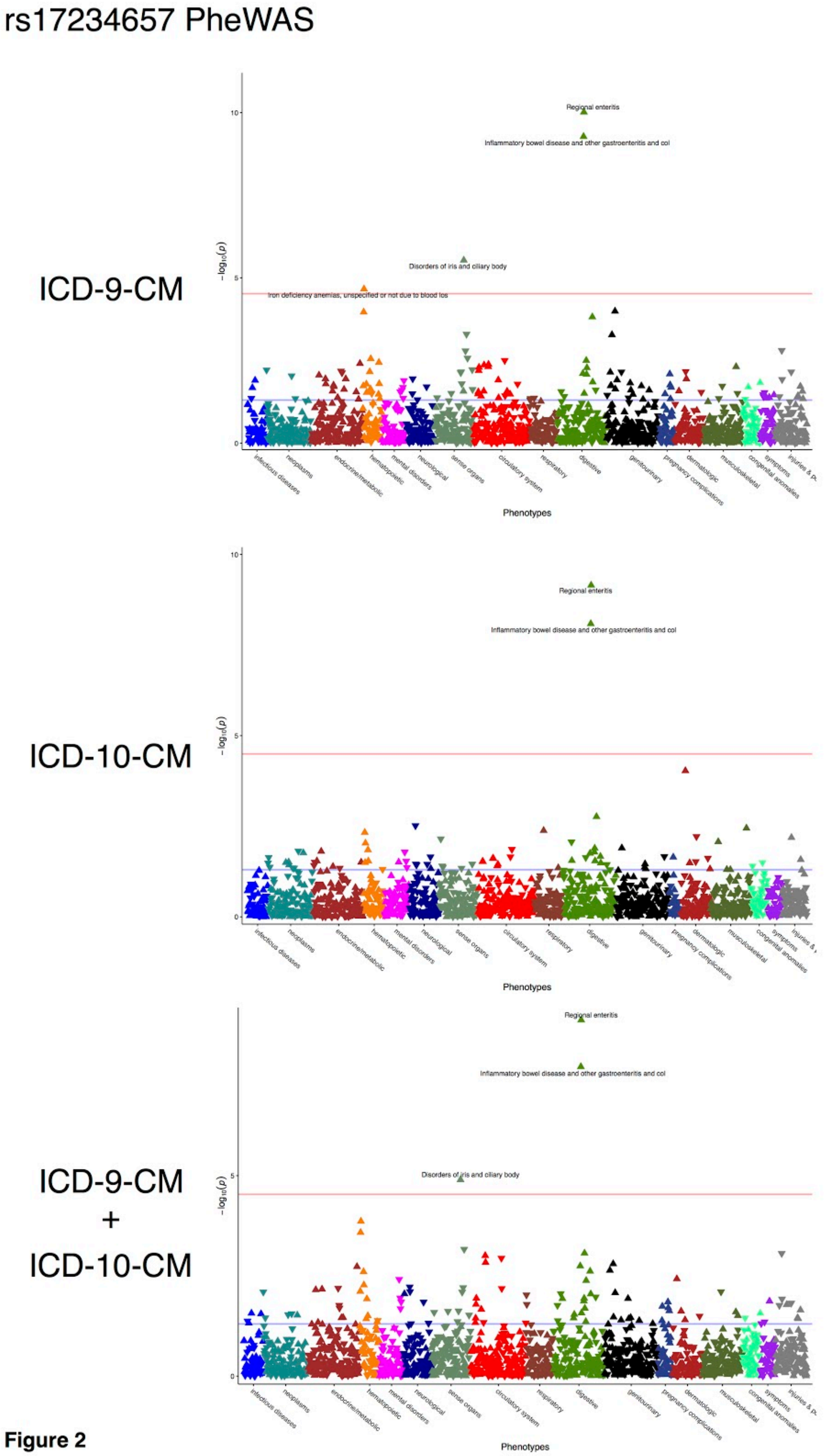

Figure 2

Figure 2. PheWAS manhattan plots for rs17234657.

