Linked electronic health records for research on a nationwide cohort including over 54 million people in England

CVD-COVID-UK consortium*

3 tables, 3 figures

6 Supplementary tables

Key Words: Electronic Health Records, England, administrative health datasets, nationwide, population-scale, trusted research environment, NHS Digital, primary care, hospital admissions, Covid-19

Word count = 4342

***Contributions to this manuscript:**

Manuscript drafting and revising: Angela Wood (writing committee chair) [4,5,8,13,19], Rachel Denholm[2,10,14], Sam Hollings[16], Jennifer Cooper[2,10,14], Samantha Ip[4], Venexia Walker[7,12], Spiros Denaxas[3,11,15,19], Ashley Akbari[18], Jonathan Sterne†[2,10,14], Cathie Sudlow†[9,21]

†equal contributions

Data wrangling, QA and analysis (including generating phenotype definitions): Angela Wood (chair), Rachel Denholm, Sam Hollings, Jennifer Cooper, Samantha Ip, Venexia Walker, Spiros Denaxas, Amitava Banerjee[1,11,20], William Whiteley[6,17]

Figures/graphics: Alvina Lai[11]

Consortium coordination (BHF Data Science Centre core team): Rouven Priedon[9], Cathie Sudlow[9], Lynn Morrice[9], Debbie Ringham[9]

Consortium membership: https://www.hdruk.ac.uk/wp-content/uploads/2021/01/210128-CVD-COVID-UK-Consortium-Members.pdf and see Annexe 1 – Consortium members, who contributed to discussions leading up to this manuscript and provided very helpful insights and comments

Public and patient advisory panel: Suzannah Power, Lynn Laidlaw, Michael Molete, John Walsh

NHS Digital (coordination, data management/provision, data access request and information governance support, Trusted Research Environment support: Garry Coleman[16], Cath Day[16], Elizabeth Gaffney[16], Tim Gentry[16], Lisa Gray[16], Sam Hollings[16], Richard Irvine[16], Brian Roberts[16], Estelle Spence[16], Janet Waterhouse[16]

Correspondence: bhfdsc@hdruk.ac.uk (Jonathan Sterne, Cathie Sudlow, Angela Wood)

Affiliations:

[1]Barts Health NHS Trust, The Royal London Hospital, Whitechapel Rd, London, UK;
[2]Bristol Medical School: Population Health Sciences, University of Bristol, Bristol, UK;
[3]British Heart Foundation Research Accelerator, University College London, UK;
[4]British Heart Foundation Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK;
[5]British Heart Foundation Centre of Research Excellence, University of Cambridge, Cambridge, UK;
[6]Centre for Clinical Brain Sciences, University of Edinburgh, Edinburgh, UK;
[7]Department of Surgery, University of Pennsylvania Perelman School of Medicine, Philadelphia, USA
[8]Health Data Research UK Cambridge, Wellcome Genome Campus and University of Cambridge, Cambridge, UK;
[9]BHF Data Science Centre, Health Data Research UK, Gibbs Building, London, UK;
[10]Health Data Research UK, South West Better Care Partnership, Bristol, UK;
[11]Institute of Health Informatics, University College London, 222 Euston Road, London, UK;
[12]MRC University of Bristol Integrative Epidemiology Unit, Bristol, UK; Bristol Medical School: Population Health Sciences, University of Bristol, Bristol, UK;
[13]National Institute for Health Research Blood and Transplant Research Unit in Donor Health and Genomics, University of Cambridge, Cambridge, UK;
[14]National Institute for Health Research Bristol Biomedical Research Centre, University of Bristol, UK;
[15]National Institute for Health Research University College London Hospitals Biomedical Research Centre, University College London, UK;
[16]NHS Digital, 1 Trevelyan Square, Leeds, UK
[17]Nuffield Department of Population Health, University of Oxford, Oxford, UK;
[18]Population Data Science and Health Data Research UK, Swansea University, UK;
[19]The Alan Turing Institute, London, UK;
[20]University College London Hospitals NHS Trust, 235 Euston Road, London, UK;
[21]Usher Institute, Edinburgh Medical School, The University of Edinburgh, Edinburgh, UK

ABSTRACT [Word count = 350]

Objectives – Describe a new England-wide electronic health record (EHR) resource enabling whole population research on Covid-19 and cardiovascular disease whilst ensuring data security and privacy and maintaining public trust.

Design – Cohort comprising linked person-level records from national healthcare settings for the English population accessible within NHS Digital's new Trusted Research Environment.

Setting – EHRs from primary care, hospital episodes, death registry, Covid-19 laboratory test results and community dispensing data, with further enrichment planned from specialist intensive care, cardiovascular and Covid-19 vaccination data.

Participants - 54.4 million people alive on 1st January 2020 and registered with an NHS general practitioner in England.

Main measures of interest – Confirmed and suspected Covid-19 diagnoses, exemplar cardiovascular conditions (incident stroke or transient ischaemic attack (TIA) and incident myocardial infarction (MI)) and all-cause mortality between 1st January and 31st October 2020.

Results – The linked cohort includes over 96% of the English population. By combining person-level data across national healthcare settings, data on age, sex and ethnicity are complete for over 95% of the population. Among 53.2M people with no prior diagnosis of stroke/TIA, 98,721 had an incident stroke/TIA, of which 30% were recorded only in primary care and 4% only in death registry records. Among 53.1M people with no prior history of MI, 62,966 had an incident MI, of which 8% were recorded only in primary care and 12% only in death records. A total of 959,067 people had a confirmed or suspected Covid-19 diagnosis (714,162 in primary care data, 126,349 in hospital admission records, 776,503 in Covid-19 laboratory test data and 48,433 participants in death registry records). While 58% of these were recorded in both primary care and Covid-19 laboratory test data, 15% and 18% respectively were recorded in only one.

Conclusions – This population-wide resource demonstrates the importance of linking person-level data across health settings to maximize completeness of key characteristics and to ascertain cardio-

vascular events and Covid-19 diagnoses. Although established initially to support research on Covid-19 and cardiovascular disease to benefit clinical care and public health and to inform health care policy, it can broaden further to enable a very wide range of research.

INTRODUCTION

The Covid-19 pandemic has increased awareness of the importance of population-wide person-level electronic health record (EHR) data from a range of sources for examining, modelling and reporting disease trends to inform healthcare and public health policy[1]. Key benefits of research using such data on nationwide cohorts include: (i) generalisability of findings across all age groups, ethnicities, geographical locations and socioeconomic, health and personal characteristics and (ii) inclusion of very large numbers of people and events, enhancing the precision of findings and enabling a wide spectrum of novel research studies (e.g., characterising shapes of relationships between risk factors and disease or studying minority sub-populations and rare disease sub-types) . Whilst EHRs for whole country cohorts for Wales, Scotland, Denmark and Sweden (populations approximately 3 to 10 million) have been used for research for several years,[2,3,4,5,6] at the start of the COVID-19 pandemic, there was no access for bona fide researchers to national linked healthcare data across the population of England to enable critical research to support healthcare decisions and public health policy. There were two main reasons for this: there was no national collection of comprehensive, linkable primary care data; and there was no secure, privacy-protecting mechanism for researchers to access and conduct population-wide research using national datasets linked across different parts of the health data system (from primary care, hospitals, death registries, laboratories etc). EHR research in England to date has, therefore, not been able to take advantage of the statistical power of studying a population of almost 60 million people, while clinical, public health and policy insights have directly represented only a subset of the population. Hence, there remains a need for accessible, nationwide health data in England for research, whilst ensuring participant safety and maintaining public trust.

Motivated by the public health importance of fully understanding the relationship between Covid-19 and cardiovascular disease (CVD), the British Heart Foundation (BHF) Data Science Centre[7] established the CVD-COVID-UK initiative[8] to partner with NHS Digital[9] in the development and secure provision for approved research of linked, nationally collated EHRs for the whole population of the UK. Here we describe key features of the new English component of this effort: a nationwide linked health data resource, provided within a new Trusted Research Environment (TRE) for England. We use descriptive analyses of the currently available data to illustrate the importance for whole population research studies of linking EHRs from across different health settings.

METHODS

Data resources

The newly established NHS Digital TRE for England provides secure, remote access for researchers to linked, person-level EHR data from national health settings. The data sources currently available include primary care data, hospital episodes (covering inpatient, outpatient, emergency department and critical care episodes), registered deaths (including cause of death), Covid-19 laboratory tests and community dispensed medicines (Table 1; CVD-COVID-UK Dataset dashboard;[10] CVD-COVID-UK Dataset TRE asset in Health Data Research Innovation Gateway)[11]. Further incorporation of specialist intensive care, cardiovascular audit and Covid-19 vaccination data is planned in the near future. Datasets from each source include the same set of unique person-level master keys (or pseudo-identifiers) to enable linkage of peoples' records between datasets.

