Linked electronic health records for research on a nationwide cohort of more than 54 million people in England: data resource

Angela Wood,1,2,3,4,5 Rachel Denholm,6,7,8 Sam Hollings,9 Jennifer Cooper,6,7,8 Samantha Ip,1 Venexia Walker,6,10,11 Spiros Denaxas,5,12,13,14 Ashley Akbari,15 Amitava Banerjee,16,13,17 William Whiteley,18,19 Alfina Lai,13 Jonathan Sterne,6,7,8 Cathie Sudlow,18,20,21 on behalf of the CVD-COVID-UK consortium

ABSTRACT
OBJECTIVE
To describe a novel England-wide electronic health record (EHR) resource enabling whole population research on covid-19 and cardiovascular disease while ensuring data security and privacy and maintaining public trust.

DESIGN
Data resource 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 1 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 and incident myocardial infarction) and all cause mortality between 1 January and 31 October 2020.

RESULTS
The linked cohort includes more than 96% of the English population. By combining person level data across national healthcare settings, data on age, sex, and ethnicity are complete for around 95% of the population. Among 53.3 million people with no previous diagnosis of stroke or transient ischaemic attack, 98,721 had a first ever incident stroke or transient ischaemic attack between 1 January and 31 October 2020, of which 30% were recorded only in primary care and 4% only in death registry records. Among 53.2 million people with no previous diagnosis of myocardial infarction, 62,966 had an incident myocardial infarction during follow-up, of which 8% were recorded only in primary care and 12% only in death registry records. A total of 959,470 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 50,504 in death registry records). Although 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 shows the importance of linking person level data across health settings to maximise completeness of key characteristics and to ascertain cardiovascular events and covid-19 diagnoses. Although this resource was initially established to support research on covid-19 and cardiovascular disease to benefit clinical care and public health and to inform healthcare policy, it can broaden further to enable a 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.2 Key benefits of research using such data on nationwide cohorts include generalisability of findings across all age groups, ethnicities, geographical locations, and socioeconomic, health, and personal characteristics, and inclusion of large numbers of people and events, enhancing the precision of findings and enabling a wide spectrum of novel research studies (eg, characterising shapes of relations between risk factors and disease, or studying minority groups and rare disease subtypes). While EHRs for whole country cohorts for Wales, Scotland, Denmark, and Sweden (populations of 3 to 10 million) have been used in research for several years,24 at the start of the covid-19 pandemic, researchers had no access to national linked healthcare data across the population of England to enable critical

WHAT IS ALREADY KNOWN ON THIS TOPIC
At the start of the covid-19 pandemic, approved researchers were unable to access national, linked health data across the whole UK population to conduct analyses that would support healthcare and public health policy

WHAT THIS STUDY ADDS
In partnership with NHS Digital, the British Heart Foundation Data Science Centre has developed a new trusted research environment for England, providing researchers with secure access to linked health data from primary and secondary care, registered deaths, covid-19 laboratory and vaccination data, and cardiovascular specialist audits.

These datasets cover almost the entire population of England (>54 million people); similar linked data have been made available in trusted research environments for Scotland and Wales (>8 million people).

Large numbers of approved researchers are now accessing health data on almost all people in the UK to address important covid-19 related research questions.
The newly established NHS Digital Trusted Research Environment (TRE) for England provides researchers with secure, remote access 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 causes of death), covid-19 laboratory tests, and community dispensed medicines (table 1, fig 1, CVD-COVID-UK Dataset dashboard.\textsuperscript{10} CVD-COVID-UK Dataset trusted research environment asset in Health Data Research Innovation Gateway).\textsuperscript{11} Further incorporation of specialist intensive care, cardiovascular audit, hospital electronic prescribing, and covid-19 vaccination data are planned soon.

Supplementary figure 1 shows how data generated in hospitals, general practices, community pharmacies, covid-19 NHS and commercial testing services, covid-19 vaccination services, and registry offices flow to NHS Digital. Regular flows to NHS Digital of specialist intensive care and cardiovascular audit data from several different providers were established for the first time during 2020. Furthermore, the new primary care dataset established during spring 2020 by the GP Extraction Service, is the most comprehensive yet to flow to NHS Digital.

**Data processing and linkage**

The processing and quality checks applied by data providers and processors before arriving into and within NHS Digital vary by dataset (see notes accompanying supplementary figure 1 for full details). Linkage between datasets is enabled by NHS Digital’s Master Person Service,\textsuperscript{12} which aims to match multiple records with 99% accuracy for each person from different clinical computer systems (eg, hospitals and general practices) to a single unique identifier, the National Health Service number representing a single person. NHS numbers that are included in records are cross checked with associated personal details, including age, sex, and postcode, within the Personal Demographics Service. If the NHS number is verified, no further processing is required. If the NHS number cannot be verified or is not provided, the Master Person Service then attempts to match the records to a single NHS number held in the Personal Demographics Service. The Master Person Service checks the personal details in the submitted data file for closeness to the data held in the Personal Demographics Service and produces an associated match confidence score. Each unique person specific NHS number is replaced with a non-identifying unique master key (or pseudo-identifier), which enables linkage of people’s records between datasets. At present this score is not provided at person level within each dataset, although work is ongoing within NHS Digital to enable this in the future. However, NHS Digital’s monthly reports for data quality maturity index indicate that 97-100% of records submitted to NHS Digital each month include information on NHS number and other key personal variables, providing confidence in the accuracy of the matching process.\textsuperscript{13}

