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, Kezhi 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 (2569 [50·7%] of 5063 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 diagnoses 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.

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

Lancet Digit Health 2022

Published Online June 8, 2022 https://doi.org/10.1016/S2589-7500(22)00091-7

*Joint first authors

Institute of Health Informatics (J H Thygesen PhD, C Tomlinson MBBS, M A Mizani PhD, A Handy MSc, Prof A Banerjee PhD, A G Lai PhD, K Li PhD, B A Mateen MBBS, Prof R Sofat PhD, A Torralbo PhD, H Wu PhD, Prof H Hemingway FMedSci, Prof S Denaxas PhD), **Clinical Operational Research Unit** (Prof C Pagel PhD), **UK Research and Innovation Centre for Doctoral Training in AI-enabled Healthcare Systems** (C Tomlinson, K Li), **British Heart Foundation Research Accelerator** (Prof S Denaxas), **and University College London Hospitals Biomedical Research Centre** (C Tomlinson, Prof H Hemingway, Prof S Denaxas), **University College London, London, UK; NHS Digital, Leeds, UK** (S Hollings PhD); **Population Data Science, Swansea University, Swansea, UK** (A Akbari MSc); **Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, UK** (J Cooper PhD, Prof J A C Sterne PhD); **The Wellcome Trust, London, UK** (B Mateen); **Institute of Cardiovascular and Medical Sciences, University of Glasgow, Glasgow, UK** (Prof N Sattar MD); **Institute of Systems, Molecular and Integrative Biology, University of Liverpool, Liverpool, UK** (Prof R Sofat); **British Heart Foundation Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, and Cambridge Centre for AI in Medicine, University of Cambridge, Cambridge, UK** (Prof A Wood PhD); **Centre for Clinical Brain Sciences,**

University of Edinburgh, Edinburgh, UK (W N Whiteley FRCP, Prof C Sudlow FMedSci)**; British Heart Foundation Data Science Centre, Health Data Research UK, London, UK** (Prof R Sofat, Prof C Sudlow, Prof S Denaxas)**; Health Data Research UK, London, UK** (Prof C Sudlow, Prof H Hemingway, Prof S Denaxas)**; Medical Research Council Population Health Research Unit, Nuffield Department of Population Health, University of Oxford, Oxford, UK** (W N Whiteley)

Correspondence to: Prof Spiros Denaxas, Institute of Health Informatics, University College London, London NW1 2DA, UK **s.denaxas@ucl.ac.uk**

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.

Added value of this study

To our knowledge, this is the first study to provide a comprehensive view of COVID-19 across pandemic waves using national data with a focus on severity, vaccination, 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 repurposable 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.

breadth of individuals captured. These issues are apparent in the existing literature, whereby studies approaching a population-scale have been restricted to primary care settings,[1] and those reporting more detailed outcomes have been limited to population subsets.[2] Previous studies have attempted to use linkage to mitigate these issues, but have fallen short of reaching a size equitable to a full population. For example, Mathur and colleagues used five linked datasets to determine COVID-19 positive tests, hospital admissions, intensive care unit (ICU) admissions, and deaths;[3] however, because the authors used a single dataset per health system, the study population was limited to 17·3 million individuals (around 30% of the population of England).

Third, there is a need for open and accessible, reproducible electronic health record (EHR) COVID-19 phenotypes that capture clinically significant treatments and outcomes, such as intensive care admission and ventilatory support. Such phenotypes need to evolve to reflect changes in clinical practice and data recording, and incorporate uncertainty—eg, clinical diagnoses in the absence of testing, or deaths occurring after infection in the absence of COVID-19 as a documented cause.

Despite the proliferation of EHR-based COVID-19 research, few studies have explicitly addressed phenotyping. Previous studies have focused on the development of binary phenotypes for COVID-19 infection and hospital admission, by comparing a single terminology (International Classification of Diseases [ICD]-10-Clinical Modification) with SARS-CoV-2 test positivity,[4] or defining severe COVID-19 using laboratory tests, medications, diagnoses, and procedures.[5] Neither of these approaches are directly transferable to national EHR data in England, and both fail to capture the full range of COVID-19 events, including primary care and deaths, and do not fully exploit the depth and breadth of available national data sources that potentially reduce the fidelity of identified phenotypes.

