UK phenomics platform for developing and validating EHR phenotypes: CALIBER

Spiros Denaxas [1,2,9,*]
Arturo Gonzalez-Izquierdo [1,2]
Kenan Direk [1,2]
Natalie K Fitzpatrick [1,2]
Ghazaleh Fatemifar [1,2]
Amitava Banerjee [1,2]
Richard Dobson [1,2, 5]
Laurence J. Howe [6]
Valerie Kuan [2,6]
R. Tom Lumbers [1,2]
Laura Pasea [1,2]
Riyaz S. Patel [1,2]
Anoop D Shah [1,2]
Aroon D. Hingorani [2,6]
Cathie Sudlow [3,4,7]
Harry Hemingway [1,2,8]

1. Institute of Health Informatics, University College London, UK
2. Health Data Research UK London (HDR UK), University College London, 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
7. Health Data Research UK Scotland (HDR UK), University of Edinburgh, Edinburg, UK
8. The National Institute for Health Research University College London Hospitals Biomedical Research Centre, University College London, London, UK
9. The Alan Turing Institute, 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

Word count (abstract): 250
Word count (main body): 6351
Tables: 2
Figures: 4
Supplementary Figures: 0
Supplementary Tables: 1
References: 95
Keywords: electronic health records, phenotyping, medical informatics, personalized medicine

ABSTRACT

Objective Electronic Health Records (EHR) 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 EHR in the United Kingdom (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 e.g. 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) International Classification of Diseases 9th/10th Revision (ICD-9, ICD-10, secondary care diagnoses and cause of mortality), c) OPCS Classification of Interventions and Procedures (OPCS-4, hospital surgical procedures) and d) Gemscript Drug Codes.

Results Using the CALIBER phenotyping framework, we created algorithms for 51 diseases, syndromes, biomarkers and lifestyle risk factors and provide up to six validation approaches. The EHR phenotypes are curated in the open-access CALIBER Portal (https://www.caliberresearch.org/portal) and have been used by 40 national/international research groups in 60 peer-reviewed publications.

Conclusion We describe a UK EHR phenomics approach within the CALIBER EHR data platform 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) National Health Service (NHS) 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 healthcare system, 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, United States (US) and other researchers for decades [2]. Furthermore, these national EHR data sources are being linked with large-scale consented genomic resources i.e. 100,000 Genomes Project (also known as Genomics England) [3] and 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 requirements [8,9]. Significant progress has been made by US initiatives (Electronic Medical Records and Genomics (eMERGE) [10], BioVU [11], Million Veteran Programme (MVP), [12] *All Of Us* [13]), Canada [14], Australia [15], Sweden [16] and Denmark [17]. In the UK however, there has been no recognized phenotyping framework or go-to resource for EHR researchers for systematically creating, curating and validating (rule-based or otherwise) EHR-derived phenotypes, obtaining information on controlled clinical terminologies, sharing algorithms, and communicating 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 (referred to as *codelists*) rather than self-contained phenotypes i.e. terms, implementation, validation evidence.

The scope of our research focuses on rule-based algorithms as the majority of research studies (with some exceptions [23,24]) using UK EHR utilize this approach for creating EHR-derived phenotypes [25]. The main use-case for CALIBER phenotypes and the approach presented in the manuscript is observational research (which is also the main stakeholder group of UK EHR) a) high-resolution clinical epidemiology

using national EHR examining disease aetiology or prognosis, or b) genetic epidemiology studies through the UK Biobank and Genomics England investigating simple and complex traits across populations. Our aspiration however is for CALIBER phenotypes to be adopted by the NHS in terms of computable knowledge which can be integrated in the healthcare system and used for interventional studies and clinical guidelines.Each of these use cases however has a different threshold on what is considered adequate performance and we adopted a systematic and robust validation approach in order to quantify phenotype performance.

EHR phenotype validation is a critical process guiding their subsequent use in research or care [26,27]. 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 [28–32]. In contrast with the US, only recently have scalable methods been developed to access the entire hospital record for expert review [33] and text corpora are not available at scale [34]. There have been few previous studies [35] of the validity of International Classification of Disease and Health Related Problems, 10th Revision (ICD-10) terms [36] in the UK against hospital records because introduction of hospital EHRs are recent (e.g. there are only three hospitals that have achieved stage six on the Healthcare Information and Management Systems Society (HIMSS) Electronic Medical Record Adoption Model (EMRAM) [37]).

