UK phenomics platform for developing and validating EHR phenotypes: CALIBER

Spiros Denaxas [1,2,*], Arturo Gonzalez-Izquierdo [1,2], Kenan Direk [1,2], Natalie Fitzpatrick [1,2], Amitava Banerjee [1,2], Richard Dobson [1,2, 5], Ghazaleh Fatemifar [1,2], Valerie Kuan [2,6], Tom Lumbers [1,2], Laura Pasea [1,2], Riyaz Patel [1,2], Aroon Hingorani [2,6], Cathie Sudlow [3,4], Harry Hemingway [1,2]

1. Institute of Health Informatics, University College London, UK
2. Health Data Research UK (HDR UK), London, UK
3. Centre for Clinical Brain Sciences, University of Edinburgh, Edinburgh, UK
4. Centre for Medical Informatics, Usher Institute of Population Health Sciences and Informatics, University of Edinburgh, Edinburgh, UK
5. Department of Biostatistics and Health Informatics, Institute of Psychiatry Psychology and Neuroscience, King's College London, London, UK
6. Institute of Cardiovascular Science, University College London, London, UK

[*] Corresponding author: Institute of Health Informatics, University College London, 222 Euston Road, NW1 2DA, United Kingdom - email: s.denaxas@ucl.ac.uk , telephone: +44 203 5495324

Keywords: electronic health records, phenotyping, medical informatics, personalized medicine

ABSTRACT

Objective Electronic health records are a rich source of information on human diseases, but the information is variably structured, fragmented, curated using different coding systems and collected for purposes other than medical research. We describe an approach for developing, validating and sharing reproducible phenotypes from national structured Electronic Health Records (EHR) in the UK with applications for translational research.

Materials and Methods We implemented a rule-based phenotyping framework, with up to six approaches of validation. We applied our framework to a sample of 15 million individuals in a national EHR data source (population based primary care, all ages) linked to hospitalization and death records in England. Data comprised continuous measurements such as blood pressure, medication information and coded diagnoses, symptoms, procedures and referrals, recorded using five controlled clinical terminologies: a) Read (primary care, subset of SNOMED-CT), b) ICD-9, ICD-10 (secondary care diagnoses and cause of mortality), c) OPCS-4 (hospital surgical procedures) and d) Gemscript Drug Codes.

Results The open-access CALIBER Portal (https://www.caliberresearch.org/portal) demonstrates phenotyping algorithms for 50 diseases, syndromes, biomarkers and lifestyle risk factors and provides up to six validation layers. These phenotyping algorithms have been used by 40 national/international research groups in 60 peer-reviewed publications.

Conclusion Herein, we describe the UK EHR phenomics approach, CALIBER, with initial evidence of validity and use, as an important step towards international use of UK EHR data for health research.

BACKGROUND AND SIGNIFICANCE

The United Kingdom (UK) health system offers international researchers opportunities to explore 'cradle to grave' longitudinal electronic health record (EHR) phenotypes at scale. It is one of the few countries which combines a single-payer-and-provider comprehensive national health service (NHS), free at the point of care, with extensive national data resources across the entire 65M population. Patients are identified by a unique healthcare-specific identifier which enables linkage of patient data across EHR sources and the creation of longitudinal phenotypes that span primary and secondary care [1]. Over 99% of people are registered with a general practitioner (GP) and structured primary care data collected electronically have been used by UK, US and other researchers for decades [2]. Furthermore, national EHR are being linked with large-scale consented genomic resources i.e. 100,000 Genomes Project [3], UK Biobank [4–6] and enable the investigation of simple/complex traits across participant populations with diverse genetic backgrounds [7].

The UK EHR landscape differs from the US and elsewhere in important ways. Although the UK, unlike the US, has the opportunity to establish a national approach, it faces the common challenge that EHR for primary care and hospital care are handled by different data providers and are kept separately, with independent access consent requirements [8,9]. Significant progress has been made by US initiatives (eMERGE [10], BioVU [11], Million Veteran Programme, [12] *All Of Us* [13]), Canada [14], Australia [15], Sweden [16] and Denmark [17]. In the UK however, there has been a lack of a recognized phenotyping framework and no go-to resource for EHR researchers for systematically creating, curating and validating EHR-derived phenotypes, obtaining information on controlled clinical terminologies, sharing algorithms, and disseminating best approaches. Structured primary care EHR have been used in >1800 published studies [18] but only 5% of studies published sufficiently reproducible phenotypes [19] while significant heterogeneity exists (one review reported 66 asthma definitions [20]). Current UK initiatives [19,21,22] for curating EHR-derived phenotypes focus on lists of controlled clinical terminology terms (*codelists*) rather than self-contained phenotypes i.e. terms, implementation, validation evidence.

EHR phenotype validation is a critical process guiding their subsequent use in research or care [23,24]. There are multiple sources of evidence/study designs that contribute to building confidence in the validity of an EHR phenotype for a particular purpose. Countries may also differ in the opportunities for validation: e.g. in the UK cross referencing against multiple EHR sources, prognostic validation and risk factor validation are all made possible by nationwide population-based records [25–29]. In contrast with the US,

only recently have scalable methods been developed to access the entire hospital record for expert review [30] and text corpora are not available [34]. There have been few previous studies [31] of the validity of International Classification of Disease and Health Related Problems, 10th Revision (ICD-10) terms [32] in the UK against hospital records because introduction of hospital EHRs are recent (e.g. three hospitals that have stage six on the HIMSS EMRAM [33]).

We have developed the CALIBER EHR platform for the UK by adopting and extending best practices from US initiatives. Here, we define a framework for enabling EHR phenotyping in a scalable and reproducible manner. We present a systematic validation framework for assessing accuracy consisting of up to six approaches of evidence (i.e. expert review to prognostic validation) and disseminating through a centralized open-access repository. We have chosen heart failure (HF), acute myocardial infarction (AMI) and bleeding as some of the medical conditions that exemplify the strengths of national linked UK EHR and the non-trivial challenges researchers encounter.

MATERIALS AND METHODS

We developed an iterative and collaborative approach for creating and validating rule-based EHR phenotyping algorithms using UK structured EHR. The approach involved expert review interwoven with data exploration and analysis. An EHR phenotyping algorithm translates the clinical requirements for a particular patient to be considered a case into queries that leverage EHR sources stored in a relational database and extracts disease onset, severity and subtype information. In the following sections we describe the platform, the algorithm development process and validation consisting of six approaches of evidence.

UK primary care EHR, hospital billing data and cause-specific mortality in the CALIBER platform

CALIBER platform [35] is currently built around four national EHR data sources **(Figure 1)** deterministically linked using the NHS number (unique ten-digit identifier assigned at birth or first interaction), gender, postcode and date of birth; 96% of patients with a valid NHS number successfully linked [36].

The baseline cohort is composed of a national primary care EHR database, the *Clinical Practice Research Datalink (CPRD) [37]*. Primary care has used computerised health records since 2000 and general practices use one of several EHR systems. CPRD contains longitudinal primary care data (Vision and EMIS) on diagnoses, symptoms, drug prescriptions, vaccinations, blood tests and risk factors (irrespective of disease status and hospitalization). CPRD uses Read [38] terms (112,806 terms, subset of SNOMED-CT [38]). Prescriptions are recorded using Gemscript [39] (72,664 entries). CPRD contains >10billion rows of data

from >15M patients (from all the contributing primary care practices, irrespective of consent to linkage) shown to be representative in terms of age, sex, mortality and ethnicity [40–42] and of high validity [43].