To address these gaps, we leveraged nationwide data from England linking laboratory testing, primary care, hospital admissions (including ventilatory support and ICU admissions), and registered deaths to define COVID-19 phenotypes reflecting clinically relevant stages of disease severity, characterised demographics and comorbidities of those individuals with these phenotypes, and compared disease severity, mortality, and trajectories stratified by pandemic wave, demographics, and vaccination status. Using data linkage, we aimed to establish an updatable framework that could be used to reconstruct an individual's COVID-19 trajectory across distinct severity states, to provide vital insight with regard to the impact of new variants, efficacy of booster doses, post-exposure prophylaxis, and emerging drug treatments.

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);[6] 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,[8] 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 10 000 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.

For more on the **CVD-COVID-UK/COVID-IMPACT Consortium** see https://www.hdruk.ac.uk/projects/cvd-covid-uk-project/

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).

See **Online** for appendix

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).

For the **Government data on COVID-19 deaths** see https://coronavirus.data.gov.uk/details/deaths

For the **UK Government data on COVID-19 testing capacity** see https://coronavirus.data.gov.uk/details/testing

For the **HDR UK CALIBER Phenotype Library** see https://phenotypes.healthdatagateway.org/phenotypes/PH1020/version/2119/detail/#home

We used the CALIBER rule-based phenotyping approach[9] to create reproducible phenotypes that are publicly available, in addition to the study protocol and analytical code, on the HDR UK CALIBER Phenotype Library and GitHub.[10] Phenotype codelists 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[11] and IMD mapped to deprivation quintiles (quintile 1 denotes the most deprivation; quintile 5 denotes the least deprivation).

We assessed 270 previously described comorbidities,[12] 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).[12] 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.[13] 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[14] 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 57 032 174 individuals who were registered with a GP in England and alive on Jan 23, 2020, among whom 13 990 423 COVID-19 events were identified in 7 244 925 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 7 244 925 individuals with an identified COVID-19 phenotype, 6 778 342 (93·6%) had a positive SARS-CoV-2 test, 3 056 132 (42·2%) had a COVID-19 diagnosis in primary care, 460 737 (6·4%) were admitted to hospital, and 76 607 (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).

6 773 299 (93·5%) of 7 244 925 individuals with COVID-19 events did not require hospital admission or died due to causes other than COVID-19. Of 46 0737 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 7 244 925 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 (714 168 [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 (9458 [86·9%]), were White (9665 [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 (138 434 [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 4 084 631 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 384 135 patients). Between