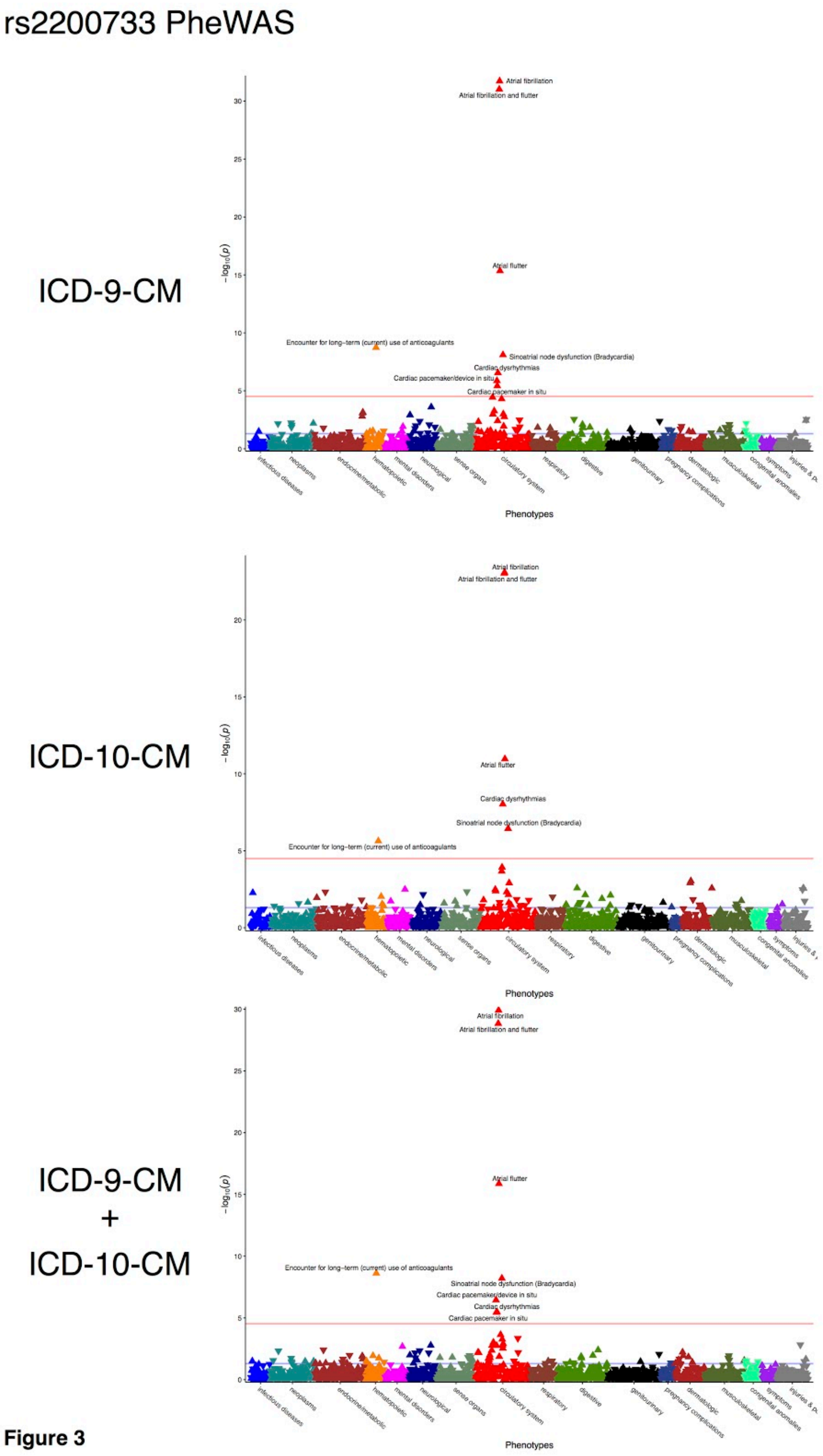

Figure 3

Figure 3. PheWAS manhattan plots for rs2200733.