Data linkage

Linkage between datasets is enabled by NHS Digital's *Master Person Service,*[12] which uses a four-stage algorithm to match multiple records for each person from different clinical computer systems (e.g., hospitals and general practices) to a single unique identifier, the National Health Service (NHS) number representing a single person. The algorithm verifies and cross-checks the NHS numbers with associated demographic details including age, gender and postcode.[13]

Data resource access – NHS Digital TRE for England

On behalf of the CVD-COVID-UK consortium, the BHF Data Science Centre requested access to the data sources via the NHS Digital online Data Access Request Service[14] and received approval for the CVD-COVID-UK research programme (Ref no: DARS-NIC-381078-Y9C5K) following discussion with NHS Digital's Independent Group Advising on the Release of Data (IGARD).[15] A data sharing agreement with NHS Digital allows approved researchers based in UK research organisations (universities and NHS bodies) that jointly sign this agreement to access the data held within the NHS Digital TRE service for England.[16] The BHF Data Science Centre coordinates an Approvals and Oversight Board (including representation from NHS Digital, participating research organisations and lay members) that ensures research projects undertaken fall within scope of the ethical and regulatory approvals for the CVD-COVID-UK consortium programme. The TRE provides secure storage and remote data access, avoiding the need for any person-level data to leave NHS Digital (Figure 1). An expanding suite of tools (currently including SQL, Python and R Studio) supports data management, visualization and analysis.

CVD-COVID-UK Consortium: Aims, Membership and Principles

The CVD-COVID-UK consortium aims to use analyses of UK population-wide linked EHR data to investigate: the effects of cardiovascular diseases, their risk factors and medications on susceptibility to and poor outcomes from Covid-19; the direct impact of SARS-CoV-2 infection on acute cardiovascular complications and longer-term cardiovascular risk; and the indirect impact of the pandemic on the presentation, diagnosis, management and outcomes of cardiovascular diseases.[8] Lay summaries of approved projects are published on the consortium's web page[8]. All consortium members (currently over 130 people from around 40 research or NHS organisations, including NHS data custodians) commit to: conducting research according to the 'Five Safes';[17] an inclusive approach that enables additional researchers to join the consortium as the work evolves; and the open sharing of research protocols and analysis code (via the BHF Data Science Centre Github repository)[18] and of phenotype code lists and algorithms (via the HDR UK Phenotype Library).[19]

Data updates

The datasets within the TRE are updated regularly from NHS Digital's internal systems (between daily and fortnightly depending on the dataset) and have a variable lag behind real time at the point of update (Table 2). The datasets are currently refreshed on a synchronized monthly schedule, but more frequent updates (e.g. daily or weekly) can be requested according to clinical, public health and health policy research needs.

Data security, privacy and confidentiality

The data within the TRE are de-identified (i.e., directly identifying data items, such as each person's name, address, NHS number and exact date of birth, are removed) and pseudonymised (i.e., each unique person-specific NHS number is replaced with a non-identifying unique master key). Postcodes are replaced with lower layer super output areas which can be converted to indices of multiple deprivation.[20] Further, NHS Digital operates a 'safe outputs' service: only summary, aggregate results can be extracted from the TRE by approved researchers, subject to approval through disclosure control processes and rules, following similar principles to those used by other established TREs, such as the Secure Anonymised Information Linkage (SAIL) Databank for Wales[21,22] and the Scottish National Data Safe Haven.[23] This ensures that no output that might be placed in the public

domain contains information that could be used either on its own or in conjunction with other data to identify a person.

Ethical approval

The North East - Newcastle & North Tyneside 2 Research Ethics Committee provided ethical approval for the CVD-COVID-UK research programme (REC number: 20/NE/0161).

Patient and Public Involvement

The lay panel of the UK National Institutes for Health Research-BHF Cardiovascular Partnership reviews the CVD-COVID-UK programme every few months and provides feedback that informs ongoing and future research. In addition, lay people directly affected by cardiovascular disease are members of the consortium and its Approvals and Oversight Board, enabling co-generations of research ideas and providing valuable perspective and input on research proposals, lay summaries and research outputs.

Derivation of participant characteristics and disease diagnoses

For descriptive analyses, we defined a linked cohort including all people in the primary care data known to be alive on 1st January 2020, excluding those who had either died before or were born on or after that date (as recorded in the death registry and in the primary care records, respectively). We censored follow-up on 31st October 2020, the latest record date common across the datasets. We defined eligible records within the hospital episodes, death registry and Covid-19 laboratory test results as those which could be linked by their unique master key to a person included in the primary care data.

We combined primary care and hospital episodes records (covering inpatient, outpatient, emergency department and critical care episodes) from before the index date of 1st January 2020 to define key characteristics, including sex, age and ethnicity (categorised into White; Mixed; Asian and Asian British, Black and Black British and other ethnic groups). For each characteristic, we extracted the most recent record from the primary care data if available, otherwise we used the most recent record from the hospital episodes records. Characteristics were classified as "unknown" for people with no records. Using previously validated phenotypes from the CALIBER resource,[24] we defined previous diagnoses of MI (yes/no), stroke/TIA (defined as ischaemic stroke, haemorrhagic stroke,

unspecified stroke or TIA) (yes/no), diabetes (yes/no) and obesity (yes/no) from Systematized Nomenclature of Medicine Clinical Terms (SNOMED-CT) concept codes in the primary care data and from ICD-10 codes in the hospital episodes (main or secondary diagnostic code position in the admitted patient care component of the hospital episode statistics data) recorded prior to 1st January 2020. For primary care phenotypes, we translated and expanded the phenotypes defined in Read Terms V2 to SNOMED-CT and cross referenced them with codes in the primary care dataset.[25] Two clinicians independently reviewed all phenotype code lists and, where applicable, classified ICD-10 terms and SNOMED-CT concepts into prevalent or incident (**Supplementary Tables 1-4**).[19]

We ascertained people with a first-ever incident MI or stroke/TIA as those with no diagnosis of MI or stroke/TIA (as defined above) prior to 1st January 2020 and with a diagnosis SNOMED-CT or ICD-10 code appearing in the primary care data, hospital episodes (main or secondary diagnostic code position in the admitted patient care component of the hospital episode statistics data) or death registry (underlying or contributing cause of death) between 1st January and 31st October 2020 (phenotype algorithms provided in **Supplementary Tables 1-2**).[19]

We ascertained people with a confirmed or suspected Covid-19 diagnosis as follows: (i) a positive PCR or antigen test from the Covid-19 laboratory test data, with specimen date on or before 31st October 2020; or (ii) a Covid-19 diagnosis SNOMED-CT concept code appearing in the primary care data, with event date on or before 31st October 2020; or (iii) a diagnosis ICD-10 code appearing in the hospital episodes (main or secondary diagnostic code position in the admitted patient care component of the hospital episode statistics), with admission date on or before 31st October 2020 or (iv) death registration including a diagnosis ICD-10 code (as underlying or contributing cause), with date of death on or before 31st October 2020. All Covid-19 phenotype definitions are provided in **Supplementary Table 5**.[19]

We followed the RECORD guidance in preparing this manuscript (Annexe 2).[26]

RESULTS

Overview of data resources

Table 2 provides an overview of the currently available primary care, hospital episodes, death registrations, Covid-19 test data and community dispensing data sources.

The primary care dataset includes healthcare information coded with SNOMED-CT concepts for all people registered with an English NHS general practice (excluding around 1.3 million people with a registered objection to their general practice records being provided to NHS Digital).[25] It includes data from 98% of all English general practices across all relevant general practice computer system suppliers (TPP, EMIS, InPracticeSystems and Microtest) and holds approximately 4.9 billion records on 54.4 million people alive on 1st January 2020 (over 96% of the total population of England based on the UK Office for National Statistics mid-2019 population estimate for England of 56,286,961).[27] Around 34,000 SNOMED codes are included (over 90% of all those currently extracted for a wide range of purposes by NHS Digital's GP Extraction Service), covering a broad range of diagnoses and procedures (from as far back in time as records exist) along with laboratory results, physical measurements, clinical referrals and prescriptions. Of note, while over 900,000 SNOMED codes are listed in UK and international releases, large numbers of these are either inactive or hardly used.

Administrative and clinical hospital episode data are available from both the Secondary Uses Service (SUS+) and Hospital Episode Statistics (HES) resources.[27] These data include information on length of stay, diagnoses and procedures during hospital admissions as well as on outpatient, emergency department and critical care episodes. Diagnoses are coded with ICD-10 codes and procedures with Operating Procedure Codes (OPCS-4).[28] The SUS+ resource contains raw data collected from NHS healthcare providers, representing the most up-to-date hospital episodes within NHS Digital (amongst hospitals making prompt and complete returns). These data are consolidated, validated and cleaned on a monthly basis to form the HES database.[29] As a result, each month's HES data become available about two months behind real time. Thereafter, a fixed update is produced for each full year of HES data. Amongst the 54.4 million people included in our linked cohort, the SUS+ dataset holds 2.5 billion records for 6.3 million people (from November 2019 onwards) and the HES dataset holds 0.2 billion records on 42.3 million people (from 1997 onwards).