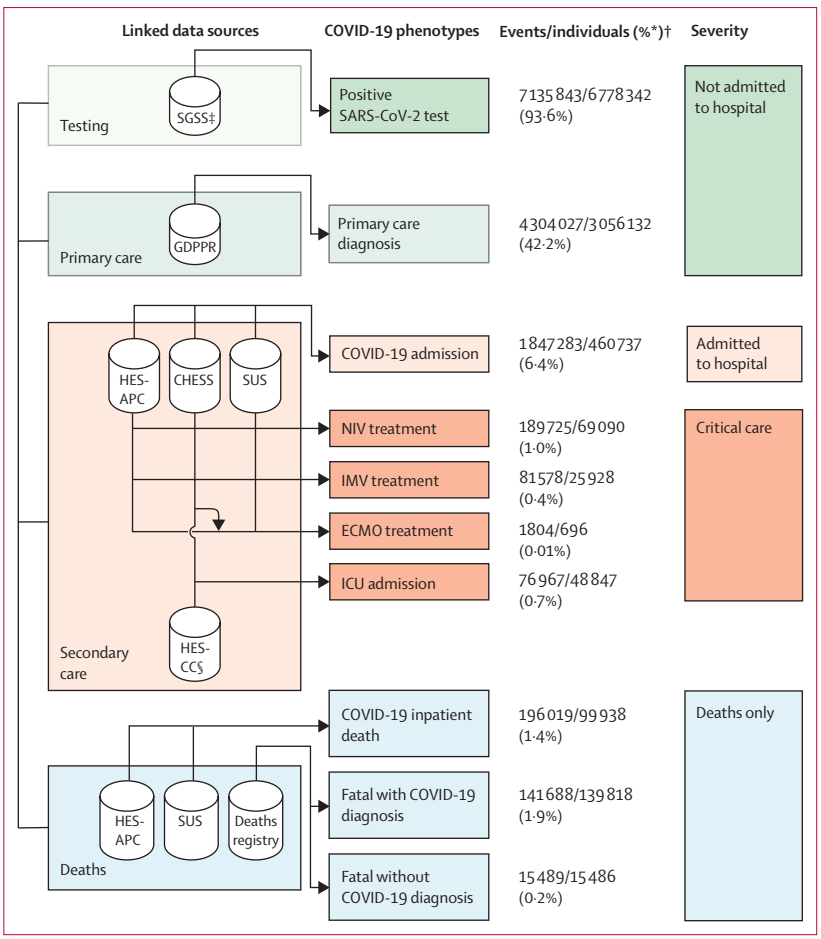


Figure 1: **Framework describing the ten COVID-19 phenotypes, and severity categories, produced using seven linked data sources to evaluate difference between COVID-19 waves and vaccination status**
For all sources, ontology terms for both suspected and confirmed diagnosis were used. CHESS=COVID-19 Hospitalisations in England Surveillance System. ECMO= extracorporeal membrane oxygenation. ICU=intensive care unit. IMV=invasive mechanical ventilation. 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. NIV=non-invasive ventilation. SGSS=Second Generation Surveillance System. SUS=Secondary Uses Service. *The proportion of individuals with a specific COVID-19 event phenotype, of all individuals with any COVID-19 event phenotype (n=7 244 925). †COVID-19 phenotypes were not mutually exclusive, thus for some phenotypes, the number of events exceeds the number of individuals (eg, individuals could have more than one positive SARS-CoV-2 test). ‡Includes SARS-CoV-2 tests from National Health Service hospitals for individuals with a clinical need and health-care workers and swab testing from the wider population. §HES-CC does not provide data on ECMO treatments.