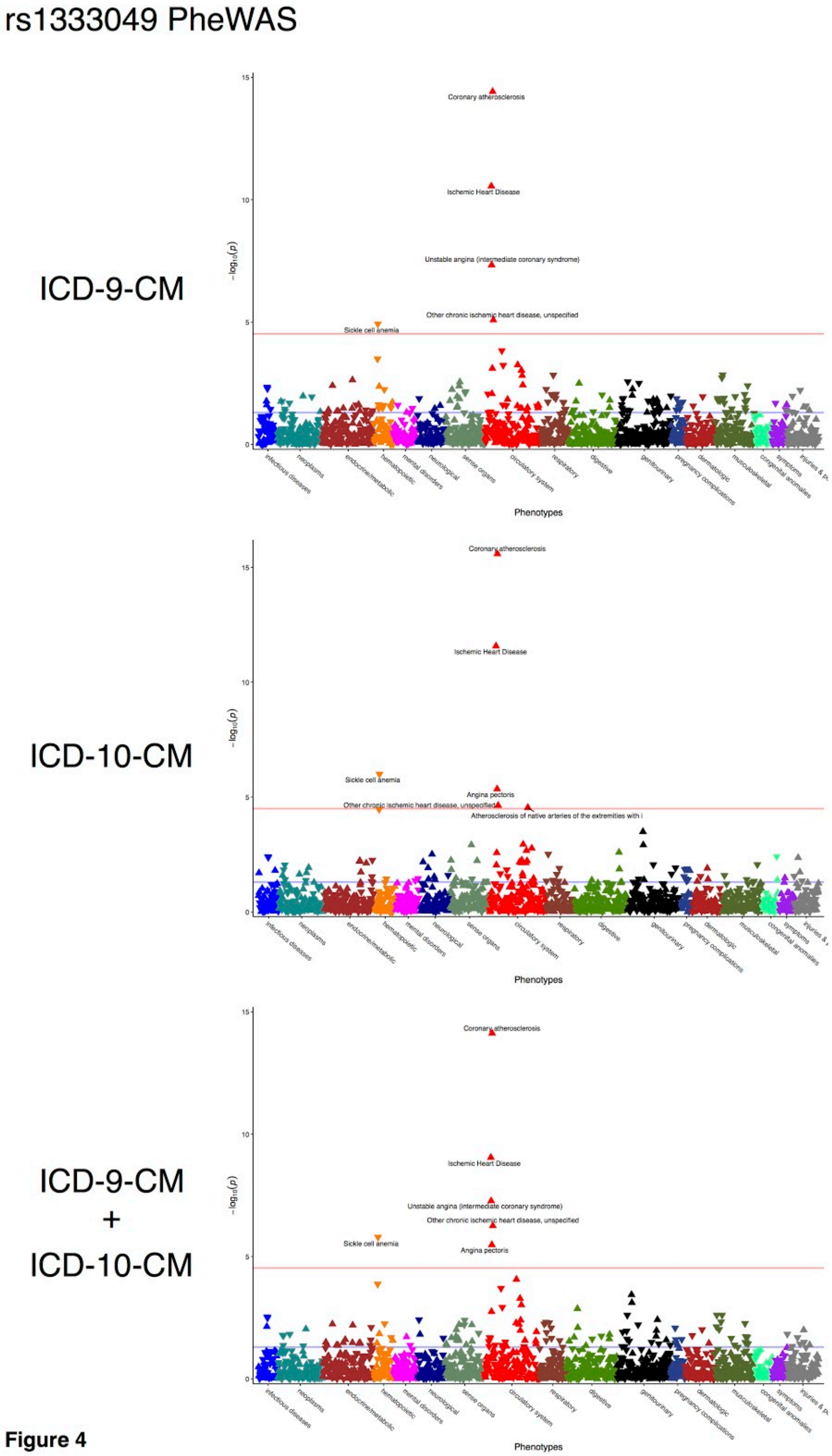

Figure 4

Figure 4. PheWAS manhattan plots for rs1333049.

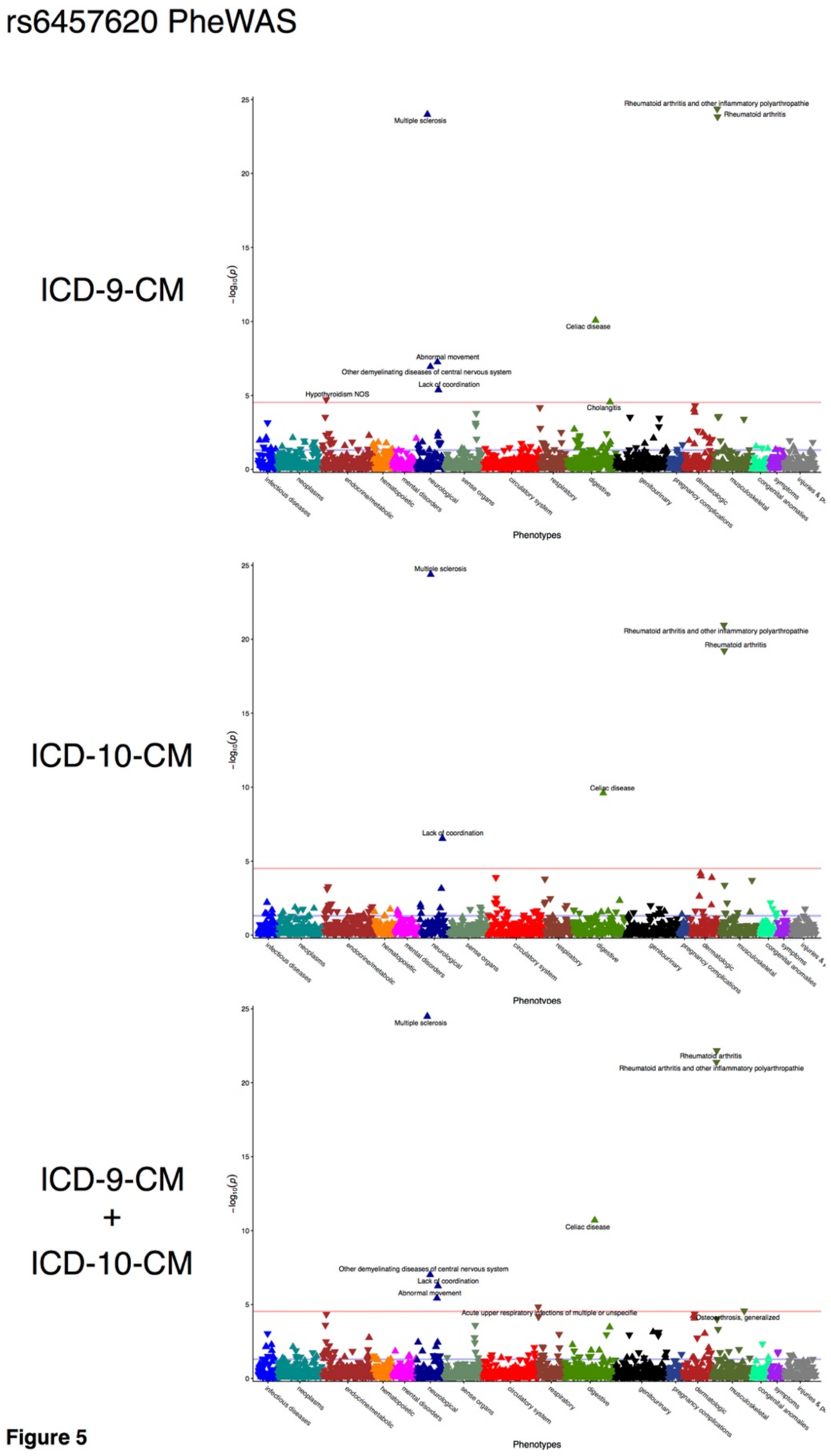

Figure 5

Figure 5. PheWAS manhattan plots for rs6457620.