Hospital Episode Statistics (HES) (https://digital.nhs.uk/) contains administrative data on diagnoses and procedures generated during hospital interactions. Diagnoses are recorded using ICD-10 and procedures using The Office of Population Censuses and Surveys Classification of Surgical Operations and Procedures, 4th Revision (OPCS-4) (10,713 terms, similar to Current Procedural Terminology [44]). Up to 20 primary and secondary discharge diagnoses are recorded per finished consultant episode. The *Myocardial Ischaemia National Audit Project (MINAP)*, is a national disease and quality improvement registry capturing all acute coronary syndrome events across England. MINAP contains diagnostic, severity and treatment information using 120 structured data fields [45]. The *Office for National Statistics* contains socioeconomic deprivation using the Index of Multiple Deprivation [46] and physician-certified cause-specific mortality (underlying and up to 14 secondary causes using ICD-9/ICD-10).

Algorithm development

The development pipeline was a collaborative and iterative process involving researchers from a diverse set of scientific backgrounds (e.g. clinicians, epidemiologists, computer scientists, public health researchers, statisticians). An iteration refers to an adjustment in the computational strategy to derive the phenotype in question, based on data-driven examinations of its internal validity and according to the clinical context. The number of development iterations was proportionate with the complexity of the clinical phenotype: algorithms leveraging multiple sources required multiple iterations and substantially more clinician input.

We initially defined search strategies for identifying relevant diagnosis terms and their synonyms which were selected based on input from clinicians, existing literature, national guidelines and by consulting medical vocabulary repositories e.g. UMLS Metathesaurus [47,48]. Two clinicians independently classified identified terminology terms (disagreements resolved by third) into non-overlapping categories: a) *prevalent* (e.g. “history of heart failure”) b) *possible* (e.g. “congestive heart failure monitoring”), and c) *incident* (potentially sub-classified e.g. “chronic congestive heart failure”, “acute left ventricular failure”, “heart failure not otherwise specified”). Similarly we identified and classified coded symptoms recorded in primary care EHR. Many CALIBER phenotyping algorithms combine coded diagnosis, symptom information, continuous measurements e.g. laboratory values or other physiological measurements and medication prescription information e.g. hypertension is defined using continuous blood pressure, coded diagnoses, blood-pressure lowering prescriptions, and comorbidities. Additionally, we generated ad-hoc rules to reconcile: a) coding differences across EHR sources with respect to the granularity of diagnosis i.e. Read terms include left/right ventricular dysfunction but ICD-10 are less specific, b) the presence of multiple

terms i.e. multiple different ethnicity entries, and c) transience in coding (e.g. ICD-9 for mortality before 2000).

Primary care EHR contain over 450 structured data items for recording continuous measurements e.g. blood markers. For continuous phenotypes (e.g. blood pressure), we normalised data quality by identifying the relevant units, specified unit conversions (where required) and defined valid value ranges. For example, the neutrophil count structured data area contained both the absolute values and percentages, and these had to be differentiated by supplementary Read terms and by checking the distribution of values by unit. Sometimes values were obviously on the wrong scale e.g. haemoglobin where some values were distributed as if measured in g/L but had (presumably incorrect) units recorded as g/dL. Zero values caused particular problems; they could be impossible and represent missing data in some cases (e.g. ferritin) but might be true zeroes representing undetectable values in other cases (e.g. basophils). Careful investigation by units and Read term was required to avoid creating Missing Not at Random data (if the zeroes were true) or false data (if the zeroes were false). Definition of valid ranges for values was also problematic, as we wanted to exclude erroneous values without excluding true physiologically extreme values.

Validation: Systematic evaluation using six approaches

We evaluated EHR-derived phenotypes across six approaches of evidence:

1) Cross-EHR source concordance

For EHR-derived cases of AMI, HF and bleeding, we quantified the percentage of cases identified in each source, the overlap between sources and evaluated per-source completeness and diagnostic validity. Additionally, we calculated the positive predictive value (PPV) and negative predictive value (NPV) by cross-validating AMI cases with a disease registry.

2) Case note review

For bleeding, two clinicians (blinded to the terms) reviewed the entire hospital record (charts, referral letters, discharge letters, imaging reports) for 283 completed patient hospital episodes across two large hospitals (University College London Hospitals, King's College Hospital). Bleeding assignments from the clinicians review was compared with those from the phenotyping algorithm and we estimated the PPV, NPV, sensitivity and specificity using the case review data as the “gold standard”. We extracted hospital data (14,364,947 words) using CogStack [49] from the consented SIGNUM study.

3) Consistency of risk factor-disease associations from non-EHR studies

For all exemplars, we produced hazard ratios (HR) of known risk factors from Cox proportional hazards models adjusted for age, sex and other covariates). We evaluated the ability of obtaining consistent estimates

(in terms of direction and magnitude) with risk factor associations derived from non-EHR research-driven studies.

4) Consistency with prior prognosis research

We produced Kaplan–Meier cumulative incidence curves at appropriate time intervals and endpoints and stratified by EHR source. We evaluated the observed prognostic profiles with previously-reported evidence for example observing different survival patterns between patients diagnosed with HF in CPRD but never hospitalized compared with patients in HES.

5) Consistency of genetic associations

Using UK Biobank, we examined consistency of genetic associations reported using non-EHR traits from published Genome-Wide Association Studies (GWAS) by comparing with results obtained using EHR-derived traits. We downloaded and mapped associations from the NHGRI-EBI catalogue [50] and CARDIoGRAMplusC4D [51]. The association between expected allelic dosages and traits was analysed using logistic regression with an underlying additive model in PLINK v2 [52] (adjusted for sex, age, genotyping array, principle components 1-15). We used LocusZoom to highlight region, linkage disequilibrium patterns and recombination rates of SNPs with the lowest p-value [53].

6) External populations

We assessed the validity of developed algorithms by implementing them in external data sources (UK or elsewhere), and examining consistency of results in the evaluation criteria.

Phenotype dissemination

We generated a textual descriptions algorithms with explicit detail on the logic behind the algorithm (pre-processing, cross-source reconciliation, quality checks) in a clinician-friendly manner. We generated flow-chart representations accompanied by pseudo-code for facilitating the translation of the algorithm to SQL. We created entries in the CALIBER Portal **(Figure 4)** describing implementation details across sources, research outputs, validation evidence and a Digital Object Identifier (DOI) [54]. We created an open-source R library for manipulating clinical terminologies (http://caliberanalysis.r-forge.r-project.org/) using a custom file format including metadata (e.g. naming, version, authors, timestamp).

Ethical approval

The CPRD has broad ethical approval for purely observational research using pseudonymised primary care data for supporting medical purposes that are in the interests of patients and the wider public. Linkages were performed by NHS Digital, the statutory body in England responsible for providing core healthcare information technology and curating many of the national datasets. This study was approved by the

Medicines and Healthcare Products Regulatory Agency (MHRA) Independent Scientific Advisory Committee (ISAC) - protocol references: 11_088, 12_153R, 16_221, 18_029R2, 18_159R.

RESULTS

We curated over 90,000 terms from five controlled clinical terminologies to create 50 validated phenotyping algorithms (35 diseases/syndromes, ten biomarkers, six lifestyle risk factors). In this manuscript, we used three exemplar phenotypes: heart failure (https://www.caliberresearch.org/portal/phenotypes/heartfailure), bleeding (https://www.caliberresearch.org/portal/phenotypes/bleeding) and acute myocardial infarction (https://www.caliberresearch.org/portal/phenotypes/acutemyocardialinfarction). **Table 1** provides a complete list of published, peer-reviewed phenotypes and the approaches of evidence supporting their validity. CALIBER phenotypes have been used by 40 national/international research groups in 60 peer-reviewed publications [55]. The CALIBER Portal (http://www.caliberresearch.org/portal) opened in Oct./2018 to the community and provides a centralized resource for curating EHR phenotypes.