Death registration data[11] flow daily to NHS Digital from the Office for National Statistics (ONS) Civil Registration dataset, including date, cause (coded with ICD-10) and place of death, and are available historically from April 1997. Deaths in England should be registered within five days of the date of

death, although registration of a death is delayed in some situations.[30,31] Amongst the 54.4 million people in our linked cohort, 417,236 died on or before 31st October 2020.

The Second Generation Surveillance System (SGSS)[11] is the national laboratory reporting system used in England to capture routine laboratory data on mainly infectious diseases and antimicrobial resistance, including the SARS-CoV-2 virus. SGSS provides reports daily to NHS Digital on positive Covid-19 results (including the test date) fed directly from Pillar 1 pathology labs (i.e., established labs in hospitals for patients as well as NHS key workers), and indirectly from Pillar 2 labs (i.e., new, centralised, mostly privately-run labs, created specifically for Covid-19 testing for the wider population). In total 884,341 participants have at least one positive Covid-19 test recorded in the SGSS Covid-19 laboratory test dataset, of which 776,503 participants (88%) are linkable to the 54.4 million person cohort.

The community dispensing dataset, provided to NHS Digital monthly by the NHS Business Services Authority, contains person-level information on NHS primary care prescriptions dispensed by community pharmacists, appliance contractors and dispensing doctors in England, including the name and strength of medication coded from the British National Formulary (BNF) Dictionary of Medicines and Devices (DM+D).[32] Amongst the 54.4 million person cohort, there are over 40.6 million with dispensed medications and approximately 2.3 billion records (from April 2018).

Demographic characteristics and cardiovascular disease incidence

Characteristics of the linked cohort of 54.4 million people alive on 1st January 2020 are shown in **Table 3**; on 1st January 2020, 51% were female and 14% aged 70 years or older, with a mean age of 40.0 years for males and 41.6 years for females. By linking and combining person-level records from primary care and hospital episodes, ethnicity information is available for over 95% of people, among whom 63% have their ethnic group recorded in primary care and 92% in hospital episodes data (**Figure 2a**). A previous diagnosis of stroke/TIA or MI is recorded for 2.2% and 2.1% of people, respectively, while 7% and 8% people have a record indicating a previous diagnosis of diabetes and obesity, respectively. Among 53.2 million people with no prior diagnosis of stroke/TIA, 98,721 had a first-ever incident stroke/TIA between 1 January and 31st October 2020, of which 30% were recorded only in primary care (i.e. not in hospital episodes or death registry data) and 4% only in death registry records (**Figure 2b**). Among 53.1M people with no prior MI, 62,966 had an incident MI during follow up, of which 8% were recorded only in primary care and 12% only in death registry records (**Figure 2c**).

Covid-19 diagnoses

Among people in the linked cohort, a total of 959,067 people had a confirmed or suspected Covid-19 diagnosis between 1st January and 31st October 2020 (714,162 in primary care data, 126,349 in hospital admission records, 776,503 in Covid-19 laboratory test data and 48,433 in death registry records). While 58% of these were recorded in both primary care and Covid-19 laboratory test data, 15% and 18% respectively were recorded in only one of these **(Figure 3).**

Whereas females are more likely to have a confirmed or suspected Covid-19 diagnosis in their primary care records (1.4% females versus 1.2% males) and in Covid-19 laboratory test data (1.5% females versus 1.3% males), they are less likely to have a Covid-19 diagnosis recorded in hospital episodes (0.21% females versus 0.26% males) or on death certificates (0.08% females versus 0.10% males). Older people are more likely to have a Covid-19 diagnosis from hospital episodes and death registrations, although young adults are more likely to have Covid-19 diagnoses recorded in Covid-19 laboratory test data and primary care. People with unknown age or sex are over 10 times more likely to have a Covid-19 diagnosis recorded in hospital episodes or on death certificates. A higher proportion of Asian and Asian British people have a Covid-19 diagnosis in primary care and in the Covid-19 laboratory tests in comparison with other ethnicities. However, such differences are not observed in information from hospital episodes or death certificates. People with a previous history of stroke/TIA, MI, obesity or diabetes are more likely to have a Covid-19 diagnosis recorded in all healthcare settings (**Table 3**).

When compared with the latest Public Health England reports of Covid-19 laboratory tests,[33] Covid-19-related hospital admissions[34] and deaths with Covid-19 on the death certificate,[35] our linked cohort produces statistics which concord with relevant cumulative counts of Covid-19 cases (**Supplementary Table 6).**

DISCUSSION

Summary

We have described the development and key features of a novel linked EHR resource comprising a range of current and future planned linked datasets covering the entire population of England and forming part of wider UK-wide initiative to accelerate UK-wide research on Covid-19 and cardiovascular disease and beyond. We include descriptive analyses of a cohort of 54.4 million people alive at the start of 2020, including over 96% of the English population. The datasets described are already being accessed through the new NHS Digital TRE service for England to enable an expanding range of research projects via the BHF Data Science Centre's CVD-COVID-UK consortium. Notably, combining person-level information across data sources delivers approximately 95% complete data on key characteristics including age, sex and ethnicity and is essential for identifying cardiovascular diseases of interest, such as stroke and myocardial infarction. Approximately 90% of people with a positive Covid-19 laboratory test have linkable primary care records, and enriching the Covid-19 laboratory test data with primary care, hospital episodes and death registry data enables ascertainment of approximately 20% additional confirmed or suspected Covid-19 cases.

Previously, research use of linked EHRs in England has been restricted to subsets of the population, according to the coverage of various data providers, including the individual primary care computer system suppliers (e.g., the Clinical Practice Research Datalink,[36] The Health Improvement Network,[37] QResearch[38] and, more recently, OpenSafely[39]). As the national provider of information, data and IT systems for commissioners, analysts and clinicians in health and social care in England, NHS Digital handles larger volumes of health data than any other organisation globally and has extremely well developed and robust processes for maintaining data security and privacy. Alignment of the new Trusted Research Environment for England with NHS Digital's systems therefore maximises security while minimising the need for transmission of large volumes of linked data to support population-scale research.

Strengths and limitations

The currently available linked data assets comprise the world's largest single population-based cohort available for research, which will be further enhanced as further datasets are added. The availability of primary care data linked to such a wide range of other data is unparalleled at this scale, while the resource is also making linked nationwide Covid-19 laboratory testing and community dispensing data available for research for the first time. Unsurprisingly, given the >96% coverage of the

English population, the linked cohort represents the English population in terms of age, sex, ethnicity, and diabetes, when compared with UK Government England official statistics,[40-42] includes the full distribution of general practices according to geographic location and size,[25] and includes large enough numbers of people with different characteristics to support a diverse range of statistically well-powered research studies. For example, the cohort includes large numbers of: people in subgroups typically under-represented in research (e.g., several tens of thousands in each of the ethnic minority subgroups); younger people for whom poor outcomes of Covid-19 are uncommon but nonetheless devastating (e.g., over 200,000 under 30 years of age, among whom 75 deaths were recorded by 31st October 2020); people experiencing the common exemplar cardiovascular outcomes of stroke/TIA and MI (many tens of thousands), suggesting substantial potential to support studies of the impact of Covid-19 on subtypes of stroke and MI as well as on a wide range of rare conditions.

The NHS Digital TRE for England ensures secure, privacy-protecting storage of and access to large volumes of data, while minimizing the expense and security risks of data travel. Provision of data in this way is enabling a broad programme of collaborative research, encompassing several projects, which would be challenging to justify under the data dissemination model but which meet the relevant ethics and data access requirements under the TRE model. Researchers from many different organisations have been able to gain rapid access to the linked datasets via the CVD-COVID-UK consortium and its data sharing agreement with NHS Digital, avoiding lengthy and costly processes for multiple separate organisational data-access approvals and agreements. The consortium is enabling collaboration amongst researchers from across the UK, with a wide range of expertise (including clinicians from many different specialist backgrounds, data managers, computer scientists, data wranglers, epidemiologists and biostatisticians). Further, it has encouraged productive interactions between researchers and NHS Digital staff (including project management, data management, data science and technical development teams), enabling joint approaches to developing the TRE service and to identifying and solving data provision and linkage challenges. The rich and diverse nature of this interdisciplinary collaboration supports clinically and methodologically informed data curation and analysis pipelines and will enhance the interpretation and clinical application of research outputs. Regular dataset updates ensure the contemporary relevance and dynamic nature of the data resource and will enable ongoing long-term follow-up of the whole population. The development of publicly shareable, validated phenotyping algorithms[19] and analytic code[18] will avoid duplication of effort by additional groups of researchers working with the same or similar datasets. Although developed with the initial intent of supporting the CVD-COVID-UK consortium research programme,

establishing the NHS Digital TRE Service for England has wider benefits, given its clear potential to expand to support research more broadly beyond Covid-19 and beyond the cardiovascular domain. In addition, the work to establish the TRE has generated knowledge about linked EHR data and routes to their access across the UK health data science research community, benefiting other UK-wide initiatives, including: the UK-wide ISARIC study of the clinical characteristics of people hospitalized with Covid-19,[43] collaborative efforts to address the determinants of Covid-19 susceptibility, severity and outcome through analyses of population-based cohorts with bio-samples linked to national EHRs[44,45]; the RECOVERY randomized trial of treatments for Covid-19[46]; the COG-UK Covid-19 viral sequencing study;[47] and the UK Government Chief Scientific Adviser's National Core Studies programme, established to coordinate the UK's Covid-19 research response (in particular its underpinning Data and Connectivity theme led by Health Data Research UK).[48,49]