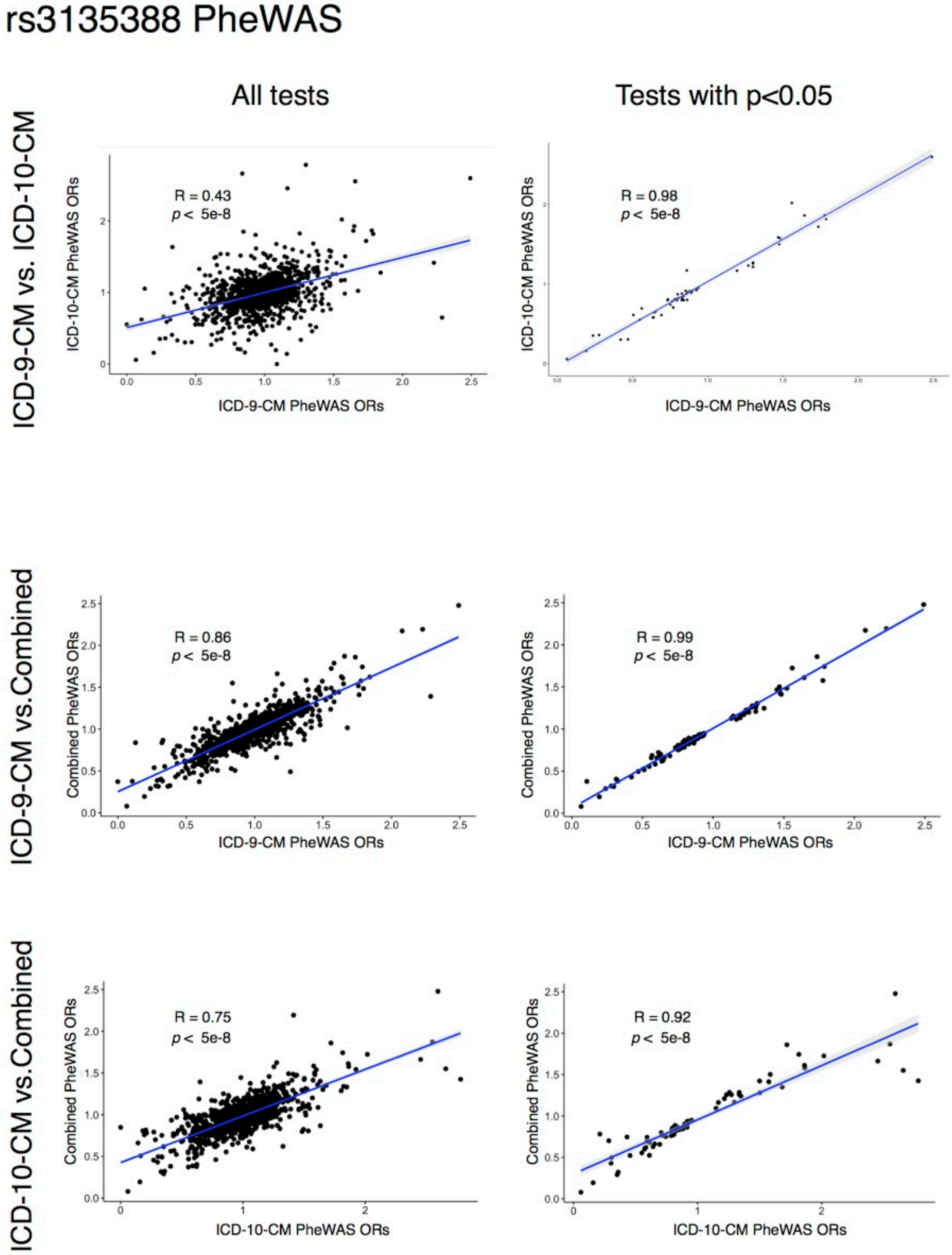

Figure 6. Correlation scatter plots plot for rs3135388 PheWAS

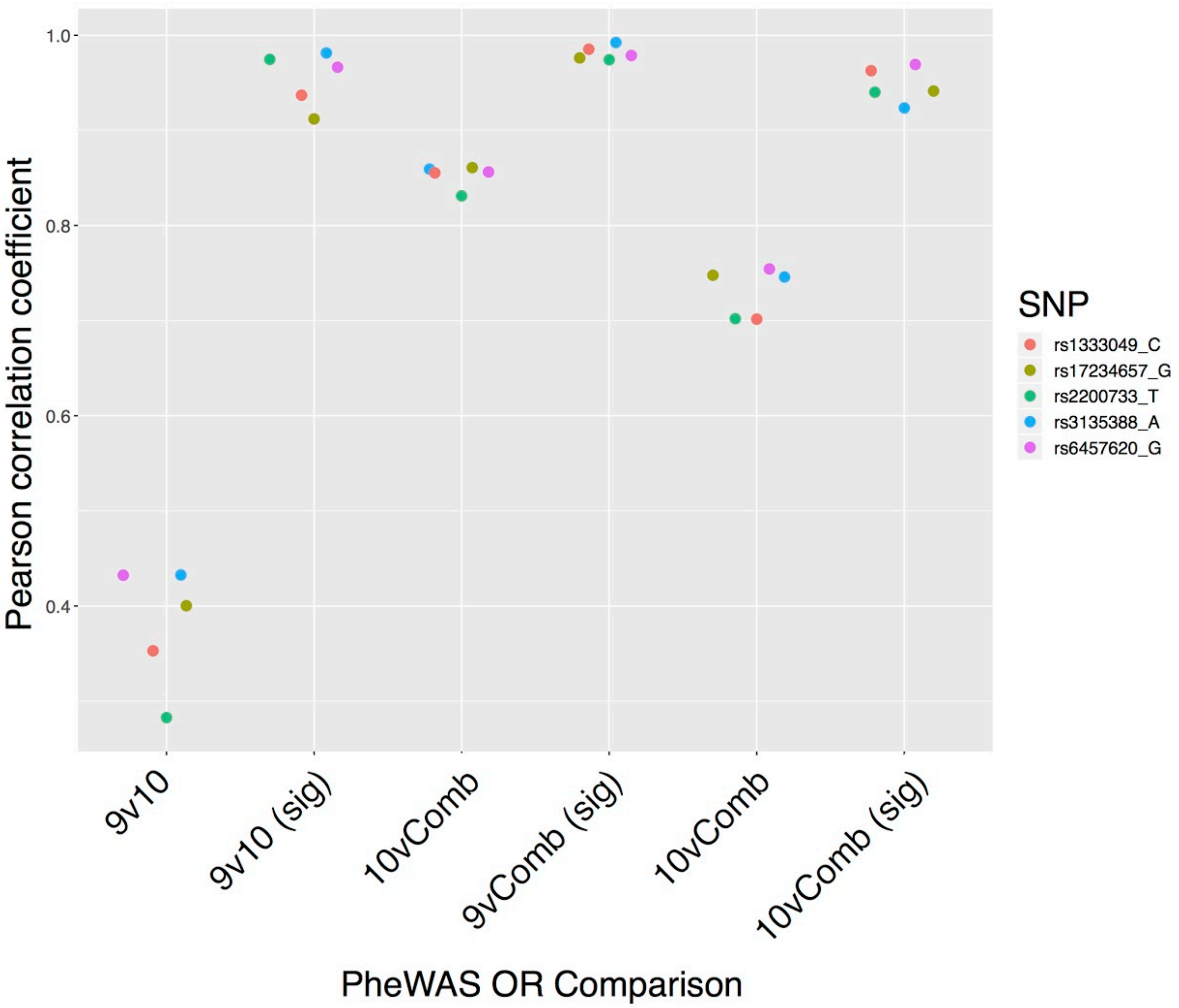

Figure 7

Figure 7. Pearson correlation coefficients for PheWAS OR comparisons

CHAPTER 4

Summary

In this thesis, I described how we mapped ICD-10 and ICD-10-CM codes to phecodes and evaluated the performance of the maps by looking at phecode coverage of all ICD codes in the UMLS (70-75% coverage) and the subset of codes used in the UKBB dataset and at VUMC (>90% coverage). I also show that a majority of patients with chronic diseases identified using the ICD-9-CM phecode map were also identified as having the same disease using the ICD-10-CM phecode map. Using PheWAS for 6 SNPs, I replicated a majority of the known genotype-phenotype associations in the GWAS catalog. Further an extensive comparison of the PheWAS effect sizes showed that ICD-10-CM sourced PheWAS phenotypes were similar to phenotypes sourced from the ICD-9-CM phecode map.

Limitations

There are limitations in this study. First, we used manually validated maps from the UMLS, CMS, and OHDSI to automate the mapping of ICD-10 and ICD-10-CM codes to phecodes. But, not all of the conversions provided by these maps are one-to-one, as some mappings are approximate (eg, one ICD-9-CM code maps to >1 ICD-10-CM code). We are currently in the process of spot checking and correcting potential mapping errors. Second, in Chapter 3, ICD-9-CM codes that were recorded after October 1, 2015 (deadline for transition to ICD-10-CM) and ICD-10-CM codes that were recorded before the same date were used to map to PheWAS phenotypes. It will be necessary to evaluate whether the dates associated with these ICD codes are true, or an artifact due date shifting in the Synthetic Derivative. Third, this study did not evaluate the effect of a shorter ICD-10-CM observation window vs. longer ICD-9-CM observation window, on identifying true associations.