	Population (n=57 032 174)	All COVID-19 cases (n=7 244 925)	Severity*					All COVID-19 deaths (n=158 020)
			Positive SARS-CoV-2 test (n=3 948 079)	Primary care diagnosis (n=2 825 220)	Hospital admission (n=384 135)	Ventilatory support (n=76 607)	Deaths only† (n=10 884)	
COVID-19 deaths	158 020 (0·3%)	158 020 (2·2%)	12 026 (0·3%)	20 903 (0·7%)	82 681 (21·5%)	31 526 (41·2%)	..	..
All deaths	962 754 (1·7%)	224 475 (3·1%)	20 638 (0·5%)	47 346 (1·7%)	112 693 (29·3%)	32 914 (43·0%)	..	..
Sex								
Male	28 279 784 (49·6%)	3 367 117 (46·5%)	1 881 770 (47·7%)	1 239 487 (43·9%)	192 556 (50·1%)	48 197 (62·9%)	5107 (46·9%)	86 094 (54·5%)
Female	28 752 390 (50·4%)	3 877 808 (53·5%)	2 066 309 (52·3%)	1 585 733 (56·1%)	191 579 (49·9%)	28 410 (37·1%)	5777 (53·1%)	71 926 (45·5%)
Age, years								
<18	11 612 932 (20·4%)	1 456 663 (20·1%)	1 094 645 (27·7%)	353 197 (12·5%)	7995 (2·1%)	823 (1·1%)	<5 (0·0%)	85 (0·1%)
18–29	8 749 196 (15·3%)	1 458 665 (20·1%)	866 886 (22·0%)	569 538 (20·2%)	20 190 (5·3%)	2036 (2·7%)	15 (0·1%)	318 (0·2%)
30–49	15 736 022 (27·6%)	2 222 207 (30·7%)	1 169 541 (29·6%)	980 092 (34·7%)	59 536 (15·5%)	12 806 (16·7%)	232 (2·1%)	3651 (2·3%)
50–69	13 615 556 (23·9%)	1 485 351 (20·5%)	655 180 (16·6%)	699 003 (24·7%)	95 869 (25·0%)	34 123 (44·5%)	1176 (10·8%)	24 828 (15·7%)
≥70	7 318 468 (12·8%)	622 039 (8·6%)	161 827 (4·1%)	223 390 (7·9%)	200 545 (52·2%)	26 819 (35·0%)	9458 (86·9%)	129 138 (81·7%)
Ethnicity								
White	45 800 803 (80·3%)	5 898 279 (81·4%)	3 291 417 (83·4%)	2 226 930 (78·8%)	313 864 (81·7%)	56 403 (73·6%)	9665 (88·8%)	138 434 (87·6%)
Asian or Asian British	4 962 875 (8·7%)	714 168 (9·9%)	315 031 (8·0%)	349 669 (12·4%)	37 733 (9·8%)	11 202 (14·6%)	533 (4·9%)	10 573 (6·7%)
Black or Black British	2 209 984 (3·9%)	241 053 (3·3%)	118 109 (3·0%)	99 736 (3·5%)	17 917 (4·7%)	4958 (6·5%)	333 (3·1%)	4771 (3·0%)
Chinese	518 709 (0·9%)	21 758 (0·3%)	11 954 (0·3%)	8218 (0·3%)	1145 (0·3%)	412 (0·5%)	29 (0·3%)	382 (0·2%)
Mixed	1 224 101 (2·1%)	153 693 (2·1%)	91 333 (2·3%)	55 647 (2·0%)	5338 (1·4%)	1300 (1·7%)	75 (0·7%)	1110 (0·7%)
Other	1 213 559 (2·1%)	126 040 (1·7%)	64 121 (1·6%)	53 581 (1·9%)	6406 (1·7%)	1809 (2·4%)	123 (1·1%)	1648 (1·0%)
Unknown	1 102 143 (1·9%)	89 934 (1·2%)	56 114 (1·4%)	31 439 (1·1%)	1732 (0·5%)	523 (0·7%)	126 (1·2%)	1102 (0·7%)
Social deprivation quintiles								
1 (most deprived)	11 762 185 (20·6%)	1 607 009 (22·2%)	826 775 (20·9%)	652 071 (23·1%)	104 070 (27·1%)	21 797 (28·5%)	2296 (21·1%)	37 636 (23·8%)
5 (least deprived)	10 910 882 (19·1%)	1 334 226 (18·4%)	775 562 (19·6%)	489 938 (17·3%)	56 392 (14·7%)	10 212 (13·3%)	2122 (19·5%)	26 162 (16·6%)