We have developed the CALIBER EHR platform for the UK by adopting and extending best practices from leading initiatives and consortia (e.g. eMERGE, MVP, BioVU and others) with regards to creating, evaluating and disseminating EHR-derived phenotypes for research. Specifically, these practices, which were previously not systematically followed in the UK EHR community prior to CALIBER, include: a) establishing a robust and iterative phenotype creation process involving multiple scientific disciplines, b) systematically curating EHR-derived phenotypes, c) using methods for enhancing reproducibility, and d) undertaking and reporting robust phenotype validation analyses. Here, we define a framework for enabling EHR phenotyping in a scalable and reproducible manner. Algorithm reproducibility was defined similarly to Goodman's "methodology reproducibility" [38] i.e. providing a systematic and precise description of the algorithm components, logic, implementation and evidence of validity that would enable national or international independent researchers to create, apply and evaluate CALIBER phenotyping algorithms in local similar data sources. 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 examples of 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

The CALIBER platform [39] is currently built around four national EHR data sources **(Figure 1)** deterministically linked using 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 [40].

The baseline cohort is composed of a national primary care EHR database, the *Clinical Practice Research Datalink (CPRD) [41]*. Primary care has used computerised health records since 2000 and general practices use one of several EHR systems. CPRD contains longitudinal primary care data (extracted from the Vision and Egton Medical Information Systems (EMIS) clinical information systems) on diagnoses, symptoms, drug prescriptions, vaccinations, blood tests and risk factors (irrespective of disease status and hospitalization). CPRD uses Read [42] terms (112,806 terms, subset of the The International Health Terminology Standards Development Organisation Systematized Nomenclature Of Medicine- Clinical Terms (SNOMED-CT) [42]) to record information. Prescriptions are recorded using Gemscript (a commercial derivative of the NHS Dictionary of Medicines and Devices (dm+d)) [43] (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 [44–46] and of high validity [47].

Hospital Episode Statistics (HES) (https://digital.nhs.uk/) [48] contains administrative data on diagnoses and procedures generated during hospital interactions. Diagnoses are recorded using the International Classification of Diseases 10th Revision (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 (CPT) [49]). 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 [50]. The *Office for National Statistics* (ONS) contains socioeconomic deprivation using the Index of Multiple Deprivation (IMD) [51] and physician-certified cause-specific mortality (underlying and up to 14 secondary causes using ICD-9/ICD-10).

Data quality

Primary care

Our analyses incorporated primary care EHR data quality metrics across two dimensions: at the patient level and at the primary care practice level [41]:

Patient-level data quality: In line with previous research using UK primary care electronic health records from the Clinical Practice Research Datalink (CPRD) and CPRD guidance, we only utilized patients which were marked as "acceptable for research" by the CPRD. Patients are labelled as acceptable through an algorithmic process which identified and excludes patients with non-continuous follow up and patients with poor data according to a predefined list of data quality metrics (e.g. empty date of first registration, first registration prior to date of birth, invalid gender, missing or incorrect dates across all recorded healthcare episodes). We additionally excluded records where the date was invalid/malformed or in the future occurring after the last date of data collection.

Practice-level: The overall quality of the data recorded in a primary care practice is algorithmically marked by an "up to standard (UTS)" date by the CPRD. The UTS date is deemed as the date at which data in the practice is considered to have continuous high-quality data fit for use in research. The algorithm used to derive this date is based on two concepts: a) gap analysis (assurance of continuity in data recording and establishing if any unexpected and prolonged gaps in recording exist) and, b) death recording (observing the expected and actual deaths recorded at a practice over time by taking into account season and geographical variation in death rates and establishing if any gaps in recording exist). In both of these cases, the UTS date is set to the latest of these dates.

Completeness patterns of key clinical covariates such as risk factors (e.g. smoking status, blood pressure, BMI) has been previously shown to have rapidly increased after the introduction of a financial incentives framework (Quality and Outcomes Framework) which encourages GPs to record key data items [41].

