# Mapping the Read2/CTV3 controlled clinical terminologies to Phecodes in UK Biobank primary care electronic health records: implementation and evaluation

**Spiros Denaxas[1,3,4,5,6], Ge Liu[2], Qiping Feng[2], Ghazaleh Fatemifar[2,3], Lisa Bastarache[1], Eric .V. Kerchberger[2], Aroon D. Hingorani[1,3,4], Tom Lumbers[1,3], Josh F. Peterson[2], Wei-Qi Wei[2], Harry Hemingway[1,3,4,6]**

**[1]University College London, London, UK; [2]Vanderbilt University Medical Center, Nashville, TN, USA; [3]Health Data Research UK, London, UK; [4]BHF Research Accelerator, London, UK; [5]The Alan Turing Institute, London, UK, [6]NIHR UCLH BRC, London, UK**

## Abstract

***Objective:*** *To establish and validate mappings between primary care clinical terminologies (Read Version 2, Clinical Terms Version 3) and Phecodes.* ***Methods:*** *We processed 123,662,421 primary care events from 230,096 UK Biobank (UKB) participants. We assessed the validity of the primary care-derived Phecodes by conducting PheWAS analyses for seven pre-selected SNPs in the UKB and compared with estimates from BioVU.* ***Results:*** *We mapped 92% of Read2 (n=10,834) and 91% of CTV3 (n=21,988) to 1,449 and 1,490 Phecodes. UKB PheWAS using Phecodes from primary care EHR and hospitalizations replicated all (n=22) previously-reported genotype-phenotype associations. When limiting Phecodes to primary care EHR, replication was 81% (n=18).* ***Conclusion:*** *We introduced a first version of mappings from Read2/CTV3 to Phecodes. The reference list of diseases provided by Phecodes can be extended, enabling researchers to leverage primary care EHR for high-throughput discovery research.*

## Introduction

Phenotype codes (Phecodes), are a hierarchical phenotype classification system offering a reference list of diseases for genetic and clinical research(1,2). Phecodes were originally created by mapping to coded electronic health records (EHR) using International Classification of Diseases (ICD) 9-Clinical Modification (CM), and more recently ICD-10 and ICD-10-CM(3). Phecodes may overcome some of the shortcomings of these and other clinical classification systems with each leaf node representing a clinically recognisable diagnosis. In the UK Biobank(4), a large scale resource of genomic sequencing and longitudinal phenotypic information, linkage has been obtained to longitudinal primary care information as well as information on hospital admissions. Although hospital episodes are coded using the ICD system, primary care data are coded using the Read version 2 (Read2) and Read version 3 (Clinical Terms 3 or CTV3) controlled clinical terminologies. Both terminologies provide a common standard vocabulary for clinicians and have been widely used for research(5,6).

Phecodes validated for the primary care setting, might offer several advantages. First, acute and chronic disease is often diagnosed within primary care settings in the UK. Patients with a disease treated in primary care may represent less severe forms of the disease than those who are treated in hospital. Phecodes can potentially act as data scaffolding providing a reference list of diseases between primary and secondary care settings. At a practical level, there is a need to reduce dimensionality of primary care EHR (with ~ 400K unique ontology concepts utilized) in a reproducible fashion to enable high-throughput phenotypic and genetic analyses(2). Controlled clinical terminologies that are used in primary care data however have not been mapped to Phecodes. As a result, the extent to which using primary care derived Phecodes offer a valid approach or whether they can provide additional statistical power or markers of disease severity is yet to be explored.

In this paper, we sought to develop and validate a mapping of clinical terminologies used in UK primary care EHR to Phecodes. Specifically, this was achieved through three objectives:

1. Create mappings between the Read controlled clinical terminology (Read2 and CTV3) and Phecodes.
2. Translate primary care EHR data in the UK Biobank to Phecodes.
3. Evaluate the mapping quality by conducting PheWAS analyses in the UK Biobank and BioVU and seeking to replicate previously-reported genotype–phenotype associations.

## Methods

Population and data sources

We used data from the UK Biobank (UKB), a prospective study of 500,000 deeply phenotyped individuals recruited from England, Scotland and Wales. Longitudinal follow up for UKB participants is achieved through linkages to national data sources covering primary care, hospitalizations, cancer registrations and mortality. Specifically, information on hospital admissions for the entire cohort is available for all admitted patient episodes and is coded using ICD-9 and ICD-10 for (primary and secondary) diagnoses and OPCS-4 for surgical procedures. Approx 50% of participants have their primary care record linked which provides information on diagnoses, symptoms, laboratory results, referrals, examination findings and prescriptions.