Nevertheless, there are some limitations: (i) it is not yet possible to bring external cohort studies or trials into the environment for linkage (although data can be linked to these through NHS Digital's standard data dissemination route); (ii) the primary care data are currently restricted to a large subset of SNOMED codes and limited to people known to be alive from November 2019 onwards, although NHS Digital is currently enacting its plans to obtain a fully comprehensive primary care dataset, to be updated daily, which will become available during 2021 and will eventually replace the current primary care dataset; (iii) the TRE currently has a relatively limited range of services and analytical tools, although NHS Digital are committed to expanding these; (iv) the descriptive results presented here provide an overview of the available resources with illustrative examples but are not designed to inform reliable conclusions about the associations between patient characteristics and COVID-19 outcomes, as the analyses are unadjusted and so prone to confounding; (v) whilst some data quality checks have been performed before creating the linked cohort, future analyses may require additional checks to minimize influential errors, outliers and inconsistent records.

Combining resources across the four nations of the UK

Similar, albeit not identical, EHR data resources are available in separate TREs provided by SAIL Databank for Wales, the National Data Safe Haven in Scotland and the Honest Broker Service in Northern Ireland. Due to differences in data structure and coding procedures between nations, we advocate the development of analysis plans which aim for maximum consistency but allow for nation-specific differences. Where appropriate, results of nation-specific analyses can be combined to produce results with UK-wide coverage. Such combined analyses will, increasingly, be able to take

advantage of Health Data Research UK's plans to provide the infrastructure, methods and tools to enable federation of analyses across TREs.[50]

Conclusion

We describe the first-ever provision for research of linked nationwide EHR data for England and demonstrate the importance of linking person-level data from different health settings for defining exemplar cardiovascular disease outcomes, Covid-19 diagnoses and key characteristics. By covering almost the entire English population, the resource includes all age groups, ethnic geographic, and socioeconomic, health and personal characteristics and can enable statistically powerful population-scale research with very large numbers of outcomes. It is accessible by approved researchers through a secure TRE hosted by NHS Digital to support research on Covid-19 and cardiovascular disease and can expand to benefit other future research initiatives beyond Covid-19 and cardiovascular disease.

TABLES AND FIGURES

Figure 1: Overview of current (in bold text) and planned (regular text) data flows into the NHS Digital Trusted Research Environment for England

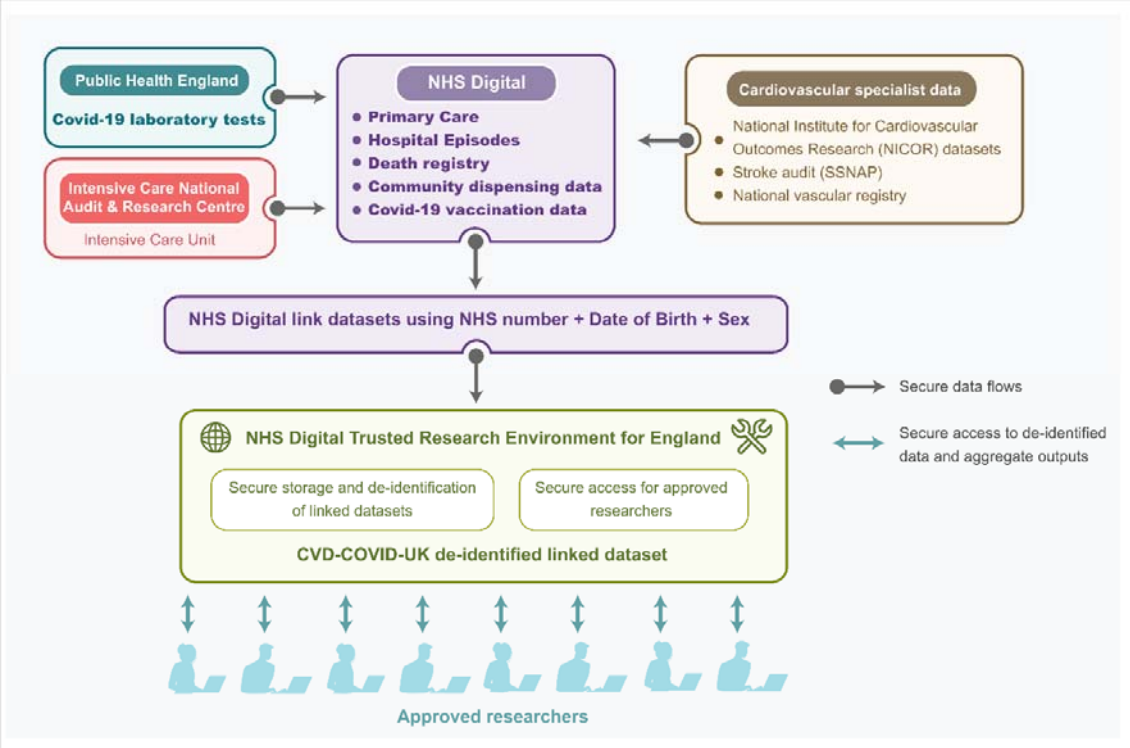

Figure 2: Data sources reporting person-level data on (a) ethnicity; (b) incident stroke/TIA (c) incident MI

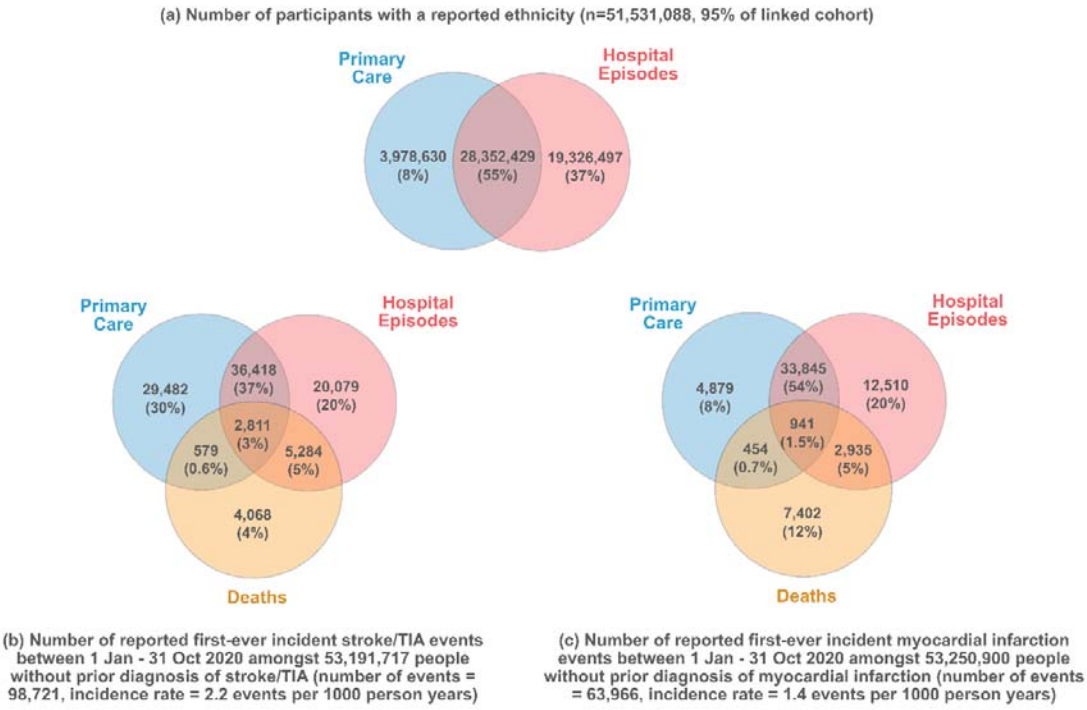

Figure 3: Data sources reporting person-level data on confirmed or suspected Covid-19 diagnoses between 1st Jan 2020 - 31st October 2020 (n=959,067). Numbers indicate distinct people with a confirmed or suspected Covid-19 diagnosis.