Secondary care

Hospital Episode Statistics (HES) Admitted Patient Care (APC) data are collected for all admissions to all National Health Service (NHS) secondary healthcare providers. The NHS funds 98-99% of hospital activity in England. HES APC are administrative data collected for reimbursement of hospital activity and are post-discharge derived by clinical coders according to standardized rules for translating information from from

discharge summaries into diagnosis (ICD-10) and surgical procedure terms (OPCS-4) terms [48]. The overarching reimbursement framework, Payment-By-Results (a fixed tariff case mix based payment system [52]), provides financial incentives for hospitals to improve their coding accuracy and depth and ensure accurate reimbursement. This has led to an increase in the number of diagnosis terms recorded and coding accuracy i.e. primary diagnoses accuracy was 96% (interquartile range (IQR): 89.3-96.3) when compared to expert review of case notes [53]. The NHS Digital Data Quality Maturity Index (DQMI) provides a per hospital overall score for clinical data quality in term of data field and hospitalization episode completeness on a quarterly basis [54].

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. Unified Medical Language System (UMLS) Metathesaurus [55,56]. 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 in a rule-based fashion e.g. hypertension is defined using continuous blood pressure, coded diagnoses, blood-pressure lowering prescriptions, and comorbidities. We generated ad-hoc rules to reconcile: a) coding differences across EHR sources with respect to the granularity of diagnosis, b) the presence of multiple terms i.e. multiple different ethnicity entries, and c) transience in coding (e.g. ICD-9 was used for recording the cause of death before 2000). In primary care EHR, identified Read terms were evaluated in terms of their information content and subsequent ability to ascertain a phenotype reliably.

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

Obtaining and curating evidence of phenotype validity is an essential component of the phenotyping process. We evaluated EHR-derived phenotypes across up to six different approaches of providing of evidence of phenotype validity, acknowledging that that the use case will inform which validation(s) are most important. For example, phenotyping algorithms developed for disease epidemiology (e.g. screening or disease surveillance) might be designed for higher sensitivity whereas those used in genetic association studies might be designed to maximize PPV. We provide details of these validation approaches below:

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 used a disease registry (MINAP) as a reference in order to derive the positive predictive value (PPV) of AMI diagnoses recorded in hospital EHR (HES) i.e. the probability that an AMI diagnosis recorded in HES was indeed an AMI as ascertained by MINAP (that contains information on AMI ascertainment such as electrocardiogram results and troponin measurements) rather than unstable angina or a non-cardiac diagnosis. We did not calculate sensitivity and specificity relative to MINAP given that MINAP does not include all cases of AMI, as it is a disease registry which requires bespoke data entry by audit staff separate from clinical care or coding. It is therefore not possible to use MINAP as a gold standard to evaluate hospital EHR (Hospital Episode Statistics) in relation to completeness of detection of AMI (sensitivity) or non-MI (specificity). However, there is a concern that HES data may be inaccurate, and MINAP can be used to evaluate its positive predictive value for the subset of cases with a MINAP record for the event, where the exact diagnosis in MINAP can be considered a “gold standard.”

2) Case note review

We evaluated the performance of the secondary care component of the bleeding phenotype by assessing the ability of the diagnosis terms (ICD-10) utilized by the phenotype to correctly identify hospitalized bleeding events in two independent hospital EHR sources. Two clinicians (blinded to the ICD-10 diagnosis 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 [57] from the consented Stroke InvestiGation Network- Understanding Mechanisms (SIGNUM) study.

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

For all exemplars, we produced and reported hazard ratios (HR) and 95% Confidence Intervals (CI) 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 (KM) 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 diagnosed in HES.

5) Consistency of genetic associations

Similar to previous studies, we attempted to replicate previously reported associations between genetic variants and diseases discovered from non-EHR studies (e.g. research-driven observational cohort studies or interventional studies). The ability of EHR-derived phenotypes to replicate previously-discovered associations derived from non-EHR studies and observing similar direction and magnitude of association reinforces the evidence towards the overall validity of the EHR phenotype [58]. Using PLINK [59], we extracted genetic variants associated with AMI reaching genome-wide significance ($P < 5x10^{-8}$) from publicly-available 1000 Genomes-based Genome Wide Association Study (GWAS) summary data ("CARDIoGRAMplusC4D - mi.additive.Oct2015") in the CARDIoGRAMplusC4D [60] consortium. In the UK Biobank, we identified AMI cases in linked hospital and mortality EHR using the CALIBER AMI phenotype and defined controls as non-case participants with no self-reported record of AMI at baseline. We estimated the association of genetic variants and AMI using logistic regression with an underlying additive model in PLINK adjusting for the first 10 principal components, age and sex. Replication was defined as the

Single Nucleotide Polymorphism (SNP) being associated with AMI in the UK Biobank (Bonferroni-adjusted P<0.0016) with a concordant direction of effect with CARDIOGRAMPlusC4D.