1) Cross-EHR source concordance

The PPV of AMI (the probability that the diagnosis recorded in MINAP was AMI rather than unstable angina or a non-cardiac diagnosis) was 92.2% (6,660/7,224, 95% CI 91.6%-92.8%) in CPRD and 91.5% (6,851/7,489, 90.8%-92.1%) in HES **(Figure 2)**. Among the 17,964 patients with at least one record of non-fatal AMI, 13,380 (74.5%) were recorded by CPRD, 12,189 (67.9%) by HES, and 9,438 (52.5%) by MINAP. Overall, 5,561 (31.0%) of patients had AMI recorded in three sources (32.0% within 90 days) and 11,482 (63.9%) in at least two sources. For 89,554 HF cases, 26% were recorded in CPRD only, 27% in both CPRD and HES, 34% in HES only, and 13% had HF as cause of death without a previous record elsewhere. In 39,804 bleeding cases 59.4% were captured in CPRD, 50.2% in HES, and 3.8% events in ONS. Allowing a 30 day window, only 13.2% of events were captured in two or more sources. Similarly, a very small proportion (13.2%) of bleeding cases identified were captured in multiple data sources.

2) Case note review

We tested the validity of ICD-10 terms used in our bleeding phenotype and found a PPV of 0.88, i.e. 88% of bleeding events identified by ICD-10 terms were indeed bleeding events according to the independent review of the entire hospital record by two clinicians, blinded to the term assignment. We found that coded terms underestimate the occurrence of bleeding, with a sensitivity estimate of 0.48, consistent with a previous study where 38% of hospitalised bleeding events were not captured coded terms [56].

3) Aetiology

Figure 3 shows age and sex adjusted HRs from Cox proportional hazards models for HF and CVD risk factors (smoking, Type-II diabetes, systolic blood pressure, heart rate) in CALIBER and non-EHR studies.

4) Prognosis

In 20,819 AMI cases we found that patients with events recorded in only one source had higher mortality than those recorded in more than one source (age and sex adjusted HR 2.29, 95% CI 2.17 to 2.42; P<0.001) [26]. Among patients with AMI recorded in only one source, those only in CPRD had the highest mortality on the first day but the lowest mortality thereafter. Among patients with AMI recorded in HES or MINAP, those in MINAP had lower coronary mortality in the first month (age and sex adjusted HR 0.33, 0.28 to 0.39, P<0.001) but similar mortality for non-coronary events (1.12, 0.90 to 1.40, P=0.3). After the first month, patients with AMI in CPRD had about half the hazard of mortality of patients with AMI recorded in one of MINAP or HES (age/sex adjusted HR 0.49, 95% CI 0.40-0.60, P<0.001). In 89,994 HF cases, we observed 51,903 deaths and generated KM curves for 90 day survival. Adjusted for age and sex, HF was strongly associated with mortality, with HRs for all cause mortality ranging from 7.01 (95% CI 6.83–7.20), 7.23 (95% CI 7.03-7.43), up to 15.38 (95% CI 15.02-15.83) for patients in CPRD with acute HF hospitalization, CPRD only, and HES only, compared with a age/sex-matched reference population. Age, concomitant COPD, and diabetes were amongst the strongest predictors of death. Compared to patients with no bleeding, patients with bleeding recorded in CPRD and HES were at increased risk of all-cause mortality and atherothrombotic events. (HR all-cause mortality 1.98 (95% CI: 1.86, 2.11) for CPRD bleeding, and 1.99 (95% CI: 1.92, 2.05) for HES bleeding).

5) Genetic associations

In the UK Biobank, we identified 3,408 affected and 108,734 unaffected participants. We discovered 69 non-independent genetic variants spanning regions on chromosomes 6 and 9 (e.g. 9p21 loci) showing genome-wide significance ($p<5x10-8$, $\lambda=1.02$). Consistent direction and magnitude of associations were replicated in 67 (97%) of previously reported genetic variants

6) External populations

We assessed the validity of the AMI, HF and bleeding phenotypes by comparing long-term outcomes (any cause death, fatal AMI/stroke, hospital bleeding) in AMI survivors in England (n=4,653), Sweden (n=5,484), USA (n=53,909) and France (n=961) [57]. We found consistent associations with 12 baseline prognostic factors (age, gender, AMI, HF, diabetes, stroke, renal disease, peripheral arterial disease, atrial fibrillation, hospital bleeding, cancer, COPD) and each outcome. In each country, we observed high 3-year crude cumulative risks of all-cause death (from 19.6% [England] to 30.2% [USA]); the composite of AMI, stroke, or death [from 26.0% (France) to 36.2% (USA)]; and hospitalized bleeding [from 3.1% (France) to 5.3% (USA)]. Adjusted risks were similar across countries, but higher in the USA for all-cause death [RR

USA vs. Sweden, 1.14 (95% CI 1.04–1.26)] and hospitalized bleeding [RR USA vs. Sweden, 1.54 (1.21–1.96)]. Similar analyses were performed for white blood cell (WBC) comparing all-cause mortality in England and New Zealand (NZ) [58,59]. High WBC within the reference range (8.65–10.05×10^9/L) was associated with significantly increased mortality compared to the middle quintile (6.25–7.25×10^9/L); adjusted HR 1.51 (95% CI 1.43-1.59) [England], 1.33 (95% CI 1.06-1.65) [NZ].

DISCUSSION

We have demonstrated the strengths and challenges of phenotyping national UK EHR data using three exemplars (HF, AMI, bleeding) and have exemplified the UK's national EHR phenomics platform. Creating nationally-applicable phenotypes leveraging multiple EHR sources has, until recently, been a challenging, time-consuming, unreplicable and somewhat opaque process without any centralised resources. In this manuscript, we presented the CALIBER platform providing a framework for using national EHR from primary and secondary health care, disease and mortality registries. CALIBER is analogous and complementary to other leading initiatives, e.g. eMERGE, in that it ensures best practice and reproducibility for creating and validating EHR-derived phenotypes [60,61]. In contrast with eMERGE however, which uses secondary care data (higher proportion of disease), CALIBER exploits primary care EHR which contain healthy and ill individuals. Importantly, the approaches described here are potentially scalable/adaptable to the entire 65M population.

CALIBER phenotyping algorithms use structured information on diagnoses, symptoms, referrals, prescriptions and procedures which are recorded using five controlled clinical terminologies and continuous measurements in order to extract disease onset and severity. Complementary initiatives exist [19] but are different from CALIBER as they focus on curating codelists. We have taken a different approach and define an EHR phenotype as a combination of three essential elements: controlled clinical terminology terms, implementation logic and validation evidence all of which are curated on the CALIBER Portal. Compared with PheKB, CALIBER contains additional approaches of validation e.g. aetiological and prognostic across population samples but lacks detailed PPV/NPV measurements which are made possible by the availability of text and access to case notes in the US.

Validation **(Table 2)** is a critical step for assessing the accuracy of EHR-derived phenotypes. Whether a phenotype is valid or not however depends on the use-case i.e. EHR phenotypes used in trials requiring regulatory approval require a different level of validation than algorithms used in observational studies. Furthermore, validation is not a finite task but a constantly evolving process due to the underlying evolving nature of EHR, local feedback loops and data generation processes, and our understanding of disease [62]. Clinician case note review is considered the "gold standard" of phenotype validation and allows PPV/NPV

calculations for quantifying the accuracy which in turn can be curated (e.g. PheKB). Due to the limited availability of text in the UK and to address the spectrum of validation views, we developed an approach which captures six levels of evidence. CALIBER EHR phenotypes have at least one evidence layer with some phenotypes having multiple e.g. HF has aetiological, prognostic, cross-EHR source evidence. Prognostic validation i.e. survival patterns (crucial for disease progression/severity phenotypes is enabled by the longitudinal nature of EHR sources and the linkage with national mortality data.

In terms of the complete hospital interaction, HES data are a snapshot of the patient journey as data are collected for reimbursement [8,63]. Multiple initiatives exist for obtaining hospital data for research e.g. Health Informatics Collaborative which links eleven ICUs in five hospitals research (>18,000 patients, >21,000 admissions, median 1,030 time-varying measures [64]). These initiatives will enable researchers to have access to hospital data, including free text, for creating and validating phenotypes. Importantly, they will create a feedback loop with clinical care that will provide detailed information on the healthcare processes generating the data (critical for phenotyping) and drive data standardization and quality.