Primary care EHR were extracted from four data sources based on three EHR vendors (two in England from Vision and TPP SystmOne, and two in Scotland and Wales combining data from EMIS and Vision). The basic unit of interaction in primary care is a consultation (similar to an admission in secondary care) during which the clinician records information directly into the EHR. In the UK, primary care healthcare concepts (except prescriptions) are recorded using Read codes, a hierarchical controlled clinical terminology. Read codes are used in two versions: Read Version 2 (Read2), also known as the 5-byte Read consists of ~100,000 concepts and Clinical Terms Version 3 (CTV3), consists of ~390,000 concepts and contains all Read 2 concepts. Both versions of Read are organized in 30 top level chapters in a fashion similar to ICD-10; e.g. chapter "G" is for Circulatory System Diseases Relationships in Read2 are derived directly through the code structure meaning only single parent-child term relationships are supported. In CTV3, relationships are defined through a separate relationship table and supports polyhierarchical relationships (e.g. 'Infective pneumonia' has two parents: 'Pneumonia' and 'Acute lower respiratory tract infection'). SNOMED-CT, which is now becoming an international standard for recording information across healthcare settings and has >400K concepts encapsulates by definition all CTV3 terms since CTV3 was one of its core components.

Mapping Read2 and CTV3 to Phecodes

We implemented a systematic data-driven approach leveraging terminology reference sets, cross-map files, and expert input to map primary care EHR to Phecodes (Figure 1). The approach, which was similar for both Read2 and CTV3 is described below:

1. Retrieve all unique Read2/CTV3 codes from the UKB *gp_clinical* file.
2. Establish eligible codes Read2/CTV3 terms for mapping
    a. Remove Read2/CTV3 codes related to drug prescription events (defined in the Read Codes Drug and Appliance Dictionary (DAAD) dictionary and identified by having a lower-case first character).
    b. Remove invalid Read2/CTV3 codes (e.g. 'XXXX', '@@A2') or local General Practitioner (GP) codes which could not be identified in the NHS terminology reference and are often used as placeholders (e.g. 'UTEST' or 'UAB00').
    c. Remove Read2/CTV3 codes related to occupations, examinations, symptoms, diagnostic and laboratory procedures (all codes in Read chapters 0 to 9 inclusive).
3. Map Read2/CTV3 codes to ICD-10 using the NHS Digital Technology Reference data Update Distribution (TRUD) (7)cross-map file and remove codes which were not not found in the map file.
    a. Remove Read2/CTV3 codes where the mapped ICD-10 chapter is X (Other external causes of accidental injury, self-harm), Y (Assault, legal interventional, complications) or Z (Factors influencing health status and contact with health services) as they were excluded in the original Phecode definitions.
    b. Exclude Read2/CTV3 codes mapping to an entire ICD-10 chapter (e.g. "H…. Respiratory diseases" maps to ICD-10 Chapter X "Diseases of Respiratory System") since they are rarely used and are too broad.
4. Apply manual refinements e.g. recode specific ICD-10 leaf nodes to parent nodes (described in next paragraph)
5. Map ICD-10 codes to Phecodes using the existing translation file (3) (version 1.2beta).

Mappings between Read codes and ICD-10 defined in the "cross map' file used in the algorithm are defined in a different approach as outlined in Table 1. Where multiple target ICD-10 concepts existed (e.g. "one-to-block" mappings in Read2 or "type A" mappings in CTV3) we chose the broadest ICD-10 concept available (often, but not always, identified by the fourth digit being equal to '9' e.g. L21.9 – Seborrheic dermatitis, unspecified). This was done in order to avoid assigning a specific diagnosis when the source Read concept was broader than the candidate target ICD-10 concept. Where asterisk, dagger or asterisk and dagger terms were specified in the map file, we retained the most specific of ICD-10 (e.g. “H36.0 Diabetic Retinopathy” was preferred over “E14 Unspecified diabetes mellitus”. In instances where one Read2/CTV3 to many target ICD-10 concepts were defined in the map file, we added all individual ICD-10 codes as new records (e.g. Read2 code “F4K0.0 Scleritis and episcleritis” was mapped to two ICD-10 codes “H15.0 Scleritis” and “H15.1 Episcleritis”). In all other cases where a one-to-one map type was supplied, we used the target ICD-10 code specified. Finally, in order to align the target ICD-10 codes with the ICD-10-Phecode map, in some cases we upcoded (specific to broad) or downcoded (broad to specific) terms. Specifically, in some cases four character ICD-10 codes were remapped to children terms (e.g. “J34X Other disorders of nose and nasal sinuses” was remapped to “J34.8 Other specified disorders of nose and nasal sinuses”). In other cases, specific children terms were remapped to broader parent terms (e.g. “I48.9 Atrial fibrillation and atrial flutter, unspecified” was remapped to “I48 Atrial fibrillation and flutter”.

**Table 1**. Map types specified by the NHS TRUD terminology cross-map files between source terms in Read2 and CTV3 to target ICD-10 codes. The Type column denotes the cardinality of the map; for example, one-to-one denotes one source term with exactly one target term while one-to-many denotes one source term with multiple potential target terms.

