COVID-19 trajectories among 57 million adults in England: a cohort study using electronic health records

Johan H Thygesen*, Christopher Tomlinson*, Sam Hollings, Mehrdad A Mizani, Alex Handy, Ashley Akbari, Amitava Banerjee, Jennifer Cooper, Alvina G Lai, Kehzi Li, Bilal A Mateen, Naveed Sattar, Reecha Sofat, Ana Torralbo, Honghan Wu, Angela Wood, Jonathan A C Sterne, Christina Pagel, William N Whiteley, Cathie Sudlow, Harry Hemingway, Spiros Denaxas, on behalf of the Longitudinal Health and Wellbeing COVID-19 National Core Study and the CVD-COVID-UK/COVID-IMPACT Consortium

Summary

Background Updatable estimates of COVID-19 onset, progression, and trajectories underpin pandemic mitigation efforts. To identify and characterise disease trajectories, we aimed to define and validate ten COVID-19 phenotypes from nationwide linked electronic health records (EHR) using an extensible framework.

Methods In this cohort study, we used eight linked National Health Service (NHS) datasets for people in England alive on Jan 23, 2020. Data on COVID-19 testing, vaccination, primary and secondary care records, and death registrations were collected until Nov 30, 2021. We defined ten COVID-19 phenotypes reflecting clinically relevant stages of disease severity and encompassing five categories: positive SARS-CoV-2 test, primary care diagnosis, hospital admission, ventilation modality (four phenotypes), and death (three phenotypes). We constructed patient trajectories illustrating transition frequency and duration between phenotypes. Analyses were stratified by pandemic waves and vaccination status.

Findings Among 57 032 174 individuals included in the cohort, 13 990 423 COVID-19 events were identified in 7 244 925 individuals, equating to an infection rate of 12.7% during the study period. Of 7 244 925 individuals, 460 737 (6.4%) were admitted to hospital and 158 020 (2.2%) died. Of 460 737 individuals who were admitted to hospital, 48 847 (10.6%) were admitted to the intensive care unit (ICU), 69 090 (15.0%) received non-invasive ventilation, and 25 928 (5.6%) received invasive ventilation. Among 384 135 patients who were admitted to hospital but did not require ventilation, mortality was higher in wave 1 (23 485 [30.4%] of 77 202 patients) than wave 2 (44 220 [23.1%] of 191 528 patients), but remained unchanged for patients admitted to the ICU. Mortality was highest among patients who received ventilatory support outside of the ICU in wave 1 (25 659 [50.7%] of 50 635 patients). 15 486 (9.8%) of 158 020 COVID-19-related deaths occurred within 28 days of the first COVID-19 event without a COVID-19 diagnosis on the death certificate. 10 884 (6.9%) of 158 020 deaths were identified exclusively from mortality data with no previous COVID-19 phenotype recorded. We observed longer patient trajectories in wave 2 than wave 1.

Interpretation Our analyses illustrate the wide spectrum of disease trajectories as shown by differences in incidence, survival, and clinical pathways. We have provided a modular analytical framework that can be used to monitor the impact of the pandemic and generate evidence of clinical and policy relevance using multiple EHR sources.

Funding British Heart Foundation Data Science Centre, led by Health Data Research UK.

Copyright © 2022 The Author(s). Published by Elsevier Ltd. This is an Open Access article under the CC BY 4.0 license.

Introduction

Understanding the population impact of COVID-19 requires consideration of how COVID-19 varies in severity (from asymptomatic to fatal) and time course (from acute infection to chronic sequelae [ie, long COVID]). These diverse clinical manifestations are reflected in a patients’ digital trace across multiple, often unconnected, health system organisations including public health, general practice, hospitals, intensive care, and civil death registration.

The trajectories of disease severity in COVID-19 are poorly understood for three reasons. First, there is an important need for scale to identify outcomes in less common demographic groups, and to determine the impact on comorbidities and treatments. Second, an unmet need exists to comprehensively link individual’s data across siloed institutional datasets. In practice, scale and linkage are intrinsically related concepts. In national health systems, such as that in England, datasets might encapsulate the population, but are restricted to an isolated component of the health system (eg, primary care); therefore, linkage is vital to capture the depth of patients’ interactions across aspects of health care. Conversely, in other health systems, the unit of the dataset might be that of a single health-care provider, encompassing rich patient data across both primary and secondary care, but for a small subset of the population. For datasets collected by a single health-care provider, case linkage between providers becomes necessary to expand the...
Research in context

Evidence before this study
We searched PubMed from database inception to Oct 14, 2021, for publications using the search terms “COVID-19” or “SARS-CoV-2”, “severity”, and “electronic health records” or “EHR”, without language restrictions. Multiple studies have explored factors associated with severity of COVID-19 infection, and model predictions of outcome for patients admitted to hospital. However, to date, most studies have focused on isolated facets of the health-care system, such as primary or secondary care only, have been done in subpopulations (eg, patients admitted to hospital), have been of small sample size, and often utilised dichotomised outcomes (eg, mortality or hospital admission) not representative of the full spectrum of disease. We identified no studies that comprehensively detailed severity of infections across pandemic waves, and by vaccination status and patient trajectories. Using data from linked electronic health records, at the national level, we reported the key demographic factors, frequency of comorbidities, impact of the two main waves in England, and effect of vaccination on COVID-19 severity on 57 million people alive and registered with a general practitioner in England. Additionally, we identified and described patient trajectory networks that illustrate the main transition pathways of patients with COVID-19 in the health-care system. We also provide reproducible COVID-19 phenotyping algorithms reflecting clinically relevant stages of disease severity (ie, positive tests, primary care diagnoses, hospital admission, ventilatory support, and mortality).

Implications of all the available evidence
The COVID-19 phenotypes and trajectory analysis framework provides a reproducible, extensible, and repurposible method to generate national-level data to support crucial policy decision making. By modelling patient trajectories as a series of interactions with health-care systems, and linking these to demographic and outcome data, we provide a method of identifying and prioritising care pathways associated with adverse outcomes and highlight tangible targets for intervention within the health-care system.

Methods

Study design and data sources
In this cohort study, we used eight linked, English National Health Service (NHS) datasets available within the NHS Digital Trusted Research Environment,
accessed through the CVD-COVID-UK/COVID-IMPACT Consortium. The following datasets were included: national laboratory COVID-19 testing data from the Public Health England Second Generation Surveillance System (SGSS); primary care data from the General Practice Extraction Service Extract for Pandemic Planning and Research (GDPPR);4 hospital admission data from Secondary Uses Service (SUS), Hospital Episode Statistics for admitted patient care (HES-APC), and Hospital Episode Statistics for adult critical care (HES-CC); COVID-19 hospital admission data from the COVID-19 Hospitalisations in England Surveillance System (CHESS; a dataset initiated by the NHS at the start of the pandemic to record information about patients admitted to hospital with COVID-19); COVID-19 vaccination status capturing vaccination details on a weekly basis; and mortality information from the Office for National Statistics (ONS) Civil Registration of Deaths.7

Datasets were linked at the individual level by NHS Digital using a pseudonymised version of the NHS number, a unique ten-digit patient identifier used in the UK health-care system that is assigned at birth or at the first interaction.