Through CALIBER we provide a framework for the consistent definition, use and re-use of EHR derived phenotypes from national UK EHR sources. One of the primary audiences of CALIBER phenotypes is international: US investigators account for a third of studies using UK primary care EHR [18], two thirds of UK Biobank studies are from US investigators. Additionally, the controlled clinical terminologies used in UK EHR are transferable to the US e.g. Read terms are part of SNOMED-CT, ICD-9-CM-ICD-10 mappings. As PheKB and other initiatives evolve, there may a need for a connection across national portals [65] and infrastructure enabling cross-biobank analyses of complex/rare phenotypes [7].

CALIBER phenotyping algorithms are rule-based, deterministic and have provided the groundwork on which future phenotypes can be created. While this approach yields robust and accurate algorithms, it does not scale with our ambition to create thousands of high-quality phenotypes that capture the entire human phenome. To do this, research is required on high-throughput phenotyping involving supervised and unsupervised learning approaches and natural language processing[66]. These methods can generate thousands of phenotypes and discover hidden associations within clinical data in a fraction of the current cost and time requirements and with minimal human intervention. Robust approaches for classifying phenotype complexity are required to ensure proportional resourcing for phenotyping [67].

The translation from phenotype definition to SQL for data extraction was manual due to the lack of an established storage format [68]. Human interpretation is potentially error-prone and we are exploring ways for storing phenotype definitions in a computable manner that can enable automatic phenotyping. We have evaluated the use of Semantic Web Technologies and openEHR for curating computable phenotyping

algorithms [69,70]. Given that the majority of UK primary care EHR data are hosted by four vendors, there is a real opportunity to create computable phenotypes which can be utilized across the NHS [71]. To do this, information exchange standards (e.g. FHIR[72]) have to be utilized and combined with approaches such as PHEMA[73] and MAT [74].

The creation of a national phenomics platform through CALIBER provides an opportunity for the UK EHR community to interact, nationally and internationally, and connects data producers and consumers. Researchers can deposit phenotyping algorithms in the Portal for others to download, refine and use. EHR users i.e. clinicians can view these algorithms and understand how the data they record is being used for research.

CONCLUSION

We have demonstrated a reproducible approach for phenotyping national UK EHR data and validating using multiple approaches of evidence. The CALIBER platform catalyzes UK EHR expertise and is aligned with similar national and international research initiatives.

FIGURES

Figure 1: The CALIBER platform (https://www.caliberresearch.org) links national structured electronic health records (EHR) across primary care, secondary care and mortality for research. EHR-derived phenotypes are created using an iterative methodology and six independent approaches of evidence are generated to assess algorithm accuracy. More than 50 phenotypes are published in an open-access resource, the CALIBER Portal (https://www.caliberresearch.org/portal), and are used in >60 publications.

Figure 2: Assessing the recording and concordance of three electronic health record (EHR)-derived phenotypes (heart failure, non-fatal acute myocardial infarction (AMI) and bleeding) across three EHR data sources: primary care (CPRD), hospital care (HES) and mortality (ONS) or disease registry data (MINAP). Only a very small proportion (9% for heart failure, 31% for AMI and <1% for bleeding) of cases are identified concurrently by all three data sources.

Figure 3: Risk factors for initial presentation of heart failure (HF) phenotype: Hazard Ratios (95% CIs) of smoking status, Type-II diabetes, systolic blood pressure and heart rate based on previously-published CALIBER studies [29,75,76] compared to estimates obtained from investigator-led studies derived using

manually curated research data [77–80]. All individual analyses have been adjusted for age and sex and other covariates.

Figure 4: CALIBER Portal entry for the heart failure phenotype (available at https://www.caliberresearch.org/portal/phenotypes/heartfailure). Each entry in the Portal contains implementation details on the logic and the terms from controlled clinical terminologies associated with the phenotyping algorithm. Additionally, the six approaches of validation evidence are presented and the research output that has used the phenotype is provided.

TABLES

Table 1: Overview of published, peer-reviewed EHR phenotypes derived from the CALIBER platform and the approaches of validation evidence - More information available on the CALIBER Portal https://www.caliberresearch.org/portal/phenotypes.

Phenotype	**EHR Data Sources**			**Validation evidence**					
	Primary care	**Secondary care**	**Death**	**Cross source**	**Case note review**	**Prognosis**	**Aetiology**	**Genetic**	**Cross country**
Disease or syndrome									
AAA	●	●	●	●		●	●		
AMI	●	●	●	●		●	●	●	●
AD	●	●	●	●		●			
AF	●	●	●	●		●	●	●	
Uveitis	●	●		●					
Bleeding	●	●	●	●	●	●	●		●
Bullous disorder	●	●		●		●			
CHD	●	●		●		●	●		
Depression	●	●		●		●			
Diabetes	●	●		●		●			
Giant cells	●	●		●		●			
HF	●	●	●	●		●	●		
HIV	●	●	●	●		●			
Hypertension	●	●		●		●	●		
HCM	●	●		●		●			
Influenza	●					●			
MS	●	●		●		●			
PAD	●	●	●	●		●	●		

Polymyalgia	●	●		●		●			
PBC	●	●		●		●			
Psoriasis	●	●		●		●			
Dementia NOS	●	●	●	●		●			
RA	●	●		●		●			
SA	●	●		●		●	●		
Intracerebral haem.	●	●	●	●		●	●		
Ischaemic stroke	●	●	●	●		●	●		
SAH	●	●	●	●		●	●		
Stroke NOS	●	●	●	●		●	●		
SCD	●	●	●	●		●	●		
Systemic sclerosis	●	●		●		●			
TIA	●	●	●	●		●	●		
UCD	●	●	●			●	●		
UA	●	●		●		●	●		
Vascular dementia	●	●	●	●		●			
Obesity	●	●		●		●			
Biomarkers									
Blood pressure	●					●			
Eosinophils	●					●			
Heart rate	●					●			
Lymphocytes	●					●			
Neutrophils	●					●			
White blood cells	●					●			●
LDL Cholesterol	●					●			
HDL Cholesterol	●					●			
Triglycerides	●					●			
BMI	●	●				●			
Lifestyle risk factors and other									
Alcohol	●					●			
Ethnicity	●	●				●			
Pregnancy	●	●				●			
Sex	●					●			
Smoking	●					●			
Deprivation	●					●			

AAA Abdominal Aortic Aneurysm; AMI Acute Myocardial Infarction; AD Alzheimer's Disease; AF Atrial Fibrillation; CHD Coronary Heart Disease; HF Heart Failure; HIV Human Immunodeficiency Virus; HCM Hypertrophic Cardiomyopathy; MS Multiple Sclerosis; PAD Peripheral Arterial Disease; PBC Primary Biliary Cirrhosis; NOS Not Otherwise Specified; RA Rheumatoid Arthritis; SA Stable Angina; SAH Subarachnoid Haemorrhage; SCD Sudden Cardiac Death; TIA Transient Ischaemic Attack; UCD Unheralded Coronary Death; UA Unstable Angina; LDL Low Density Lipoprotein; HDL High Density Lipoprotein; BMI Body Mass Index

Table 2: Systematic validation of the CALIBER EHR-derived phenotypes for a) heart failure, b) acute myocardial infarction, and c) bleeding across six approaches of evidence: cross-EHR concordance, case-note review, aetiology, prognosis, genetic associations and external populations.