Unknown	54 905 (0·1%)	5272 (0·1%)	2565 (0·1%)	2299 (0·1%)	340 (0·1%)	43 (0·1%)	25 (0·2%)	136 (0·1%)
Comorbidities								
Shielded Patient List	4 084 631 (7·2%)	563 515 (7·8%)	173 799 (4·4%)	227 966 (8·1%)	135 327 (35·2%)	24 322 (31·7%)	2101 (19·3%)	51 113 (32·3%)
Circulatory (35)	10 853 569 (19·0%)	1 116 173 (15·4%)	373 739 (9·5%)	462 290 (16·4%)	228 120 (59·4%)	43 490 (56·8%)	8534 (78·4%)	126 310 (79·9%)
Respiratory (16)	9 930 315 (17·4%)	1 407 172 (19·4%)	687 359 (17·4%)	566 344 (20·0%)	124 530 (32·4%)	25 677 (33·5%)	3262 (30·0%)	55 695 (35·2%)
Genitourinary (17)	3 876 594 (6·8%)	530 319 (7·3%)	197 903 (5·0%)	219 146 (7·8%)	93 613 (24·4%)	16 006 (20·9%)	3651 (33·5%)	50 915 (32·2%)
Digestive (25)	5 195 635 (9·1%)	664 291 (9·2%)	284 903 (7·2%)	274 936 (9·7%)	85 601 (22·3%)	16 416 (21·4%)	2435 (22·4%)	40 303 (25·5%)
Endocrine (7)	7 625 521 (13·4%)	941 512 (13·0%)	357 992 (9·1%)	398 041 (14·1%)	147 747 (38·5%)	33 415 (43·6%)	4317 (39·7%)	71 917 (45·5%)
Haematological or immunological (18)	663 176 (1·2%)	93 402 (1·3%)	36 693 (0·9%)	36 952 (1·3%)	16 017 (4·2%)	3376 (4·4%)	364 (3·3%)	7349 (4·7%)
Infectious diseases (26)	8 619 925 (15·1%)	1 328 220 (18·3%)	672 065 (17·0%)	492 033 (17·4%)	137 849 (35·9%)	20 739 (27·1%)	5534 (50·8%)	71 595 (45·3%)
Neurological (12)	1 826 233 (3·2%)	235 511 (3·3%)	88 328 (2·2%)	93 773 (3·3%)	44 333 (11·5%)	7625 (10·0%)	1452 (13·3%)	21 802 (13·8%)
Cancers (41)	5 848 962 (10·3%)	722 801 (10·0%)	347 013 (8·8%)	295 587 (10·5%)	66 772 (17·4%)	11 567 (15·1%)	1862 (17·1%)	31 857 (20·2%)
Benign neoplasms (7)	2 731 961 (4·8%)	328 575 (4·5%)	129 357 (3·3%)	145 444 (5·1%)	44 432 (11·6%)	8284 (10·8%)	1058 (9·7%)	19 528 (12·4%)
Mental health (13)	12 214 921 (21·4%)	1 566 420 (21·6%)	694 957 (17·6%)	677 292 (24·0%)	159 053 (41·4%)	28 006 (36·6%)	7112 (65·3%)	78 549 (49·7%)
Musculoskeletal [19]	4 264 373 (7·5%)	504 216 (7·0%)	181 960 (4·6%)	207 463 (7·3%)	96 128 (25·0%)	14 399 (18·8%)	4266 (39·2%)	55 333 (35·0%)
Skin (9)	1 519 639 (2·7%)	216 893 (3·0%)	112 218 (2·8%)	83 282 (2·9%)	17 748 (4·6%)	3143 (4·1%)	502 (4·6%)	7959 (5·0%)
Eye (12)	4 730 278 (8·3%)	509 135 (7·0%)	148 608 (3·8%)	196 015 (6·9%)	135 043 (35·2%)	24 269 (31·7%)	5200 (47·8%)	79 071 (50·0%)
Ear (3)	179 756 (0·3%)	21 340 (0·3%)	6592 (0·2%)	8254 (0·3%)	5588 (1·5%)	686 (0·9%)	220 (2·0%)	3185 (2·0%)
Perinatal (9)	1 595 554 (2·8%)	204 830 (2·8%)	141 840 (3·6%)	58 499 (2·1%)	3752 (1·0%)	680 (0·9%)	59 (0·5%)	860 (0·5%)
Median number of comorbidities (IQR)	1·00 (0·00–3·00)	1·00 (0·00–3·00)	1·00 (0·00–2·00)	1·00 (0·00–3·00)	5·00 (2·00–9·00)	4·00 (2·00–8·00)	7·00 (4·00–11·00)	8·00 (4·00–11·00)