A Reporting of studies Conducted using Observational Routinely-collected Data statement is included in the appendix (p 35). The North East-Newcastle and North Tyneside 2 research ethics committee provided ethical approval for the CVD-COVID-UK/COVID-IMPACT research programme (REC No 20/NE/0161) to access, within secure trusted research environments, unconsented, whole-population, de-identified data from electronic health records collected as part of patients’ routine health care, which has been described previously.7

Study population and pandemic waves

We collected data between Jan 23, 2020, the date of the first recorded COVID-19 case in the UK,4 and Nov 30, 2021. We included individuals of any age who were alive at the start of the study; registered with a general practitioner (GP) in England (minimum one patient record in GDPPR), associated with a valid person pseudo-identifier, enabling data linkage; had a minimum of 28 days of follow-up time (in accordance with death estimates reported on the UK Government COVID-19 dashboard) across the eight linked data sources between the index non-fatal infection event and the study end date; and resided in England, as defined using lower-layer super output areas (LSOAs; appendix p 14).

We defined pandemic waves using a data-driven approach, in the absence of a consensus definition. We defined the first wave as the period in which more than 1000 cases per day were reported by Public Health England (Feb 20–May 29, 2020) and the second wave as the period in which more than 10000 cases per day were reported (Sept 30, 2020 to Feb 12, 2021), accounting for the increase in testing capacity (as reported on the UK Government COVID-19 dashboard). Individuals were assigned to waves on the basis of the date of their first identified COVID-19 phenotype.

COVID-19 phenotypes

To identify COVID-19 from EHR spanning all health-care settings, we combined all relevant COVID-19 events—ie, diagnosis codes in primary or secondary care, SARS-CoV-2 laboratory testing, disease outcomes, and the provision of ventilatory support (within and outside of the ICU). To improve the generalisability and reproducibility of the COVID-19 phenotypes, we created modular algorithms that can be adapted and applied in other datasets to make use of all available information for event ascertainment. We defined ten COVID-19 phenotypes reflecting clinically relevant stages of disease severity and encompassing five categories: (1) positive SARS-CoV-2 tests, (2) COVID-19 diagnosis recorded in primary care, (3) hospital admissions with a COVID-19 diagnosis, (4) ventilatory support (four phenotypes) related to COVID-19, and (5) deaths (three phenotypes). COVID-19 phenotypes were not mutually exclusive and thus it was possible for an individual to have more than one phenotype or more than one event of a specific phenotype. Positive SARS-CoV-2 tests were defined as a positive result from national testing data (SGSS), encompassing tests from NHS hospitals for individuals with a clinical need and health-care workers and swab testing from the wider population. Primary care diagnoses were ascertained using SNOMED CT concepts, and hospital admissions were identified using ICD-10 terms in any diagnostic position in HES-APC and SUS, or the presence of a patient in CHESS, a COVID-19 specific hospital dataset (appendix p 24).

We defined four ventilatory support phenotypes on the basis of ventilatory modalities: non-invasive ventilation, invasive mechanical ventilation, extracorporeal membrane oxygenation, and ICU admission. These phenotypes were defined using proprietary fields from HES-CC and CHESS in addition to OPCS-4 procedure codes (analogous to Current Procedural Terminology codes used in the USA). Ventilatory support outside the ICU was defined as the presence of a non-invasive ventilation, invasive mechanical ventilation, or extracorporeal membrane oxygenation phenotype in the absence of an ICU admission phenotype. Fatal COVID-19 events were identified from ONS Civil Registration of Deaths and secondary care (HES-APC, SUS) and defined as: (1) a suspected or confirmed COVID-19 diagnosis ICD-10 term listed anywhere on the death certificate, (2) death within 28 days of the first recorded COVID-19 event (positive test, diagnosis, or admission; consistent with UK Government and Public Health England reported threshold [UK Government COVID-19 dashboard]), where a COVID-19 diagnosis was not listed anywhere on the death certificate, or (3) a COVID-19 hospital admission with a discharge method or discharge destination denoting death, irrespective of cause and duration after the index event (appendix p 24).
We used the CALIBER rule-based phenotyping approach 7 to create reproducible phenotypes that are publicly available, including the study protocol and analytical code, on the HDR UK CALIBER Phenotype Library and GitHub. 8 Phenotype coders and further details on phenotyping are in the appendix (pp 3, 24).

Age, sex, and ethnicity were derived from the most recent non-missing value across primary care (GDPPR) and secondary care (HES-APC), with preference given to primary care in the event of a match on the same date. Ethnicity was categorised by adapting the ONS census categories to: White, Asian or Asian British, Black or Black British, Chinese, Mixed, or Other. Socioeconomic deprivation information was derived using the 2011 LSOA from GDPPR to index the 2019 English indices of deprivation 9 and IMD mapped to deprivation quintiles (quintile 1 denotes the most deprivation; quintile 5 denotes the least deprivation).

We assessed 270 previously described comorbidities, 10 across 16 clinical specialities or organ systems, using validated CALIBER phenotypes and data records from Jan 1, 1996 to Dec 31, 2019 from primary care, hospital admission, and procedure data (appendix p 27). 11 A multimorbidity variable was created as the binary sum across all 270 conditions.

At the start of the pandemic the UK Government implemented a so-called shielding policy, whereby patients without COVID-19 who had specified underlying conditions considered to make them clinically extremely vulnerable to the development of severe COVID-19 infections were advised to remain at home. 12 Patients included on the national Shielded Patient List, were identified by the presence of SNOMED CT code 1300561000000107 in their primary care record from May 4, 2020 onwards.

Vaccination status was determined from the COVID-19 vaccination dataset, including all vaccinations administered after Dec 12, 2020 (when the first official dose was administered in England). Patients were denoted as vaccinated after 14 days had elapsed since their second dose. To examine effects, vaccinated patients were matched 1:1 with unvaccinated individuals for age (5-year age groups), sex, and ethnicity. Unvaccinated individuals were defined as those who had neither received a COVID-19 vaccine nor had previously been infected with SARS-CoV-2. Survival analysis was performed from a single timepoint (Feb 1, 2021), with a minimum of 28 days follow-up.

Statistical analysis
We used descriptive statistics to summarise patient populations and characteristics. UpSet plots were used to illustrate the congruence of COVID-19 phenotype ascertainment between data sources.