Validation domain	**Description**	**What has been done**		
		Heart failure	**Acute Myocardial Infarction**	**Bleeding**
Cross EHR source concordance	To what extent is the phenotype concordant across EHR sources?	The proportion of HF cases recorded in primary care and hospital care EHR was 27% [28].	The proportion of non-fatal AMI defined across primary care, hospital care and disease registry was 32% [26].	The proportion of bleeding events recorded in primary care and hospital care was 12% with 47% of bleeding events recorded only in primary care and 12% only in hospital care.
Case-note review	What is the positive predictive value and the negative predictive value when comparing the algorithm with clinician review of case notes or "gold standard" source of information?		Compared with AMI defined in the disease registry, the PPV of AMI recorded in primary care was 92.2% (95% CIal 91.6% to 92.8%) and in hospital admissions was 91.5% (90.8% to 92.1%) [26]..	Compared through independent review by two clinicians, the PPV of bleeding events identified through the phenotyping algorithm was 0.88.
Aetiology	Are the prospective associations with risk actors consistent with previous evidence?	Type 2 diabetes [76], systolic/diastolic blood pressure [29], heart rate	Type 2 diabetes [76], systolic/diastolic blood pressure [29], heart rate	At 5 years 29.1% (95% CI: 28.2, 29.9%) of atrial fibrillation patients, 21.9% (21.2, 22.5%)

		[81], socioeconomic deprivation [82], alcohol consumption [83], smoking [75], ethnicity [40], acute myocardial infarction [26], depression [84], neutrophil counts [85] , eosinophil/lymphocyte counts[86], atrial fibrillation [27], sex [87]	[81], socioeconomic deprivation [82], alcohol consumption [83], smoking [75], ethnicity [40], acute myocardial infarction [26], depression [84], neutrophil counts [85] , eosinophil/lymphocyte counts[86], atrial fibrillation [27], influenza infection [88], ischaemic presentations [89], sex [87]	of myocardial infarction patients, 25.3% (24.2, 26.3%) of unstable angina patients and 23.4% (23.0, 23.8%) of stable angina had bleeding of any kind
Prognosis	Are the risks of subsequent events plausible?	Corrected for age and sex, HF was strongly associated with mortality, with HRs for all□cause mortality ranging from 7.01 (95% CI 6.83–7.20), 7.23 (95% CI 7.03–7.43), up to 15.38 (95% CI 15.02–15.83) for patients in primary care with acute HF hospitalization, primary care only, and patients hospitalized but no PC record [28].	Patients with myocardial infarction identified in the disease registry had lower crude 30 day mortality (10.8%, 95% CI 10.2% to 11.4%) than those identified in hospital care (13.9%, 13.3% to 14.4%) or in primary care (14.9%, 14.4% to 15.5%, fig 2□). At one year, however, mortality was similar in all three groups, at around 20% [26]. Of the 24 479 patients with AMI, 5775 (23.6%) developed HF during a median follow-up of 3.7 years (incidence rate per 1000 person-years: 63.8, 95% CI 62.2 to	The hazard ratio for all-cause mortality was 1.98 (95% CI: 1.86, 2.11) for primary care bleeding with markers of severity, and 1.99 (95% CI: 1.92, 2.05) for hospitalised bleeding without markers of severity, compared to patients with no bleeding.

			65.5) [90]	
Genetic associations	Are the observed genetic associations plausible and concordance with previous evidence?		Consistent direction and magnitude of associations were replicated in 67 (97%) of previously reported genetic variants [4].	
External populations	Has the algorithm been tested (in any of the above validation domains) in different countries?		We observed high 3-year crude cumulative risks of all-cause death (from 19.6% [England] to 30.2% [USA]); the composite of MI, stroke, or death [from 26.0% (France) to 36.2% (USA)]; and hospitalized bleeding [from 3.1% (France) to 5.3% (USA)]. [57]	

ACKNOWELDGEMENTS

This study was approved by the Medicines and Healthcare Products Regulatory Agency (MHRA) Independent Scientific Advisory Committee (ISAC) - protocol references: 11_088, 12_153R, 16_221, 18_029R2, 18_159R.

This study is based in part on data from the Clinical Practice Research Datalink obtained under licence from the UK Medicines and Healthcare products Regulatory Agency. The data is provided by patients and collected by the NHS as part of their care and support. The interpretation and conclusions contained in this study are those of the author/s alone.

Hospital Episode Statistics Copyright © (2019), re-used with the permission of The Health & Social Care Information Centre. All rights reserved.

The OPCS Classification of Interventions and Procedures, codes, terms and text is Crown copyright (2016) published by Health and Social Care Information Centre, also known as NHS Digital and licensed under the Open Government Licence available at www.nationalarchives.gov.uk/doc/open-government-licence/open-government-licence.htm

This study was carried out as part of the CALIBER programme (https://www.ucl.ac.uk/health-informatics/caliber). CALIBER, led from the UCL Institute of Health Informatics, is a research resource consisting of linked electronic health records phenotypes, methods and tools, specialised infrastructure, and training and support.

The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

FUNDING

This study is part of the BigData@Heart programme that has received funding from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement no. 116074. This Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA.
This work was supported by Health Data Research UK, which receives its funding from HDR UK Ltd (NIWA1) funded by the UK Medical Research Council, Engineering and Physical Sciences Research Council, Economic and Social Research Council, Department of Health and Social Care (England), Chief Scientist Office of the Scottish Government Health and Social Care Directorates, Health and Social Care Research and Development Division (Welsh Government), Public Health Agency (Northern Ireland), British Heart Foundation (BHF) and the Wellcome Trust.
This study was supported by National Institute for Health Research (RP-PG-0407-10314), Wellcome Trust (086091/Z/08/Z).
This study was supported by the Farr Institute of Health Informatics Research at UCL Partners, from the Medical Research Council, Arthritis Research UK, British Heart Foundation, Cancer Research UK, Chief Scientist Office, Economic and Social Research Council, Engineering and Physical Sciences Research Council, National Institute for Health Research, National Institute for Social Care and Health Research, and Wellcome Trust (MR/K006584/1).
This paper represents independent research part funded (AGI, KD, NF) by the National Institute for Health Research (NIHR) Biomedical Research Centre at University College London Hospitals.
HH is an NIHR Senior Investigator.
GF is supported by an American Heart Association Fellowship [17IF33890001].

COMPETING INTERESTS

None.

REFERENCES

1. Williams T, van Staa T, Puri S, Eaton S. Recent advances in the utility and use of the General Practice Research Database as an example of a UK Primary Care Data resource. Ther Adv Drug Saf. journals.sagepub.com; 2012;3: 89–99. doi:10.1177/2042098611435911

2. Ludwick DA, Doucette J. Adopting electronic medical records in primary care: lessons learned from health information systems implementation experience in seven countries. Int J Med Inform. 2009;78: 22–31. doi:10.1016/j.ijmedinf.2008.06.005

3. Turnbull C, Scott RH, Thomas E, Jones L, Murugaesu N, Pretty FB, et al. The 100 000 Genomes Project: bringing whole genome sequencing to the NHS. BMJ. 2018;361: k1687. doi:10.1136/bmj.k1687

4. Denaxas SC, Fatemifar G, Patel R, Hemingway H. Deriving research-quality phenotypes from national electronic health records to advance precision medicine: a UK Biobank case-study. Proceedings of the BHI-2017 International Conference on Biomedical and Health Informatics. Orlando, FL, USA: IEEE Engineering in Medicine and Biology Society (EMBS); 2017. Available: http://discovery.ucl.ac.uk/1542412/

5. Schnier C, Denaxas S, Eggo R, Patel R, Zhang Q, Woodfield R, et al. Identification and validation of myocardial infarction and stroke outcomes at scale in UK Biobank. IJPDS. 2017;1. doi:10.23889/ijpds.v1i1.358

6. Wei W-Q, Denny JC. Extracting research-quality phenotypes from electronic health records to support precision medicine. Genome Med. 2015;7: 41. doi:10.1186/s13073-015-0166-y

7. Casey JA, Schwartz BS, Stewart WF, Adler NE. Using Electronic Health Records for Population Health Research: A Review of Methods and Applications. Annu Rev Public Health. 2016;37: 61–81. doi:10.1146/annurev-publhealth-032315-021353

