This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by-nc/3.0/), which permits non-commercial reuse, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com.
Published by Oxford University Press on behalf of the International Epidemiological Association *International Journal of Epidemiology* 2012;**41**:1625–1638
© The Author 2012; all rights reserved. Advance Access publication 5 December 2012 doi:10.1093/ije/dys188

DATA RESOURCE PROFILE

Data Resource Profile: Cardiovascular disease research using linked bespoke studies and electronic health records (CALIBER)

Spiros C Denaxas,[1]*† Julie George,[1]† Emily Herrett,[2] Anoop D Shah,[1] Dipak Kalra,[3] Aroon D Hingorani,[1] Mika Kivimaki,[1] Adam D Timmis,[4] Liam Smeeth[2] and Harry Hemingway[1]

[1]Department of Epidemiology and Public Health, Clinical Epidemiology, University College London, London, UK, [2]Department of Non-Communicable Disease Epidemiology, London School of Hygiene and Tropical Medicine, London, UK, [3]Centre for Health Informatics and Multi-professional Education, University College London, London, UK and [4]National Institute for Health Research Biomedical Research Unit, Barts Health, London, UK

*Corresponding author. Department of Epidemiology and Public Health, Clinical Epidemiology, University College London, 1-19 Torrington Place, London, WC1E 6BT, UK. E-mail: s.denaxas@ucl.ac.uk
†These authors contributed equally to this work.

Accepted 4 October 2012

The goal of cardiovascular disease (CVD) research using linked bespoke studies and electronic health records (CALIBER) is to provide evidence to inform health care and public health policy for CVDs across different stages of translation, from discovery, through evaluation in trials to implementation, where linkages to electronic health records provide new scientific opportunities. The initial approach of the CALIBER programme is characterized as follows: (i) Linkages of multiple electronic heath record sources: examples include linkages between the longitudinal primary care data from the Clinical Practice Research Datalink, the national registry of acute coronary syndromes (Myocardial Ischaemia National Audit Project), hospitalization and procedure data from Hospital Episode Statistics and cause-specific mortality and social deprivation data from the Office of National Statistics. Current cohort analyses involve a million people in initially healthy populations and disease registries with $\sim 10^5$ patients. (ii) Linkages of bespoke investigator-led cohort studies (e.g. UK Biobank) to registry data (e.g. Myocardial Ischaemia National Audit Project), providing new means of ascertaining, validating and phenotyping disease. (iii) A common data model in which routine electronic health record data are made research ready, and sharable, by defining and curating with meta-data >300 variables (categorical, continuous, event) on risk factors, CVDs and non-cardiovascular comorbidities. (iv) Transparency: all CALIBER studies have an analytic protocol registered in the public domain, and data are available (safe haven model) for use subject to approvals. For more information, e-mail s.denaxas@ucl.ac.uk

Keywords electronic heath records, linkages, cardiovascular

Data resources basics

Using linked electronic health records for translational research

The goal of cardiovascular disease (CVD) research using linked bespoke studies and electronic health records (CALIBER) is to provide evidence across different stages of translation, from discovery, through evaluation to implementation where electronic health records provide new scientific opportunities. The expanding role of such record research with clinical and public health impact has been extensively reviewed.[1–4] With healthy cohort sample sizes $\times 10^6$ and disease registries 10^5, such records offer a larger scale and higher degree of phenotypic resolution[5] than has been available in investigator led studies.

Opportunity

The UK is the only country in the world that has both detailed electronic primary care records, and CVD and procedure registries at a national scale, as well as more standard sources such as cause-specific hospitalization and mortality records and census data (see Table 1 and Table 2). Neither Sweden nor Denmark has comparably rich national primary care data in their respective linkage programmes. Scotland[14,15] and Wales[16,17] have established record-linkage programmes that have focused on diabetes and public health research, respectively. Some centres have had prominent research programmes in CVD with a focus on prognosis and quality of care and drug safety research in all phases of translational research in smaller (10^5) populations such as the Institute of Clinical and Evaluative Sciences in Canada[18] and Yale, Duke, Intermountain Heart Centre and the Mayo Clinic in the USA.[19] There have been few, if any, attempts to establish an Electronic Health Record (EHR)-based population-based cohort research platform in CVDs.

Who set CALIBER up and how was it funded?

In 2012, the UK Government launched four new centres of EHR research in London, Manchester, Dundee and Swansea. CALIBER is led from the London centre, Centre for Health Service and Academic Partnership in Translational Electronic health record Research, Director Hemingway (CHAPTER) incorporating the National Institute for Cardiovascular Outcomes Research (NICOR). CHAPTER is led from University College London (UCL) and Partners, which include UCL, the London School of Hygiene and Tropical Medicine and Queen Mary University of London. CALIBER investigators represent a collaboration between epidemiologists, clinicians, statisticians, health informaticians and computer scientists with initial funding from the Wellcome Trust and the National Institute for Health Research. The first tranche of raw-linked data was received in 2011. Third party linkage was coordinated by the Medicines and Healthcare products Regulatory Agency.

Challenges

Researchers wishing to harness large-scale national data sources or rich regional resources on imaging and bioresources, on CVDs in the UK, face major challenges:

(i) Lack of CVD–EHR research programmes spanning the entire translational cycle.[20]

Table 1 Availability of primary care data for research in different countries

Country	National or regional	Primary and ambulatory care data available for research linkages
UK	National	CPRD, access through Academic Health Sciences Networks. See Table 2
Sweden	National	Primary care is organized regionally; national initiatives in SwedeHeart[6] and the National Registry of Secondary Prevention (SEPHIA)
Denmark	National	Register of Medicinal Product Statistics[7,8]
Canada	Regional	Ontario, Institute for Clinical Evaluative Sciences[9]
		Ontario Health Insurance Plan Physician claims database
USA	National	Medicare (for people aged $\geqslant$65 years)
	National	Million Veteran Programme
	Regional	Mayo Clinic[10]
		Rochester Epidemiology Project, Olmsted County
	Regional	Kaiser Permanente California Research Program on Genes, Environment, and Health[11]
	Regional	Intermountain Healthcare[12]
South Korea	National	National health insurance claims database from the Health Insurance Review & Assessment Service[13]

Table 2 Linked electronic health record sources in CALIBER: types of data, coding system used and data recording details

Sources	Types of data	Coding system	When and by whom data is coded?
Primary care: CPRD and other sources	Longitudinal primary care data Diagnoses and symptoms irrespective of hospitalization, drug prescriptions, vaccinations, blood test results, risk factors	Data recorded using the Read clinical terminology system, version 3 contains ~99 000 codes	Data recoded by the general practitioner in real time during the consultation Hospital discharge letters coded by a practice administrator
Social deprivation: ONS	Small area patient social deprivation data	Index of Multiple Deprivation (2007) and Townsend score	Derived from multiple national administrative data sets
Disease registry: MINAP	National registry of Acute Coronary Syndrome admissions Phenotype (ST Elevation Myocardial Infarction, Non-ST Elevation Myocardial Infarction, Unstable Angina), severity and treatment data	In all, 120 fields most with multiple response categories, as defined by the MINAP steering group	Recorded usually by audit nurse, days or weeks after admission, by abstracting data from hospital records
Secondary care: HES	National data warehouse of hospitalizations recorded for administrative purposes Inpatient, outpatient, emergency, critical care and maternity admissions[a] Operations and surgical procedures	Up to 20 primary and secondary discharge diagnoses recorded using ICD-10 Up to 24 codes using the Office of Population, Censuses and Surveys Classification of Surgical Operations and Procedures and used for operations The 4th revision (OPCS-4) contains ~10 000 codes	Recorded by non-clinical trained coders based on the discharge summary weeks after discharge
Mortality: ONS	National census of all deaths Primary and underlying cause of death	The primary, underlying and up to 14 secondary causes of death are recorded using ICD-10	Doctor (general practitioner or hospital) completes death certificate with cause of death. ICD codes added by trained non-clinical coders

[a]Emergency, critical care and maternity data not included in CALIBER for now.