**Methods**

**Data resources**

The newly established NHS Digital Trusted Research Environment (TRE) for England provides researchers with secure, remote access 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 causes of death), covid-19 laboratory tests, and community dispensed medicines (table 1, fig 1, CVD-COVID-UK Dataset dashboard.\textsuperscript{10} CVD-COVID-UK Dataset trusted research environment asset in Health Data Research Innovation Gateway).\textsuperscript{11} Further incorporation of specialist intensive care, cardiovascular audit, hospital electronic prescribing, and covid-19 vaccination data are planned soon.

Supplementary figure 1 shows how data generated in hospitals, general practices, community pharmacies, covid-19 NHS and commercial testing services, covid-19 vaccination services, and registry offices flow to NHS Digital. Regular flows to NHS Digital of specialist intensive care and cardiovascular audit data from several different providers were established for the first time during 2020. Furthermore, the new primary care dataset established during spring 2020 by the GP Extraction Service, is the most comprehensive yet to flow to NHS Digital.

**Data processing and linkage**

The processing and quality checks applied by data providers and processors before arriving into and within NHS Digital vary by dataset (see notes accompanying supplementary figure 1 for full details). Linkage between datasets is enabled by NHS Digital’s Master Person Service,\textsuperscript{12} which aims to match multiple records with 99% accuracy for each person from different clinical computer systems (eg, hospitals and general practices) to a single unique identifier, the National Health Service number representing a single person. NHS numbers that are included in records are cross checked with associated personal details, including age, sex, and postcode, within the Personal Demographics Service. If the NHS number is verified, no further processing is required. If the NHS number cannot be verified or is not provided, the Master Person Service then attempts to match the records to a single NHS number held in the Personal Demographics Service. The Master Person Service checks the personal details in the submitted data file for closeness to the data held in the Personal Demographics Service and produces an associated match confidence score. Each unique person specific NHS number is replaced with a non-identifying unique master key (or pseudo-identifier), which enables linkage of people’s records between datasets. At present this score is not provided at person level within each dataset, although work is ongoing within NHS Digital to enable this in the future. However, NHS Digital’s monthly reports for data quality maturity index indicate that 97-100% of records submitted to NHS Digital each month include information on NHS number and other key personal variables, providing confidence in the accuracy of the matching process.\textsuperscript{13}

**Methods**

**Data resources**

The newly established NHS Digital Trusted Research Environment (TRE) for England provides researchers with secure, remote access 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 causes of death), covid-19 laboratory tests, and community dispensed medicines (table 1, fig 1, CVD-COVID-UK Dataset dashboard.\textsuperscript{10} CVD-COVID-UK Dataset trusted research environment asset in Health Data Research Innovation Gateway).\textsuperscript{11} Further incorporation of specialist intensive care, cardiovascular audit, hospital electronic prescribing, and covid-19 vaccination data are planned soon.

Supplementary figure 1 shows how data generated in hospitals, general practices, community pharmacies, covid-19 NHS and commercial testing services, covid-19 vaccination services, and registry offices flow to NHS Digital. Regular flows to NHS Digital of specialist intensive care and cardiovascular audit data from several different providers were established for the first time during 2020. Furthermore, the new primary care dataset established during spring 2020 by the GP Extraction Service, is the most comprehensive yet to flow to NHS Digital.

**Data processing and linkage**

The processing and quality checks applied by data providers and processors before arriving into and within NHS Digital vary by dataset (see notes accompanying supplementary figure 1 for full details). Linkage between datasets is enabled by NHS Digital’s Master Person Service,\textsuperscript{12} which aims to match multiple records with 99% accuracy for each person from different clinical computer systems (eg, hospitals and general practices) to a single unique identifier, the National Health Service number representing a single person. NHS numbers that are included in records are cross checked with associated personal details, including age, sex, and postcode, within the Personal Demographics Service. If the NHS number is verified, no further processing is required. If the NHS number cannot be verified or is not provided, the Master Person Service then attempts to match the records to a single NHS number held in the Personal Demographics Service. The Master Person Service checks the personal details in the submitted data file for closeness to the data held in the Personal Demographics Service and produces an associated match confidence score. Each unique person specific NHS number is replaced with a non-identifying unique master key (or pseudo-identifier), which enables linkage of people’s records between datasets. At present this score is not provided at person level within each dataset, although work is ongoing within NHS Digital to enable this in the future. However, NHS Digital’s monthly reports for data quality maturity index indicate that 97-100% of records submitted to NHS Digital each month include information on NHS number and other key personal variables, providing confidence in the accuracy of the matching process.\textsuperscript{13}
Request Service14 and received approval for the CVD-COVID-UK research programme (reference 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), which ensures that research projects undertaken fall within the scope of the ethical and regulatory approvals for the CVD-COVID-UK consortium programme. The trusted research environment provides secure storage and remote data access, avoiding the need for any person level data to leave NHS Digital (fig 1).