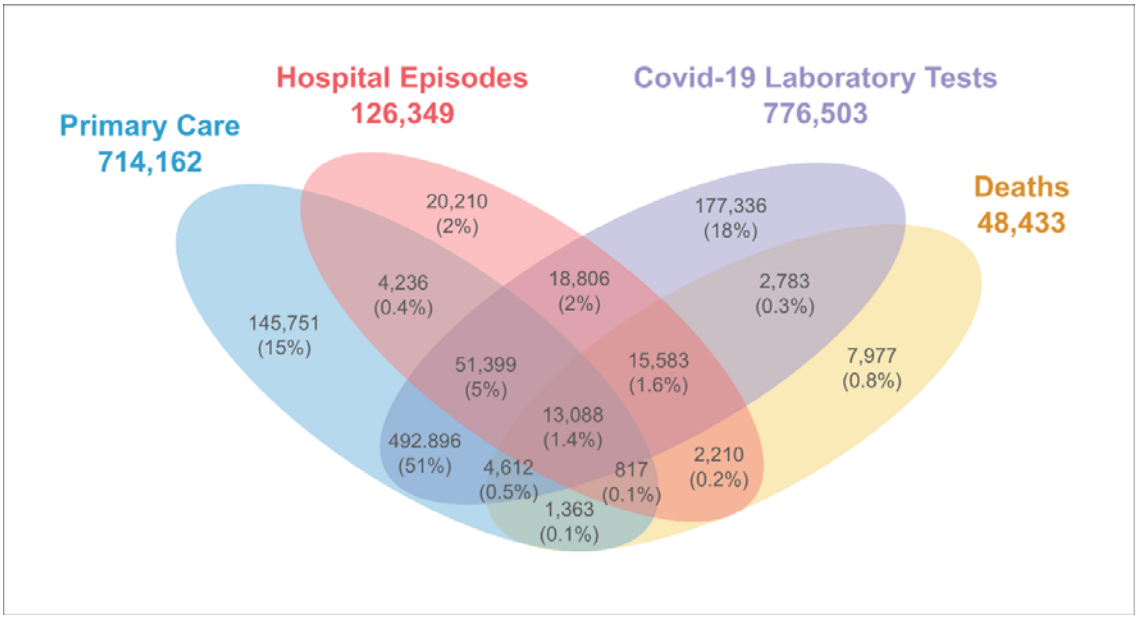

Table 1: Overview of available and planned linked resources

Availability of population-wide linked data	Data description	Data resource
Available Jan 2021	Primary Care	GDPPR: General Practice Extraction Service [GPES] data for pandemic planning and research
	Hospital Episodes	Secondary Uses Service (SUS+) and Hospital Episode Statistics (HES) including: Emergency Care Dataset (ECDS) Admitted Participant Care (APC) Adult Critical Care (ACC) Outpatients (OP)
	Death registry	Office for National Statistics (ONS) death registrations
	Covid-19 laboratory tests	Public Health England (PHE) Second Generation Surveillance System (SGSS) Covid-19 test data results (Pillars 1 & 2)
	Community dispensing data	NHS Business Services Authority (BSA) community dispensing data
To become available during 2021	Intensive Care Unit	Intensive Care National Audit and Research Centre (ICNARC) data
	Cardiovascular specialist audit/registry data	National Institute for Cardiovascular Outcomes Research (NICOR) datasets including: Myocardial Infarction National Audit Programme Adult Percutaneous Coronary Interventions National Heart Failure Audit Cardiac Rhythm Management Audit Congenital Heart Disease in Children and Adults Adult Cardiac Surgery Audit NICOR Health Technology Registries
		Sentinel Stroke National Audit Programme (SSNAP) data
		National Vascular Registries (NVR) data
	National Covid-19 vaccination data	National Immunisation Management System Covid-19 vaccination dataset
		Covid-19 vaccination Adverse Reactions dataset

Table 2: Key details of main data resources

	Primary Care	Hospital Episodes		Death registry	Covid-19 laboratory tests	Community dispensing
Name of resource	GDPPR: General Practice Extraction Service [GPES] data for pandemic planning and research	Secondary Uses Service (SUS+)	Hospital Episode Statistics (HES) including: Emergency Care Dataset (ECDS), Admitted Participant Care (APC), Adult Critical Care (ACC), Outpatients (OP)	Civil Registration – Deaths (Office of National Statistics [ONS] asset)	Public Health England (PHE) Second Generation Surveillance System (SGSS) Covid-19 test results	NHS Business Services Authority (BSA) community dispensing data
Who is included?	People registered with a general practice in England, without a registered objection to sharing of data with NHS Digital, alive on 1st November 2019‡	People receiving treatment or care at an NHS hospital in England	People receiving treatment or care at an NHS hospital in England	All people with a registered death in England	People with a laboratory confirmed polymerase chain reaction (PCR) positive test under pillar 1 or 2 testing guidelines	People with at least one prescription dispensed in the community
What is recorded?	Demographics, diagnoses, symptoms, signs, prescriptions, referrals, immunisations, behavioural factors, tests	Diagnoses, procedures, personal demographics (including ethnicity and area-level deprivation), admission and discharge dates, hospital and other variables.	Diagnoses, procedures, personal demographics (including ethnicity and area-level deprivation), admission and discharge dates, hospital and other variables.	Date of death, date death registered, sex, underlying cause of death, district, subdistrict, place of death (code, establishment and type), neonate flag	Demographics (age, sex, ethnicity, Lower-layer Super Output Areas [LSOA]), date of specimen, laboratory report, reporting laboratory	Information on dispensed medications (name, strength, substance, quantity)
How are records coded?	SNOMED-CT	ICD-10; OPCS-4	ICD-10; OPCS-4; proprietary emergency care codes	ICD-10	Not coded	BNF Dictionary of Medicines and Devices (DM+D)
Period of record dates	From the earliest record for each person topresent	November 2019to present	April 1997 to present	April 1997 to present	March 2020 to present	April 2018 to present
Frequency of provision and time lag	Extracted fortnightly; up-to-date at time of each extract	Daily flows into NHS Digital; up-to-date on submission for completed episodes of care from submitting trusts	Updated monthly (from SUS) within NHS Digital; about 2 months behind real time	Weekly flows into NHS Digital; up-to-date at time of provision	Provided daily to NHS Digital; up-to-date at time of provision	Updated monthly; about 7-11 weeks behind real time.
Number of people with records (before quality assurance exclusions)	57,908,487	7,153,569	61,958,690	14,643,921	884,311	44,546,519
Total number of records	4,937,121,423	2,781,364,103	365,438,996	18,815,693	1,160,138	2,796,440,797
Number of People known to be on alive 1 January 2020	54,388,181	6,251,673	42,582,312	417,236	776,503	40,623,625

Total number of records among people alive on 1 January 2020	4,870,642,482	2,491,646,379	228,933,294	457,412	988,174	2,329,914,169

[†]Current restriction to scientific research relevant to the Covid-19 pandemic

Table 3: Characteristics of linked cohort and of people with a confirmed or suspected Covid-19 diagnosis, by data resource.

Characteristic	Subgroup	Total N, *% of population* (N=54,388,181)		Number of confirmed or suspected Covid-19 diagnosis, *% of subgroup*							
				Recorded in primary care records (N=714,162)		Recorded in Covid-19 laboratory test data (N=776,503)		Recorded in hospital episodes[1] (N=126,349)		Recorded in death registration data (N=48,433)	
Sex	Female	**27,718,313**	*50.96*	400,448	*1.44*	424,870	*1.53*	57,789	*0.21*	21,565	*0.08*
	Male	**26,661,385**	*49.02*	313,558	*1.18*	351,383	*1.32*	68,329	*0.26*	26,617	*0.10*
	Unknown	**8,483**	*0.02*	156	*1.84*	250	*2.95*	231	*2.62*	251	*2.96*
Age group(years)	0-17	**11,188,814**	*20.57*	60,571	*0.54*	68,028	0.61	1,827	*0.02*	8	*0.00*
	18-29	**7,925,142**	*14.57*	151,304	*1.91*	184,885	2.33	4,081	*0.05*	67	*0.00*
	30-49	**14,701,289**	*27.03*	207,672	*1.41*	226,179	1.54	15,828	*0.11*	853	*0.01*
	50-69	**13,026,860**	*23.95*	179,977	*1.38*	181,399	1.39	35,070	*0.27*	6,692	*0.05*
	70+	**7,543,288**	*13.87*	114,505	*1.52*	115,796	1.54	69,317	*0.92*	40,562	*0.54*
	Unknown	**2,788**	*0.01*	133	*4.77*	216	7.75	226	*8.11*	251	*9.00*
Ethnicity	White	**41,786,891**	*76.83*	556,489	*1.33*	588,550	*1.41*	99,629	*0.24*	41,444	*0.10*
	Mixed	**1,156,060**	*2.13*	11,810	*1.02*	14,053	*1.22*	1,748	*0.15*	420	*0.04*
	Asian and Asian British	**4,589,778**	*8.44*	95,752	*2.09*	108,455	*2.36*	13,317	*0.29*	3,071	*0.07*
	Black and Black British	**1,860,340**	*3.42*	20,863	*1.12*	25,051	*1.35*	6,540	*0.35*	1,644	*0.09*
	Other Ethnic Groups	**2,138,019**	*3.93*	14,334	*0.67*	18,962	*0.89*	3,221	*0.15*	963	*0.05*
	Unknown	**2,857,093**	*5.25*	14,914	*0.52*	21,432	*0.75*	1,894	*0.07*	891	*0.03*
Previous diagnosis of Stroke/TIA	No	**53,191,717**	*97.80*	685,197	*1.29*	745,978	1.40	108,596	*0.20*	37,805	*0.07*
	Yes	**1,196,464**	*2.20*	28,965	*2.42*	30,525	2.55	17,753	*1.48*	10,628	*0.89*
Previous diagnosis of myocardial infarction	No	**53,250,900**	*97.91*	689,909	*1.30*	750,466	1.41	110,106	*0.21*	39,692	*0.07*
	Yes	**1,137,281**	*2.09*	24,253	*2.13*	26,037	2.29	16,243	*1.43*	8,741	*0.77*
Previous diagnosis of obesity	No	**49,827,060**	*91.61*	625,911	*1.26*	687,445	1.38	97,905	*0.20*	39,084	*0.08*
	Yes	**4,561,121**	*8.39*	88,251	*1.93*	89,058	1.95	28,444	*0.62*	9,349	*0.20*
Previous diagnosis of diabetes	No	**50,778,499**	*93.36*	642,095	*1.26*	700,133	1.38	88,262	*0.17*	31,867	*0.06*
	Yes	**3,609,682**	*6.64*	72,067	*2.00*	76,370	2.12	38,087	*1.06*	16,566	*0.46*