(ii) Major national EHR sources (e.g. in NICOR) remain unlinked among themselves or to investigator led studies.

(iii) Lack of understanding of the potential of different record sources. Although the research community is familiar with death and hospitalization records coded with the International Classification of Disease, other EHR data sources, e.g. primary care, have been much less used by cardiovascular researchers. Poor adoption of meta-data standards[21] has led to little information being available on how raw data might be reliably converted into a useable form for researchers, and data inconsistency problems, both of which hinder replication of results and data sharing.

(iv) Lack of transparency initiatives, which might address well-recognized problems of selective non-publication, selective reporting and significance chasing biases.[22]

(v) Data quality has been inadequately tied to the quality of the clinical care received.

CALIBER principles

CALIBER was established within CHAPTER, a centre seeking to address these challenges. We set out below our approach for:

- Linkage of multiple EHR data sources. (See Table 2 and Figure 1 for key features of five sources of data currently linked in CALIBER.)
- Linkage with bespoke investigator led cohort studies.
- Establishing a common data model with reproducible data variables and meta-data (see Figure 2 and Figure 3).
- Transparency—CALIBER studies are registered in the public domain on clinicaltrials.gov, and their analytic protocol is made available for download (see Figure 4).

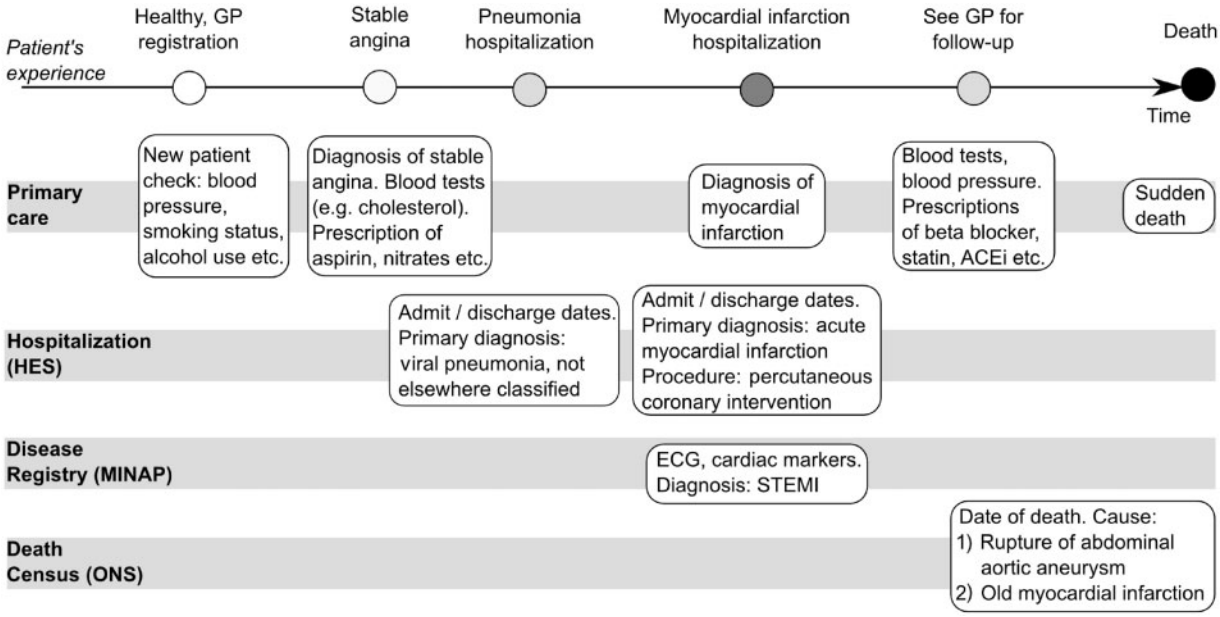

Figure 1 Longitudinal nature of multiple linked data sources in CALIBER. ECG = Electrocardiography, STEMI = ST-segment elevation Myocardial Infarction, ACEI = Angiotensin-converting-enzyme Inhibitor