Table 1 | Overview of available and planned linked resources

<table>
<thead>
<tr>
<th>Availability of population-wide linked data, and data description</th>
<th>Data resource</th>
</tr>
</thead>
<tbody>
<tr>
<td>Available January 2021</td>
<td></td>
</tr>
<tr>
<td>Primary care</td>
<td>GPES: General Practice Extraction Service (GPES) data for pandemic planning and research</td>
</tr>
<tr>
<td>Hospital episodes</td>
<td>Secondary Uses Service (SUS+) and Hospital Episode Statistics (HES) including: Emergency Care Dataset (ECDS), admitted participant care, adult critical care, and outpatients</td>
</tr>
<tr>
<td>Death registry</td>
<td>Office for National Statistics (ONS) death registrations</td>
</tr>
<tr>
<td>Laboratory tests for covid-19</td>
<td>Public Health England (PHE) Second Generation Surveillance System (SGSS) test data results for covid-19 (pillars 1 and 2)</td>
</tr>
<tr>
<td>Community dispensing data</td>
<td>NHS Business Services Authority (BSA) community dispensing data</td>
</tr>
</tbody>
</table>

To become available during 2021

<table>
<thead>
<tr>
<th>Intensive care unit</th>
<th>Intensive Care National Audit and Research Centre (ICNARC) data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cardiovascular specialist audit and registry data</td>
<td>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</td>
</tr>
<tr>
<td>National covid-19 vaccination data</td>
<td>National immunisation management system covid-19 vaccination dataset</td>
</tr>
<tr>
<td>Hospital electronic prescribing data</td>
<td>Adverse reactions dataset for covid-19 vaccination</td>
</tr>
</tbody>
</table>

The NHS Digital TRE team provides named, approved researchers with secure log in details for remote access. Users of the trusted research environment can access data within NHS Digital’s data platform, where they can interrogate, manage, analyse, and visualise data using Databricks, a collaborative analytics platform that supports SQL and Python, and R Studio, a data analysis environment for R. Other analytical tools (eg, Stata) are to be made available soon. Flexible cloud compute is available through Amazon Web Services. The CVD-COVID-UK consortium’s data management and methods group has created user guides that explain how to navigate the environment; show where to find the data, data dictionaries, collaborative and personal work spaces and folders, and spaces for imported files (eg, phenotype code lists from external code libraries); and provide rules for database etiquette. NHS Digital hosts a regular user forum to consult with users on further developments in the NHS Digital TRE and

<table>
<thead>
<tr>
<th>Secure storage and deidentification of linked datasets</th>
<th>Secure access for approved researchers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Secure data flows</td>
<td>Secure access to deidentified data and aggregate outputs</td>
</tr>
</tbody>
</table>

Fig 1 | Overview of current (in bold) and planned data flows into NHS Digital Trusted Research Environment (TRE) for England
runs a service to deal with problems and queries from users.

**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 CVDs, related risk factors, and treatments 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 CVDs. Lay summaries of approved projects are published on the consortium’s web page. All consortium members (currently more than 130 people from 40 research or NHS organisations, including NHS data custodians) commit to: conducting research according to the “five safes”; an inclusive approach that enables additional researchers to join the consortium as work evolves; and the open sharing of research protocols, analysis code, and phenotype code lists and algorithms (through the BHF Data Science Centre Github repository and, for phenotypes, the Health Data Research UK Phenotype Library).

Researchers based in UK research organisations, including universities, NHS bodies, and charities, can join the consortium provided they are willing and able to agree to its principles, and they can access the data within the NHS Digital TRE provided their organisation is willing and able to sign the joint data sharing agreement with NHS Digital. A streamlined mechanism for requesting information and potential data access is available through the Health Data Research UK Innovation Gateway. The consortium does not currently have any members based in non-UK organisations and has not yet explored data access for such researchers with NHS Digital, but in the meantime it welcomes contributions from the international research community through collaboration with

---

**Table 2 | Key details of main data resources**

<table>
<thead>
<tr>
<th>Name of resource</th>
<th>Primary care</th>
<th>Hospital episodes</th>
<th>Death registry</th>
<th>Laboratory tests for covid-19</th>
<th>Community dispensing</th>
</tr>
</thead>
<tbody>
<tr>
<td>GDPPR: General Practice Extraction Service (GPES) data for pandemic planning and research</td>
<td>Secondary Uses Service (SUS+)</td>
<td>Hospital Episode Statistics (HES) including: Emergency Care Dataset (ECDs), admitted participant care, adult critical care, outpatients</td>
<td>Civil Registration-Deaths (Office for National Statistics) asset</td>
<td>Public Health England (PHE) Second Generation Surveillance System (SGSS) covid-19 test results</td>
<td>NHS Business Services Authority (BSA) community dispensing data</td>
</tr>
</tbody>
</table>

**Who is included?**

People registered with a general practice in England, without a registered objection to sharing of data with NHS Digital, alive on 1 November 2019 and retained on the resource. 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 positive test result under pillar 1 or pillar 2 testing guidelines. People with at least one prescription dispensed in the community.