COVID-19 event trajectory networks were based on research by Siggard and colleagues 13 and created by extracting and chronologically sorting COVID-19 phenotype events across each of the five categories: positive SARS-CoV-2 test, primary care diagnosis, hospital admission, ICU admission, and death. Only the first date for each phenotype event was considered. If multiple events were recorded on the same day (eg, positive SARS-CoV-2 test and hospital admission), events were ordered in the following hierarchy: positive SARS-CoV-2 test, primary care diagnosis, hospital admission, ICU admission, and death. Deterministic trajectory network plots were developed, from all the individual patient trajectories, by extracting and aggregating all pairwise transitions between events (a to b to c becomes; a to b and b to c) and calculating the frequency of transition between events and the median number of days elapsed between those two events across all the patients with observations of this transition.

We used Kaplan-Meier plots to estimate 28-day COVID-19 mortality, stratified by worst health-care presentation as a proxy for COVID-19 severity. Trajectory and Kaplan-Meier plots were stratified by pandemic waves, sex, age, ethnicity, and socioeconomic deprivation quintile. Kaplan-Meier plots were also used to compare the cumulative event frequency of the five main COVID-19 phenotypes stratified by vaccination status.

Data were accessed through the NHS Digital Trusted Research Environment. Data cleaning, exploratory analysis, phenotype creation, and cohort assembly were performed using Python (version 3.7) and Spark SQL (version 2.4.5) on Databricks (version 6.4). Analysis was performed in RStudio (version 1.3.1093.1) and R (version 4.0.3). Summary statistics were created using tableOne (version 0.12.0). Figures were constructed using ggplot2 (version 3.3.3), UpSetR (version 1.4.0), igraph (version 1.2.6), survival (version 3.2.7), and survminer (version 0.4.8) packages.

Role of the funding source
The funders had no role in study design, data collection, data analysis, data interpretation of data, or writing of the report.

Results
Our cohort comprised 57032174 individuals who were registered with a GP in England and alive on Jan 23, 2020, among whom 13990423 COVID-19 events were identified in 7244925 individuals (figure 1), equating to an infection rate of 12.7% during the study period (appendix p 14). The mean follow-up time from the first COVID-19 event was 225 days (SD 151). Of 7244925 individuals with an identified COVID-19 phenotype, 6778342 (93.6%) had a positive SARS-CoV-2 test, 3056132 (42.2%) had a COVID-19 diagnosis in primary care, 460737 (6.4%) were admitted to hospital, and 76607 (1.1%) received ventilatory support or were admitted to the ICU. A sensitivity analysis of primary admission diagnoses for COVID-19 hospital admissions identified from HES-APC showed COVID-19 was present in a
primary diagnostic position in the first admission episode in 61% of individuals with a total of 3726 unique ICD-10 codes identified (appendix p 34).

673,299 (93.5%) of 724,925 individuals with COVID-19 events did not require hospital admission or died due to causes other than COVID-19. Of 46,073 individuals admitted to hospital, 69,090 (15.0%) received non-invasive ventilation, 48,847 (10.6%) were admitted to an ICU, 25,928 (5.6%) received invasive mechanical ventilation, 21,717 (4.7%) patients received both non-invasive ventilation and invasive mechanical ventilation, and 696 (0.2%) received extracorporeal membrane oxygenation.

Demographic characteristics of the entire dataset and individuals with COVID-19 are reported in table 1. Of 724,925 individuals with COVID-19, COVID-19 infections were highest among individuals who were female (3,877,808 [53.5%]), White individuals (5,898,279 [81.4%]), Asian or Asian British individuals (7,141,689 [9.9%]), and individuals in the most deprived quintile (1,607,009 [22.2%]). The overall frequency of COVID-19 events was consistent with the distribution reported by the UK Government COVID-19 dashboard (appendix pp 16–17).

158,020 individuals died, equating to a mortality rate of 2.2%. Of these deaths, the majority occurred among patients who were admitted to hospital (114,206 [72.3%]); however, 43,814 (27.7%) individuals died without admission to hospital (appendix p 33). We identified 10,884 deaths among individuals for whom COVID-19 was a recorded cause of death without previous evidence of infection; most individuals were aged 70 years or older (9,458 [86.9%]), while White (9,965 [88.8%]), and had multiple comorbidities (median 7 conditions [IQR 4.00–11.00]; table 1).

15,486 individuals died within 28 days of a COVID-19 event without a confirmed or suspected COVID-19 diagnosis listed on the death certificate. Compared with individuals who died with COVID-19 as a recorded cause of death, these individuals were less likely to have a positive SARS-CoV-2 test and had fewer admissions and ventilatory treatments (appendix p 33). The most frequent primary causes of death were unspecified dementia (1008 [6.5%] of 15,486 deaths), cancer of bronchus and lung (809 [5.2%]), and pneumonia (794 [5.1%]; appendix p 34).

Among individuals with COVID-19 who died, the proportion of individuals who were male (86,094 [54.5%]) of 158,020 individuals, aged 70 years or older (129,138 [81.7%]), White (184,434 [87.6%]), and were in the most deprived quintile (37,636 [23.8%]) was higher than that for the overall population who had COVID-19 (table 1). Multimorbidity was substantially higher among individuals who died of COVID-19 than those who did not: individuals who died of COVID-19 had a median of 8 comorbidities (IQR 4–11) compared with 1 (0–3) for all individuals with COVID-19, five (2–4) among individuals admitted to hospital, and 4 (2–8) among individuals who received ventilatory support. Among the 40,843,631 individuals with a SNOMED-CT code indicating a high risk category for developing complications with COVID-19 due to a condition included in the Patient Shielding List, 563,515 (13.8%) contracted COVID-19, of whom 51,113 died, equating to a mortality rate of 9.1%.

Overall, mortality was higher among patients who received ventilatory support outside the ICU (12,503 [45.0%] of 27,760 patients) than among patients admitted to the ICU (19,023 [38.9%] of 48,847 patients) and patients admitted to hospital who did not receive ventilatory support (82,681 [21.5%] of 384,113 patients). Between
**Table 1: Sample characteristics of full population and individuals with COVID-19, by most severe COVID-19 phenotype**