[1]From Hospital Episode Statistics – Admitted Patient Care data.

ACKNOWLEDGEMENTS

The BHF Data Science Centre (BHF Grant no. SP/19/3/34678, awarded to Health Data Research UK) funded co-development (with NHS Digital) of the TRE, provision of linked datasets, data access, user software licences, computational usage and data management and wrangling support, with additional contributions from the HDR UK Data and Connectivity component of the UK Chief Scientific Adviser's National Core Studies programme to coordinate national Covid-19 priority research. Consortium partner organisations funded the time of contributing data analysts, biostatisticians, epidemiologists and clinicians.

The results described are based on data from patients, collected by the NHS as part of their care and support. We would also like to acknowledge all data providers who make anonymised data available for research.

AA is supported by Health Data Research UK [HDR-9006] which receives its funding from 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; and Administrative Data Research UK which is funded by the Economic and Social Research Council [grant ES/S007393/1]. AB is supported by research funding from NIHR, British Medical Association, Astra-Zeneca and UK Research and Innovation. AB, AW and SD are part of the BigData@Heart Consortium, funded by the Innovative Medicines Initiative-2 Joint Undertaking under grant agreement No. 116074. AW and SI are supported by the BHF-Turing Cardiovascular Data Science Award (BCDSA\100005) and by core funding from: UK Medical Research Council (MR/L003120/1), British Heart Foundation (RG/13/13/30194; RG/18/13/33946) and NIHR Cambridge Biomedical Research Centre (BRC-1215-20014). JC, JS and RD are supported by the HDRUK South West Better Care Partnership and NIHR Bristol Biomedical Research Centre. SD is supported by Health Data Research UK London, which receives its funding from Health Data Research UK Ltd 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, and the Wellcome Trust; Alan Turing Fellowship (EP/N510129/1); National Institute for Health Research (NIHR) Biomedical Research Centre (BRC) at University College London Hospital NHS Trust (UCLH). VW is supported by the University of Bris-

tol Medical Research Council Integrative Epidemiology Unit (MC_UU_00011/4). WW is supported by a Scottish Senior Clinical Fellowship, CSO (SCAF/17/01)

The views expressed are those of the author(s) and not necessarily those of the organisations listed.

Annexe 1: CVD-COVID-UK consortium members

Institution	Member Name
Addenbrooke's Hospital	Jon Boyle
British Heart Foundation	Dan O'Connell
British Heart Foundation	Kate Cheema
British Heart Foundation	Naomi Herz
British Heart Foundation	Nilesh Samani
British Heart Foundation	Sonya Babu-Narayan
European Bioinformatics Institute	Ewan Birney
European Bioinformatics Institute	Moritz Gerstung
Great Ormond Street Hospital	Katherine Brown
Health Data Research UK / BHF Data Science Centre	Cathie Sudlow
Health Data Research UK / BHF Data Science Centre	Debbie Ringham
Health Data Research UK / BHF Data Science Centre	Jackie MacArthur
Health Data Research UK / BHF Data Science Centre	Lydia Martin
Health Data Research UK / BHF Data Science Centre	Lynn Morrice
Health Data Research UK / BHF Data Science Centre	Rouven Priedon
Health Data Research UK	Sinduja Manohar
Healthcare Quality Improvement Partnership	Caroline Rogers
Healthcare Quality Improvement Partnership	Mirek Skrypak
Imperial College London	Alun Davies
Imperial College London	Safa Salim
Imperial College London	Sarah Onida
Keele University	Mamas Mamas
King's College London	Abdel Douiri
King's College London	Ajay Shah
King's College London	Ben Bray
King's College London	Charles Wolfe
King's College London	Elena Nikiphorou
London School of Hygiene & Tropical Medicine	Qiuju Li
NHS Digital	Brian Roberts
NHS Digital	Sam Hollings
NHS England	Deborah Lowe
NHS Lanarkshire	Mark Barber
NHS Scotland	Carole Morris
NICE	Adrian Jonas
NICE	Brett Doble
NICE	Felix Greaves
NICE	Jennifer Beveridge
NICE	Seamus Kent
NICE	Thomas Lawrence
Office for National Statistics	Ben Humberstone
Office for National Statistics	Myer Glickman
Office for National Statistics	Vahé Nafilyan
Queen's University Belfast	Abdul Qadr Akinoso-Imran
Queen's University Belfast	Frank Kee
Royal College of Surgeons of England	David Cromwell
Royal Papworth Hospital NHS Foundation Trust	Florian Falter
Swansea Bay University Health Board	Daniel Harris

Institution	Member Name
Swansea University	Ashley Akbari
Swansea University	Fatemeh Torabi
Swansea University	Gareth Davies
Swansea University	Hoda Abbasizanjani
Swansea University	Jane Lyons
Swansea University	Julian Halcox
Swansea University	Laura North
Swansea University	Libby Ellins
Swansea University	Mike Gravenor
Swansea University	Ronan Lyons
Swansea University	Rowena Griffiths
University College London	Alex Handy
University College London	Alvina Lai
University College London	Ami Banerjee
University College London	Ashkan Dashtban
University College London	Caroline Dale
University College London	Christopher Tomlinson
University College London	Eloise Withnell
University College London	Harry Hemingway
University College London	Honghan Wu
University College London	Johan Thygesen
University College London	Ken Li
University College London	Laura Pasea
University College London	Mehrdad Mizani
University College London	Michalis Katsoulis
University College London	Paula Lorgelly
University College London	Pedro Machado
University College London	Reecha Sofat
University College London	Rohan Takhar
University College London	Spiros Denaxas
University of Aberdeen	Mary Joan Macleod
University of Bristol	Deborah Lawler
University of Bristol	Jennifer Cooper
University of Bristol	Jonathan Sterne
University of Bristol	Livia Pierotti
University of Bristol	Massimo Caputo
University of Bristol	Neil Davies
University of Bristol	Rachel Denholm
University of Bristol	Rupert Payne
University of Bristol	Tom Palmer
University of Bristol	Venexia Walker
University of Cambridge	Angela Wood
University of Cambridge	David Brind
University of Cambridge	Emanuele Di Angelantonio
University of Cambridge	Fabian Falck
University of Cambridge	Haoting Zhang
University of Cambridge	Howard Tang