**What is recorded?**

Personal characteristics, diagnoses, symptoms, signs, prescriptions, referrals, immunisations, behavioural factors, tests. Diagnoses, procedures, personal characteristics (including ethnicity and area level deprivation), admission and discharge dates, hospital and other variables. Diagnoses, procedures, personal characteristics (including ethnicity and area level deprivation), admission and discharge dates, hospital and other variables. Date of death, date death was registered, sex, underlying cause of death, district, subdistrict, place of death (code, establishment, and type). Personal characteristics (age, sex, ethnicity, lower layer super output area), date of specimen, laboratory report, reporting laboratory. Information on dispensed drugs (name, strength, substance, quantity).

**How are records coded?**


**Period of record dates**

From the earliest record for each person to present. November 2019 to 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, up to 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.

**No 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 No of records**

4 937 121 423. 2 781 364 103. 365 438 996. 18 815 693. 1 160 138. 2796 440 797.

**No of people known to be alive on 1 January 2020**

54 388 181. 6 251 673. 42 582 312. 417 236. 776 503. 406 236 265.

**Total No of records among people alive on 1 January 2020**

4 870 642 482. 2 491 646 379. 2289 312 294. 457 412. 988 174. 2 329 914 169.
UK based consortium members (see supplementary annexe 1).

As a result of demand from researchers, in particular those addressing questions of critical policy relevance for the UK government chief scientific adviser’s national core studies programme, established to coordinate the UK’s covid-19 research response, the BHF Data Science Centre is currently seeking to extend the ethics and regulatory approvals for the consortium’s research programme. This will enable it to cover a comprehensive range of covid-19-relevant research that can be conducted using national linked EHRs.

Data updates
All the datasets are updated regularly within 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 in the NHS Digital TRE on a synchronised monthly schedule, but more frequent updates (eg, weekly or daily) can be requested according to clinical, public health, and health policy research needs.

Data security, privacy, and confidentiality
The data within the NHS Digital TRE are deidentified (ie, directly identifying data items, such as each person’s name, address, NHS number, and exact date of birth, are removed) and pseudonymised (ie, each unique person specific NHS number is replaced with a non-identifying unique master key). Postcodes are replaced with lower layer super output areas that can be converted to indices of multiple deprivation. NHS Digital also operates a “safe outputs” service: although approved researchers work with record level and person level data within the trusted research environment, only summary, aggregate results can be extracted, subject to approval through disclosure control processes and rules, which follow similar principles to those used by other established trusted research environments, such as the Secure Anonymised Information Linkage (SAIL) Databank for Wales and the Scottish National Data Safe Haven. 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.

Derivation of participant characteristics and disease diagnoses
For descriptive analyses, we defined a linked cohort as all people in the primary care data known to be alive on 1 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 31 October 2020, the latest common record date across the datasets. We defined eligible records within the hospital episodes, death registry, and covid-19 laboratory test results as those that 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 1 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, we defined previous diagnoses of myocardial infarction (yes or no), stroke or transient ischaemic attack (defined as ischaemic stroke, haemorrhagic stroke, unspecified stroke, or transient ischaemic attack) (yes or no), diabetes (yes or no), and obesity (yes or no) from Systematized Nomenclature of Medicine Clinical Terms (SNOMED-CT) concept codes in the primary care data and from ICD-10 (international classification of diseases, 10th revision) codes in the hospital episodes (main or secondary diagnostic code position in the admitted patient care component of the Hospital Episode Statistics data) recorded before 1 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. Two clinicians independently reviewed all phenotype code lists and, when applicable, classified ICD-10 terms and SNOMED-CT concepts into prevalent or incident diagnoses (supplementary tables 1-4).

We ascertained people with a first ever incident myocardial infarction or stroke or transient ischaemic attack as those with no diagnosis of myocardial infarction or stroke or transient ischaemic attack before 1 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 1 January and 31 October 2020 (phenotype algorithms provided in supplementary tables 1 and 2).

We ascertained people with a confirmed or suspected covid-19 diagnosis as those with a positive polymerase chain reaction or antigen test result from the covid-19 laboratory test data, with a specimen date on or before 31 October 2020; or with a covid-19 diagnosis SNOMED-CT concept code appearing in the primary care data, with event date on or before 31 October 2020; or with 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 31 October 2020; or with a death registration including a diagnosis ICD-10 code (as underlying or contributing cause), with date of death on or before 31 October 2020. Supplementary table 5 provides definitions for all covid-19 phenotypes.
Table 3 | Characteristics of linked cohort and of people with a confirmed or suspected covid-19 diagnosis, by data resource. Values are numbers (percentages) unless stated otherwise