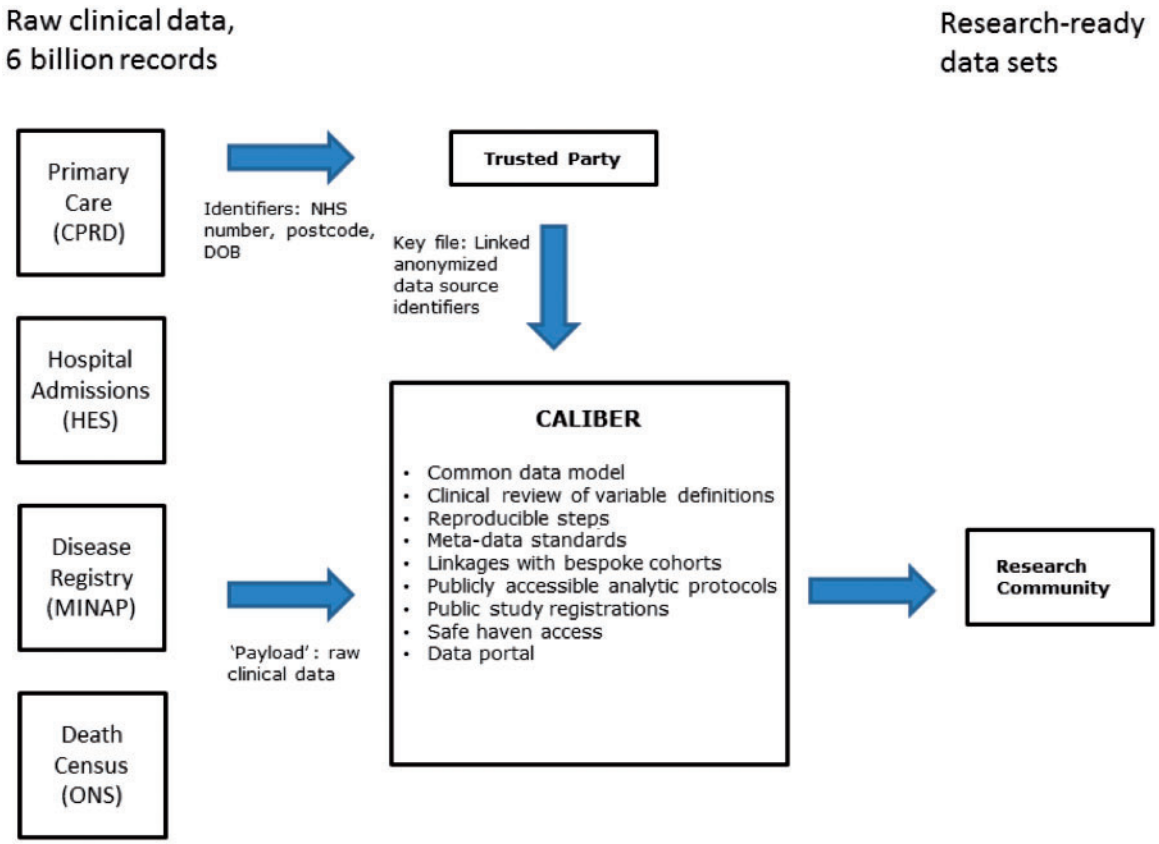

Figure 2 The CALIBER framework of transforming raw electronic health record data into usable research-ready data sets

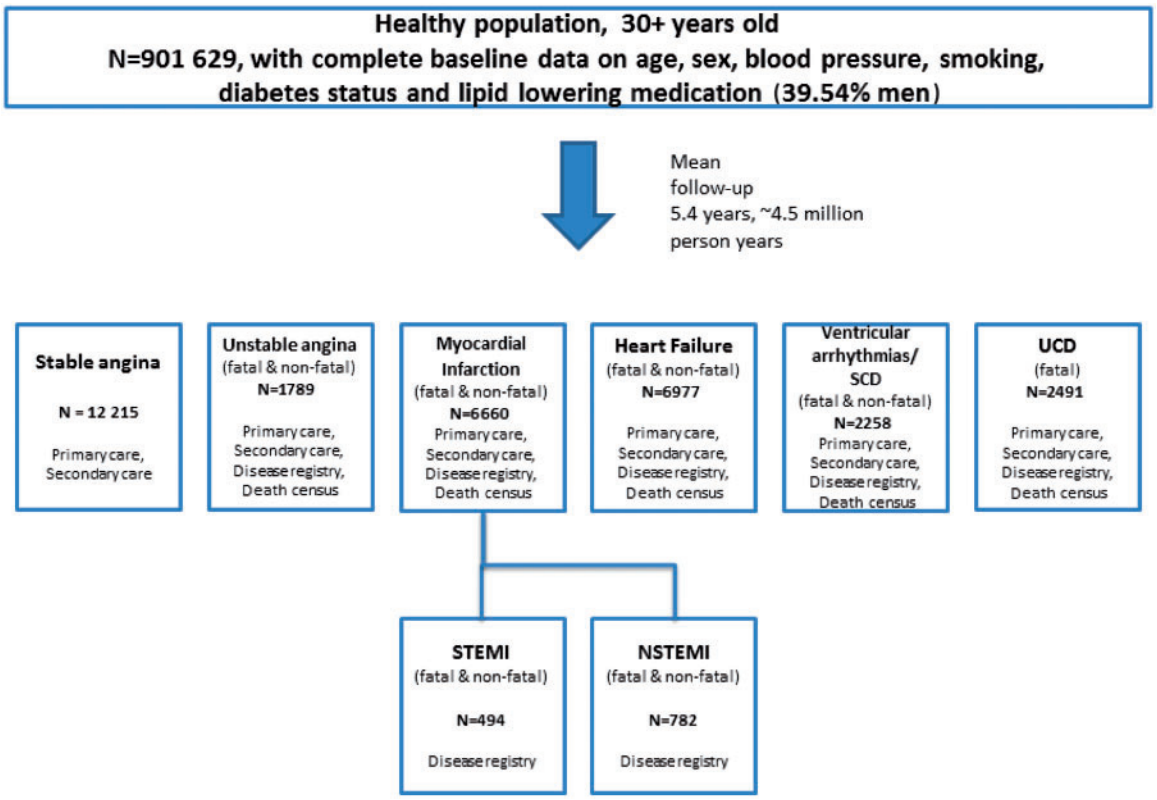

Figure 3 Example of a CALIBER cohort showing initial presentation of specific cardiac endpoints ($n = 32\,390$) with counts and sources. Appendix A illustrates the approach to defining cardiovascular diseases using multiple record sources in CALIBER

Data resource area and population coverage

Figure 1 illustrates the longitudinal, population-based nature of CALIBER data sets and how each source captures a different aspect of the patient journey. Current linkages in CALIBER involve the following: the Clinical Practice Research Datalink (CPRD),[23] the Myocardial Ischaemia National Audit Project (MINAP),[24] Hospital Episodes Statistics (HES)[25] and the Office for National Statistics (ONS) mortality[26] and social deprivation data (Table 2).

CPRD is a longitudinal primary care data set and records data on symptoms, diagnosis prescriptions, investigations, referrals and vaccinations. Roughly 97% of the population in the UK are registered with a general practitioner (GP).[27] Diagnostic data are coded using Read codes,[28] used in several countries in Europe and which map to Systematic Nomenclature of Medicine – Clinical Terms.[29] The population of all CPRD practices has been shown to be broadly representative of the population of UK;[23] the subset of practices consenting to data-linkage shares the demographic profile of the full CPRD data set. The CPRD has two key methods of ensuring quality of the data made available to the research community: the 'acceptable research quality' flag, a patient-level quality marker; and the Up To Standard date, a practice-level quality marker.

MINAP is a national registry of patients admitted to 230 hospitals in the UK with acute coronary syndrome (ACS) events. For each admission, MINAP records patient demographic characteristics, limited medical history, clinical features and investigations, drug treatment (pre-, during and post-admission) and final diagnosis, including differentiation into ST elevation MI, non-ST elevation MI and unstable angina.