| **Type** | **Example: Read V2/CTV3 => ICD-10 code** | **N (%)** |
|---|---|---|
| **Read2** | | |
| **one to one** | One source code can be mapped to exactly one ICD-10 code.<br>G5730 Atrial Fibrillation => I48 Atrial fibrillation and atrial flutter | 11,079<br>94% |
| **one to block** | One source code can be mapped to any one code from a specific ICD-10 block.<br>G30..Acute myocardial infarction => I21.0-I21.9 | 334<br>2.8% |
| **one to one of many** | One source code can be mapped to one of multiple potential ICD-10 codes.<br>AB20. Candidiasis of mouth and oesophagus => B37.0 Candidal stomatitis *or* B37.8 Candidiasis of other sites | 230<br>1.95% |
| **asterisk** | Asterisk codes relate to disease manifestations (as opposed to disease aetiology)<br>Cyu4K [X]Disorders of adrenal glands => E35.1* Disorders of adrenal glands | 57<br>0.5% |
| **asterisk & dagger** | This map type combines both asterisk and dagger codes.<br>A053. Amoebic liver abscess => A06.4† Amoebic liver abscess (K77.0*) K77.0* Liver disorders in infectious and parasitic diseases classified elsewhere | 49<br>0.5% |
| **dagger** | Dagger codes relate to disease aetiology.<br>N042. Other rheumatoid arthropathy => M05.3† Rheumatoid arthritis with involvement of other organs and systems | 22<br>0.1% |
| **one to many** | One source code maps to multiple ICD-10 codes that need to be combined<br>SP08Z Thrombosis of artery of transplanted kidney => N280 Ischemia and infarction of kidney + Z940 Kidney transplant status | 4<br>0.03% |
| **CTV3** | | |
| **G** | **G**eneric mapping, target ICD-10 code broader than source concept.<br>X101u Late onset asthma => J45.9 Asthma, unspecified | 11,425<br>47% |
| **D** | **D**efault mapping - most acceptable amongst alternatives given absence of other information<br>G5730 Atrial fibrillation => I48.9 Atrial fibrillation and atrial flutter | 11,064<br>46% |
| **A** | **A**lternative mapping - semantically similar to “D” mappings, multiple potential target | |

| | | |
|---|---|---|
| | concepts exist.<br>G5730 Atrial fibrillation => I48.0 Paroxysmal atrial fibrillation, I48.1 Persistent atrial fibrillation, I48.2 Chronic atrial fibrillation | |
| **E** | **E**xact one to one mapping, no alternatives exist.<br>H322. Centrilobular emphysema => J43.2 Centrilobular emphysema | 1,392<br>6% |

Assigning Phecodes, separately in primary care and hospitalisation data

All unique codes from primary care (Read2, CTV3) and hospitalisations (ICD-10) from each UKB participant were extracted and translated into Phecodes. For hospitalization data, we included all ICD-10 codes (recorded as either a primary or secondary diagnosis) and used an existing ICD-10-Phecode translation file (3). In order to enhance the specificity and positive predictive value, individuals were considered a "case" if they had at least two occurrences (irrespective of when) of an ICD-10 code that mapped to a Phecode. Individuals were marked as "controls" if they did not have any ICD-10 codes belonging to the exclusion codes defined in the Phecode definition files. Individuals with ICD-10 codes in the Phecode exclusion range were excluded from the respective analyses. We excluded Phecodes that occurred in less than 100 participants (0.05 prevalence) in order to reduce data sparsity.

Comparative PheWAS analysis

To evaluate the quality of the Phecode data derived from primary care EHR, we performed a PheWAS analysis in the UK Biobank seeking to replicate existing known GWAS associations identified through previous analyses (8,9) in seven Single Nucleotide Polymorphisms (SNPs): rs3135388, rs17234657, rs2200733, rs1333049, rs6457620, rs8050136, and rs7903146. We obtained up to date estimates between the seven SNPs and 25 phenotypes which we used as comparators by conducting a PheWAS analysis in BioVU(10), the Vanderbilt University Medical Center biobank.

The main analysis used phenotypes derived from both primary care and hospitalization billing data in the UK Biobank. We performed a subsequent sensitivity analysis by analyzing phenotype data derived only from primary care EHR. SNP-phenotype associations were assessed using a logistic regression model assuming an additive genetic model. UK Biobank models were adjusted for sex, genotyping array and the first 11 principal components. BioVU models were adjusted for age at most recent medical encounter, sex, 10 principal components and length of EHR (defined as the time between each patient's first and most recent medical encounter). We calculated and reported the expected effect sizes as odds ratios (OR) and assumed a SNP-phenotype association was significant if it's P-value surpassed Bonferroni significance.

UK Biobank and Bio VU Genotyping and Quality Control (QC)

487,409 UKB participants were genotyped using one of two custom genome-wide arrays and data were imputed to a combination of the UK10K, 1000 Genomes Phase 3 and the Haplotype Reference Consortium (HRC) reference panels resulting in 93,095,623 variants. We applied additional variant level QC and excluded variants with: a) Fisher's exact test <0.3, b) minor allele frequency (MAF) <1% and, c) a missing call rate of ≥5%. We applied individual-level QC and excluded participants with: a) excessive or minimal heterozygosity, b) more than 10 putative third-degree relatives as per the kinship matrix, c) no consent to extract DNA, d) sex mismatches between self-reported and genetic sex, and d) non-European ancestry (more details on QC provided elsewhere (11)). BioVU samples were genotyped on the Infinium Multi-Ethnic Genotyping Array (MEGA). To curate the genotyping data, a quality control step was conducted. Briefly, the quality control has been performed with following steps: a) the samples with per-individal call rate less than 95% were removed, b) the samples with wrongly assinged sex were deleted, c) the samples from related individiuals (PI_HAT ≥ 0.25) were removed, d) other unexpected duplications were eliminated. A genome imputation process was applied to increase coverage of the GWAS, using the Michigan Imputation Server(12) and referring to the HRC.