| Ethnicity                  | Unknown | Other | Mixed | Chinese | Black or Black British | Asian or Asian British | White | All COVID-19 deaths (n=25 820) |
|----------------------------|---------|-------|-------|---------|------------------------|------------------------|-------|-------------------------------|
| Sex                       | 101 143 (1.2%) | 8 939 (1.2%) | 112 204 (2.1%) | 51 789 (10.9%) | 9 209 994 (3.8%) | 4 962 875 (8.7%) | 45 800 803 (80.3%) | 5 898 279 (81.4%) |
| Age, years                 | <18     | 18–29 | 30–49 | 50–69   | ≥70                    |                       |                   |                  |
| Sex                       |         |       |       |         |                       |                       |                   |                  |
| Male                      | 11 612 932 (20.4%) | 1 456 663 (20.1%) | 1 688 770 (27.4%) | 1 224 053 (20.5%) | 713 170 (8.4%) | 209 209 (3.5%) | 45 800 803 (80.3%) | 5 898 279 (81.4%) |
| Female                    | 28 752 390 (50.4%) | 3 877 808 (53.5%) | 3 066 099 (49.5%) | 2 655 180 (43.6%) | 1 718 468 (21.8%) | 515 599 (25.0%) | 45 800 803 (80.3%) | 5 898 279 (81.4%) |
| Sex                       |         |       |       |         |                       |                       |                   |                  |
| Male                      | 11 612 932 (20.4%) | 1 456 663 (20.1%) | 1 688 770 (27.4%) | 1 224 053 (20.5%) | 713 170 (8.4%) | 209 209 (3.5%) | 45 800 803 (80.3%) | 5 898 279 (81.4%) |
| Female                    | 28 752 390 (50.4%) | 3 877 808 (53.5%) | 3 066 099 (49.5%) | 2 655 180 (43.6%) | 1 718 468 (21.8%) | 515 599 (25.0%) | 45 800 803 (80.3%) | 5 898 279 (81.4%) |
| Sex                       |         |       |       |         |                       |                       |                   |                  |
| Male                      | 11 612 932 (20.4%) | 1 456 663 (20.1%) | 1 688 770 (27.4%) | 1 224 053 (20.5%) | 713 170 (8.4%) | 209 209 (3.5%) | 45 800 803 (80.3%) | 5 898 279 (81.4%) |
| Female                    | 28 752 390 (50.4%) | 3 877 808 (53.5%) | 3 066 099 (49.5%) | 2 655 180 (43.6%) | 1 718 468 (21.8%) | 515 599 (25.0%) | 45 800 803 (80.3%) | 5 898 279 (81.4%) |

**Data are n (%), unless otherwise specified. For comorbidities, numbers in parentheses show the number of conditions included under each clinical specialty, counts of individuals identified for each of the 270 CALIBER phenotypes are shown in the appendix (pp 27–31). Multimorbidity represents the binary sum across all 270 CALIBER phenotypes. Counts of less than 5 are presented as <5 to maintain anonymity.**

*70% of the 270 CALIBER phenotypes are shown in the appendix (pp 27–31). Multimorbidity represents the binary sum across all 270 CALIBER phenotypes. Counts of less than 5 are presented as <5 to maintain anonymity. *Severity was defined as the worst presentation of a patient in their disease course and was mutually exclusive, thus individuals were only assigned a single severity.

**Represents patients who did not have any other COVID-19 related health-care presentation before death, for whom COVID-19 was recorded on the death certificate.**
wave 1 and 2, mortality among all patients admitted to hospital decreased from 32.6% (30034 of 92108 individuals) to 26.5% (60853 of 230033 individuals; table 2). The overall proportion of patients admitted to hospital who received ventilatory support and were admitted to the ICU was similar between wave 1 and wave 2; however, an increase in the proportion of patients who received non-invasive ventilation (13.9% [12837 of 92108 individuals] to 15.4% [35516 of 230033 individuals]), and a corresponding decrease in the proportion of patients who required invasive mechanical ventilation (7.1% [6584 of 92108 individuals] to 5.4% [12436 of 230033 individuals]) was observed in wave 2, coupled with a small increase in the proportion of patients who received ventilatory support outside of the ICU (5.5% [5063 of 92108 individuals] to 6.4% [14752 of

![Table](https://www.thelancet.com/digital-health)
230033 individuals]. Despite these changes, no significant difference in mortality was observed among patients admitted to the ICU between waves 1 and 2 (40·4% [3980 of 9843 individuals] vs 41·2% [9781 of 23753 individuals]; p=0·21). By contrast, mortality decreased between waves 1 and 2 among inpatients who did not receive ventilatory support (30·4% [23 485 of 77 202 individuals]) vs 23·1% [44 220 of 191 528 individuals]; p<0·0001) and those who received ventilatory support outside the ICU (50·7% [25 69 of 50 63 individuals] vs 46·5% [68 54 of 147 52 individuals]; p<0·0001).

Kaplan-Meier curves for 28-day mortality corroborated the finding that no differences in overall mortality were observed among patients who were admitted to hospital and those not admitted to the ICU. Trajectory analysis provided further insight into the temporal progression of phenotypes and was consistent with our other findings, demonstrating an increase in the median number of days between a positive SARS-CoV-2 test and death (4 days), primary care diagnosis and death (5 days), and between hospital admission and death (3 days) in wave 2 when compared with wave 1 (figure 3). Trajectories and Kaplan-Meier curves stratified by age, sex, ethnicity, and deprivation are available in the appendix (pp 27–31). Counts of less than 5 are presented as <5 to maintain anonymity.

No single source captured all COVID-19 cases: 2641 682 (39·0%) of 6 778 342 individuals with a positive SARS-CoV-2 test also received a primary care diagnosis, whereas 3936 053 (54·3%) had a positive SARS-CoV-2 test but no other record. Of the 7 244 925 individuals identified with COVID-19 events, 397 255 (5·5%) were identified exclusively from primary care records, 31 525 (0·4%) from secondary care records alone, and 10 884 (0·2%) exclusively from mortality data (figure 4). A small number of individuals were identified from Public Health England hospital surveillance data only (CHESS; 726 126 individuals [0·2% of all hospital admissions]). Details of data source overlap are summarised in the appendix (p 15).

Comparison of 28-day cumulative event frequency of the five main phenotype categories between fully vaccinated individuals and unvaccinated controls showed that the proportion of individuals with a COVID-19 positive test, primary care diagnosis, hospital admission, requirement for ventilatory support, or death was lower
in the vaccinated group than the unvaccinated group (log-rank p< 0·0001 for both plots). Crosses denote censoring. ICU=intensive care unit.

Figure 2: Cumulative COVID-19 mortality events in wave 1 (A) and wave 2 (B), stratified by most severe COVID-19 phenotype

Shaded areas show 95% CIs. Log-rank p< 0·0001, for both plots. Crosses denote censoring. ICU=intensive care unit.

Discussion
In this study, we provided a comprehensive examination of COVID-19 disease recording patterns, severity, and patient trajectories, across pandemic waves, using linked EHRs for 57 million people in England. In contrast with existing research, which has been undertaken in clinical populations, or that has solely used administrative data, our study utilised eight complementary datasets spanning health-care settings at the national level and captured a diverse set of disease exposures and outcomes. We defined and evaluated ten COVID-19 phenotypes, associated with five severity categories, and explored disease trajectories and mortality between pandemic waves. COVID-19 testing and treatment has been carried out via a number of different routes that have rapidly evolved during the pandemic and our findings illustrate the importance of linking and triangulating information across multiple sources to maximise event ascertainment, to fully capture the spectrum of potential health outcomes and to identify patient transitions through the health-care system. The phenotypes presented have already enabled other analyses with highly relevant public health policy implications, such as assessing the association of COVID-19 vaccines ChAdOx1 and BNT162b2 with reduced hospitalisation or death. The framework can be used to monitor the impact of the current, and any subsequent pandemics, and inform policy in a systematic manner.