Future Directions

Given the adoption of SNOMED CT as one of the standard vocabularies in UMLS and OHDSI, it may be worthwhile to create a map to translate SNOMED CT concepts to phecodes. This work is partially completed through the mapping of ICD-10-CM and ICD-10 codes to phecodes, but to create accurate maps, a more thorough review of mappings may be necessary. It may also be of interest to develop a framework to look at the temporal relationship between phenotypes, like varying observation windows to better understand the causal phenotype driving the other observed phenotypes/traits, such as development of Atrial fibrillation (phecode 427.21) leading to exposure to anticoagulant drugs (phecode 286.2). Quantifying the correlation between phenotypes may also help the researcher to better understand and interpret the genotype-phenotype associations, such as is done with region plots in GWAS.[91] Last, it may be worthwhile to discuss whether providing synonyms for some of the PheWAS phenotype descriptions would be helpful, to better match disease names that are more familiar to clinicians and biomedical researchers (e.g. phecode 555.1 “regional enteritis” → “Crohn’s disease”).

In 2019, the WHO adopted ICD-11.[92] ICD-11 contains more than >55k unique codes, compared to <15k codes in ICD-10. ICD-11 has more chapters than ICD-10 (27 vs. 22). The five new chapters in ICD-11 are “Diseases of the immune system”, “Diseases of the blood or blood-forming organs”, “Sleep-wake disorders”, “Conditions related to sexual health”, and “Traditional medicine”. With the release of ICD-11, the WHO provided a map to convert ICD-10 codes to ICD-11. The U.S. will most likely adopt ICD-11 for tracking morbidity and mortality data after 2023.[93] Results from this thesis shows that using validated translation tools in an automated pipeline is feasible to map ICD-11 codes to phecodes.

Conclusion

In this thesis, I describe the process of creating a map to allow translation of ICD-10-CM and ICD-10 (non-CM) codes to phecodes that will allow researchers to perform PheWAS in the EHR. With the recent adoption of EHRs and ICD-10-CM in the U.S. and initiatives to create Big Data resources (i.e. AoU and OHDSI), these maps will be a valuable resource for investigators aiming to conduct genetic studies using EHR data. The initial evaluation of the ICD-10-CM phecode map for PheWAS provides evidence for using validated vocabulary mappings in an automated pipeline to translate future versions of ICD codes to phecodes.

REFERENCES

1 Gold M, McLAUGHLIN C. Assessing HITECH Implementation and Lessons: 5 Years Later. *Milbank Q* 2016;**94**:654–87.

2 Adler-Milstein J, Jha AK. HITECH Act Drove Large Gains In Hospital Electronic Health Record Adoption. *Health Aff* 2017;**36**:1416–22.

3 Office of the National Coordinator for Health Information Technology. Percent of Hospitals, By Type, that Possess Certified Health IT. Health IT Quick-Stat #52. 2018.https://dashboard.healthit.gov/quickstats/pages/certified-electronic-health-record-technology-in-hospitals.php (accessed 23 Dec 2019).

4 Hripcsak G, Ryan PB, Duke JD, *et al.* Characterizing treatment pathways at scale using the OHDSI network. *Proc Natl Acad Sci U S A* 2016;**113**:7329–36.

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

6 Wetterstrand KA. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP). www.genome.gov/sequencingcostsdata (accessed 6 Jan 2020).

7 Mardis ER. A decade’s perspective on DNA sequencing technology. *Nature* 2011;**470**:198–203.

8 Toga AW, Foster I, Kesselman C, *et al.* Big biomedical data as the key resource for discovery science. *J Am Med Inform Assoc* 2015;**22**:1126–31.

9 Kho AN, Pacheco JA, Peissig PL, *et al.* Electronic medical records for genetic research: results of the eMERGE consortium. *Sci Transl Med* 2011;**3**:79re1.

10 McCarty CA, Chisholm RL, Chute CG, *et al.* The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. *BMC Med Genomics* 2011;**4**:13.

11 Bycroft C, Freeman C, Petkova D, *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* 2018;**562**:203–9.

12 welcome to eMerge > Collaborate. https://emerge-network.org/ (accessed 13 Jan 2020).

13 McCarthy MI, Abecasis GR, Cardon LR, *et al.* Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet* 2008;**9**:356–69.

14 Manolio TA. Genomewide association studies and assessment of the risk of disease. *N Engl J Med* 2010;**363**:166–76.

15 Sladek R, Rocheleau G, Rung J, *et al.* A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* 2007;**445**:881–5.

16 LaFramboise T. Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. *Nucleic Acids Res* 2009;**37**:4181–93.

17 Cutler DJ, Zwick ME, Carrasquillo MM, *et al.* High-throughput variation detection and genotyping using microarrays. *Genome Res* 2001;**11**:1913–25.

18 Wheeler DA, Srinivasan M, Egholm M, *et al.* The complete genome of an individual by massively parallel DNA sequencing. *Nature* 2008;**452**:872–6.

19 Klein RJ, Zeiss C, Chew EY, *et al.* Complement factor H polymorphism in age-related macular degeneration. *Science* 2005;**308**:385–9.

20 Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 2007;**447**:661–78.

21 Visscher PM, Brown MA, McCarthy MI, *et al.* Five years of GWAS discovery. *Am J Hum Genet* 2012;**90**:7–24.

22 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.

23 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**:e147–54.

24 Hennekam RCM, Biesecker LG. Next-generation sequencing demands next-generation phenotyping. *Hum Mutat* 2012;**33**:884–6.