HES is a national data set of all admissions to National Health Service (NHS) hospitals in the UK. Diagnoses are recorded using the International Statistical Classification of Diseases and Health-Related Problems, 10th revision (ICD-10)[30] and procedures using the Office of Population Censuses and Surveys Classification of Interventions and Procedures (OPCS) Version 4.[31]

The ONS provides data on cause-specific mortality for all residents of the UK. Cause-specific mortality data are extracted from death certificates and recorded using ICD-10. Small-area measures of social deprivation are recorded using the Index of Multiple Deprivation 2007[32] and Townsend scores.[33]

Linkage

A 10-digit numeric identifier, known as the NHS number, uniquely identifies patients in the UK. Number generation is centrally managed and not

Registration reference number	Project title	Validation	Aetiology	Prognosis	Qual. Care	Pharm.
NCT01569139	Comparison of Information recorded in MINAP, CPRD and HES	●	○	●	○	○
NCT01163513	Coronary mortality in south Asians: aetiological and prognostic effects	○	●	●	●	○
NCT01162187	Secondary prevention in Acute Coronary Syndromes	○	○	●	●	●
NCT01111071	Variation between Hospitals in Short-term mortality after Acute Coronary Syndromes	○	○	●	●	○
NCT01164371	Gender Differences in the Development, Treatment and Prognosis of Coronary Disease	○	●	○	●	○
NCT01168869	Myocardial Infarction as the First Manifestation of Coronary Heart Disease: Rates of heralded and Unheralded Myocardial Infarction	○	●	○	●	○
NCT01106196	The Role of Influenza as a Trigger for Acute Myocardial Infarction	○	●	○	○	○
NCT01231867	Cohort Study of Clopidogrel and Proton Pump Inhibitors (PPIs)	○	○	●	○	●
NCT01240798	Depression and Anxiety in the Aetiology and Prognosis of Specific Cardiovascular Disease Syndromes	○	●	●	○	○
NCT01236274	The Risk of Myocardial Infarction in Users of Antipsychotic Agents	○	●	○	○	●
NCT01335672	Survival After First Myocardial Infarction in Patients With and Without Chronic Obstructive Pulmonary Disease	○	○	●	○	●
NCT01609465	Prognostic Models for People With Stable Coronary Artery Disease	○	○	●	○	○
NCT01704300	Body Mass Index and Initial Presentations of Cardiovascular Diseases (CALIBER)	○	○	●	○	○
NCT01687686	Differential Effects of Lipids on Cardiovascular Diseases: A CALIBER Study	○	○	●	○	○

Figure 4 Example of CALIBER research projects registered in the public domain

derived from demographic information, enabling the identification of patients throughout the NHS in an unambiguous and unique way, making NHS numbers a reliable and robust identifier for record linkage and data sharing while preserving patient confidentiality. Roughly 45% of 592 CPRD practices in the UK consented to linkage of their registered patients. For these participating practices, all patients registered at the practice were included. Following the 'separation principle',[16] data were linked on NHS number, date of birth, sex and post code by a Trusted Party (see Figure 2).

Measures

Converting raw data to research-ready data: curated common data model

Figure 2 illustrates how CALIBER curates data from multiple EHR sources, generating research-ready data from raw data. CALIBER contains >3 billion records of raw data originating from multiple sources. The recorded data are of variable quality, completeness and specificity and unusable for research without proper processing. We have created >300 variables on medical history, diagnosis, investigations, procedures and prescriptions. Data from all of the source data sets were used in the creation of these variables, and each definition has been reviewed and agreed by both clinical and non-clinical researchers in a transparent and reproducible manner. Coding lists containing Read codes, ICD-10 and OPCS-4 codes were compiled for each variable in a reproducible manner.[34] Variables contain detailed response categories to provide individual researchers with the flexibility to adjust the sensitivity and specificity of their definitions according to their needs.

Variables are curated using established meta-data standards[35] and are versioned using a distributed source control repository system (http://git-scm.com/), which facilitates accurate tracking of changes over time as it keeps all previous revisions of the variables and records the changes between versions. CALIBER researchers have access to the CALIBER online data portal, which includes individual coding lists and programming scripts used to extract both data and clinical coding lists. For each variable, the data portal entry contains the type, source data files, valid range (if applicable), response categories, revision, implementation details and any other relevant information such as references to published literature or descriptive notes. This transparent and reproducible manner of operation ensures that knowledge is fed back into the platform and subsequently becomes available for researchers to make further use of. For example, if a potential researcher is working on an exposure previously not defined within CALIBER, the new variables they define would be incorporated back into the data portal (following peer review) and made available to other future researchers for use. Metadata enable the automatic generation of the centralized data portal and documentation, keeping the variable definition process synchronized with the research data set creation and curation.

Figure 5 illustrates how CALIBER exploits multiple EHR sources to create a robust data variable (CALIBER hypertension). In total, >700 clinical terms were combined from three EHR sources to create this composite variable, which defines a patient's hypertensive status at a given time point. A patient is considered hypertensive if they have been diagnosed by a GP or within the past year have had at least three measurements of raised systolic or diastolic blood pressure or at least three records of raised blood pressure or at least two prescriptions of blood pressure lowering medication.

CALIBER data include all diagnosed co-existing conditions irrespective of hospital attendance and values of clinically collected circulating biomarkers such as haemoglobin and prescriptions issued, including medication type, timing and quantity. Detailed mental health data on diagnosis and hospitalization for numerous morbidities such as depression, anxiety, bipolar disorder, schizophrenia and psychosis have been included in the data portal (D Batty, personal communication).

Frequency of data collection on patients

Unlike traditional cohorts, in EHR cohorts follow-up is continuous and initiated by capture within one or more EHR source (Figure 1). In the example cohort (Figure 3), the median observation time from study entry date is 5.5 years, with an inter-quartile range (IQR) of 2.1–9.1. Roughly 75.5% of people have at least two blood pressure measurements, and the median number of measurements in people with at least one measurement is 9 (IQR 4–20). The median number of BMI measurements in people with at least one measurement is 3 (IQR 2–7), and similarly the median number of total cholesterol measurements are 3 (IQR 1–6). The mean annual consultation rate at baseline is 7.9.

Biobanks and bespoke cohort linkages

Further linkages have also been established between bespoke investigator led cohort studies and MINAP. Such linkages provide traditional cohort studies with higher-resolution ACS endpoints. Specifically, we have coordinated the linkage between MINAP and (i) the UK Collaborative Trial of Ovarian Cancer Screening, a randomized trial of 200 000 post-menopausal women aged 50–74 years;[36] (ii) Whitehall II, a cohort of ~10 000 civil servants aged 35–55 years;[37] and (iii) UK Biobank, a cohort of 500 000 middle-aged adults with a range of physical measurements and biological samples.[38] Cross-referencing of acute myocardial infarction in four sources is under way.