We utilised the CALIBER phenotyping approach and performed methods of validation including replication of findings consistent with existing literature. We internally validated our COVID-19 phenotypes through demonstration of cross-source concordance with EHRs. For example, using HES-CC, which incorporates mandatory reporting of basic and advanced respiratory support, mapping to non-invasive ventilation and invasive mechanical ventilation phenotypes respectively, 73·0% (377 of 22 618) and 77·5% (17 740 of 22 618) of these individuals received a corresponding OPCS-4 code in HES-APC or SUS (appendix p 15).

To date, phenotype algorithm validation has involved manual review of case notes by experts and generation of
positive and negative predictive value estimates. However, such an approach does not feasibly scale to the large sample sizes or phenotypes used in this study. Furthermore, due to information governance legislation, patient records are not made available at scale for research. To address these challenges, we have previously developed and applied a robust phenotyping framework that generates multiple layers of evidence towards algorithmic validity, including replication of aetiological, prognostic, and descriptive characteristics from published literature. We applied the CALIBER phenotyping framework to generate such evidence and observed consistent patient demographics and comorbidity patterns with widely reported associations between sex, age, ethnicity, and deprivation and outcomes for primary care, secondary care, and the ICU. Additionally, the ascertained infection rate of 12.7%, was comparable with official estimates from Public Health...
England (ascertained from the UK Government COVID-19 Dashboard).\textsuperscript{20} Comparison of 28-day cumulative event frequency of the five main phenotype categories between fully vaccinated individuals and unvaccinated controls reproduced the known and expected protective effects of vaccination (ie, reduced frequency of COVID-19 events).\textsuperscript{2}

A key challenge of using multiple data sources is the harmonisation of information across each source in the absence of a gold standard since each dataset contained information at different levels of resolution and reflected variations in health-care delivery across the duration of the pandemic. For example, when creating phenotypes for ventilatory support, we exploited linkage across multiple EHR sources (HES-APC, HES-CC, SUS, CHESS), in addition to OPCS-4 procedure codes for ventilatory support modalities. This approach allowed us to identify 27153 individuals who received non-invasive ventilation outside of the ICU, representing 39% of all patients treated with non-invasive ventilation, and showed that patients receiving ventilatory support outside of the ICU had the highest mortality. These important findings are consistent with observations that COVID-19 would overwhelm pre-existing critical care capacity and thus necessitated expansion of services to new areas, such as operating theatres and recovery wards.\textsuperscript{21} Furthermore, such findings illustrate the value of linkage and our rigorous phenotypes for maximising data capture, particularly in identifying groups that might otherwise be missed when comparing between datasets.

The increased median duration between COVID-19 phenotypes observed in wave 2, when compared with wave 1, has several potential explanations, including the increased availability of testing, leading to individuals being identified earlier in their infection, and changes in inpatient management, such as the creation of robust clinical management protocols and the widespread adoption of dexamethasone following the results of the RECOVERY trial.\textsuperscript{22}

Stratification of trajectories by demographics identified patterns including fewer days between positive SARS-CoV-2 test and primary care diagnosis to death among individuals of non-White ethnicity. These patterns might suggest these groups are accessing health care later in the disease course, for reasons that are likely to be multifactorial, but associated with existing socioeconomic health inequalities exacerbated by the pandemic.\textsuperscript{23}

The networks presented provide a succinct and interpretable view, at the national level, of how and when patients with COVID-19 interact with the healthcare system and the individual disease trajectories they follow. Previously, such estimates have been drawn from smaller samples that are potentially not generalisable to the entire population. Stratifying and flagging individuals who belong to a specific trajectory can additionally be utilised to recruit patients into relevant clinical trials. Additionally, we showed that significant

---

**Figure 4:** UpSet plots of individuals with one or more COVID-19 event phenotypes across seven datasets. Vertical bars report unique individuals in the intersection denoted by the intersection matrix below. Empty intersections are not shown. Horizontal bars report unique individuals identified from each dataset. Datasets were the SGSS (COVID-19 testing), GDPPR (primary care), HES-APC, HES-CC, SUS, CHESS, and the ONS Civil Registration of Deaths. CHESS=COVID-19 Hospitalisations in England Surveillance System. GDPPR=General Practice Extraction Service Extract for Pandemic Planning and Research. HES-APC=Hospital Episode Statistics for admitted patient care. HES-CC=Hospital Episode Statistics for adult critical care. ONS=Office of National Statistics. SGSS=Second Generation Surveillance System. SUS=Secondary Uses Service.
variation exists between waves, and within demographic strata, underlying the requirement for a systematic approach for monitoring. The trajectory networks can be used to monitor the impact of the pandemic on the health-care system itself and flag potential issues (eg, increases in the time between state transitions) to enable remedial action.

The COVID-19 phenotypes and trajectory analysis outlined produce a reproducible, extensible, and repurposable method to generate national-scale data to support crucial policy decision making. By modelling patient trajectories as a series of interactions within the health-care system, and linking these to demographic and outcome data, we provide a means to identify and

Figure 5: Kaplan-Meier curve of cumulative COVID-19 events in vaccinated and unvaccinated individuals. Cumulative proportion of individuals with a positive SARS-CoV-2 test (A), primary care diagnosis (B), hospital admission (C), ventilatory support (D), and COVID-19 mortality (E) in matched groups of fully vaccinated (two doses administered at least 14 days before Feb 1, 2021) and unvaccinated individuals (no doses administered before or during follow-up). Analysis was run from Feb 1, 2021 to March 15, 2021 and individuals were matched on the basis of sex, 5-year age groups, and ethnicity. Shaded areas show 95% CIs. Log-rank p<0·001, for all subplots.
prioritise care pathways associated with adverse outcomes and identify so-called patient touch points with the health-care system that might act as tangible targets for intervention—eg, access to testing for the most deprived and non-White ethnicities. Beyond the pandemic, we believe that trajectory analysis has the potential to transform analysis of complex conditions and multimorbidity by utilising linked data to disentangle the progression of individuals through the health-care system and disease states over time.

In sharing fully reproducible analytical code and phenotypes, we envisage that this work will facilitate other researchers to produce high quality and consistent outputs across a diverse range of topics. Linkage to additional datasets, as illustrated by vaccination data, allows extension to address new research questions, such as the emergence of novel variants of concern, or assessing the efficacy of booster or third primary vaccine doses on outcomes at both the patient and health-care system levels. Thus we provide a framework in which real world health-care data can be used to address questions of crucial policy relevance and positively impact the care and safety of populations at the national level.