25 Murphy S, Churchill S, Bry L, *et al.* Instrumenting the health care enterprise for discovery research in the genomic era. *Genome Res* 2009;**19**:1675–81.

26 Wilke RA, Berg RL, Peissig P, *et al.* Use of an electronic medical record for the identification of research subjects with diabetes mellitus. *Clin Med Res* 2007;**5**:1–7.

27 Kho AN, Hayes MG, Rasmussen-Torvik L, *et al.* Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study. *J Am Med Inform Assoc* 2012;**19**:212–8.

28 Turchin A, Kohane IS, Pendergrass ML. Identification of patients with diabetes from the text of physician notes in the electronic medical record. *Diabetes Care* 2005;**28**:1794–5.

29 Ritchie MD, Denny JC, Crawford DC, *et al.* Robust replication of genotype-phenotype associations across multiple diseases in an electronic medical record. *Am J Hum Genet* 2010;**86**:560–72.

30 Wei W-Q, Leibson CL, Ransom JE, *et al.* The absence of longitudinal data limits the accuracy of high-throughput clinical phenotyping for identifying type 2 diabetes mellitus subjects. *Int J Med Inform* 2013;**82**:239–47.

31 Lasko TA, Denny JC, Levy MA. Computational phenotype discovery using unsupervised feature learning over noisy, sparse, and irregular clinical data. *PLoS One* 2013;**8**:e66341.

32 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.

33 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.

34 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.

35 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* 2017;**12**:e0175508.

36 Phecode Map 1.2 with ICD-9 Codes. PheWAS Catalog. https://phewascatalog.org/phecodes (accessed 17 Jul 2019).

37 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.

38 Steindel SJ. International classification of diseases, 10th edition, clinical modification and procedure coding system: descriptive overview of the next generation HIPAA code sets. *J Am Med Inform Assoc* 2010;**17**:274–82.

39 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.

40 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.

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

42 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.

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

44 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.

45 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.

46 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.

47 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.

48 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.

49 Li X, Meng X, He Y, *et al.* Genetically determined serum urate levels and cardiovascular and other diseases in UK Biobank cohort: A phenome-wide mendelian randomization study. *PLoS Med* 2019;**16**:e1002937.

50 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.

51 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 (Wash DC)* 2016;**4**:1211.

52 Wilder V. UMLS 2018AA Release Available. NLM Technical Bulletin. 2018.https://www.nlm.nih.gov/pubs/techbull/mj18/mj18_umls_2018aa_release.html (accessed 17 Jul 2019).

53 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.

54 Centers for Medicare & Medicaid Services. 2018 ICD-10 CM and GEMs. CMS.gov. 2017.https://www.cms.gov/Medicare/Coding/ICD10/2018-ICD-10-CM-and-GEMs.html (Archived by WebCite® at http://www.webcitation.org/77SlBvhUD) (accessed 26 Nov 2018).

55 Fung KW, Bodenreider O. Utilizing the UMLS for semantic mapping between terminologies. *AMIA Annu Symp Proc* 2005;:266–70.

56 Hripcsak G, Duke JD, Shah NH, *et al.* Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. *Stud Health Technol Inform* 2015;**216**:574–8.

57 documentation:vocabulary:icd9cm [Observational Health Data Sciences and Informatics]. Observational Health Data Sciences and Informatics. 2016.https://www.ohdsi.org/web/wiki/doku.php?id=documentation:vocabulary:icd9cm (accessed 17 Jul 2019).

58 SNOMED CT to ICD-9-CM Rule Based Mapping to Support Reimbursement. US National Library of Medicine. 2018.https://www.nlm.nih.gov/research/umls/mapping_projects/snomedct_to_icd9cm_reimburse.html (accessed 6 Apr 2019).

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

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

61 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.

62 Zhao J, Feng Q, Wu P, *et al.* Using topic modeling via non-negative matrix factorization to identify relationships between genetic variants and disease phenotypes: A case study of Lipoprotein(a) (LPA). *PLoS One* 2019;**14**:e0212112.

63 Carroll R. PheWAS R Package, GitHub Repository. GitHub. https://github.com/PheWAS/PheWAS

64 Zhou W, Nielsen JB, Fritsche LG, *et al.* Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat Genet* 2018;**50**:1335–41.

65 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.

66 Chaganti S, Mawn LA, Kang H, *et al.* Electronic Medical Record Context Signatures Improve Diagnostic Classification using Medical Image Computing. *IEEE J Biomed Health Inform* Published Online First: 28 December 2018. doi:10.1109/JBHI.2018.2890084

67 Zhao J, Feng Q, Wu P, *et al.* Learning from Longitudinal Data in Electronic Health Record and Genetic Data to Improve Cardiovascular Event Prediction. *Sci Rep* 2019;**9**:717.

68 Huang M, ElTayeby O, Zolnoori M, *et al.* Public Opinions Toward Diseases: Infodemiological Study on News Media Data. *J Med Internet Res* 2018;**20**:e10047.

69 Shi X, Li X, Cai T. Spherical Regression under Mismatch Corruption with Application to Automated Knowledge Translation. arXiv [stat.ME]. 2018.http://arxiv.org/abs/1810.05679

70 Bastarache L, Hughey JJ, Hebbring S, *et al.* Phenotype risk scores identify patients with unrecognized Mendelian disease patterns. *Science* 2018;**359**:1233–9.