Institution	Member Name
University of Cambridge	Jessica Barrett
University of Cambridge	John Danesh
University of Cambridge	Mike Inouye
University of Cambridge	Samantha Ip
University of Cambridge	Spencer Keene
University of Cambridge	Tianxiao Wang
University of Dundee	David Moreno Martos
University of Dundee	Huan Wang
University of Dundee	Ify Mordi
University of Edinburgh	Annemarie Docherty
University of Edinburgh	Gwenetta Curry
University of Edinburgh	Tim Wilkinson
University of Edinburgh	William Whiteley
University of Exeter	John Dennis
University of Glasgow	Clea du Toit
University of Glasgow	Colin Berry
University of Glasgow	Sandosh Padmanabhan
University of Leeds	Jianhua Wu
University of Leicester	Anna Hansell
University of Leicester	Claire Lawson
University of Leicester	Francesco Zaccardi
University of Leicester	Kamlesh Khunti
University of Leicester	Tom Norris
University of Liverpool	David Hughes
University of Liverpool	Munir Pirmohamed
University of Liverpool	Ruwanthi Kolamunnage-Dona
University of Manchester	Craig Smith
University of Manchester	Maya Buch
University of Oxford	Ben Goldacre
University of Oxford	Ben Cairns
University of Oxford	Eva Morris
University of Oxford	George Nicholson
University of Oxford	Lucy Wright
University of Oxford	Nick Hall
University of Oxford	Olena Seminog
University of Oxford	Raph Goldacre
University of Oxford	Seb Bacon
University of Strathclyde	Amanj Kurdi
University of Strathclyde	Kim Kavanagh
University of Strathclyde	Marion Bennie
University of Strathclyde	Raymond Carragher
University of Warwick	Harry Wilde
University Hospital of North Midlands	Arun Pherwani
Wellcome Trust	Bilal Mateen

Institutions party to the Data Sharing Agreement for the NHS Digital CVD Trusted Research Environment for England are: NICE, Swansea University, University College London, University of Bristol, University of Cambridge, University of Leicester, University of Liverpool, University of Oxford

Annexe 2:
CVD-COVID-UK RECORD statement – checklist of items, extended from the STROBE statement, that should be reported in observational studies using routinely collected health data.

	Item No.	STROBE items	Location in manuscript where items are reported	RECORD items	Location in manuscript where items are reported
Title and abstract					
	1	(a) Indicate the study's design with a commonly used term in the title or the abstract (b) Provide in the abstract an informative and balanced summary of what was done and what was found	Title and abstract	RECORD 1.1: The type of data used should be specified in the title or abstract. When possible, the name of the databases used should be included. RECORD 1.2: If applicable, the geographic region and timeframe within which the study took place should be reported in the title or abstract. RECORD 1.3: If linkage between databases was conducted for the study, this should be clearly stated in the title or abstract.	Title and abstract Title and abstract Title and abstract
Introduction					
Background rationale	2	Explain the scientific background and rationale for the investigation being reported	Page 4		
Objectives	3	State specific objectives, including any prespecified hypotheses	Page 4		
Methods					
Study Design	4	Present key elements of study design early in the paper	Pages 6 and 7 Figure1, Table 1		
Setting	5	Describe the setting, locations, and relevant dates, including periods of	Pages 6 and 7 Table 2		

		recruitment, exposure, follow-up, and data collection			
Participants	6	*(a) Cohort study* - Give the eligibility criteria, and the sources and methods of selection of participants. Describe methods of follow-up *Case-control study* - Give the eligibility criteria, and the sources and methods of case ascertainment and control selection. Give the rationale for the choice of cases and controls *Cross-sectional study* - Give the eligibility criteria, and the sources and methods of selection of participants *(b) Cohort study* - For matched studies, give matching criteria and number of exposed and unexposed *Case-control study* - For matched studies, give matching criteria and the number of controls per case	Pages 8 and 9, Supplementary Tables 1-5	RECORD 6.1: The methods of study population selection (such as codes or algorithms used to identify subjects) should be listed in detail. If this is not possible, an explanation should be provided. RECORD 6.2: Any validation studies of the codes or algorithms used to select the population should be referenced. If validation was conducted for this study and not published elsewhere, detailed methods and results should be provided. RECORD 6.3: If the study involved linkage of databases, consider use of a flow diagram or other graphical display to demonstrate the data linkage process, including the number of individuals with linked data at each stage.	Page 9, Supplementary Tables 1-5 Page 9 Table 2
Variables	7	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable.	Pages 8 and 9	RECORD 7.1: A complete list of codes and algorithms used to classify exposures, outcomes, confounders, and effect modifiers should be provided. If these cannot be reported, an explanation should be provided.	Pages 8 and 9, Supplementary Tables 1-5
Data sources/ measurement	8	For each variable of interest, give sources of data and details of	Pages 8 and 9		

		methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group			
Bias	9	Describe any efforts to address potential sources of bias	Not applicable		
Study size	10	Explain how the study size was arrived at	Page 10, Table 2		
Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen, and why	Not applicable		
Statistical methods	12	(a) Describe all statistical methods, including those used to control for confounding (b) Describe any methods used to examine subgroups and interactions (c) Explain how missing data were addressed (d) *Cohort study* - If applicable, explain how loss to follow-up was addressed *Case-control study* - If applicable, explain how matching of cases and controls was addressed *Cross-sectional study* - If applicable, describe analytical methods taking account of sampling strategy (e) Describe any sensitivity analyses	Not applicable		

Data access and cleaning methods		..		RECORD 12.1: Authors should describe the extent to which the investigators had access to the database population used to create the study population. RECORD 12.2: Authors should provide information on the data cleaning methods used in the study.	Pages 7, 8 and 9 Page 8
Linkage		..		RECORD 12.3: State whether the study included person-level, institutional-level, or other data linkage across two or more databases. The methods of linkage and methods of linkage quality evaluation should be provided.	Page 6
Results					
Participants	13	(a) Report the numbers of individuals at each stage of the study (*e.g.*, numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed) (b) Give reasons for non-participation at each stage. (c) Consider use of a flow diagram	Pages 10 and 11 Table 2 Pages 10 nd 11 In Table 2	RECORD 13.1: Describe in detail the selection of the persons included in the study (*i.e.,* study population selection) including filtering based on data quality, data availability and linkage. The selection of included persons can be described in the text and/or by means of the study flow diagram.	Pages 10 and 11 Table 2
Descriptive data	14	(a) Give characteristics of study participants (*e.g.*, demographic, clinical, social) and information on exposures and potential confounders (b) Indicate the number of participants with missing data for each variable of interest	Table 3 Table 3		

		(c) *Cohort study* - summarise follow-up time (*e.g.*, average and total amount)	Page 11, Figure 2		
Outcome data	15	*Cohort study* - Report numbers of outcome events or summary measures over time *Case-control study* - Report numbers in each exposure category, or summary measures of exposure *Cross-sectional study* - Report numbers of outcome events or summary measures	Table 3, Figure 2, Figure 3		
Main results	16	(a) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (e.g., 95% confidence interval). Make clear which confounders were adjusted for and why they were included (b) Report category boundaries when continuous variables were categorized (c) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period	Table 3 Table 3 NA		
Other analyses	17	Report other analyses done—e.g., analyses of subgroups and interactions, and sensitivity analyses	Page 12 Table 3 Supplementary Table 6		
Discussion					
Key results	18	Summarise key results with reference to study objectives	Page 13		

Limitations	19	Discuss limitations of the study, taking into account sources of potential bias or imprecision. Discuss both direction and magnitude of any potential bias	Page 15	RECORD 19.1: Discuss the implications of using data that were not created or collected to answer the specific research question(s). Include discussion of misclassification bias, unmeasured confounding, missing data, and changing eligibility over time, as they pertain to the study being reported.	Page 15
Interpretation	20	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence	Page 16		
Generalisability	21	Discuss the generalisability (external validity) of the study results	Page 14, 15		
Other Information					
Funding	22	Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based	In Acknowledgements		
Accessibility of protocol, raw data, and programming code		..		RECORD 22.1: Authors should provide information on how to access any supplemental information such as the study protocol, raw data, or programming code.	Page 6

*Reference: Benchimol EI, Smeeth L, Guttmann A, Harron K, Moher D, Petersen I, Sørensen HT, von Elm E, Langan SM, the RECORD Working Committee. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) Statement. *PLoS Medicine* 2015; 2:e1001885.