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Total population (n=54 388 181)</th>
<th>Source of data on confirmed or suspected covid-19 diagnoses (% of subgroup)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Primary care records (n=714 162)</td>
</tr>
<tr>
<td><strong>Sex:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Women</td>
<td>27 718 313 (51.0)</td>
<td>400 468 (1.4)</td>
</tr>
<tr>
<td>Men</td>
<td>26 661 365 (49.0)</td>
<td>313 558 (1.2)</td>
</tr>
<tr>
<td>Unknown</td>
<td>8483 (0.0)</td>
<td>156 (1.8)</td>
</tr>
<tr>
<td><strong>Age group (years):</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0-17</td>
<td>11 188 814 (20.6)</td>
<td>60 571 (0.5)</td>
</tr>
<tr>
<td>18-29</td>
<td>7 925 142 (14.6)</td>
<td>151 304 (1.9)</td>
</tr>
<tr>
<td>30-49</td>
<td>16 701 289 (27.0)</td>
<td>207 672 (1.4)</td>
</tr>
<tr>
<td>50-69</td>
<td>13 026 680 (24.0)</td>
<td>179 977 (1.4)</td>
</tr>
<tr>
<td>≥70</td>
<td>7 543 288 (13.9)</td>
<td>114 505 (1.5)</td>
</tr>
<tr>
<td>Unknown</td>
<td>2 888 (0.0)</td>
<td>133 (4.8)</td>
</tr>
<tr>
<td><strong>Ethnicity:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>41 786 891 (76.8)</td>
<td>556 489 (1.3)</td>
</tr>
<tr>
<td>Mixed</td>
<td>11 566 062 (2.1)</td>
<td>11 810 (1.0)</td>
</tr>
<tr>
<td>Asian and Asian British</td>
<td>4 589 775 (8.4)</td>
<td>95 752 (2.1)</td>
</tr>
<tr>
<td>Black and black British</td>
<td>1 860 340 (3.4)</td>
<td>20 863 (1.1)</td>
</tr>
<tr>
<td>Other</td>
<td>2 138 019 (3.9)</td>
<td>14 334 (0.7)</td>
</tr>
<tr>
<td>Unknown</td>
<td>2 857 093 (5.3)</td>
<td>14 916 (0.5)</td>
</tr>
</tbody>
</table>

**Previous diagnoses:**

Stroke or transient ischaemic attack:

- No: 51 917 717 (97.8) 685 197 (1.3) 745 978 (1.4) 108 959 (0.2) 39 427 (0.1)
- Yes: 11 964 662 (2.2) 28 985 (2.6) 30 525 (2.6) 17 753 (1.5) 11 077 (0.9)

Myocardial infarction:

- No: 51 250 900 (97.9) 689 909 (1.3) 750 466 (1.4) 110 106 (0.2) 41 415 (0.1)
- Yes: 1 137 281 (2.1) 24 253 (2.1) 26 037 (2.3) 16 243 (1.4) 9089 (0.8)

Obesity:

- No: 49 827 060 (91.6) 625 911 (1.3) 687 445 (1.4) 97 905 (0.2) 40 778 (0.1)
- Yes: 4 561 121 (8.4) 88 251 (1.9) 89058 (2.0) 28 444 (0.6) 9726 (0.2)

Diabetes:

- No: 50 778 499 (93.4) 642 095 (1.3) 700 133 (1.4) 88 262 (0.2) 33 283 (0.1)
- Yes: 3 609 682 (6.6) 72 067 (2.0) 76 570 (2.1) 38 087 (1.1) 17 221 (0.5)

*From Hospital Episode Statistics-admitted patient care data.

This manuscript was prepared in accordance with the REporting of studies Conducted using Observational Routinely-collected Data (RECORD) guidance (see supplementary annexe 2).26

**Patient and public involvement**

The lay panel of the UK National Institute 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 CVD are members of the consortium and its approvals and oversight board, enabling cogeneration of research ideas and providing valuable perspective and input on research proposals, lay summaries, and research outputs.

**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. Supplementary table 6 provides information on the data fields available, and definitions of these within each dataset, and supplementary table 7 summarises the codes included in the primary care dataset.

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 The dataset includes data from 98% of all English general practices across all relevant general practice computer system suppliers (TPP, EMIS, Practice Systems, and Microtest) and holds about 4.9 billion records on 54.4 million people alive on 1 January 2020 (>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 (>90% of all currently extracted for a wide range of purposes by NHS Digital’s GP Extraction Service), covering a broad range of diagnoses and procedures (from the start of each person’s records) along with laboratory results, physical measurements, clinical referrals, and prescriptions (supplementary table 7). Although more than 90 000 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 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 OPCS-4 (operating procedure codes, version 4). The SUS+ resource contains raw data collected from NHS healthcare providers, representing the most up-to-date hospital episodes within NHS Digital (among hospitals making prompt and complete returns). These data are consolidated, validated, and cleaned monthly to form the Hospital Episode Statistics database. As a result, each month’s Hospital Episode Statistics data become available about two months behind real time. Thereafter, a fixed update is produced for each full year of Hospital Episode Statistics data. Among 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 Hospital Episode Statistics dataset holds 0.2 billion records for 42.3 million people (from 1997 onwards).