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 textual descriptions of 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 Structured Query Language (SQL) queries. We created entries in the CALIBER Portal **(Figure 4)** describing implementation details across sources, research outputs, validation evidence and a Digital Object Identifier (DOI) [61]. 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 linked primary/secondary 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

Using the CALIBER EHR phenotyping approach described here, we curated over 90,000 terms from five controlled clinical terminologies to create 51 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 [62]. The CALIBER Portal (http://www.caliberresearch.org/portal) opened in October/2018 to the community and provides a centralized resource for curating EHR-derived 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 an NPV of 0.94 (0.90, 0.97) and a PPV of 0.88, i.e. 88% of bleeding events identified by the ICD-10 terms utilized in the CALIBER bleeding phenotype 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 ICD-10 coded events 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 by coded terms [63]. Specificity was found to be 0.99 (0.97, 1.00) indicating a very low number of false positive bleeding events.

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) [29]. 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 CARDIoGRAMplusC4D GWAS summary data, we identified 31 independent variants associated with AMI by linkage disequilibrium (LD) clumping (R^2<0.001, 250 kb) genetic variants reaching genome-wide significance (P <5x10-8). In the UK Biobank, we identified 8,281 AMI cases, 394,933 controls and excluded 5,266 participants from the analysis due to self-reported AMI at baseline. From 31 previously-reported SNPs, 31 (100%) had P<0.05 with same direction, with 26 (83.8%) passing Bonferroni correction (P<0.0016)(Supplementary Table 1).

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), US (n=53,909) and France (n=961) [64]. 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, Chronic Obstructive Pulmonary Disease (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% [US]); the composite of AMI, stroke, or death [from 26.0% (France) to 36.2% (US)]; and hospitalized bleeding [from 3.1% (France) to 5.3% (US)]. Adjusted risks were similar across countries, but higher in the US for all-cause death [RR US vs. Sweden, 1.14 (95% CI 1.04–1.26)] and hospitalized bleeding [RR US 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) [65,66]. 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

In this study, we describe the CALIBER phenotyping approach which has been used to produce 51 validated phenotyping algorithms which capture disease status, biomarker values and lifestyle risk factors across four

UK national EHR data sources spanning primary care, hospital admissions, a disease registry and a mortality register. 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. Research studies require precise phenotype definitions but phenotypic information found in EHR is typically inconsistent and of variable data quality. These problems are exacerbated when linking data across healthcare settings (i.e. primary care and hospital admissions) as each source records information using different healthcare processes, disparate formats and terminologies and interact with fundamentally different patient populations. Complementary initiatives exist [19] but these 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 the Phenotype Knowledgebase (PheKB) developed by the eMERGE consortium, CALIBER includes additional approaches of validation e.g. aetiological and prognostic across population samples but lacks comprehensive detailed PPV/NPV measurements which are made possible by the availability of text and access to case notes at scale in the US.

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. The actual phenotyping algorithm production was lengthy and labor intensive and usually involved a large number of iterations although the exact number of person hours was difficult to ascertain. A particular challenge was the need to reconcile differences in granularity of diagnosis terms used in primary care and secondary care EHR as each healthcare setting utilizes different clinical terminologies to record information (Read in primary care, ICD-10 in secondary care). For example, in HF, the Read controlled clinical terminology allowed us to potentially distinguish between the two main congestive heart failure types: heart failure with normal/preserved ejection fraction (HFpEF) (i.e. Read term “G583.11 HFNEF - heart failure with normal ejection fraction”) and heart failure with reduced ejection function (HFrEF) / left ventricular systolic dysfunction (i.e. Read term “G5yy900 - Left ventricular systolic dysfunction”). Conversely, ICD-10 terms are substantially less specific (i.e. ICD terms “I50.0 Congestive heart failure” and “I50.9 Heart failure, unspecified”) and do not allow for this distinction. As a rule, for overlapping diagnoses across multiple sources, CALIBER phenotypes utilize the source with the highest clinical resolution to ascertain disease status.