Implementation

Phenotypic data extraction and translation was performed in Python 3.7.7, genetic data extraction and QC was performed using bgenix, PLINK 1.9 and qctool v2. PheWAS analyses were conducted in R v. 4.0.0 using the PheWAS library (https://github.com/PheWAS/PheWAS(. The Read-Phecode map file, programming scripts, and documentation files will be made available under an open source license at the PheWAS catalogue(13) and the HDR UK Phenotype Portal(14).

**Results**

In the UKB, we processed 123,662,421 primary care clinical events: 3,6175,235 using Read2 and 87,487,186 using CTV3. We extracted 38,228 unique Read2 terms and 80,994 unique CTV3 terms which were recorded among 230,096 participants. After applying the mapping algorithm (described previously), we identified 11,775 and 23,881 unique Read2 and CTV3 terms respectively which were eligible for mapping. We mapped 10,834 (92%) of Read2 terms to 1,449 Phecodes, and 21,988 (91%) of CTV3 terms to 1,490 unique Phecodes (Table 2**)**. We additionally processed 3,541,618 admitted patient hospitalization events from 435,632 patients and translated 6,758 unique ICD-10 primary and secondary diagnoses from 9,493,039 hospitalizations to Phecodes.

**Table 2.** Translation of Read Version 2 and Clinical Terms Version 3 terms to Phecodes.

| | **Read Version 2** | **Clinical Terms Version 3** |
|---|---|---|
| **Events** | 36,175,235 | 87,487,186 |
| **Unique codes** | 38,228 | 80,994 |
| **Codes eligible for mapping** | 11,775 | 23,881 |
| **Unique Read codes mapped (%)** | 10,834 (92%) | 21,988 (91%) |
| **Unique ICD-10 codes** | 4,655 | 5,407 |
| **Unique Phecodes** | 1,449 | 1,490 |

In Read2, the top five unmapped terms were: a) *G84.. Hemorrhoids* (n=9,601), b) *Eu32z Depressive episode* (n=7,702), unspecified, c) *Eu32. Depressive episode* (n=3,684), d) *G843. External haemorrhoids, simple* (n=1701), and e) *N224. Ganglion and cyst of synovium, tendons, bursa* (n=1,553). In CTV3, the most common unmapped terms were: a) *XE2q5 Serum creatinine level* (n=1,194,627), b) *XE2q0 Serum sodium level* (n=1,142,496), c) *XE2pz Serum potassium level* (n=1,140,859), d) *XM0lt Serum urea level* (n=1,041,244) and e) *XE2eA Serum albumin level* (n=1,014,201). In Read2, laboratory test results are in chapter 4 while in CTV3 concepts on laboratory measurements exist across multiple chapter chapters (e.g. “44M4. Serum albumin” and “XE2eA Serum albumin”).

The five most commonly recorded leaf Phecodes in hospital administrative data were: a) 401.10 Essential hypertension (n=100,940), b) *272.11 Hypercholesterolemia* (n=35,033), c) *716.90 Arthropathy* NOS (n=32,345), d) *550.20 Diaphragmatic hernia* (n=21,203) and e) *562.10 Diverticulosis* (n=20,763). In primary care EHR data, the most frequently-recorded leaf Phecodes were: a) *745.0 Pain in joint* (n=46,343), b) *401.1 Essential hypertension* (n=40,645), c) *716.9 Arthropathy Not Elsewhere Specified (NOS)* (n=30,516), d) *519.8 Other diseases of respiratory system, Not Elsewhere Classified (NEC)* (n=29,146) and e) 760.0 *Back pain* (n=25,562).

In the primary analyses (Figure 2., Table 3) using Phecodes from hospitalization and primary care events, we performed a PheWAS in 408,415 UKB participants (54.4% female, mean age 59 SD 7.96) and 65,561 BioVU participants (55.6% female, mean age 57.87 SD 22.78) using 1,851 Phecodes respectively. We replicated 22 (100%) of 22 previously-reported genotype-phenotype associations with adequate statistical power across both of the datasets. In the secondary analyses, we performed a PheWAS in 185,648 UKB participants (~50% of the cohort is linked to primary care EHR, 54.5% female, mean age 58 SD 7.96) and 1,851 Phecodes and replicated 18 (81%) of 22 previously-reported PheWAS associations.