Death registration data flow daily to NHS Digital from the ONS civil registration dataset, including date, cause (coded with ICD-10), and place of death, and are available 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. Among the 54.4 million people in our linked cohort, 417236 died on or before 31 October 2020. The Second Generation Surveillance System (SGSS) is the national laboratory reporting system used in England to capture routine laboratory data on mainly infectious diseases, including SARS-CoV-2, and antimicrobial resistance. SGSS provides reports daily to NHS Digital on positive covid-19 test results (including the test date) fed directly from pillar 1 pathology laboratories (ie, established laboratories in hospitals for patients as well as NHS key workers) and indirectly from pillar 2 laboratories (ie, new, centralised, mostly privately run laboratories, created specifically for covid-19 testing for the wider population). By 31 October 2020, a total of 884341 people had at least one positive covid-19 test result recorded in the SGSS covid-19 laboratory test dataset, with 776503 (88%) linkable to the 54.4 million person cohort.

The community dispensing dataset, provided monthly to NHS Digital 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 drug coded from the British National Formulary Dictionary of Medicines and Devices. Among the 54.4 million person cohort, there were dispensed drug records for more than 40.6 million people and about 2.3 billion records (from April 2018).

**Personal characteristics and CVD incidence**

Table 3 shows the characteristics of the linked cohort of 54.4 million people alive on 1 January 2020; 51% were women and 14% were aged 70 years or older, with a mean age of 40.0 years for men and 41.6 years for women. By linking and combining person level records from primary care and hospital episodes, ethnicity information is available for approximately 95% of people, among whom 63% have their ethnic group recorded in primary care and 92% in hospital episodes data (fig 2). A previous diagnosis of stroke or transient ischaemic attack or of myocardial infarction is recorded for 2.2% and 2.1% of people, respectively, whereas 7% and 8% people have a record indicating a previous diagnosis of diabetes and obesity, respectively. Among 53.3 million people with no previous diagnosis of stroke or transient ischaemic attack, 98721 had a first ever incident stroke or transient ischaemic attack between 1 January and 31 October 2020, 30% of which were recorded only in primary care (ie, not in hospital episodes or death registry data) and 4% only in death registry records (fig 2). Among 53.2 million people with no previous diagnosis of myocardial infarction, 62966 had an incident myocardial infarction during follow-up, 8% of which were recorded only in primary care and 12% only in death registry records (fig 2).

**Diagnoses of covid-19**

Among people in the linked cohort, 959470 had a confirmed or suspected covid-19 diagnosis between 1 January and 31 October 2020 (714162 in primary care data, 126349 in hospital admission records, 776503 in covid-19 laboratory test data, and 50504 in death registry records). Although 58% of these were recorded in both primary care and covid-19 laboratory test data, 15% and 18%, respectively, were recorded in only one (fig 3).

Although women were more likely to have a confirmed or suspected covid-19 diagnosis in their primary care records (1.4% women v 1.2% men and in covid-19 laboratory test data (1.5% v 1.3%), they were less likely to have a covid-19 diagnosis recorded in hospital episodes (0.21% v 0.26%) or on death certificates (0.08% v 0.10%). Older people were more likely to have a covid-19 diagnosis from hospital episodes and death registrations, although young adults were more likely to have covid-19 diagnoses recorded in covid-19 laboratory test data and primary care. People with unknown age or sex were more than 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 had a covid-19 diagnosis in primary care and in the covid-19 laboratory test data compared with other ethnicities. Such differences were not, however, observed in information from hospital episodes or death certificates. People with a previous diagnosis of stroke or transient ischaemic attack, myocardial infarction, obesity, or diabetes were 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 test results, covid-19 related hospital admissions, and deaths...
with covid-19 on the death certificate, the data from our linked cohort concurs with those from relevant cumulative counts of people with covid-19 (supplementary table 8).

Discussion
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 a UK initiative to accelerate UK-wide research on covid-19 and CVD and beyond. We include descriptive analyses of a cohort of 54.4 million people alive at the start of 2020, including more than 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 through the BHF Data Science Centre’s CVD-COVID-UK consortium. Notably, combining person level information across data sources delivers about 95% complete data on key characteristics, including age, sex, and ethnicity, and is essential for ascertaining CVDs of interest, such as stroke and myocardial infarction. About 90% of people with a positive covid-19 laboratory test result 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 about 20% additional people with confirmed or suspected covid-19.

Comparison with other resources
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 (eg, Clinical Practice Research Datalink, Oxford-RCGP Research and Surveillance Centre, QResearch, and, more recently, OpenSafely). Population coverage for each of these has increased during the pandemic. Supplementary table 9 provides a summary of the population based linked healthcare resources for England, which cover a population of more than five million people and include primary care data as an anchoring component. 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 in the UK and has extremely well developed and robust processes for maintaining data security and privacy. Alignment of the new NHS Digital TRE 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.