We found that diagnosis terms in primary care using Read terms were not always informative and could not directly be used to ascertain particular phenotypes. For example, when attempting to create a dietary phenotype, we identified 173 Read terms related to nutrition which were recorded across 5.6M diagnosis events. Only 8% of these however were sufficiently informative to infer a particular nutritional diet i.e low

fat diet, gluten free diet, diabetic diet or low sodium diet. The majority of terms used were generic terms which carried little information (i.e. “8CA4.00 Patient advised re diet” or “9N0H.00 Seen in dietician clinic”) and could not be used for ascertaining a phenotype with reasonable performance. While such terms could potentially be utilized as supporting information for other phenotypes (i.e. diabetic diet as evidence of diabetes) they cannot be used to ascertain a phenotype directly.

We observed that clinically-informed combinations of information across EHR sources improves case detection. All disease and syndrome phenotypes (n=35) utilized information sourced from primary care and hospital care EHR and roughly half (n=18) utilizing cause-specific mortality information recorded in the national death register. In general, we considered EHR sources complementary to each other with each providing a different aspect of a patient’s disease state (chronic vs. acute) rather than denote one as the authoritative source of information. One notable exception to this is mortality where we used the ONS date of death as the “gold standard” as studies have shown that discrepancies exist between the recorded death dates in primary care EHR and the date recorded on the death certificate (ONS). A previous study [67] of 118,571 deaths in England showed that a discrepancy existed in almost a quarter of deaths. Considerable variation was observed between GP practices on the degree of such discrepancies (in the majority of cases, the date of death recorded by the GP was after the date of death recorded on the death certificate). This is because GPs do not routinely see the death certificate (which is the definitive record of time and cause of death) and there is no legal obligation for them to record the date of death accurately. If there is a delay in their receipt of notification of death, they might record the date of death as the date of notification, or the date the patient’s record was closed, rather than the actual date of death. In line with previous literature we therefore used the ONS as the “gold standard” for ascertaining mortality.

A major effort of CALIBER has been to create longitudinal disease phenotypes that capture early and late manifestations of disease. We observed that the proportion of cases contributed by each EHR source differed by age at diagnosis: patients identified in secondary care EHR were substantially older than those identified in primary care. For example, substantially more, atrial fibrillation cases were identified in secondary care EHR rather than in primary care (32,930 cases compared to 11,068 from primary care) and using only a single source would have introduced bias and underestimated the incidence of disease. Conversely type 2 diabetes cases were exclusively identified through primary care EHR with no cases identified exclusively in hospital EHR due to the fact that, like other diseases such as hypertension, diagnosis and management is almost entirely performed in primary care.

Validation **(Table 2)** was a critical step for assessing the accuracy of EHR-derived phenotypes. We did not consider algorithm validation as a finite task but as a constantly evolving process due to the underlying complexity of EHR data and the processes which generate them [68]. We sought to address the spectrum of

validation views and developed an approach which captures six levels of evidence. The majority of disease and syndrome phenotypes examined incidence estimates across different EHR sources and consistency with previous associations in terms of disease aetiology and prognosis. Validation was more restricted in biomarker and lifestyle risk factor phenotypes since information was derived from only one particular source (in the case of biomarkers, measurements were exclusively identified in primary care). Clinician case note review was considered the "gold standard" of phenotype validation that enables PPV/NPV calculations but access to medical records was not widely available and thus we could only perform this in a single instance. Prognostic validation was one of the main validation approaches where consistency with previously-reported findings provided a degree of confidence in terms of phenotype validity (for exposures, outcomes and covariates used in the analyses). Inconsistent results however were possible and could be explained due to multiple factors such as different health settings and sources of data, healthcare process factors and workflows and uncomparable definitions.

In terms of the complete hospital interaction, HES data are a snapshot of the patient journey as data are collected for hospital reimbursement [8,52]. Hospital records are converted into diagnosis and procedure codes locally (following an existing protocol) at each hospital and submitted to the NHS. HES data are provided to researchers with identifiers for hospitals removed in order to protect patient anonymity as a common identifier is used across HES and CPRD GP practices which have a substantially smaller catchment area. As such, we were unable to assess the effect of site-level variability in terms of data capture and quality and phenotype validity. Multiple initiatives however exist for obtaining raw hospital records for research e.g. National Institute for Health Research Health Informatics Collaborative (NIHR HIC) which links eleven Intensive Care Units (ICUs) in five hospitals for research (>18,000 patients, >21,000 admissions, median 1,030 time-varying measures [69]). Crucially, these initiatives will enable researchers to have access to raw hospital data, including free text, for creating and validating phenotypes and 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.