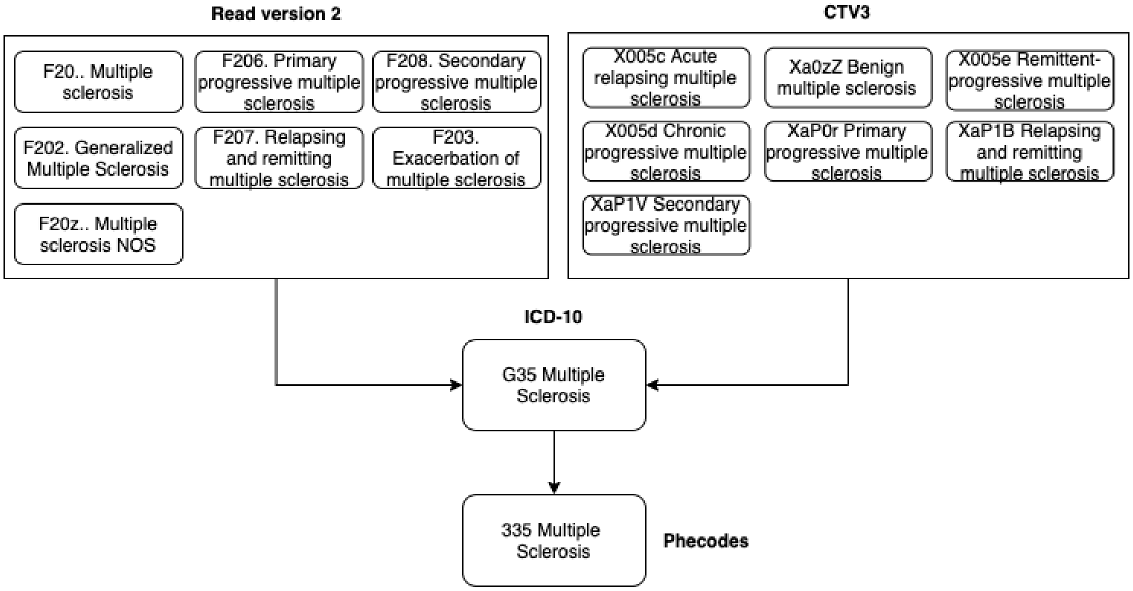
Read version 2
F20.. Multiple sclerosis
F206. Primary progressive multiple sclerosis
F208. Secondary progressive multiple sclerosis
F202. Generalized Multiple Sclerosis
F207. Relapsing and remitting multiple sclerosis
F203. Exacerbation of multiple sclerosis
F20z.. Multiple sclerosis NOS
CTV3
X005c Acute relapsing multiple sclerosis
Xa0zZ Benign multiple sclerosis
X005e Remittent-progressive multiple sclerosis
X005d Chronic progressive multiple sclerosis
XaP0r Primary progressive multiple sclerosis
XaP1B Relapsing and remitting multiple sclerosis
XaP1V Secondary progressive multiple sclerosis
ICD-10
G35 Multiple Sclerosis
335 Multiple Sclerosis
Phecodes

**Figure 1.** Exemplar flowchart of the process of mapping Read 2 and CTV3 terms, to ICD-10 and to Phecodes. CTV3 contains all Read2 terms but these were not duplicated in the display. CTV3=Clinical Terms Version 3.

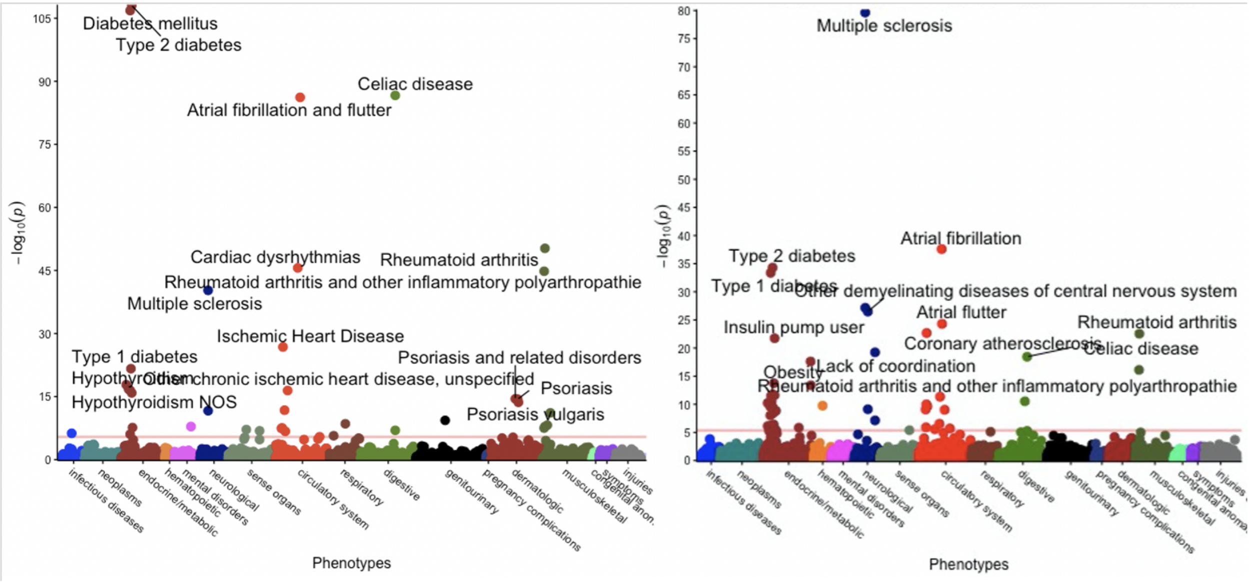

**Figure 2**. Manhattan plot of phenome-wide association analyses in 185,648 participants (UKB, left) and 65,561 participants (BioVU, right) and 1,851 Phecodes derived from electronic health records. The red line shows the Bonferroni level of significance. Only phenotypes that cross the Bonferroni level of significance are annotated.