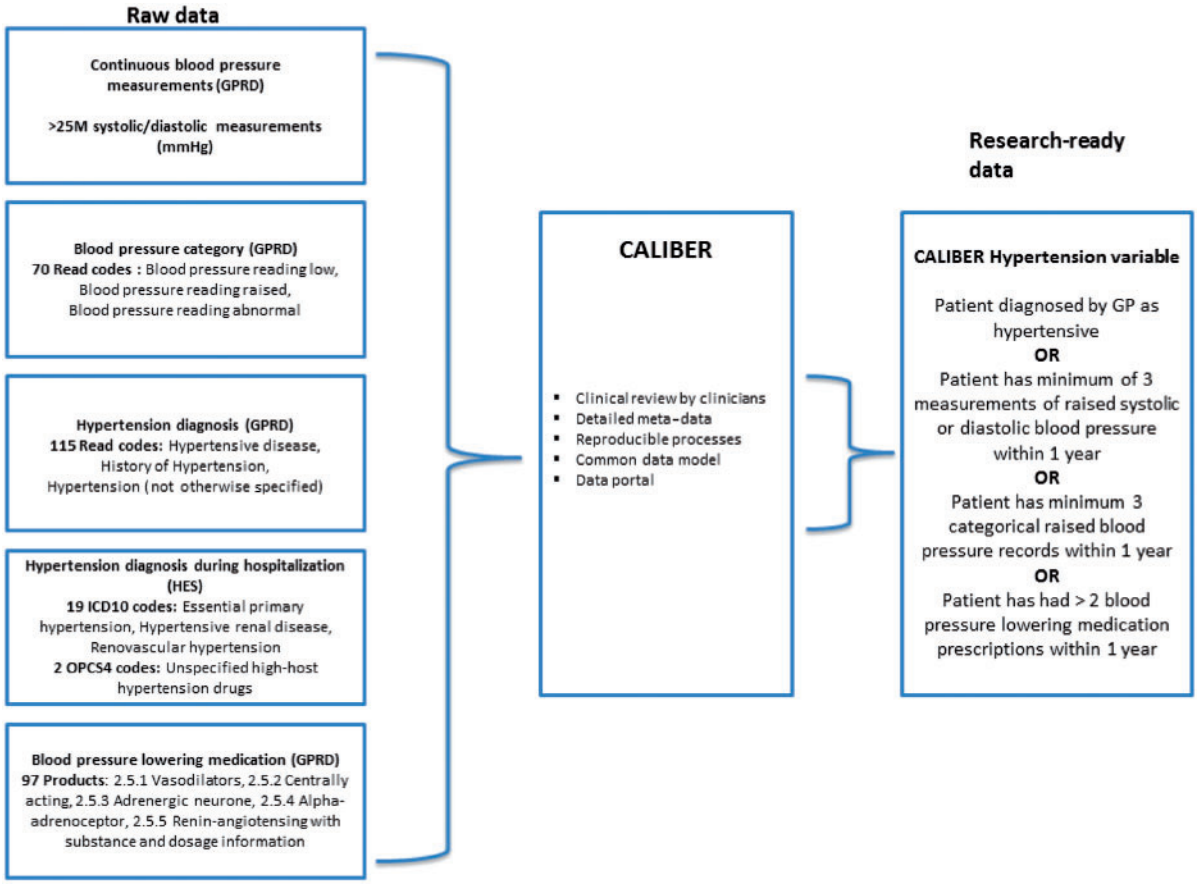

Figure 5 Example of one CALIBER research variable, hypertension, created from multiple raw electronic health record sources. The variable uses a combination of (i) repeat continuous blood pressure measurements; (ii) categorical data on measured blood pressure (over 130 Read codes); (iii) hypertension diagnosis in primary care (over 180 Read codes); and (iv) prescription of blood pressure lowering medications

Data resource use

One important use is to define cohorts using the primary care as denominator population. An example of such a CALIBER cohort (entry between years 2001–10) is shown in Figure 3. The sample consists of 901 629 adults aged $\geqslant$30 years (39.54% men) registered with a CPRD practice, which participates in the data linkage and that have data of acceptable research quality, no prior atherosclerotic disease and complete baseline data on age, sex, blood pressure, smoking, diabetes status and lipid lowering medication information. CALIBER resources are contributing new knowledge at different stages in translational pathways, e.g.:

- Risk factors for initial presentation of CVDs: CALIBER cohorts allow investigation of initial presentation of a wide range of CVDs, which has hitherto not been possible in smaller, less-well clinically phenotyped cohorts. For example smoking has heterogeneous associations with different CVD endpoints, with no association with initial presentation with ventricular arrhythmia or cardiac arrest, a modest association with chronic stable angina and heart failure and strong associations with ACSs.[39]
- Prognosis: A new prognostic model for people with stable coronary artery disease (CAD)[40] has been developed and validated in a population of 100 000 patients with stable angina or who have become stable after an ACS. We show that clinically collected information, including biomarkers, provides significant improvements in discrimination, risk reclassification and clinical cost-effectiveness beyond that achieved based on age, sex, CAD type and clinical history alone.
- Quality of care and outcomes research: Two countries in the world have MI registries, which include data on consecutive patients in all hospitals. In comparisons of over half a million patients in Sweden and the UK, there were important treatment differences, and the 30-day mortality was lower in Sweden.[41]
- Pharmacoepidemiology: Although primary care data have long been used for pharmacoepidemiology, new CALIBER linkages with hospital event data from MINAP and HES allowed CALIBER to contribute new policy-relevant evidence on the potential harms of clopidogrel discontinuation[42] and the lack of robust evidence of interaction between clopidogrel use and proton pump inhibitors.[43]

Future developments

The data resources accessible via CALIBER will grow in CHAPTER in several respects including increasing the scale of population coverage to design and implement randomized trials within the primary care record and within disease and procedure registries (NICOR), to incorporate bioresources allowing genomic approaches to discovery and to use natural language processing of text recorded in primary care. Planned extensions include wider coverage of primary care data (current linked CPRD data cover 8% and aims to include 25% by end of 2012, maximizing the coverage by March 2016) and linkage with wider disease and procedure registries in NICOR. A key next step is to incorporate linkages with integrated hospital records and integrate clinical research with medical care on a large scale by unlocking the full potential of data collected in such settings. Existing platforms such as the National Institute for Health Research Cardiovascular Biomedical Research Unit, an integrated cardiovascular hospital records platform spanning cardiac data, general trust-level systems, radiology and genomics systems for 25 000 patients will be expanded.

Strengths and weaknesses

CALIBER has potential strengths in the degree of clinical, longitudinal phenotyping spanning ambulatory and hospitalized care in large population samples. This allows opportunities for a 'higher resolution' approach to cardiovascular epidemiology in three respects. First, most existing investigator led (bespoke) cardiovascular cohorts study broad aggregates [CVD, coronary heart disease (CHD)], or at most distinguish a narrow range (typically two or three) of different diseases of interest. As the causes, treatment and prognosis vary across a range of disease phenotypes, CALIBER data sets have the statistical size and clinical phenotyping allowing the opportunity to 'resolve' these broad aggregates into distinct disease phenotypes in cerebral, peripheral and coronary circulations. Figure 3 illustrates that cardiac diseases include the following: chronic stable angina, unstable angina, ST elevation myocardial infarction (STEMI), non-ST elevation myocardial infarction (NSTEMI), heart failure, ventricular arrhythmias and cardiac arrest, sudden cardiac death and unheralded coronary death. Use of four sources of data to define endpoints is illustrated in Appendix A. Second, unlike many investigator led cardiovascular cohorts, CALIBER data sets have the temporal resolution to distinguish whether a first event (e.g. heart attack or stroke) was the first manifestation of CVD or whether it was preceded by any of a range of non-fatal (and often diagnosed in primary care) manifestations of CVD. Third, most previous studies have been too small to understand the interplay between different causal domains. For example, even addressing the simple question of gender differences in the smoking effect across different specific cardiovascular endpoints necessitates large samples. CALIBER cohorts have a sufficient number of events and patients on which accurate estimates of risk can be based.

The main weaknesses of any EHR research platform relate to data quality.