CALIBER phenotyping algorithms are rule-based, deterministic, and provide a framework on which future phenotypes can be created. While our approach yields robust and accurate algorithms, it does not scale with our ambition to create and curate thousands of high-quality phenotypes (and their validation) 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[70]. Such 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 [71]. Finally, a key

next step is to use the six sites of the recently funded Health Data Research UK (HDR UK) national institute in order to scale up the number of phenotypes created and curated using UK EHR.

The use of a common data model to map between the clinical terminologies used across EHR sources, such as the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) can potentially address some of the labour-intensive tasks associated with phenotyping. For example, the translation from phenotype definition to SQL for data extraction was manual due to the lack of an established storage format [72] for the algorithms and variable schema across EHR sources. OMOP CDM can potentially act as Relational Database Management System (RDBMS) agnostic schema which standardized analytical tools can be deployed on and has been shown to be robust [73,74] and we are currently in the process of evaluating the fidelity of the data transformation. We have additionally evaluated different approaches (Semantic Web Technologies, openEHR [75,76]) for storing phenotype definitions in a computable format that can enable high-throughput phenotyping and eliminate the need for manual human-driven translation to SQL queries. Given that all of UK primary care EHR data are hosted on four clinical information systems vendors, there is a real opportunity to create computable phenotypes which can be utilized across the NHS [77]. To accomplish this, information exchange standards (e.g. Fast Healthcare Interoperability Resources (FHIR) [78]) have to be utilized and combined with approaches such as the Phenotype Execution and Modeling Architecture (PHEMA)[79] and the Measure Authoring Tool (MAT) [80].

CONCLUSION

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. In this manuscript, we presented the CALIBER platform as a framework for using national EHR from primary and secondary health care, disease and national 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 [81,82]. 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 UK population and is work in underway to create a chronological map of human disease spanning early and late life by curating over 300 diseases [96].

Through CALIBER we provide a framework for the consistent definition, use and re-use of EHR derived phenotypes from national UK EHR sources for observational research: a) high-resolution clinical epidemiology using national EHR examining disease aetiology or prognosis, or b) genetic epidemiology studies through the UK Biobank and Genomics England investigating simple and complex traits across populations. One of the primary audiences of CALIBER phenotypes is international: US investigators

account for a third of studies using UK primary care EHR [18] and two thirds of UK Biobank studies are carried out by US investigators. Additionally, the controlled clinical terminologies used in UK EHR are directly translatable and transferable to the US e.g. Read terms (CTV3 (Clinical Terms Version 3) are part of SNOMED-CT, and ICD-9- Clinical Modification (CM) to ICD-10 mappings exist. As PheKB and other initiatives evolve, establishing links across national portals [83] and infrastructure can enable cross-biobank analyses of complex/rare phenotypes [7].

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.

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 cell arteritis	●	●		●		●			
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% [31].	The proportion of non-fatal AMI defined across primary care, hospital care and disease registry was 32% [29].	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.
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% CI 91.6% to 92.8%) and in hospital admissions	Compared through independent review by two clinicians, the PPV of bleeding events identified through the phenotyping algorithm was 0.88.

			was 91.5% (90.8% to 92.1%) [29]	
Aetiology	Are the prospective associations with risk actors consistent with previous evidence?	Type 2 diabetes [84], systolic/diastolic blood pressure [32], heart rate [85], socioeconomic deprivation [86], alcohol consumption [87], smoking [88], ethnicity [44], acute myocardial infarction [29], depression [89], neutrophil counts [90] , eosinophil/lymphocyte counts[91], atrial fibrillation [30], sex [92]	Type 2 diabetes [84], blood pressure [32], heart rate [85], socioeconomic deprivation [86], alcohol consumption [87], smoking [88], ethnicity [44], acute myocardial infarction [29], depression [89], neutrophils [90] , eosinophil/lymphocyte counts[91], atrial fibrillation [30], influenza infection [93], ischaemic presentations [94], sex [92]	At 5 years 29.1% (95% CI: 28.2, 29.9%) of atrial fibrillation patients, 21.9% (21.2, 22.5%) 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 [31].	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% [29]. 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	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.

			person-years: 63.8, 95% CI 62.2 to 65.5) [95]	
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 AMI, stroke, or death [from 26.0% (France) to 36.2% (USA)]; and hospitalized bleeding [from 3.1% (France) to 5.3% (USA)]. [64]	

CI Confidence Interval; EHR Electronic Health Records; AMI Acute Myocardial Infarction; PPV Positive Predictive Value; HF Heart Failure; HR Hazard Ratio.