References

1. Cavallaro F, Lugg-Widger F, Cannings-John R, Harron K. Open Letter: Reducing barriers to data access for research in the public interest—lessons from covid-19. BMJ Opin
2. Jones KH, Ford DV, Lyons RA. The SAIL Databank: 10 years of spearheading data privacy and research utility, 2007-2017. Swansea University. doi: http://dx.doi.org/10.23889/ [Internet] 2017. [cited 2021 Feb 19]. Available from: https://saildatabank.com/
3. McGurnaghan SJ, Weir A, Bishop J, et al. Public Health Scotland COVID-19 Health Protection Study Group; Scottish Diabetes Research Network Epidemiology Group. Risks of and risk factors for COVID-19 disease in people with diabetes: a cohort study of the total population of Scotland. Lancet Diabetes Endocrinol. 2021 Feb;9(2):82-93.
4. Shah ASV, Wood R, Gribben C, et al. Risk of hospital admission with coronavirus disease 2019 in healthcare workers and their households: nationwide linkage cohort study. BMJ. 2020 Oct 28;371:m3582.
5. Siggaard T, Reguant R, Jørgensen IF, et al. Disease trajectory browser for exploring temporal, population-wide disease progression patterns in 7.2 million Danish patients. Nature Commun. 2020 Oct 2;11(1):4952
6. Ludvigsson, J.F., Almqvist, C., Bonamy, AK.E. *et al.* Registers of the Swedish total population and their use in medical research. Eur J Epidemiol. 2016 Feb;31(2):125-3
7. BHF Data Science Centre [Internet]. [cited 2021 Feb 18]. Available from: https://www.hdruk.ac.uk/helping-with-health-data/bhf-data-science-centre/
8. CVD-COVID-UK initiative [Internet]. [cited 2021 Feb 18]. Available from: https://www.hdruk.ac.uk/projects/cvd-covid-uk-project/
9. NHS Digital [internet]. [cited 2021 Feb 18]. Available from: https://digital.nhs.uk/
10. CVD-COVID-UK Dataset dashboard [Internet]: [cited 2021 Feb 18]. Available from: https://www.hdruk.ac.uk/wp-content/uploads/2021/02/210215-CVD-COVID-UK-TRE-Dataset-Dashboard_CLMS.pdf
11. CVD-COVID-UK TRE asset in Health Data Research Innovation Gateway [Internet]. [cited 2021 Feb 18]. Available from: https://web.www.healthdatagateway.org/dataset/7e5f0247-f033-4f98-aed3-3d7422b9dc6d
12. NHS Digital. Master Person Service [Internet]. [cited 2021 Jan 14];Available from: https://digital.nhs.uk/services/master-person-service
13. NHS Digital. Data Quality Assurance [Internet]. [cited 2021 Jan 14]. Available from: https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/data-quality#current-data-quality-maturity-index-dqmi-
14. NHS Digital. Data Access Request Service (DARS) [Internet]. [cited 2021 Feb 9]. Available from: https://digital.nhs.uk/services/data-access-request-service-dars
15. NHS Digital. Independent Group Advising on the Release of Data [Internet]. [cited 2021 Feb 8]. Available from: https://digital.nhs.uk/about-nhs-digital/corporate-information-and-documents/independent-group-advising-on-the-release-of-data
16. NHS Digital. Data Access Environment [Internet]. [cited 2021 Feb 18]. Available from: https://digital.nhs.uk/services/data-access-environment-dae
17. Five Safes framework [Internet]. [cited 2021 Feb 18]. Available from: http://www.fivesafes.org/

18. BHF Data Science Centre GitHub repository [Internet]. [cited 2021 Feb 18]; Available from: https://github.com/BHFDSC

19. HDR UK CALIBER Phenotype Library [internet]. [cited 2021 Feb 19]; Available from: https://portal.caliberresearch.org/collections/bhf-data-science-centre

20. Noble S, McLennan D, Noble M, Plunkett E, Gutacker N, Silk M, Wright G. The English Indices of Deprivation 2019 Research Report. Ministry of Housing, Communities and Local Government. 2019 Sept [Internet]. [cited 2021 Feb 19]. Available from: https://www.gov.uk/government/publications/english-indices-of-deprivation-2019-research-report

21. Lyons RA, Jones KH, John G, et al. The SAIL databank: Linking multiple health and social care datasets. BMC Med Inform Decis Mak. 2009 Jan 16;9:3

22. SAIL Databank Team. SAIL DATABANK Data Privacy and Security [Internet]. [cited 2021 Jan 20];Available from: https://saildatabank.com/saildata/data-privacy-security/

23. Public Health Scotland. Use of the National Safe Haven [Internet]. [cited 2021 Jan 20]. Available from: https://www.isdscotland.org/Products-and-Services/EDRIS/Use-of-the-National-Safe-Haven/

24. Kuan V, Denaxas S, Gonzalez-Izquierdo A, et al. A chronological map of 308 physical and mental health conditions from 4 million individuals in the English National Health Service. Lancet Digit Health 2019; May 20;1(2):e63-e77

25. NHS Digital. General Practice Extraction Service (GPES) Data for pandemic planning and research: a guide for analysts and users of the data [Internet]. [cited 2021 Feb 8];Available from: https://digital.nhs.uk/coronavirus/gpes-data-for-pandemic-planning-and-research/guide-for-analysts-and-users-of-the-data

26. Benchimol EI, Smeeth L, Guttmann A, et al. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. PLoS Med. 2015;12:e1001885.

27. Office for National Statistics. Population estimates for the UK, England and Wales, Scotland and Northern Ireland: mid-2019 [Internet]. [cited 2021 Feb 8];Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/bulletins/annualmidyearpopulationestimates/mid2019estimates

28. NHS Digital. Hospital Episode Statistics Data Dictionary [Internet]. [cited 2021 Feb 8];Available from: https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/hospital-episode-statistics/hospital-episode-statistics-data-dictionary

29. NHS Digital. Hospital Admissions Submission timetables [Internet]. [cited 2021 Feb 8];Available from: https://digital.nhs.uk/services/secondary-uses-service-sus/payment-by-results-guidance#submission-timetables

30. Bird SM. End late registration of fact-of-death in England and Wales. Lancet 2015 May 9;385(9980):1830-1.

31. Office for National Statistics. Impact of registration delays on mortality statistics in England and Wales: 2019 [Internet]. [cited 2021 feb 19]. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/articles/impactofregistrationdelaysonmortalitystatisticsinenglandandwales/2019

32. Medicines dispensed in Primary Care NHS Business Services Authority data [Internet]. [cited 2021 Feb 19]. Available from: https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/medicines-dispensed-in-primary-care-nhsbsa-data

33. Public Health England. Coronavirus cases in United Kingdom [Internet]. [cited 2021 Jan 20]. Available from:

https://coronavirus.data.gov.uk/details/cases?areaType=overview&areaName=United Kingdom

34. Public Health England. Coronavirus cases admitted to hospital [Internet]. [cited 2021 Jan 20]. Available from: https://coronavirus.data.gov.uk/details/healthcare

35. Public Health England. Coronavirus deaths in England. [internet]. [cited 2021 Feb 19]. Available from: https://coronavirus.data.gov.uk/details/deaths

36. Herrett E, Gallagher AM, Bhaskaran K, et al. Data Resource Profile: Clinical Practice Research Datalink (CPRD). Int J Epidemiol 2015 Jun;44(3):827-36.

37. Blak BT, Thompson M, Dattani H, Bourke A. Generalisability of the Health Improvement Network (THIN) database: Demographics, chronic disease prevalence and mortality rates. Inform Prim Care 2011;19(4):251-5

38. QResearch. Generating new knowledge to improve patient care. [Internet]. [cited 2021 Feb 22]. Available from: https://www.qresearch.org/

39. OpenSafely. [Internet] [cited 2021 Feb 22].Available from: https://opensafely.org/

40. UK Government. 3.8 million people in England now have diabetes. 2016 Sept [Internet]. [cited 2021 Feb 19]. Available from: https://www.gov.uk/government/news/38-million-people-in-england-now-have-diabetes

41. Diabetes UK. Diabetes prevalence 2019 [Internet]. [cited 2021 Feb 19]. Available from: https://www.diabetes.org.uk/professionals/position-statements-reports/statistics/diabetes-prevalence-2019

42. UK Government. UK population by ethnicity - Population statistics and 2011 Census data.[Internet]. [cited 2021 Feb 19]. Available from: https://www.ethnicity-facts-figures.service.gov.uk/uk-population-by-ethnicity

43. Docherty AB, Harrison EM, Green CA, et al. Features of 20 133 UK patients in hospital with covid-19 using the ISARIC WHO Clinical Characterisation Protocol: Prospective observational cohort study. BMJ 2020; May 22;369:m1985

44. HDR UK multiomics initiative [Internet]. [cited 2021 feb 18]. Available from: https://www.hdruk.ac.uk/case-studies/a-national-multi-omics-consortium-to-inform-disease-aetiology-and-prediction/

45. COVIDITY-COHORT [Internet]. [cited 2021 Feb 18]; Available from: https://www.uclhospitals.brc.nihr.ac.uk/news/research-projects-understand-link-between-covid-19-and-cardiovascular-diseases

46. Wilkinson E. RECOVERY trial: The UK covid-19 study resetting expectations for clinical trials. BMJ. 2020; Apr 28;369:m1626

47. COVID-19 Genomics UK (COG-UK) consortium- viral sequencing study [Internet]. [cited 2021 Feb 18]; Available from: https://www.cogconsortium.uk/

48. HDR UK National Core Studies - Data and Connectivity [Internet]. [cited 2021 Feb 18]; Available from https://www.hdruk.ac.uk/covid-19/covid-19-national-core-studies/

49. UK Government Office for Science [Internet]. [cited 2021 Feb 18]; Available from: https://www.gov.uk/government/organisations/government-office-for-science

50. UK Health Data Research Alliance. Trusted Research Environments (TRE). A strategy to build public trust and meet changing health data science needs. 2020 July [Internet], [cited 2021 Feb 19]. Available from: https://ukhealthdata.org/projects/aligning-approach-to-trusted-research-environments/