Data quality in single sources

Data quality in single sources requires evaluation on a case by case basis with a range of methods: cross-referencing against chart review, against data collected under research conditions and cross-sectional and prospective tests of validity.

Missing data

Although some measurements, such as BMI, are reasonably complete (82.6% of people with at least one BMI measurement), others, such as total cholesterol, are less complete (44.9% with at least one total cholesterol measurement). Developments in imputation of missing data within such primary care longitudinal cohorts are promising and might offer a partial solution to the problem.[44] Variation in coding used across practices and over time, as well the use of broader diagnostic categories used (e.g. CHD Not Otherwise Specified),[45] may prove more difficult to resolve.

Conflicting data

Fatal MI may be recorded in up to four different sources, which differ in their diagnostic granularity and timing accuracy. Multiple records at similar time points may reflect the same event or subsequent events—e.g. an MI is recorded in CPRD, and, after 30 days in MINAP, could point to a single or recurrent event (E Herrett, personal communication).

Linkage quality

There is empirical evidence[46] that errors in data linkage can lead to significantly different conclusions about the associations of risk factors with outcomes. The linkage method used for linking CALIBER constituent data sets has been used before, and it has been shown to yield reasonably high quality matches.

Data resource access

Raw data are available for use by researchers subject to approval of the protocol by, and payment to, the bodies governing access to the constituent data sets. For CPRD, this involves scientific approval of the protocol by the Independent Scientific Advisory Committee and a signed licence outlining scope and data confidentiality of use of CPRD data. For MINAP, applications are made to the MINAP Academics Group, and to the NHS Information Centre[47] for HES and ONS. Following such approvals, an application to the CALIBER Scientific Advisory Committee should be submitted.

CALIBER researchers register their study and publish their protocol in the public domain before receiving data.[48] Study registration and protocol publication are of complementary nature; registration records present short, standardized elements of the full protocol in an accessible manner.

CALIBER welcomes collaborative research. Access to data for external researchers (those who are not affiliated with CALIBER investigators) is provided within the CALIBER 'safe haven' environment, which currently requires researchers to be physically based in either UCL (Clinical Epidemiology Group) or the London School of Hygiene and Tropical Medicine (Smeeth). Given the diverse clinical origins, complexity and size of the data sets, training with the CALIBER team on data sources, coding, quality and management is available.

The CALIBER data portal is available for consultation online at http://www.caliberresearch.org/. Existing and potential collaborators are encouraged to provide feedback on potential enhancements to the portal to facilitate the wider use of these data resources.

For more information, visit http://www.caliberresearch.org/ or contact Dr Spiros Denaxas at s.denaxas@ucl.ac.uk.

Acknowledgements

Tjeerd van Staa coordinated linkages involving the Clinical Practice Research Datalink; David Cunningham, National Institute for Cardiovascular Outcomes Research, coordinated linkages involving MINAP; Marina Daskalopoulou Clinical Epidemiology Group, University College London.

Funding

CALIBER is funded by a Wellcome Trust project grant (086091/Z/08/Z), a National Institute of Health Research (NIHR) programme grant (RP-PG-0407-10314) and a consortium of 10 UK government and charity funders, led by the Medical Research Council (MRC), which funds the Centre for Health service and Academic Partnership in Translational Electronic health records Research (CHAPTER). The funders are: Cancer Research UK (CRUK), Chief Scientist Office, Scottish Government Health Directorates (CSO), Engineering and Physical Sciences Research Council (EPSRC), Economic and Social Research Council (ESRC), National Institute for Health Research (NIHR), National Institute for Social Care and Health Research (NISCHR) and The Wellcome Trust.

A.D.H., A.T., and H.H. are partners in the Medical Research Council PROGnosis RESearch Strategy (PROGRESS) Partnership (www.progress-partnership.org), supported by the Medical Research Council [grant number G0902393/1] which supports research using CALIBER. M.K. is supported by an Economic and Social Research Council (ESRC) professorial fellowship. L.S. is supported by a Senior Clinical Fellowship from the Wellcome Trust. A.S. is supported by a Wellcome Trust Clinical Research Training Fellowship [0938/30/Z/10/Z] E.H. is supported by an MRC studentship, and J.G. was funded by a NIHR Doctoral Fellowship [DRF-2009-02-50].

The views and opinions expressed therein are those of the authors and do not necessarily reflect those of the NIHR PHR Programme or the Department of Health.

Conflict of interest: None declared.

KEY MESSAGES

CALIBER data resources provide a common framework for addressing questions of:

- Risk factors for initial presentation of a wide range of pathologically diverse CVDs: We have shown that smoking has heterogeneous associations with different CVD endpoints.
- Prognosis: We have developed a new prognostic model for people with stable CAD, which makes use of clinically collected information, including biomarkers.
- Quality of care and outcomes research: We observed significant treatment differences and lower 30-day mortality in Sweden when comparing over half a million MI patients with the UK ACS registry.
- Pharmacoepidemiology: We have illustrated the potential harms of clopidogrel discontinuation and the lack of robust evidence of interaction between clopidogrel use and proton pump inhibitors.

References

[1] UK Clinical Research Collaboration. (UKCRC) Advisory Group to Connecting for Health. *Report of Research Simulations*. London: UKCRC, 2007.

[2] Medical Research Council (MRC). *UK E-Health Records Research Capacity and Capability*. London: MRC, 2011.

[3] Department of Health. *The Power of Information: Putting All of Us in Control of the Health and Care Information We Need*. London: Department of Health, 2012.

[4] Department of Business Innovation & Skills. *Strategy for UK Life Sciences*. London: Department of Business, Innovation and Skills, 2011.

[5] Timmis AD, Feder G, Hemingway H. Prognosis of stable angina pectoris: why we need larger population studies with higher endpoint resolution. *Heart* 2007;**93:**786–91.

[6] Jernberg T, Attebring MF, Hambraeus K *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:**1617–21.

[7] Lindhardsen J, Ahlehoff O, Gislason GH *et al*. Risk of atrial fibrillation and stroke in rheumatoid arthritis: Danish nationwide cohort study. *BMJ* 2012;**344:**e1257.

[8] Sørensen R, Hansen ML, Abildstrom SZ *et al*. Risk of bleeding in patients with acute myocardial infarction treated with different combinations of aspirin, clopidogrel, and vitamin K antagonists in Denmark: a retrospective analysis of nationwide registry data. *Lancet* 2009;**374:** 1967–74.

[9] Gershon AS, Warner L, Cascagnette P, Victor JC, To T. Lifetime risk of developing chronic obstructive pulmonary disease: a longitudinal population study. *Lancet* 2011;**378:** 991–96.

[10] St Sauver JL, Jacobson DJ, McGree ME *et al*. Associations between longitudinal changes in serum estrogen, testosterone, and bioavailable testosterone and changes in benign urologic outcomes. *Am J Epidemiol* 2011;**173:**787–96.

[11] The Research Program on Genes, Environment, and Health. http://www.dor.kaiser.org/external/DORExternal/rpgeh/index.aspx (14 November 2012, date last accessed).