ACKNOWLEDGEMENTS

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

The BigData@Heart Consortium is funded by 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; it is chaired, by DE Grobbee and SD Anker, partnering with 20 academic and industry partners and ESC.

This work was supported by Health Data Research UK, which receives its funding from HDR UK Ltd (LOND1) 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 a National Institute for Health Research (NIHR Senior Investigator). ADS is a THIS Institute postdoctoral fellow. VK is supported by the Wellcome Trust (WT 110284/Z/15/Z). SD is supported by an Alan Turing Fellowship.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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

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.

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.

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.

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.

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

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.

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.

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.

23. Jammeh EA, Carroll CB, Pearson SW, Escudero J, Anastasiou A, Zhao P, et al. Machine-learning based

identification of undiagnosed dementia in primary care: a feasibility study. BJGP Open. 2018;2: bjgpopen18X101589.

24. Zhou S-M, Fernandez-Gutierrez F, Kennedy J, Cooksey R, Atkinson M, Denaxas S, et al. Defining Disease Phenotypes in Primary Care Electronic Health Records by a Machine Learning Approach: A Case Study in Identifying Rheumatoid Arthritis. PLoS One. 2016;11: e0154515.

25. Williams R, Kontopantelis E, Buchan I, Peek N. Clinical code set engineering for reusing EHR data for research: A review. J Biomed Inform. 2017;70: 1–13.

26. 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.

27. 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

28. 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.

29. 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.

30. 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.

31. 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.

32. 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.

33. 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.

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.

35. 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.

36. 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.

37. 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.

38. Goodman SN, Fanelli D, Ioannidis JPA. What does research reproducibility mean? Sci Transl Med. 2016;8: 341ps12.

39. 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.

40. 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.

41. 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.

42. O'Neil M, Payne C, Read J. Read Codes Version 3: A User Led Terminology. Methods Inf Med. Schattauer GmbH; 2018;34: 187–192.

43. 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.

44. 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.

45. 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.

46. 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.

47. 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.

48. Herbert A, Wijlaars L, Zylbersztejn A, Cromwell D, Hardelid P. Data Resource Profile: Hospital Episode Statistics Admitted Patient Care (HES APC). Int J Epidemiol. 2017;46: 1093–1093i.

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

50. 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.

51. 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.

52. 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.

53. Burns EM, Rigby E, Mamidanna R, Bottle A, Aylin P, Ziprin P, et al. Systematic review of discharge coding accuracy. J Public Health . 2012;34: 138–148.

54. NHS Data Quality Maturity Index. In: NHS Digital [Internet]. [cited 13 Mar 2019]. Available: https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/data-quality

55. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp. ncbi.nlm.nih.gov; 2001; 17–21.

56. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res. academic.oup.com; 2004;32: D267–70.

57. 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.

58. 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;31: 1102–1110.

59. 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.

60. 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.

61. 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

62. 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

63. 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.

64. 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.

65. 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.

66. 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.

67. Harshfield A, Abel GA, Barclay S, Payne RA. Do GPs accurately record date of death? A UK observational analysis. BMJ Support Palliat Care. 2018; doi:10.1136/bmjspcare-2018-001514

68. 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.

69. 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.

70. Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. J Am Med Inform Assoc. 2013;20: 117–121.

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

72. 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.

73. Schuemie MJ, Gini R, Coloma PM, Straatman H, Herings RMC, Pedersen L, et al. Replication of the OMOP experiment in Europe: evaluating methods for risk identification in electronic health record databases. Drug Saf. 2013;36 Suppl 1: S159–69.

74. Matcho A, Ryan P, Fife D, Reich C. Fidelity assessment of a clinical practice research datalink conversion to the OMOP common data model. Drug Saf. 2014;37: 945–959.

75. 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.

76. 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.

77. 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

78. 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.

79. 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.

80. 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.

81. 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.

82. 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.

83. 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.

84. 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.

85. 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.

86. 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.

87. 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.

88. 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.

89. 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.

90. 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.

91. 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.

92. 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.

93. 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.

94. 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.

95. 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.

96. 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 National Health Service. The Lancet Digital Health 2019 (in press).