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. *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.

Table 1: **Sample characteristics of full population and individuals with COVID-19, by most severe COVID-19 phenotype**

wave 1 and 2, mortality among all patients admitted to hospital decreased from 32·6% (30 034 of 92 108 individuals) to 26·5% (60 855 of 230 033 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% [12 837 of 92 108 individuals] to 15·4% [35 516 of 230 033 individuals]), and a corresponding decrease in the proportion of patients who required invasive mechanical ventilation (7·1% [6584 of 92 108 individuals] to 5·4% [12 436 of 230 033 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 92 108 individuals] to 6·4% [14 752 of

	All COVID-19 cases (n=7 244 925)	All hospital admissions		Hospitalised with no ventilatory support		ICU admission		Ventilatory support outside of the ICU	
		Wave 1 (n=92 108)	Wave 2 (n=230 033)	Wave 1 (n=77 202)	Wave 2 (n=191 528)	Wave 1 (n=9843)	Wave 2 (n=23 753)	Wave 1 (n=5063)	Wave 2 (n=14 752)
COVID-19 phenotypes									
Positive SARS-CoV-2 test	6 778 342 (93·6%)	71 538 (77·7%)	216 434 (94·1%)	58 947 (76·4%)	179 730 (93·8%)	8428 (85·6%)	22 812 (96·0%)	4163 (82·2%)	13 892 (94·2%)
Primary care diagnosis	3 056 132 (42·2%)	52 648 (57·2%)	127 650 (55·5%)	43 385 (56·2%)	105 352 (55·0%)	6318 (64·2%)	14 006 (59·0%)	2945 (58·2%)	8292 (56·2%)
Hospital admission	460 737 (6·4%)	..	..	..	..	..	..	..	..
Ventilatory support	76 607 (1·1%)	14 906 (16·2%)	38 505 (16·7%)	..	..	..	..	..	..
ICU admission	48 847 (0·7%)	9843 (10·7%)	23 753 (10·3%)	..	..	..	..	..	..
Non-invasive ventilation	69 090 (1·0%)	12 837 (13·9%)	35 516 (15·4%)	..	..	7976 (81·0%)	21 000 (88·4%)	4861 (96·0%)	14516 (98·4%)
Invasive ventilation	25 928 (0·4%)	6584 (7·1%)	12 436 (5·4%)	..	..	6245 (63·4%)	11 810 (49·7%)	339 (6·7%)	626 (4·2%)
Extracorporeal membrane oxygenation	696 (0·0%)	206 (0·2%)	289 (0·1%)	..	..	204 (2·1%)	278 (1·2%)	<5 (0·0%)	11 (0·1%)
COVID-19 deaths (%)									
All deaths	158 020 (2·2%)	30 034 (32·6%)	60 855 (26·5%)	23 485 (30·4%)	44 220 (23·1%)	3980 (40·4%)	9781 (41·2%)	2569 (50·7%)	6854 (46·5%)
Inpatient deaths	99 938 (1·4%)	25 959 (28·2%)	53 488 (23·3%)	19 628 (25·4%)	37 324 (19·5%)	3865 (39·3%)	9546 (40·2%)	2466 (48·7%)	6618 (44·9%)
Deaths with COVID-19 diagnosis	139 818 (1·9%)	26 874 (29·2%)	56 284 (24·5%)	20 784 (26·9%)	40 276 (21·0%)	3708 (37·7%)	9361 (39·4%)	2382 (47·0%)	6647 (45·1%)
Deaths without COVID-19 diagnosis	15 486 (0·2%)	2728 (3·0%)	3182 (1·4%)	2366 (3·1%)	2817 (1·5%)	206 (2·1%)	214 (0·9%)	156 (3·1%)	151 (1·0%)
Sex									