It is important to note that the processes and clinical workflows that generate health-care data will vary in each country and would have been affected differently by the pandemic. Our approach was not restricted to specific disease groups or processes and could therefore be replicated and used in other scenarios as a basis to create frameworks tailored to individual health-care systems. The framework we provide makes use of the most popular clinical terminologies (such as ICD-10 and SNOMED CT) and thus can be used in other countries with comparable national or regional EHR data sources, to triangulate evidence across multiple sources; define COVID-19 severity phenotypes (aligned with the WHO Clinical Progression Scale24), and monitor and assess the ongoing progression of individuals through the health-care system and disease states over time.

The use of EHR and administrative data for research is associated with known challenges in terms of variable data quality, missingness, and consistency.5 The user interface of EHR systems and local clinical coding practises can also influence recording patterns. We have mitigated these challenges by triangulating information across multiple sources and by following a robust established approach for creating and evaluating EHR-derived phenotyping algorithms.6 A key strength of this study using national-scale data is that by definition, it is representative of the general population across all age groups, ethnicities, deprivation levels, and demographic characteristics. To our knowledge, this is the largest population-wide research study of COVID-19 phenotypes that includes: (1) participant information linked across multiple health-care settings at the population level, (2) detailed identification of specific ventilatory treatments, (3) classification of COVID-19 related deaths, and (4) exploration of transitions between COVID-19 events. Using multiple EHR sources spanning different health-care settings maximised infection ascertainment and reduced the effects of variable testing and data recording patterns (especially during the first wave). The phenotypes presented broadly align with the WHO Clinical Progression Scale24 and can be used to measure patient illness by tracking disease, from the absence of infection to death, and patient progression through the health-care system (ie, ambulatory illness, hospital admission, hospital treatment with ventilation, and death). Such an approach, although potentially confounded by the capacity and data variability of health-care systems, can be used as a framework to quantify disease burden, inform policy makers, and potentially identify individuals for inclusion into clinical trials.

Since the aim of this study was to create COVID-19-related phenotypes and describe the characteristics of individuals with such phenotypes, we did not use multivariable regression analyses to control for confounders and present only unadjusted crude mortality. The findings presented are therefore not associative statements and should not be interpreted as causal. However, by sharing reproducible phenotype definitions we hope to facilitate further work addressing the questions raised in this study and other COVID-19 studies exploring national-level data, as exemplified by previous research.15–17

Although our definitions of the pandemic waves differ from previous studies, we believe using non-contiguous dates enabled a more balanced comparison across periods of increased health-care system strain than including the period of low cases during the summer in the first wave. The recording of dates in EHRs will not always be fully accurate. We sought to mitigate inaccuracies by reporting the median number of days between phenotypes, and by only reporting time differences between transitions that occurred in more than 0·01% of all transitions. Data were of low granularity: for example, we could not delineate whether a patient received non-invasive ventilation followed by invasive mechanical ventilation, representing an escalation of ventilatory support, or invasive mechanical ventilation followed by non-invasive ventilation, and therefore focused on ICU admissions, for which accurate start dates were available. Our analysis did not include any physiological or pathology data associated with COVID-19 infections, since laboratory data within hospitals is not systematically captured and made available at scale in the UK. This study did not include information on medication patterns and changes in treatment approaches across waves because secondary care prescriptions are not routinely captured in electronic form. We were unable to ascertain which patients resided in care homes due to information governance restrictions of sensitive data fields and as a result we could not study this population who were disproportionately impacted by the pandemic.26
Exploiting linkage across EHRs at the national scale highlighted the health-care trajectories of individuals with COVID-19, to identify who has been affected and how.

Defining new phenotypes empowers analysts to look beyond binary outcomes (such as mortality) to clinically significant interim events (such as ventilatory treatments and ICU admissions), and enables characterisation of an individual's progression through these disease states. Furthermore, trajectory analysis provides a method to link previously disaggregated datasets to provide insight on behaviour at a national scale and enables insights from populations that might be missed by other analysis methods, for example individuals who died outside of hospital, or who received ventilatory support outside of the ICU.

As demonstrated for vaccination efficacy, this study provided an adaptable framework that could be rapidly repurposed to answer questions of crucial clinical and policy relevance, such as those around the emergence of a new variant, the need for booster doses in the context of waning immunity, or simply maximising the value of existing health-care data to understand individuals' progression through complex chronic disease trajectories.

Contributors
JHT, CT, and SD were responsible for conceptualisation, data curation, methodology, investigation, formal analysis, and data visualisation. JHT, CT, and SD wrote the first draft. SH, MAM, and AH assisted with data curation, methodology, and analysis. AW, CP, CS, and HH assisted with conceptualisation and methodology. All authors were responsible for writing, reviewing, and editing the manuscript. CS is the Director of the British Heart Foundation (BHF) Data Science Centre and coordinated approvals for and access to data within the NHS Digital Trusted Research Environment for England for CVD-COVID-UK/COVID-IMPACT. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted. The corresponding author confirms that all authors have seen and approved the final manuscript. JHT, CT, and SD are the guarantors and had full access to the raw and derived study data.