8. Denaxas SC, Asselbergs FW, Moore JH. The tip of the iceberg: challenges of accessing hospital electronic health record data for biological data mining. BioData Min. 2016;9: 29. doi:10.1186/s13040-016-0109-1

9. Hemingway H, Asselbergs FW, Danesh J, Dobson R, Maniadakis N, Maggioni A, et al. Big data from electronic

health records for early and late translational cardiovascular research: challenges and potential. Eur Heart J. 2018;39: 1481–1495. doi:10.1093/eurheartj/ehx487

10. Gottesman O, Kuivaniemi H, Tromp G, Faucett WA, Li R, Manolio TA, et al. The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. Genet Med. 2013;15: 761–771. doi:10.1038/gim.2013.72

11. 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;84: 362–369. doi:10.1038/clpt.2008.89

12. Gaziano JM, Concato J, Brophy M, Fiore L, Pyarajan S, Breeling J, et al. Million Veteran Program: A mega-biobank to study genetic influences on health and disease. J Clin Epidemiol. 2016;70: 214–223. doi:10.1016/j.jclinepi.2015.09.016

13. Collins FS, Varmus H. A new initiative on precision medicine. N Engl J Med. 2015;372: 793–795. doi:10.1056/NEJMp1500523

14. Doiron D, Raina P, Fortier I, Linkage Between Cohorts and Health Care Utilization Data: Meeting of Canadian Stakeholders workshop participants. Linking Canadian population health data: maximizing the potential of cohort and administrative data. Can J Public Health. 2013;104: e258–61. Available: https://www.ncbi.nlm.nih.gov/pubmed/23823892

15. Holman CDJ, Bass AJ, Rosman DL, Smith MB, Semmens JB, Glasson EJ, et al. A decade of data linkage in Western Australia: strategic design, applications and benefits of the WA data linkage system. Aust Health Rev. CSIRO; 2008;32: 766–777. Available: https://www.ncbi.nlm.nih.gov/pubmed/18980573

16. Jernberg T, Attebring MF, Hambraeus K, Ivert T, James S, Jeppsson A, et al. The Swedish Web-system for enhancement and development of evidence-based care in heart disease evaluated according to recommended therapies (SWEDEHEART). Heart. 2010;96: 1617–1621. doi:10.1136/hrt.2010.198804

17. Jensen AB, Moseley PL, Oprea TI, Ellesøe SG, Eriksson R, Schmock H, et al. Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients. Nat Commun. 2014;5: 4022. doi:10.1038/ncomms5022

18. Vezyridis P, Timmons S. Evolution of primary care databases in UK: a scientometric analysis of research output. BMJ Open. 2016;6: e012785. doi:10.1136/bmjopen-2016-012785

19. Springate DA, Kontopantelis E, Ashcroft DM, Olier I, Parisi R, Chamapiwa E, et al. ClinicalCodes: an online clinical codes repository to improve the validity and reproducibility of research using electronic medical records. PLoS One. 2014;9: e99825. doi:10.1371/journal.pone.0099825

20. Al Sallakh MA, Vasileiou E, Rodgers SE, Lyons RA, Sheikh A, Davies GA. Defining asthma and assessing asthma outcomes using electronic health record data: a systematic scoping review. Eur Respir J. 2017;49. doi:10.1183/13993003.00204-2017

21. Lyons RA, Jones KH, John G, Brooks CJ, Verplancke J-P, Ford DV, et al. The SAIL databank: linking multiple health and social care datasets. BMC Med Inform Decis Mak. bmcmedinformdecismak …; 2009;9: 3. doi:10.1186/1472-6947-9-3

22. Ford DV, Jones KH, Verplancke J-P, Lyons RA, John G, Brown G, et al. The SAIL Databank: building a national architecture for e-health research and evaluation. BMC Health Serv Res. bmchealthservres.biomedcentral …; 2009;9: 157. doi:10.1186/1472-6963-9-157

23. Newton KM, Peissig PL, Kho AN, Bielinski SJ, Berg RL, Choudhary V, 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. doi:10.1136/amiajnl-2012-000896

24. Banda JM, Seneviratne M, Hernandez-Boussard T, Shah NH. Advances in Electronic Phenotyping: From Rule-Based Definitions to Machine Learning Models. Annu Rev Biomed Data Sci. Annual Reviews; 2018;

doi:10.1146/annurev-biodatasci-080917-013315

25. Boggon R, van Staa TP, Chapman M, Gallagher AM, Hammad TA, Richards MA. Cancer recording and mortality in the General Practice Research Database and linked cancer registries. Pharmacoepidemiol Drug Saf. 2013;22: 168–175. doi:10.1002/pds.3374

26. Herrett E, Shah AD, Boggon R, Denaxas S, Smeeth L, van Staa T, et al. Completeness and diagnostic validity of recording acute myocardial infarction events in primary care, hospital care, disease registry, and national mortality records: cohort study. BMJ. 2013;346: f2350. doi:10.1136/bmj.f2350

27. Morley KI, Wallace J, Denaxas SC, Hunter RJ, Patel RS, Perel P, et al. Defining disease phenotypes using national linked electronic health records: a case study of atrial fibrillation. PLoS One. journals.plos.org; 2014;9: e110900. doi:10.1371/journal.pone.0110900

28. Koudstaal S, Pujades-Rodriguez M, Denaxas S, Gho JM, Shah AD, Yu N, et al. Prognostic burden of heart failure recorded in primary care, acute hospital admissions, or both: a population-based linked electronic health record cohort study in 2.1 million people. Eur J Heart Fail. Wiley Online Library; 2017;19: 1119–1127. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/ejhf.709

29. Rapsomaniki E, Timmis A, George J, Pujades-Rodriguez M, Shah AD, Denaxas S, et al. Blood pressure and incidence of twelve cardiovascular diseases: lifetime risks, healthy life-years lost, and age-specific associations in 1·25 million people. Lancet. 2014;383: 1899–1911. doi:10.1016/S0140-6736(14)60685-1

30. Wu H, Toti G, Morley KI, Ibrahim ZM, Folarin A, Jackson R, et al. SemEHR: A general-purpose semantic search system to surface semantic data from clinical notes for tailored care, trial recruitment, and clinical research. J Am Med Inform Assoc. 2018;25: 530–537. doi:10.1093/jamia/ocx160

31. Rubbo B, Fitzpatrick NK, Denaxas S, Daskalopoulou M, Yu N, Patel RS, et al. Use of electronic health records to ascertain, validate and phenotype acute myocardial infarction: A systematic review and recommendations. Int J Cardiol. 2015;187: 705–711. doi:10.1016/j.ijcard.2015.03.075

32. Organization WH, Others. ICD-10: The ICD-10 Classification of Mental and Behavioural Disorders: diagnostic criteria for research. ICD-10: the ICD-10 classification of mental and behavioural disorders: diagnostic criteria for research. 1993.