**Table 3**: PheWAS comparison of Phecodes derived from UK Biobank electronic health records (primary care EHR exclusively and from primary care and hospitalization data combined) to BioVU PheWAS analyses for seven variants with known associations. For each variant, the mapped genes and Genome Aggregation Database (gnomAD) identifiers are provided. P-Value columns with gray shading indicate results did not survive Bonferroni correction. Phenotypes with no identifier (in first column) are reported to illustrate differences in coding between data sources.

OR = Odds Ration, PheWAS = Phenome-Wide Association Study. "n/a" denotes analyses which were not run due to a very low number of cases or no of cases due to a Phecode not being present in the data.

| | | **UK Biobank (all sources)** | | | **UK Biobank (primary care)** | | | **BioVU** | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **#** | **Phenotype (phecode)** | **Cases** | **OR** | **P Value** | **Cases** | **OR** | **P Value** | **Cases** | **OR** | **P Value** |
| | **rs3135388 (HLA-DRB1, gnomAD: 6_32413051_A_G)** | | | | | | | | | |
| 1 | Multiple sclerosis (335) | 1486 | 2.42 | 2.79E-105 | 544 | 2.46 | 5.41E-41 | 1174 | 2.44 | 2.28E-80 |
| 2 | Type 1 diabetes (250.1) | 2538 | 0.45 | 1.90E-48 | 664 | 0.28 | 2.33E-22 | 2082 | 0.49 | 4.38E-34 |
| 3 | Hypothyroidism NOS (244.4) | 19943 | 0.82 | 3.33E-35 | 7135 | 0.8 | 1.76E-17 | 7328 | 0.84 | 5.89E-11 |
| 4 | Celiac disease (557.1) | 2056 | 0.64 | 6.06E-18 | 665 | 0.61 | 1.06E-07 | 394 | 0.55 | 5.35E-06 |
| | **rs17234657 (AC108105.1, gnomAD: 5_40401407_T_G)** | | | | | | | | | |
| 5 | Regional enteritis (555.1) | 1639 | 1.34 | 3.01E-10 | 432 | 1.41 | 1.50E-04 | 2065 | 1.33 | 2.96E-11 |
| | **rs2200733 (PITX2, gnomAD: 4_110789013_C_T)** | | | | | | | | | |
| 6 | Cardiac dysrhythmias (427.0) | 30440 | 1.35 | 2.68E-124 | 8933 | 1.39 | 2.49E-46 | 762 | 1.65 | 4.50E-12 |
| 7 | Atrial fibrillation (427.21) | n/a | n/a | n/a | n/a | n/a | n/a | 6399 | 1.49 | 2.54E-38 |
| | Atrial flutter (427.22) | n/a | n/a | n/a | n/a | n/a | n/a | 1939 | 1.64 | 5.03E-25 |
| | Atrial fibrillation and flutter (427.2) | 20962 | 1.53 | 2.78E-185 | 4143 | 1.82 | 6.65E-87 | 29 | 2.35 | 8.49E-03 |
| | **rs1333049 (CDKN2B-AS1, gnomAD: 9_22125504_G_C)** | | | | | | | | | |
| 8 | Coronary atherosclerosis (411.4) | 16934 | 1.25 | 3.02E-88 | 588 | 1.07 | 2.53E-01 | 10109 | 1.19 | 2.12E-23 |
| 9 | Hyperlipidemia (272.1) | 33679 | 1.06 | 1.14E-12 | 4800 | 1.02 | 3.30E-01 | 13752 | 1.08 | 1.43E-06 |
| 10 | Angina pectoris (411.3) | 15164 | 1.19 | 1.78E-48 | 2713 | 1.15 | 1.82E-07 | 2239 | 1.23 | 1.03E-10 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 11 | Unstable angina (411.1) | 3458 | 1.19 | 4.64E-13 | 221 | 1.26 | 1.67E-02 | 1930 | 1.23 | 8.53E-10 |
| 12 | Myocardial infarction (411.2) | 13755 | 1.2 | 8.80E-51 | 2259 | 1.24 | 1.72E-12 | 3943 | 1.13 | 1.25E-06 |
| 13 | Other chronic IHD, unspecified (411.8) | 16934 | 1.18 | 6.78E-51 | 4060 | 1.21 | 3.63E-17 | 3381 | 1.18 | 2.32E-10 |
| | **rs6457620 (HLA-DQB1, gnomAD: 6_32696222_G_C)** | | | | | | | | | |
| 14 | Rheumatoid arthritis (714.1) | 5006 | 1.57 | 3.96E-106 | 1349 | 1.82 | 5.49E-51 | 1633 | 0.70 | 2.87E-23 |
| 15 | Multiple sclerosis (335.0) | 1486 | 0.63 | 3.14E-33 | 544 | 0.64 | 2.35E-12 | 1174 | 1.60 | 6.87E-28 |
| 16 | Celiac disease (557.1) | 2056 | 0.27 | 3.07E-248 | 665 | 0.25 | 2.25E-87 | 394 | 1.97 | 3.54E-19 |
| | **rs8050136 (FTO, gnomAD: 16_53782363_C_A)** | | | | | | | | | |
| 17 | Obesity (278.1) | 11088 | 1.15 | 6.60E-24 | 1223 | 1.05 | 2.77E-01 | 5812 | 1.19 | 2.46E-18 |
| 18 | Type 2 diabetes (250.2) | 26517 | 1.11 | 8.15E-28 | 9816 | 1.13 | 1.35E-16 | 9532 | 1.12 | 1.72E-12 |
| 19 | Essential hypertension (401.1) | 100106 | 1.04 | 4.63E-12 | 33304 | 1.05 | 3.06E-08 | 25958 | 1.06 | 1.08E-04 |
| | Obstructive sleep apnea (327.32) | n/a | n/a | n/a | n/a | n/a | n/a | 5118 | 1.09 | 2.24E-05 |
| 20 | Sleep apnea (327.3) | 4890 | 1.12 | 1.98E-08 | 740 | 1.17 | 2.68E-03 | 2104 | 1.09 | 6.02E-03 |
| | **rs7903146 (TCF7L2, gnomAD: 10_112998590_C_T)** | | | | | | | | | |
| 21 | Type 2 diabetes (250.2) | 26517 | 1.34 | 3.48E-210 | 9816 | 1.41 | 9.18E-109 | 9532 | 1.24 | 4.88E-35 |
| 22 | Type 2 diabetes ophthalmic compl. (250.23) | 1494 | 1.34 | 2.46E-14 | n/a | n/a | n/a | 722 | 1.33 | 3.80E-07 |