Declaration of interests
AB reports grants from the National Institute for Health Research (NIHR), British Medical Association, AstraZeneca, and UK Research and Innovation, outside the submitted work. BAM is an employee of the Wellcome Trust and reports grants from Health Data Research UK (HDR UK), UK Medical Research Council (MRC), and Diabetes UK. SH works as a data scientist and data curator for NHS Digital, which holds and processes the data. MAM is supported by research funding from AstraZeneca, outside the submitted work. AH is employed by Institute of Health Informatics, University College London. CS reports grants from the Wellcome Trust, MRC, HDR UK, University of Edinburgh, UK Research and Innovation (UKRI), and the BHF, outside the submitted work; participates on the data safety monitoring board for TARDIS; and has leadership or fiduciary roles with Cancer Research UK Early Detection and Diagnosis Research Committee, UKRI Expert Review Panel for Longitudinal Health & Wellbeing National Core Study, NIHR/UKRI Long COVID call Funding Review Panel, Accelerated Access Collaborative/NIHR/NHSX Artificial Intelligence (AI) in Healthcare Awards Funding Panel, Wellcome Trust Biomedical Resources Award Funding Panel, MRC strategic review Advisory Group for Maximising the opportunities from data science for innovative biomedical research, MRC Data Science Strategy Advisory Group, UKRI Digital Health Research and Innovation Strategy Expert Group, MRC Strategic Review of Units and Centres Main Panel & Population Health Panel, Wellcome Trust Science Funding Review External advisory group, MRC Methodology Research (Better Methods Better Research) Panel. REF 2021 Subpanel - Public Health, Health Services and Primary Care, Longitudinal Health & Wellbeing COVID-19 National Core Study Strategic Advisory Board, UK Government Clinical Research Recovery Resilience and Growth programme Clinical Trials Expert Group, UK Government Scientific Advisory Group & SAGE Task and Finish Advisory Group on mass population testing for COVID-19, Scottish Government Covid-19 Data Taskforce, BHF Data Science Centre Steering Group, Our Future Health Scientific Advisory Board, Imperial College UKRI Centre for Doctoral Training in AI for Healthcare External advisory board, Swansea University UKRI Centre for Doctoral Training in AI, Machine Learning & Advanced Computing for HDR UK Science Strategy Board / Science and Infrastructure Delivery Group, University of Bristol MRC Integrative Epidemiology Unit Scientific Advisory Board, International evaluation panel for Danish National Biobank, H2020 IMI ROADMAP Steering Committee, and the STAT-PD Steering Committee. NS reports grants from AstraZeneca, Boehringer Ingelheim, Novartis, and Roche Diagnostics; and has received consulting fees from Alimmune, Amgen, AstraZeneca, Boehringer Ingelheim, Eli Lilly, Hammi Pharmaceuticals, Merck Sharp & Dohme, Novartis, Novo Nordisk, Pfizer, and Sanofi, outside of the submitted work. WNW is supported by a Scottish senior clinical fellowship, Chief Scientist Office (SCAF/17/01), and the Stroke Association (SA CV 20/100018), has received consulting fees from Bayer; payment for expert testimony from UK courts; participates on the data safety monitoring or advisory board for PROTECT-U, CATIS, INTERACT-4, MOSES, and Bayer; has leadership of fiduciary roles with BiASP Scientific Committee; and is associate editor of Stroke. SD has received research funding from GlaxoSmithKline, AstraZeneca, Bayer, and BenevolentAI. All other authors declare no competing interests.

Data sharing
CALIBER disease phenotyping algorithms are available on the HDR Phenotype Library (https://phenotypes.healthdatagateway.org/) and machine-readable versions of the phenotypes can be obtained from a GitHub repository (https://github.com/spiros/chronological-map-phenotypes). The analytical code used to define these phenotypes in the study data sources within the NHS Digital Trusted Research Environment is available in the project repository (https://github.com/BHFDS/CCU013_01_ENG-COVID-19_event-phenotyping). This research used anonymised electronic health record and administrative data collected and curated by NHS Digital in a Trusted Research Environment. Due to information governance restrictions, the authors are unable to share the data directly, but access is available for research to third parties after approval of a research proposal and protocol, signature of data access agreements with NHS Digital, and other information governance requirements. The authors and colleagues across the CVD-COVID-UK/COVID-IMPACT consortium have invested considerable time and energy in developing this data resource and would like to ensure that it is used widely to maximise its value. For inquiries about data access, please see https://web.www.healthdatagateway.org/dataset?c=6b247-9938-4b8-ae3d-3d74226b7d6c. Results will be disseminated through the 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. Data access approval was granted to the CVD-COVID-UK consortium (under project proposal CCU013 High-throughput electronic health record phenotyping approaches) through the NHS Digital online Data Access Request Service (DARS-NIC-381078-19C5). NHS Digital data have been made available for research under the Control of Patient Information notice, which mandated the sharing of national electronic health records for COVID-19 research (https://digital.nhs.uk/coronavirus/coronavirus-covid-19-response-information-governance-hub/control-of-patient-information-copi-notice). For further details see the appendix (p 4).

Acknowledgments
The authors would like to thank the BHF Data Science Centre’s lay members panel for their input and NHS data access environment output.
checkers Lisa Gray and James Walker. This work was supported by the BHF Data Science Centre led by HDR UK (grant SP/19/3/34678). This study makes use of the identified data held in NHS Digital’s Trusted Research Environment for England and made available via the BHF Data Science Centre’s CVD-COVID-UK/COVID-IMPACT consortium. This work uses data provided by patients and collected by the NHS as part of their care and support. We would also like to acknowledge all data providers who make health relevant data available for research. The views expressed are those of the authors and not necessarily those of the organisations listed. This study was supported by a BHF Data Science Centre grant (SP/19/3/34678), awarded to HDR UK, which 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 obtaining support, with additional contributions from the HDR UK Data and Connectivity component of the UK Government Chief Scientific Adviser’s National Core Studies programme to coordinate national COVID-19 priority research. Consortium partner organisations enabled data analysts, bio-statisticians, epidemiologists, and clinicians to contribute their time to the study. This study was also funded by the Longitudinal Health and Wellbeing COVID-19 National Core Study, which was established by the UK Chief Scientific Officer in October, 2020, and funded by UK Research and Innovation (grants MC_PC_20030 and MC_PC_20059), by the Data and Connectivity National Core Study, led by Health Data Research UK in partnership with the Office for National Statistics and funded by UK Research and Innovation (grant MC_PC_20058), and by the CONVALESCE study of long COVID, which is funded by the NIHR and UKRI. This study was also supported by Health Data Research UK, which receives its funding from HDR-UK (H9006) 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 Directorate, Health and Social Care Research and Development Division (Welsh Government), Public Health Agency (Northern Ireland), BHF, and the Wellcome Trust. AA is supported by HDR UK (H9006); and Administrative Data Research UK, which is funded by the Economic and Social Research Council (grant ES/S007393/1). AGL is supported by funding from the Wellcome Trust (204814/Z/16/Z), the NIHR University College London Hospitals Biomedical Research Centre (BRC741/HH/RW/101440), NIHR Great Ormond Street Hospital Biomedical Research Centre (19RX02), and the Academy of Medical Sciences (SBF0061084). AH is supported by research funding from the HDR UK text analytics implementation project. AB, AW, HH, and SD are part of the BigData@Heart Consortium, funded by the Innovative Medicines Initiative-2 Joint Undertaking under grant agreement 116074. AW is supported by the BHF-Turing Cardiovascular Data Science Award (BCD5A100005) and by core funding from UK MRC (MR/L001102/1), BHF (RG/11/13/30194; RG/18/13/33946), and NIHR Cambridge Biomedical Research Centre (BRC1215–2004). JACS and JC are supported by the HDR UK South West Better Care Partnership and the NIHR Bristol Biomedical Research Centre at University Hospitals Bristol, Weston NHS Foundation Trust, and the University of Bristol. JACS is additionally supported by UKRI and the MRC. SD and HH are supported by HDR UK London. HH and SD are supported by the NIHR Biomedical Research Centre at University College London (UCL) Hospital NHS Trust. SD is supported by an Alan Turing Fellowship (EP/N510129/1), the BHF Data Science Centre, and the NIHR-UKRI CONVALESCE study. HH is an NIHR Senior Investigator. SD and HH are supported by the BHF Accelerator Award (AA/18/6/34223). CT is supported by a UCL UKRI Centre for Doctoral Training in AI-enabled Healthcare (EP/S021612/1), MRC Clinical Top-Up, and a studentship from the NIHR Biomedical Research Centre at University College London Hospital NHS Trust. HW is supported by the MRC (MR/S004149/2), NIHR (grant NIHR202639), and the Advanced Care Research Centre Programme at the University of Edinburgh. KL is supported by University College London and Rosetrees Trust (UCL-IHE-2020-102). NIHR, the NHS (AI_AWARD0786), the NIHR University College London Hospitals NHS Foundation Trust Biomedical Research Centre (BRC731/HH/RW/101440), and the UCL Higher Education Innovation Fund (KE1201-03-16).