[12] Intermountain Healthcare Cardiovascular Research. http://intermountainhealthcare.org/services/heart/research/pages/home.aspx (14 November 2012, date last accessed).

[13] Jang MJ, Bang SM, Oh D. Incidence of pregnancy-associated venous thromboembolism in Korea: from the Health Insurance Review and Assessment Service database. *J Thromb Haem* 2011;**9:**2519–21.

[14] Colhoun HM. Use of insulin glargine and cancer incidence in Scotland: a study from the Scottish Diabetes Research Network Epidemiology Group. *Diabetologia* 2009;**52:**1755–65.

[15] Scottish Health Informatics Programme. *A Blueprint for Health Records Research in Scotland*. Dundee: Scottish Health Informatics Programme, 2011.

[16] Ford DV, Jones KH, Verplancke JP *et al*. The SAIL Databank: building a national architecture for e-health research and evaluation. *BMC Health Serv Res* 2009;**9:**157.

[17] Lyons RA, Jones KH, John G *et al*. The SAIL databank: linking multiple health and social care datasets. *BMC Med Inf Dec Mak* 2009;**9:**3.

[18] Institute for Clinical Evaluative Sciences (ICES), http://www.ices.on.ca/ (14 November 2012, date last accessed).

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

[20] Westfall JM, Mold J, Fagnan L. Practice-based research–''Blue Highways'' on the NIH roadmap. *JAMA* 2007;**297:** 403–06.

[21] International Organization for Standardization (ISO). *Health Informatics - Good Principles and Practices for a Clinical Data Warehouse* (ISO/TR 22221:2006). Geneva: ISO, 2006.

[22] Ioannidis JP. Why most published research findings are false. *PLoS Med* 2005;**2:**e124.

[23] Walley T, Mantgani A. The UK General Practice Research Database. *Lancet* 1997;**350:**1097–99.

[24] Herrett E, Smeeth L, Walker L, Weston C. The Myocardial Ischaemia National Audit Project (MINAP). *Heart* 2010; **96:**1264–67.

[25] Centre TH and SCI. *Hospital Episodes Statistics (HES)*. 2011. Available from: http://www.hesonline.nhs.uk/Ease/servlet/ContentServer?siteID=1937 (14 November 2012, date last accessed).

[26] Office for National Statistics. *Mortality Statistics: Metadata 2010 Statistics*. London, 2011; (14 November 2012, date last accessed).

[27] Simon C. Overview of the GP contract. *InnovAiT* 2008;**1:** 134–39.

[28] Chisholm J. The Read clinical classification. *BMJ* 1990; **300:**1092.

[29] International Health Terminology Standards Development Organization. *Systematized Nomenclature of Medicine—Clinical Terms* (SNOMED-CT) [Internet]. http://www.ihtsdo.org/snomed-ct/ (14 November 2012, date last accessed).

[30] ICD. *International Classification of Diseases, 10th Revision, Clinical Modification* (ICD-10-CM). 2011 release. Available from: http://www.cdc.gov/nchs/icd/icd10cm.htm (14 November 2012, date last accessed).

[31] OPCS-4 Classification — NHS Connecting for Health. Available from: http://www.connectingforhealth.nhs.uk/systemsandservices/data/clinicalcoding/codingstandards/opcs4/index_html (27 November 2012, date last accessed).

[32] Noble M, Mclennan D, Wilkinson K, Whitworth A. *The English Indices of Deprivation 2007*. London: Communities, 2007.

[33] Townsend P. *Health and Deprivation: Inequality and the North*. London: Routledge, 1988.

[34] Dave S, Petersen I. Creating medical and drug code lists to identify cases in primary care databases. *Pharm Drug Saf* **18:**704–07.

[35] Data Documentation Initiative (DDI) website. http://www.ddialliance.org/Specification/ (14 November 2012, date last accessed).

[36] UK Collaborative Trial of Ovarian Cancer Screening (UKCTOCS), 2012. http://www.ukctocs.org.uk (14 November 2012, date last accessed).

[37] Marmot M, Brunner E. Cohort Profile: the Whitehall II study. *Int J Epidemiol* 2005;**34:**251–56.

[38] Manolio TA, Weis BK, Cowie CC *et al*. New models for large prospective studies: is there a better way? *Am J Epidemiol* 2012;**175:**859–66.

[39] George J, Herrett E, Denaxas S *et al*. *Differential Effects of Smoking on Specific Cardiovascular Presentations in Men and Women: Prospective Cohort Study in 900,000 Patients Using CALIBER Linked Electronic Health Records*. Los Angeles: American Heart Association Scientific Sessions, 2012.

[40] Rapsomaniki E, Shah AD, Denaxas S *et al*. *Prognostic Models for People with Stable Coronary Artery Disease Based on 115,500 Patients from the CALIBER Study*. Munich: European Society of Cardiology (ESC), 2012.

[41] Chung SC, Gedeborg R, Nicholas O *et al*. *Comparative Effectiveness of Acute Myocardial Infarction Care Delivered in Sweden and the United Kingdom Using National Outcome Registries*. Los Angeles: American Heart Association Scientific Sessions, 2012.

[42] Boggon R, van Staa TP, Timmis A *et al*. Clopidogrel discontinuation after acute coronary syndromes: frequency, predictors and associations with death and myocardial infarction–a hospital registry-primary care linked cohort (MINAP-GPRD). *Eur Heart J* 2011;**32:**2376–86.

[43] Douglas IJ, Evans SJW, Hingorani AD *et al*. Clopidogrel and interaction with proton pump inhibitors: comparison between cohort and within person study designs. *BMJ* 2012;**345:**e4388.

[44] Marston L, Carpenter JR, Walters KR, Morris RW, Nazareth I, Petersen I. Issues in multiple imputation of missing data for large general practice clinical databases. *Pharmacoepidemiol Drug Saf* 2010;**19:**618–26.

[45] Bhattarai N, Charlton J, Rudisill C, Gulliford MC. Coding, recording and incidence of different forms of coronary heart disease in primary care. *PLoS ONE* 2012;**7:**e29776.

[46] Lariscy JT. Differential record linkage by Hispanic ethnicity and age in linked mortality studies: implications for the epidemiologic paradox. *J Aging and Health* 2011;**23:** 1263–84.

[47] NHS Information Centre website. http://www.ic.nhs.uk/ (14 November 2012, date last accessed).

[48] CALIBER. *Cardiovascular Disease Research using Linked Bespoke Studies and Electronic Records Data Portal*. 2010. http://caliberresearch.org/ (14 November 2012, date last accessed).