Fig 2 | Data sources reporting person level data on ethnicity, incident stroke or transient ischaemic attack (TIA), and incident myocardial infarction
The currently available linked data assets comprise, to our knowledge, 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 for the first time making linked nationwide covid-19 laboratory testing and community dispensing data available for research. Given that the linked cohort comprises more than 96% coverage of the English population, it represents the English population in terms of age, sex, ethnicity, and diabetes when compared with UK government official statistics for England,40-42 includes the full distribution of general practices according to geographical 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 underrepresented in research (eg, several tens of thousands in each of the ethnic minority groups); younger people for whom poor outcomes of covid-19 are uncommon but nonetheless devastating (eg, almost 20 million people younger than 30 years, among whom 79 covid-19 related deaths were recorded by 31 October 2020); people experiencing the common exemplar cardiovascular outcomes of stroke or transient ischaemic attack and myocardial infarction (many tens of thousands), suggesting substantial potential to support studies on the impact of covid-19 on subtypes of stroke and myocardial infarction 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 minimising 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 trusted research environment model. Researchers from many different organisations have been able to gain rapid access to the linked datasets through 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 among 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). Furthermore, the consortium 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 trusted research environment 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 it 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 algorithms19 and analytical code18 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, 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 trusted research environment has generated knowledge about linked EHR data and routes to data access across the UK health
data science research community, benefiting other UK-wide initiatives, including the International Severe Acute Respiratory and Emerging Infection Consortium (ISARIC) study on the clinical characteristics of people admitted to hospital with covid-19,53 collaborative efforts to address the determinants of covid-19 susceptibility, severity, and outcome through analyses of population based cohorts with biosamples linked to national EHRs44 45; the Randomised Evaluation of COVID-19 Therapy (RECOVERY) trial of treatments for covid-1996; the COVID-19 Genomics UK (COG-UK) covid-19 viral sequencing study67; 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).20

Nevertheless, limitations do exist: 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); 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; the trusted research environment currently has a relatively limited range of services and analytical tools, although NHS Digital is committed to expanding these; 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; information on the accuracy of the Master Patient Service matching at the level of each record within each dataset is needed to provide assurance of high linkage quality and to allow assessment of whether this varies by important patient characteristics, such as age, ethnicity, and deprivation; several previous studies have shown that nationally collated, coded health data from primary and secondary healthcare settings in the UK are sufficiently accurate for many research purposes, using a range of different validation approaches,48 49 but detailed, methodologically robust validation studies comparing contemporary linked UK routine data sources with expert adjudication based on the complete medical records for particular conditions are relatively rare, because they are time consuming, resource intensive, and challenging to perform50 51; while data quality checks have been performed on each dataset before creating the linked data resource, these vary by dataset (see supplementary figure 1, notes on data processing and quality checks) and future analyses might require additional checks to detect duplicates and minimise influential errors, outliers, and inconsistent records.

Sustainability for the future

NHS Digital has already invested in the infrastructure of its Data Access Environment that supports the new trusted research environment service. Because this service is new, ongoing costs remain uncertain but will operate on the basis of cost recovery and will vary depending on numbers of datasets being accessed, number of researchers, software licensing costs, and compute usage. The costs of establishing and maintaining the trusted research environment and associated services for the CVD-COVID-UK consortium and research aligned to the UK government’s chief scientific officer’s national core studies programme are being met by the BHF Data Science Centre and funding from the national core studies programme. We anticipate that data access costs for future research will be met through a range of research funding routes (potentially including the National Institutes of Health Research, UK research councils, charities, industry, and others) given the considerable value and research potential of these data.

Extending coverage across the four nations of the UK

The BHF Data Science Centre has also enabled access for members of the CVD-COVID-UK consortium to similar, albeit not identical, linked health data in separate, national trusted research environments in Scotland (Scottish National Data Safe Haven) and Wales (SAIL Databank), with plans to extend further to Northern Ireland.10 22 23 Thus we have already extended coverage to enable studies across a population of more than 65 million people, including the majority of people in the UK and increasing possibilities for geographical comparisons across the UK. Owing to differences in data structure and coding procedures between nations, we advocate the development of analysis plans that aim for maximum consistency but allow for these nation specific differences. When appropriate, results of nation specific analyses can be combined to produce UK-wide results. Such combined analyses will, increasing, 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 trusted research environments.52

Conclusion

We describe provision for research of linked nation-wide EHR data for England and show the importance of linking person level data from different health settings for defining exemplar CVD outcomes, covid-19 diagnoses, and key characteristics. By covering almost the entire population of England, the resource includes all age groups, ethnic groups, geographical locations, and socioeconomic, health, and personal characteristics, and it can enable statistically powerful population scale research with large numbers of outcomes. The resource is accessible to approved researchers through a secure trusted research environment hosted by NHS Digital to support research on
covid-19 and CVD, with plans to expand to benefit a broad range of research.

**AUTHOR AFFILIATIONS**

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

**Contributors:** All authors drafted and revised the manuscript. AL prepared the figures and graphics. JS and CS contributed equally. AW, RD, SH, JC, SI, VW, SD, AB, and WW were responsible for data wrangling, quality assurance, and analysis (including generating phenotype definitions). CS is the guarantor. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

In addition, the following people are acknowledged:

**Consortium coordination (BHF Data Science Centre core team):** Rouven Priedon, Cathie Sudlow, Lynn Morrice, and John Walsh.

**Consortium membership:** www.hdruk.ac.uk/wp-content/uploads/2021/01/21O128-CVD-COVID-UK-Consortium-Members.pdf (also see supplementary annex 1)—consortium members who contributed to discussions leading up to this manuscript and provided helpful insights and comments.

**Public and patient advisory panel:** Suzannah Power, Lynn Laidlaw, Michael Molere, and John Walsh.

NHS Digital (coordination, data management or provision, data access request, and information governance support, trusted research environment support: Garry Coleen, Cath Day, Elizabeth Gaffney, Tim Gentry, Lisa Gray, Sam Hollings, Richard Irvine, Brian Roberts, Estelle Spence, and Janet Waterhouse.