## Discussion

In this study, we described the process of mapping primary care EHR data recorded using the Read2 and CTV3 controlled clinical terminologies to Phecodes. We performed a PheWAS in two contemporary international biobanks, the UK Biobank (UK) and BioVU (US) and provided evidence towards the validity of Phecodes derived from longitudinal primary care EHR by showing concordant findings. Similar to the US, UK hospital data are collected for billing purposes which may influence the coding process however billing codes in BioVU additionally include ambulatory care. UK primary care has a true EHR generated and captured during clinical care by healthcare professionals; billing and funding may also have a modest effect (e.g. through the Quality and Outcomes Framework(15)). Reassuringly, the distribution of most frequently mapped Phecodes between sources reflected these differences with Phecodes on symptoms (e.g. back pain, joint pain) being more prevalent in primary care EHR.

Our study is the first, to our knowledge, to develop and evaluate Phecodes sourced from primary care EHR data and show the validity of the translation file through a PheWAS analyses in two international resources. Comparison between UKB and BioVU uncovered challenges with regards to the resolution of diagnosis codes used in the UK and the US. ICD-10-CM used in the US has a higher fidelity than ICD-10 used in the UK which directly influenced the level at which Phecodes were assigned. For example, in BioVU, the availability of ICD-10-CM codes meant that sleep apnea was recorded using a lower-level Phecode "Obstructive sleep apnea (327.32)" whereas in the UK, a higher level one "Sleep apnea (327.3)". As a result, while the PheWAS analyses failed to show concordant results between the resources, each individual analyses on the Phecodes that were actually available in the data sources was significant. The same effect was observed in atrial fibrillation and flutter which are recorded using a single ICD-10 code in UK data but are split across two different codes in ICD-10-CM (and as a result two different Phecodes).

The main limitation of our study is the fact that only approximately 50% of the population in the UK Biobank has primary care data available. As a result, our PheWAS analyses using Phecodes derived from primary care EHR failed to replicate previous associations due to low statistical power (e.g. 740 cases of sleep apnoea recorded in primary care EHR). Some diagnoses in primary care were not mapped to Phecodes. This could be due to to the fact that the target ICD-10 code was not in the ICD-10-Phecode map; for example in the case of haemorrhoids we observed a mismatch between the ICD-10 code available in the cross-map file “"K64 Hemorrhoids and perianal venous thrombosis” and the ICD-10 code available in the ICD-10-Phecode map “"I84 Haemorrhoids". Another reason for missed mappings was that source Read2/CTV3 concepts were too generic (e.g. S3z.. Fracture of unspecified bones) and could not be mapped. In some cases, such as Read code Eu32z "[X]Depressive episode, unspecified", the target term was ICD-10 F32 "Major depressive disorder, single episode" which is a non-billable code with no associated Phecode. Finally, some Read codes (e.g. E2273 Impotence) were not defined in the Read-ICD-10 cross-map files and as a result were not mapped to a relevant Phecode during data extraction.