71 Clinical Classifications Software (CCS) for ICD-10-PCS (beta version). Agency for Healthcare Research and Quality. Healthcare Cost and Utilization Project (HCUP). 2019.https://www.hcup-us.ahrq.gov/toolssoftware/ccs10/ccs10.jsp (accessed 5 Jul 2019).

72 Sabbatini AK, Kocher KE, Basu A, *et al.* In-Hospital Outcomes and Costs Among Patients Hospitalized During a Return Visit to the Emergency Department. *JAMA* 2016;**315**:663–71.

73 Hu Z, Hao S, Jin B, *et al.* Online Prediction of Health Care Utilization in the Next Six Months Based on Electronic Health Record Information: A Cohort and Validation Study. *J Med Internet Res* 2015;**17**:e219.

74 Hripcsak G, Levine ME, Shang N, *et al.* Effect of vocabulary mapping for conditions on phenotype cohorts. *J Am Med Inform Assoc* 2018;**25**:1618–25.

75 Gaziano JM, Concato J, Brophy M, *et al.* Million Veteran Program: A mega-biobank to study genetic influences on health and disease. *J Clin Epidemiol* 2016;**70**:214–23.

76 All of Us Research Program Investigators, Denny JC, Rutter JL, *et al.* The ‘All of Us’ Research Program. *N Engl J Med* 2019;**381**:668–76.

77 Rabbee N, Speed TP. A genotype calling algorithm for affymetrix SNP arrays. *Bioinformatics* 2006;**22**:7–12.

78 Nielsen R, Paul JS, Albrechtsen A, *et al.* Genotype and SNP calling from next-generation sequencing data. *Nat Rev Genet* 2011;**12**:443–51.

79 Teixeira PL, Wei W-Q, Cronin RM, *et al.* Evaluating electronic health record data sources and algorithmic approaches to identify hypertensive individuals. *J Am Med Inform Assoc* 2017;**24**:162–71.

80 Wu P, Gifford A, Meng X, *et al.* Mapping ICD-10 and ICD-10-CM Codes to Phecodes: Workflow Development and Initial Evaluation. *JMIR Med Inform* 2019;**7**:e14325.

81 Buniello A, MacArthur JAL, Cerezo M, *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res* 2019;**47**:D1005–12.

82 McKinney W, Others. Data structures for statistical computing in python. In: *Proceedings of the 9th Python in Science Conference*. Austin, TX 2010. 51–6.

83 van der Walt S, Colbert SC, Varoquaux G. The NumPy Array: A Structure for Efficient Numerical Computation. *Computing in Science Engineering* 2011;**13**:22–30.

84 Seabold S, Perktold J. Statsmodels: Econometric and statistical modeling with python. *of the 9th Python in Science Conference* Published Online First: 2010.https://www.researchgate.net/profile/Josef_Perktold/publication/264891066_Statsmodels_Econometric_and_Statistical_Modeling_with_Python/links/5667ca9308ae34c89a0261a8/Statsmodels-Econometric-and-Statistical-Modeling-with-Python.pdf

85 Jones E, Oliphant T, Peterson P, *et al.* SciPy: Open source scientific tools for Python. http://www.scipy.org/

86 Kluyver T, Ragan-Kelley B, Pérez F, *et al.* Jupyter Notebooks-a publishing format for reproducible computational workflows. In: *ELPUB*. 2016. 87–90.

87 Gretarsdottir S, Thorleifsson G, Manolescu A, *et al.* Risk variants for atrial fibrillation on chromosome 4q25 associate with ischemic stroke. *Ann Neurol* 2008;**64**:402–9.

88 Malik R, Chauhan G, Traylor M, *et al.* Multiancestry genome-wide association study of 520,000 subjects identifies 32 loci associated with stroke and stroke subtypes. *Nat Genet* 2018;**50**:524–37.

89 University of Michigan. UKBiobank ICD PheWeb - Analysis of 1403 ICD-based traits using SAIGE. UKBiobank ICD PheWeb. http://pheweb.sph.umich.edu/SAIGE-UKB/ (accessed 19 Sep 2012).

90 Solovieff N, Hartley SW, Baldwin CT, *et al.* Ancestry of African Americans with sickle cell disease. *Blood Cells Mol Dis* 2011;**47**:41–5.

91 Tardif Jean-Claude, Rhéaume Eric, Lemieux Perreault Louis-Philippe, *et al.* Pharmacogenomic Determinants of the Cardiovascular Effects of Dalcetrapib. *Circ Cardiovasc Genet* 2015;**8**:372–82.

92 World Health Organization. ICD-11 Implementation and Transition Guide. https://icd.who.int/docs/ICD-11%20Implementation%20or%20Transition%20Guide_v105.pdf (accessed 23 Dec 2019).

93 Pickett D. Update on ICD-11: The WHO Launch and Implications for U.S. Implementation. https://www.cdc.gov/nchs/data/icd/ICD-11-WHOV-CM-2018-V3.pdf (accessed 23 Dec 2019).