The results described are based on data from patients, collected by the NHS as part of its care and support. We also thank the data providers who make anonymised data available for research.

**Funding:** The British Heart Foundation Data Science Centre (grant No SP/19/3/34678, awarded to Health Data Research (HOR) UK) funded co-development (with NHS Digital) of the trusted research environment, provision of linked datasets, data access, user software licences, computational usage, and data management and wrangling support, with additional contributions from the HOR UK data and connectivity component of the UK governments’ 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.

AA is supported by Health Data Research UK (HDR-9006), which receives its funding from the UK Medical Research Council (MRC), Engineering and Physical Sciences Research Council (EPSRC), Economic and Social Research Council (ESRC), Department of Health and Social Care (England), Chief Scientist Office of the Scottish Government Health and Social Care Directors, Health and Social Care Research and Development Division (Welsh government), Public Health Agency (Northern Ireland), British Heart Foundation (BHF), and Wellcome Trust; and Administrative Data Research UK, which is funded by the ESRC (grant ES/S007393/1). AB is supported by research funding from the National Institute for Health Research (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 (BCDSAI/100005) and by core funding from UK MRC (MR/L003120/1), BHF (RG/13/13/30194; RG/18/13/31946), and NIHR Cambridge Biomedical Research Centre (BRC-1215-20014). JC, JS, and RD are supported by the Health Data Research (HDR) UK South West Better Care Partnership and NIHR Bristol Biomedical Research Centre. SD is supported by HDR UK London, which receives its funding from HDR UK funded by the UK MRC, EPSRC, ESRC, Department of Health and Social Care (England), Chief Scientist Office of the Scottish Government Health and Social Care Directors, Health and Social Care Research and Development Division (Welsh government), Public Health Agency (Northern Ireland), BHF, and Wellcome Trust; Alan Turing Fellowship (EP/NS10129/1). NIHR Biomedical Research Centre at University College London Hospital NHS Trust. WW is supported by the University of Bristol Medical Research Council Integrative Epidemiology Unit (MC/UI_00011/4). WW is supported by a Scottish senior clinical fellowship, CSO (SCAF/17/01).

The views expressed are those of the authors and not necessarily those of the organisations listed. The funders of this work played no role in the collection, analysis, or interpretation of data; in the writing of the report; or in the decision to submit the article for publication.

**Competing interests:** All authors have completed the ICMJE uniform disclosure form at www.icmje.org/coi_disclosure.pdf and declare: support from the funders listed above; no financial relationships with any organisations that might have an interest in the submitted work in the previous three years; no other relationships or activities that could appear to have influenced the submitted work. SH works as a data scientist and data curator for NHS Digital, which holds and processes the data. For inquiries about data access, please see www.healthdatagateway.org/dataset/7e5f0247-f033-4f98-aed3-3d7422b9dc6d or email bhfdsc@hdruk.ac.uk.

The three lead authors, AW, JS, and CS (the manuscript’s guarantors), affirm that the manuscript is an honest, accurate, and transparent account of the research performed. No other potential conflicts of interest relevant to this article are reported.

**Data sharing:** The authors and colleagues across the CVD-COVID UK consortium have invested considerable time and energy in developing the data resource described here and are keen to ensure that it is used widely to maximise its value. For inquiries about data access, please see www.healthdatagateway.org/dataset/7e5f0247-f033-4f98-aed3-3d7422b9dc6d or email bhfdsc@hdruk.ac.uk.

The three lead authors, AW, JS, and CS (the manuscript’s guarantors), affirm that the manuscript is an honest, accurate, and transparent account of the resource and analyses being described and that no important aspects have been omitted.

**Ethical approval:** The North East-Newcastle and North Tyneside 2 research ethics committee provided ethical approval for the CVD-COVID-UK research programme (REC No 20/NE/0161).

**Provenance:** The authors and colleagues across the CVD-COVID UK consortium have invested considerable time and energy in developing the data resource described here and are keen to ensure that it is used widely to maximise its value. For inquiries about data access, please see www.healthdatagateway.org/dataset/7e5f0247-f033-4f98-aed3-3d7422b9dc6d or email bhfdsc@hdruk.ac.uk.

The three lead authors, AW, JS, and CS (the manuscript’s guarantors), affirm that the manuscript is an honest, accurate, and transparent account of the resource and analyses being described and that no important aspects have been omitted.

**Dissemination to participants and related patient and public communities:** Results will be disseminated through the British Heart Foundation (BHF) Data Science Centre and CVD-COVID UK webpages on the Health Data Research UK website, BHF communication channels, the BHF Data Science Centre’s lay members panel, and NHS Digital communications channels.

**Provenance and peer review:** Not commissioned; externally peer reviewed.

This is an Open Access article distributed in accordance with the terms of the Creative Commons Attribution (CC BY 4.0) license, which permits others to distribute, remix, adapt and build upon this work.
Supplementary information:
Additional figures, tables, and information

Supplementary information: Annexes 1 and 2