References
1 The OpenSAFEly Collaborative, Walker AJ, MacKenna B, et al. Clinical coding of long COVID in England’s primary care: a federated analysis of 58 million patient records in situ using OpenSAFEly. medRxiv 2021; published online May 13. https://doi.org/10.1101/2021.05.06.21225675 (preprint).
2 Dagan N, Barda N, Kepten E, et al. BNT162b2 mRNA Covid-19 vaccine in a nationwide mass vaccination setting. N Engl J Med 2021; 384:1412–23.
3 Madur R, Rentsch CT, Morton CE, et al. Ethnic differences in SARS-CoV-2 infection and COVID-19-related hospitalisation, intensive care unit admission, and death in 17 million adults in England: an observational cohort study using the OpenSAFEly platform. Lancet 2021; 397:1711–24.
4 Khera R, Mortazavi BJ, Sangha V, et al. Accuracy of computable phenotyping approaches for SARS-CoV-2 infection and COVID-19 hospitalizations from the electronic health record. medRxiv 2021; published online May 13. https://doi.org/10.1101/2021.03.16.21223770 (preprint).
5 Kiann JC, Weber GM, Estiri H, et al. Validation of an internationally derived patient severity phenotype to support COVID-19 analytics from electronic health record data. J Am Med Inform Assoc 2021; 28:1411–20.
6 NHS Digital. General Practice Extraction Service (GPES) data for pandemic planning and research. April 27, 2021. https://digital.nhs.uk/coronavirus/gpes-data-for-pandemic-planning-and-research/ guide-for-analysts-and-users-of-the-data (accessed May 6, 2021).
7 Wood A, Denholm R, Hollings S, et al. Linked electronic health records for research on a nationwide cohort of more than 54 million people in England: data resource. BMJ 2021; 373: n826.
8 Lillie PJ, Samson A, Li A, et al. Novel coronavirus disease (COVID-19): the first two patients in the UK with person to person transmission. J Infect 2020; 80: 578–606.
9 Denaxas S, Gonzalez-Izarbe A, Direk K, et al. UK phenomics platform for developing and validating electronic health record phenotypes: CALIBER. J Am Med Inform Assoc 2019; 26: 1545–59.
10 Thygesen J, Tomlinson C, Denaxas S, CCU013_01_ENG-COVID-19_event_phenotyping. CVD-COVID-UK project: characterising COVID-19 related events in a nationwide electronic health record cohort of 55.9 million people in England. https://github.com/BHF/DSC/CCU013_01_ENG-COVID-19_event_phenotyping (accessed Aug 11, 2023).
11 UK Government. England indices of deprivation 2019. https://www.gov.uk/government/statistics/english-indices-of-deprivation-2019 (accessed May 13, 2021).
12 Kuan V, Denaxas S, Gonzalez-Izarbe 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; 1:e63–77.
13 NHS Digital. Shielded patient list: guidance for general practice. Jan 4, 2021. https://digital.nhs.uk/coronavirus/shielded-patient-list/guidance-for-general-practice (accessed June 5, 2021).
14 Siggard T, Reguant R, Jørgensen IF, et al. Disease trajectory browser for exploring temporal, population-wide disease progression patterns in 7·2 million Danish patients. Nat Commun 2020; 11:4952.
15 Whiteley WN, Ip S, Cooper JA, et al. Association of COVID-19 vaccines ChAdOx1 and BNT162b2 with major venous, arterial, and thrombocytopenic events: a whole population cohort study in 46 million adults in England. PLoS Med 2022; published online Feb 22. https://doi.org/10.1371/journal.pmed.1003926.
16 Handy A, Banerjee A, Wood AM, et al. Evaluation of antithrombotic use and COVID-19 outcomes in a nationwide atrial fibrillation cohort. Heart 2022; published online March 10. https://doi.org/10.1136/heartjnl-2021-130235.
17 Knight R, Walker V, Ip S, et al. Association of COVID-19 with arterial and venous vascular diseases: a population-wide cohort study of 48 million adults in England and Wales. medRxiv 2021; published on Nov 23. https://doi.org/10.1101/2021.11.22.21226652 (preprint).
18 Docherty AB, Harrison EM, Green CA, et al. Features of 201135 UK patients in hospital with covid-19 using the ISARIC WHO Clinical Characterisation Protocol: prospective observational cohort study. BMJ 2020; 369: m1985.
19 Richards-Belle A, Orzechowska I, Gould DW, et al. COVID-19 in critical care: epidemiology of the first epidemic wave across England, Wales and Northern Ireland. *Intensive Care Med* 2020; 46: 2035–47.

20 Ward H, Atchison C, Whitaker M, et al. SARS-CoV-2 antibody prevalence in England following the first peak of the pandemic. *Nat Commun* 2021; 12: 905.

21 Pett E, Leung HL, Taylor E, et al. Critical care transfers and COVID-19: managing capacity challenges through critical care networks. *Pediatr Crit Care Med* 2020; published online Dec 16. [https://doi.org/10.1177/1751143720980270](https://doi.org/10.1177/1751143720980270).

22 Horby P, Lim WS, Emberson JR, et al. Dexamethasone in Hospitalized Patients with Covid-19. *N Engl J Med* 2021; 384: 693–704.

23 Public Health England. COVID-19: review of disparities in risks and outcomes. June 2, 2020. [https://www.gov.uk/government/publications/covid-19-review-of-disparities-in-risks-and-outcomes](https://www.gov.uk/government/publications/covid-19-review-of-disparities-in-risks-and-outcomes) (accessed June 22, 2021).

24 WHO Working Group on the Clinical Characterisation and Management of COVID-19 infection. A minimal common outcome measure set for COVID-19 clinical research. *Lancet Infect Dis* 2020; 20: e192–97.

25 Wei W-Q, Denny JC. Extracting research-quality phenotypes from electronic health records to support precision medicine. *Genome Med* 2015; 7: 41.

26 Schultze A, Nightingale E, Evans D, et al. Mortality among Care Home Residents in England during the first and second waves of the COVID-19 pandemic: an observational study of 4.3 million adults over the age of 65. *Lancet Reg Health Eur* 2022; 14: 100295.