Appendix A Overview of codes used to define a range of common CVDs. Details of how these codes are combined are given in the data portal. We provide a list of indicative diagnostic codes used across the data sources and not an exhaustive list

	Primary care	Disease Registry	Hospital procedures	Hospital diagnoses[a]	Causes of death[b]
	CPRD	MINAP	HES	HES	ONS
Endpoint	Read codes	Registry specific	OPCS-4	ICD-10	ICD-10
Acute myocardial infarction	G30X000 Acute ST segment EMI G307100 Acute non-ST segment EMI G30.14 Heart attack, G30.15 MI Acute myocardial infarction + 60 other codes as acute myocardial infarction Not Otherwise Specified	MI with or without ST elevation based on initial ECG findings, raised troponins and clinical diagnosis	Not used (there is no code that is specific to primary coronary intervention)	Acute myocardial infarction I21, Current complications of acute myocardial infarction I23	Acute myocardial infarction I21, Current complications of acute myocardial infarction I23
Unstable angina	G311.13/G311100 Unstable angina, G233200 Angina at rest, G311400 Worsening angina + 13 other codes	Discharge diagnosis of unstable angina, no raised ST elevationNo raised troponin levels	nu	Unstable or worsening angina I20.0 Acute ischaemic heart disease I24, Coronary thrombosis not resulting in myocardial infarction I24.0, Other forms of ischaemic heart disease I24.8, Acute ischaemic heart disease, unspecified I24.9	nu
Stable angina	G33..00 Stable Angina, G33z.00 Angina pectoris NOS + 25 other codes for diagnosis of stable angina pectoris 30 codes for evidence of coronary artery disease at angiography (CT,MR, invasive or not specified) 151 Read codes for evidence of myocardial ischaemia (Resting ECG, exercise ECG, stress echo, radioisotope scan) Two or more successive prescriptions for anti-anginals	nu	Coronary Artery Bypass Graft (CABG) K40-K46 or Percutaneous Coronary Intervention (PCI) K49,K50,K75, not within 30 days of an ACS	Stable angina pectoris I20 excluding unstable angina I20.0	nu
Coronary heart disease not otherwise specified	G3. . .00 Ischaemic Heart Disease + 90 other codes including CHD NOS, chronic ischaemic heart disease, silent myocardial infarction	nu	nu	CHD NOS, chronic ischaemic heart disease, silent myocardial infarction I25 excluding I25.2, old myocardial infarction	nu
Heart failure	G58.00 Heart Failure + 92 other Read codes for heart failure diagnosis	nu	nu	I50 Heart failure(including all sub, I11.0 Hypertensive heart disease with (congestive) heart failure, I13.0 Hypertensive heart and renal disease with (congestive) heart failure, I13.2 Hypertensive heart and renal disease with both (congestive) heart failure and renal disease	I50 Heart failure I11.0 Hypertensive heart disease with (congestive) heart failure, I13.0 Hypertensive heart and renal disease with (congestive) heart failure, I13.2 Hypertensive heart and renal disease with both (congestive) heart failure and renal disease
Ventricular arrhythmias, cardiac arrest and sudden cardiac death	G574.00 Ventricular fibrillation and flutter, G757.00 Cardiac arrest + 35 other Read codes for ventricular fibrillation, asystole, cardiac arrest, cardiac resuscitation, electro-mechanical dissociation, G575100 Sudden cardiac death, so described	nu	Implanted cardiac defibrillation device X50, Implantation, revision and renewal of cardiac defibrillator K59	I46 (cardiac arrest) I47.0 (re-entry ventricular arrhythmia) I47.2 (ventricular tachycardia)	I46 (cardiac arrest, includes I46.1 sudden cardiac death) I47.0 (re-entry ventricular arrhythmia) I47.2 (ventricular tachycardia)
Unheralded coronary death	Any CVD excluded	Any CVD excluded	Any CVD excluded	Any CVD excluded	I20 Angina Pectoris, I21 Acute myocardial infarction, I22 Subsequent myocardial infarction, I23 Certain current complications following acute myocardial infarction, I24 Other acute ischaemic heart diseases and I25 Chronic ischaemic heart disease not preceded by any other CVD presentation

(continued)

Appendix A Continued

Endpoint	Primary care CPRD Read codes	Disease Registry MINAP Registry specific	Hospital procedures HES OPCS-4	Hospital diagnoses[a] HES ICD-10	Causes of death[b] ONS ICD-10
Ischaemic stroke	G64.11 CVA – cerebral artery occlusion, G64.13 Stroke due to cerebral arterial occlusion, G6W.00 Cereb infarct due unspecified occlusion/stenosis of pre-cerebral arteries, G6X.00 cerebral infarction due to unspecified occlusion or stenosis of cerebral arteries plus 8 other codes	nu	nu	I63 cerebral infarction	I63 cerebral infarction
Haemorrhagic stroke	93 codes for subarachnoid haemorrhage, intracerebral haemorrhage, and intracranial haemorrhage not otherwise specified	nu		I60 Subarachnoid haemorrhage, I61.0 Intracerebral haemorrhage in hemisphere, subcortical, I61.1 Intracerebral haemorrhage in hemisphere, cortical, I61.2 Intracerebral haemorrhage in hemisphere unspecified, I61.3 Intracerebral haemorrhage in brain stem, I61.4 Intracerebral hemorrhage in cerebellum, I61.5 Intracerebral haemorrhage intraventricular, I61.6 Intracerebral haemorrhage, multiple localized, I61.8 Other intracerebral haemorrhage, I61.9 Intracerebral haemorrhage	nu
Peripheral arterial disease	72 codes for Lower limb peripheral arterial disease diagnosis (including diabetic peripheral arterial disease, gangrene and intermittent claudication Evidence of atherosclerosis of iliac and lower limb arteries based on angiography or Dopplers	nu	L50-L54 Bypass, reconstruction and other open operations on iliac artery L58-L60, L62 Bypass, reconstruction, transluminal operations or other open operations of femoral artery, L65 Revision of reconstruction of artery	I70.2 atherosclerosis of arteries of extremities, I73.9 peripheral vascular disease intermittent claudication Peripheral complications of diabetes including gangrene 0.5 suffix of E10 Insulting dependent diabetes mellitus, E11 Non-insulin-dependent diabetes mellitus, E12 Malnutrition-related diabetes mellitus, E13 Other specified diabetes mellitus, E14 Unspecified diabetes mellitus	I70.2 atherosclerosis of arteries of extremities, I73.9 peripheral vascular disease intermittent claudication, Peripheral complications of diabetes including gangrene 0.5 suffix of E10 Insulting dependent diabetes mellitus, E11 Non-insulin-dependent diabetes mellitus, E12 Malnutrition-related diabetes mellitus, E13 Other specified diabetes mellitus, E14 Unspecified diabetes mellitus
Abdominal aortic aneurysm (AAA)	G714.00 Abdominal aortic aneurysm without mention of rupture + 11 more codes for AAA diagnosis. 13 codes for evidence of AAA on ultrasound or CT scan	nu	L16 Extra anatomic bypass of aorta, L18-L23 Replacement of aneurysmal segment of aorta, bypass of segment of aorta, plastic repair of aorta, L25-L28 Transluminal or endovascular insertion of stent on aneurysmal segment of aorta	I71.3 Abdominal aortic aneurysm, ruptured. 171.4 AAA, without rupture	I71.3 Abdominal aortic aneurysm, ruptured. 171.4 AAA, without rupture

[a]Primary cause of admission.
[b]Underlying cause of death.
Nu, not used in definition.