33. Jarvis B, Johnson T, Butler P, O'Shaughnessy K, Fullam F, Tran L, et al. Assessing the impact of electronic health records as an enabler of hospital quality and patient satisfaction. Acad Med. 2013;88: 1471–1477. doi:10.1097/ACM.0b013e3182a36cab

34. Ford E, Carroll JA, Smith HE, Scott D, Cassell JA. Extracting information from the text of electronic medical records to improve case detection: a systematic review. J Am Med Inform Assoc. 2016;23: 1007–1015. doi:10.1093/jamia/ocv180

35. Denaxas SC, George J, Herrett E, Shah AD, Kalra D, Hingorani AD, et al. Data resource profile: cardiovascular disease research using linked bespoke studies and electronic health records (CALIBER). Int J Epidemiol. 2012;41: 1625–1638. doi:10.1093/ije/dys188

36. Gallagher AM, Puri S, van Staa TP. 528. Linkage of the General Practice Research Database (gprd) with Other Data Sources. Pharmacoepidemiol Drug Saf. Pharmacoepidemiology and Drug Safety; 2011;20: 230–231. Available: https://insights.ovid.com/pharmacoepidemiology-drug-safety/pads/2011/08/001/528-linkage-general-practice-research-database/528/01445491

37. Herrett E, Gallagher AM, Bhaskaran K, Forbes H, Mathur R, van Staa T, et al. Data Resource Profile: Clinical Practice Research Datalink (CPRD). Int J Epidemiol. 2015;44: 827–836. doi:10.1093/ije/dyv098

38. O'Neil M, Payne C, Read J. Read Codes Version 3: A User Led Terminology. Methods Inf Med. Schattauer GmbH; 2018;34: 187–192. doi:10.1055/s-0038-1634585

39. Datta-Nemdharry P, Thomson A, Beynon J. Opportunities and Challenges in Developing a Cohort of Patients with Type 2 Diabetes Mellitus Using Electronic Primary Care Data. PLoS One. 2016;11: e0162236.

doi:10.1371/journal.pone.0162236

40. George J, Mathur R, Shah AD, Pujades-Rodriguez M, Denaxas S, Smeeth L, et al. Ethnicity and the first diagnosis of a wide range of cardiovascular diseases: Associations in a linked electronic health record cohort of 1 million patients. PLoS One. 2017;12: e0178945. doi:10.1371/journal.pone.0178945

41. Bhaskaran K, Forbes HJ, Douglas I, Leon DA, Smeeth L. Representativeness and optimal use of body mass index (BMI) in the UK Clinical Practice Research Datalink (CPRD). BMJ Open. 2013;3: e003389. doi:10.1136/bmjopen-2013-003389

42. Mathur R, Bhaskaran K, Chaturvedi N, Leon DA, vanStaa T, Grundy E, et al. Completeness and usability of ethnicity data in UK-based primary care and hospital databases. J Public Health . 2014;36: 684–692. doi:10.1093/pubmed/fdt116

43. Herrett E, Thomas SL, Schoonen WM, Smeeth L, Hall AJ. Validation and validity of diagnoses in the General Practice Research Database: a systematic review. Br J Clin Pharmacol. 2010;69: 4–14. doi:10.1111/j.1365-2125.2009.03537.x

44. Association AM. Current procedural terminology: CPT. American Medical Association; 2007.

45. Herrett E, Smeeth L, Walker L, Weston C, MINAP Academic Group. The Myocardial Ischaemia National Audit Project (MINAP). Heart. heart.bmj.com; 2010;96: 1264–1267. doi:10.1136/hrt.2009.192328

46. Jordan H, Roderick P, Martin D. The Index of Multiple Deprivation 2000 and accessibility effects on health. J Epidemiol Community Health. jech.bmj.com; 2004;58: 250–257. Available: https://www.ncbi.nlm.nih.gov/pubmed/14966241

47. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp. ncbi.nlm.nih.gov; 2001; 17–21. Available: https://www.ncbi.nlm.nih.gov/pubmed/11825149

48. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res. academic.oup.com; 2004;32: D267–70. doi:10.1093/nar/gkh061

49. Jackson R, Kartoglu I, Stringer C, Gorrell G, Roberts A, Song X, et al. CogStack - experiences of deploying integrated information retrieval and extraction services in a large National Health Service Foundation Trust hospital. BMC Med Inform Decis Mak. bmcmedinformdecismak …; 2018;18: 47. doi:10.1186/s12911-018-0623-9

50. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. academic.oup.com; 2014;42: D1001–6. doi:10.1093/nar/gkt1229

51. Nikpay M, Goel A, Won H-H, Hall LM, Willenborg C, Kanoni S, et al. A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease. Nat Genet. 2015;47: 1121–1130. doi:10.1038/ng.3396

52. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. Elsevier; 2007;81: 559–575. doi:10.1086/519795

53. Pruim RJ, Welch RP, Sanna S, Teslovich TM, Chines PS, Gliedt TP, et al. LocusZoom: regional visualization of genome-wide association scan results. Bioinformatics. academic.oup.com; 2010;26: 2336–2337. doi:10.1093/bioinformatics/btq419

54. Kraker P, Lex E, Gorraiz J, Gumpenberger C, Peters I. Research Data Explored II: the Anatomy and Reception of figshare [Internet]. arXiv [cs.DL]. 2015. Available: http://arxiv.org/abs/1503.01298

55. UCL. CALIBER publications. In: UCL Institute of Health Informatics [Internet]. 26 Apr 2018 [cited 21 Jan 2019]. Available: https://www.ucl.ac.uk/health-informatics/caliber/publications

56. Li L, Geraghty OC, Mehta Z, Rothwell PM, Oxford Vascular Study. Age-specific risks, severity, time course, and outcome of bleeding on long-term antiplatelet treatment after vascular events: a population-based cohort study. Lancet. Elsevier; 2017;390: 490–499. doi:10.1016/S0140-6736(17)30770-5

57. Rapsomaniki E, Thuresson M, Yang E, Blin P, Hunt P, Chung S-C, et al. Using big data from health records from four countries to evaluate chronic disease outcomes: a study in 114 364 survivors of myocardial infarction. Eur Heart J Qual Care Clin Outcomes. 2016;2: 172–183. doi:10.1093/ehjqcco/qcw004

58. Pylypchuk R, Wells S, Kerr A, Poppe K, Riddell T, Harwood M, et al. Cardiovascular disease risk prediction equations in 400 000 primary care patients in New Zealand: a derivation and validation study. Lancet. 2018;391: 1897–1907. doi:10.1016/S0140-6736(18)30664-0

59. Shah AD, Thornley S, Chung S-C, Denaxas S, Jackson R, Hemingway H. White cell count in the normal range and short-term and long-term mortality: international comparisons of electronic health record cohorts in England and New Zealand. BMJ Open. 2017;7: e013100. doi:10.1136/bmjopen-2016-013100

60. Denaxas S, Direk K, Gonzalez-Izquierdo A, Pikoula M, Cakiroglu A, Moore J, et al. Methods for enhancing the reproducibility of biomedical research findings using electronic health records. BioData Min. biodatamining.biomedcentral.com; 2017;10: 31. doi:10.1186/s13040-017-0151-7

61. Denaxas S, Gonzalez-Izquierdo A, Pikoula M, Direk K, Fitzpatrick N, Hemingway H, et al. Methods for Enhancing the Reproducibility of Observational Research Using Electronic Health Records: Preliminary Findings from the CALIBER Resource. 2017 IEEE 30th International Symposium on Computer-Based Medical Systems (CBMS). ieeexplore.ieee.org; 2017. pp. 506–508. doi:10.1109/CBMS.2017.74

62. Denaxas SC, Morley KI. Big biomedical data and cardiovascular disease research: opportunities and challenges. Eur Heart J Qual Care Clin Outcomes. 2015;1: 9–16. doi:10.1093/ehjqcco/qcv005

63. Farrar S, Yi D, Sutton M, Chalkley M, Sussex J, Scott A. Has payment by results affected the way that English hospitals provide care? Difference-in-differences analysis. BMJ. bmj.com; 2009;339: b3047. doi:10.1136/bmj.b3047

64. Harris S, Shi S, Brealey D, MacCallum NS, Denaxas S, Perez-Suarez D, et al. Critical Care Health Informatics Collaborative (CCHIC): Data, tools and methods for reproducible research: A multi-centre UK intensive care database. Int J Med Inform. Elsevier; 2018;112: 82–89. doi:10.1016/j.ijmedinf.2018.01.006

65. Kirby JC, Speltz P, Rasmussen LV, Basford M, Gottesman O, Peissig PL, et al. PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability. J Am Med Inform Assoc. 2016;23: 1046–1052. doi:10.1093/jamia/ocv202

66. Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. J Am Med Inform Assoc. 2013;20: 117–121. doi:10.1136/amiajnl-2012-001145

67. Hripcsak G, Albers DJ. High-fidelity phenotyping: richness and freedom from bias. J Am Med Inform Assoc. 2017; doi:10.1093/jamia/ocx110

68. Xu J, Rasmussen LV, Shaw PL, Jiang G, Kiefer RC, Mo H, et al. Review and evaluation of electronic health records-driven phenotype algorithm authoring tools for clinical and translational research. J Am Med Inform Assoc. academic.oup.com; 2015;22: 1251–1260. doi:10.1093/jamia/ocv070

69. Papez V, Denaxas S, Hemingway H. Evaluating OpenEHR for Storing Computable Representations of Electronic Health Record Phenotyping Algorithms. 2017 IEEE 30th International Symposium on Computer-Based Medical Systems (CBMS). IEEE Computer Society; 2017. pp. 509–514. doi:10.1109/CBMS.2017.73

70. Papež V, Denaxas S, Hemingway H. Evaluation of Semantic Web Technologies for Storing Computable Definitions of Electronic Health Records Phenotyping Algorithms. AMIA Annu Symp Proc. 2017;2017: 1352–1361. Available: https://www.ncbi.nlm.nih.gov/pubmed/29854204

71. Mo H, Thompson WK, Rasmussen LV. Desiderata for computable representations of electronic health records-driven phenotype algorithms. Journal of the. academic.oup.com; 2015; Available:

https://academic.oup.com/jamia/article-abstract/22/6/1220/2357938

72. Bender D, Sartipi K. HL7 FHIR: An Agile and RESTful approach to healthcare information exchange. Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems. ieeexplore.ieee.org; 2013. pp. 326–331. doi:10.1109/CBMS.2013.6627810

73. Jiang G, Kiefer RC, Rasmussen LV, Solbrig HR, Mo H, Pacheco JA, et al. Developing a data element repository to support EHR-driven phenotype algorithm authoring and execution. J Biomed Inform. 2016;62: 232–242. doi:10.1016/j.jbi.2016.07.008

74. Thompson WK, Rasmussen LV, Pacheco JA, Peissig PL, Denny JC, Kho AN, et al. An evaluation of the NQF Quality Data Model for representing Electronic Health Record driven phenotyping algorithms. AMIA Annu Symp Proc. ncbi.nlm.nih.gov; 2012;2012: 911–920. Available: https://www.ncbi.nlm.nih.gov/pubmed/23304366

75. Pujades-Rodriguez M, George J, Shah AD, Rapsomaniki E, Denaxas S, West R, et al. Heterogeneous associations between smoking and a wide range of initial presentations of cardiovascular disease in 1937360 people in England: lifetime risks and implications for risk prediction. Int J Epidemiol. 2015;44: 129–141. doi:10.1093/ije/dyu218

76. Shah AD, Langenberg C, Rapsomaniki E, Denaxas S, Pujades-Rodriguez M, Gale CP, et al. Type 2 diabetes and incidence of cardiovascular diseases: a cohort study in 1·9 million people. The Lancet Diabetes & Endocrinology. 2015;3: 105–113. doi:10.1016/S2213-8587(14)70219-0

77. de Simone G, Devereux RB, Chinali M, Lee ET, Galloway JM, Barac A, et al. Diabetes and incident heart failure in hypertensive and normotensive participants of the Strong Heart Study. J Hypertens. 2010;28: 353–360. doi:10.1097/HJH.0b013e3283331169

78. Gopal DM, Kalogeropoulos AP, Georgiopoulou VV, Smith AL, Bauer DC, Newman AB, et al. Cigarette smoking exposure and heart failure risk in older adults: the Health, Aging, and Body Composition Study. Am Heart J. 2012;164: 236–242. doi:10.1016/j.ahj.2012.05.013

79. Khan H, Kunutsor S, Kalogeropoulos AP, Georgiopoulou VV, Newman AB, Harris TB, et al. Resting heart rate and risk of incident heart failure: three prospective cohort studies and a systematic meta-analysis. J Am Heart Assoc. 2015;4: e001364. doi:10.1161/JAHA.114.001364

80. Haider AW, Larson MG, Franklin SS, Levy D, Framingham Heart Study. Systolic blood pressure, diastolic blood pressure, and pulse pressure as predictors of risk for congestive heart failure in the Framingham Heart Study. Ann Intern Med. 2003;138: 10–16. Available: https://www.ncbi.nlm.nih.gov/pubmed/12513039

81. Archangelidi O, Pujades-Rodriguez M, Timmis A, Jouven X, Denaxas S, Hemingway H. Clinically recorded heart rate and incidence of 12 coronary, cardiac, cerebrovascular and peripheral arterial diseases in 233,970 men and women: A linked electronic health record study. Eur J Prev Cardiol. 2018;25: 1485–1495. doi:10.1177/2047487318785228

82. Pujades-Rodriguez M, Timmis A, Stogiannis D, Rapsomaniki E, Denaxas S, Shah A, et al. Socioeconomic deprivation and the incidence of 12 cardiovascular diseases in 1.9 million women and men: implications for risk prediction and prevention. PLoS One. journals.plos.org; 2014;9: e104671. doi:10.1371/journal.pone.0104671

83. Bell S, Daskalopoulou M, Rapsomaniki E, George J, Britton A, Bobak M, et al. Association between clinically recorded alcohol consumption and initial presentation of 12 cardiovascular diseases: population based cohort study using linked health records. BMJ. 2017;356: j909. doi:10.1136/bmj.j909

84. Daskalopoulou M, George J, Walters K, Osborn DP, Batty GD, Stogiannis D, et al. Depression as a Risk Factor for the Initial Presentation of Twelve Cardiac, Cerebrovascular, and Peripheral Arterial Diseases: Data Linkage Study of 1.9 Million Women and Men. PLoS One. 2016;11: e0153838. doi:10.1371/journal.pone.0153838

85. Shah AD, Denaxas S, Nicholas O, Hingorani AD, Hemingway H. Neutrophil Counts and Initial Presentation of 12 Cardiovascular Diseases: A CALIBER Cohort Study. J Am Coll Cardiol. 2017;69: 1160–1169. doi:10.1016/j.jacc.2016.12.022

86. Shah AD, Denaxas S, Nicholas O, Hingorani AD, Hemingway H. Low eosinophil and low lymphocyte counts and the incidence of 12 cardiovascular diseases: a CALIBER cohort study. Open Heart. 2016;3: e000477. doi:10.1136/openhrt-2016-000477

87. George J, Rapsomaniki E, Pujades-Rodriguez M, Shah AD, Denaxas S, Herrett E, et al. How Does Cardiovascular Disease First Present in Women and Men? Incidence of 12 Cardiovascular Diseases in a Contemporary Cohort of 1,937,360 People. Circulation. ncbi.nlm.nih.gov; 2015;132: 1320–1328. doi:10.1161/CIRCULATIONAHA.114.013797

88. Warren-Gash C, Hayward AC, Hemingway H, Denaxas S, Thomas SL, Timmis AD, et al. Influenza infection and risk of acute myocardial infarction in England and Wales: a CALIBER self-controlled case series study. J Infect Dis. 2012;206: 1652–1659. doi:10.1093/infdis/jis597

89. Herrett E, Bhaskaran K, Timmis A, Denaxas S, Hemingway H, Smeeth L. Association between clinical presentations before myocardial infarction and coronary mortality: a prospective population-based study using linked electronic records. Eur Heart J. 2014;35: 2363–2371. doi:10.1093/eurheartj/ehu286

90. Gho JMIH, Schmidt AF, Pasea L, Koudstaal S, Pujades-Rodriguez M, Denaxas S, et al. An electronic health records cohort study on heart failure following myocardial infarction in England: incidence and predictors. BMJ Open. 2018;8: e018331. doi:10.1136/bmjopen-2017-018331