Despite the relatively small number of events not mapped to Phecodes, in subsequent studies, it will be important to further refine these mappings and ensure higher fidelity by establishing new Phecodes where needed. Primary care EHR contains a wealth of information which is not captured in administrative hospital records. For example, Read contains terms on examination findings, laboratory tests, risk factors and symptoms. A potential future expansion of Phecodes could include a custom set of codes to capture these events despite not having a valid ICD-9-CM or ICD-10 target code. This would be relevant not only for UKB participants but also to other deeply phenotyped/genotyped resources e.g. Genomics England(16) or the NHS Digital (17).

## Conclusion

In this paper, we introduced our work of mapping the Read2 and CTV3 clinical terminologies used in primary care EHR to Phecodes. We validated our results by replicating previously-reported PheWAS genotype-phenotype associations by performing analyses in the UK Biobank and BioVU. We provide an initial version of the mapping file that can be used by researchers to leverage primary care data for high-throughput translational research.

**Acknowledgements**
SD, HH are supported by a) Health Data Research UK, b) BigData@Heart, funded by the IMI-2 No. 116074, c) the NIHR UCL Hospitals Biomedical Research Centre. HH, ADH are NIHR Senior Investigators. The dataset(s) used for the analyses described were obtained from Vanderbilt UMC BioVU, supported by institutional funding, private agencies and federal grants incl. NIGMS R01 GM139891, NIA R01AG069900, NIH-funded Shared Instrumentation Grant S10RR025141, and CTSA grants UL1TR002243, UL1TR000445, UL1RR024975. LB is supported by R01-LM010685. Work was supported by the BHF Accelerator Award AA/18/6/24223. ADH supported by Rosetrees and Stoneygate Trust.

## References

**1.** Denny JC, Bastarache L, Ritchie MD, Carroll RJ, Zink R, Mosley JD, et al. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. Nat Biotechnol. 2013 Dec 1;31(12):1102–11.

**2.** Wei W-Q, Bastarache LA, Carroll RJ, Marlo JE, Osterman TJ, Gamazon ER, 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 Jul 7;12(7):e0175508.

**3.** Wu P, Gifford A, Meng X, Li X, Campbell H, Varley T, et al. Mapping ICD-10 and ICD-10-CM Codes to Phecodes: Workflow Development and Initial Evaluation. JMIR Med Inform. 2019 Nov 29;7(4):e14325.

**4.** Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, 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(3):e1001779.

**5.** Denaxas S, Gonzalez-Izquierdo A, Direk K, Fitzpatrick NK, Fatemifar G,, Banerjee A, et al. UK phenomics platform for developing and validating electronic health record phenotypes: CALIBER. J Am Med Inform Assoc. 2019 Dec 1;26(12):1545–59.

**6.** Kuan V, Denaxas S, Gonzalez-Izquierdo A, Direk K, Bhatti O, Husain S, et al. A chronological map of 308 physical and mental health conditions from 4 million individuals in the English National Health Service. Lancet Digit Health. 2019 Jun;1(2):e63–77.

**7.** NHS TRUD . [cited 2021 Mar 5]. Available from: https://isd.digital.nhs.uk/trud3/

**8.** Denny JC, Ritchie MD, Basford MA, Pulley JM, Bastarache L, Brown-Gentry K, et al. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. Bioinformatics. 2010 May 1;26(9):1205–10.

**9.** Zheng NS, Feng Q, Kerchberger VE, Zhao J, Edwards TL, Cox NJ, et al. PheMap: a multi-resource knowledge base for high-throughput phenotyping within electronic health records. J Am Med Inform Assoc. 2020 Nov 1;27(11):1675–87.

**10.** Roden DM, Pulley JM, Basford MA, Bernard GR, Clayton EW, Balser JR, et al. Development of a large-scale de-identified DNA biobank to enable personalized medicine. Clin Pharmacol Ther. 2008 Sep;84(3):362–9.

**11.** Garfield V, Farmaki A-E, Fatemifar G, Eastwood SV, Mathur R, Rentsch CT, et al. The Relationship Between Glycaemia, Cognitive Function, Structural Brain Outcomes and Dementia: A Mendelian Randomisation Study in the UK Biobank. Diabetes [Internet]. 2021 Feb 25;

**12.** Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, et al. Next-generation genotype imputation service and methods. Nat Genet. 2016 Oct;48(10):1284–7.

**13.** Phenome Wide Association Studies (PheWAS) Catalogue. [cited 2021 Mar 10]. Available from: https://phewascatalog.org/

**14.** HDR UK CALIBER Phenotype Library. [cited 2021 Mar 10]. Available from: https://portal.caliberresearch.org/

**15.** Walker S, Mason AR, Claxton K, Cookson R, Fenwick E, Fleetcroft R, et al. Value for money and the Quality and Outcomes Framework in primary care in the UK NHS. Br J Gen Pract. 2010 May;60(574):e213–20.

**16.** Turnbull C. Introducing whole-genome sequencing into routine cancer care: the Genomics England 100 000 Genomes Project. Ann Oncol. 2018 Apr 1;29(4):784–7.

**17.** Wood A, Denholm R, Hollings S, Cooper J, Ip S, Walker V, et al. Linked electronic health records for research on a nationwide cohort of more than 54 million people in England: data resource. BMJ. 2021 Apr 7;373:n826.