Male	3 367 117 (46·5%)	50 426 (54·7%)	119 785 (52·1%)	40 503 (52·5%)	95 582 (49·9%)	6713 (68·2%)	15359 (64·7%)	3210 (63·4%)	8844 (60·0%)
Female	3 877 808 (53·5%)	41 682 (45·3%)	110 248 (47·9%)	36 699 (47·5%)	95 946 (50·1%)	3130 (31·8%)	8394 (35·3%)	1853 (36·6%)	5908 (40·0%)
Age, years									
<18	1 456 663 (20·1%)	1135 (1·2%)	2709 (1·2%)	949 (1·2%)	2478 (1·3%)	164 (1·7%)	179 (0·8%)	22 (0·4%)	52 (0·4%)
18–29	1 458 665 (20·1%)	2309 (2·5%)	8333 (3·6%)	2050 (2·7%)	7651 (4·0%)	208 (2·1%)	538 (2·3%)	51 (1·0%)	144 (1·0%)
30–49	2 222 207 (30·7%)	10 893 (11·8%)	31 461 (13·7%)	8666 (11·2%)	25 968 (13·6%)	1732 (17·6%)	4043 (17·0%)	495 (9·8%)	1450 (9·8%)
50–69	1 485 351 (20·5%)	25 596 (27·8%)	67 051 (29·1%)	18 481 (23·9%)	49 505 (25·8%)	5318 (54·0%)	12 102 (50·9%)	1797 (35·5%)	5444 (36·9%)
≥70	622 039 (8·6%)	52 175 (56·6%)	120 479 (52·4%)	47 056 (61·0%)	105 926 (55·3%)	2421 (24·6%)	6891 (29·0%)	2698 (53·3%)	7662 (51·9%)
Ethnicity									
White	5 898 279 (81·4%)	74 157 (80·5%)	186 341 (81·0%)	63 862 (82·7%)	157 501 (82·2%)	6426 (65·3%)	16 858 (71·0%)	3869 (76·4%)	11 982 (81·2%)
Asian or Asian British	714 168 (9·9%)	8671 (9·4%)	25 031 (10·9%)	6474 (8·4%)	19 213 (10·0%)	1626 (16·5%)	4131 (17·4%)	571 (11·3%)	1687 (11·4%)
Black or Black British	241 053 (3·3%)	5434 (5·9%)	9939 (4·3%)	4017 (5·2%)	7943 (4·1%)	1032 (10·5%)	1491 (6·3%)	385 (7·6%)	505 (3·4%)
Chinese	21 758 (0·3%)	331 (0·4%)	753 (0·3%)	227 (0·3%)	561 (0·3%)	80 (0·8%)	133 (0·6%)	24 (0·5%)	59 (0·4%)
Mixed	153 693 (2·1%)	1294 (1·4%)	2954 (1·3%)	989 (1·3%)	2391 (1·2%)	233 (2·4%)	392 (1·7%)	72 (1·4%)	171 (1·2%)
Other	126 040 (1·7%)	1757 (1·9%)	3939 (1·7%)	1293 (1·7%)	3100 (1·6%)	354 (3·6%)	587 (2·5%)	110 (2·2%)	252 (1·7%)
Unknown	89 934 (1·2%)	464 (0·5%)	1076 (0·5%)	340 (0·4%)	819 (0·4%)	92 (0·9%)	161 (0·7%)	32 (0·6%)	96 (0·7%)
Social deprivation quintiles									
1 (most deprived)	1 607 009 (22·2%)	23 405 (25·4%)	60 145 (26·1%)	19 591 (25·4%)	49 552 (25·9%)	2593 (26·3%)	6689 (28·2%)	1221 (24·1%)	3904 (26·5%)
5 (least deprived)	1 334 226 (18·4%)	14 109 (15·3%)	34 219 (14·9%)	11 930 (15·5%)	28 969 (15·1%)	1440 (14·6%)	3154 (13·3%)	739 (14·6%)	2096 (14·2%)
Unknown	5272 (0·1%)	70 (0·1%)	194 (0·1%)	63 (0·1%)	169 (0·1%)	<5 (0·0%)	14 (0·1%)	<5 (0·1%)	11 (0·1%)

(Table 2 continues on next page)

	All COVID-19 cases (n=7 244 925)	All hospital admissions		Hospitalised with no ventilatory support		ICU admission		Ventilatory support outside of the ICU	
		Wave 1 (n=92 108)	Wave 2 (n=230 033)	Wave 1 (n=77 202)	Wave 2 (n=191 528)	Wave 1 (n=9843)	Wave 2 (n=23 753)	Wave 1 (n=5063)	Wave 2 (n=14 752)
(Continued from previous page)									
Comorbidities									
Shielded Patient List	563 515 (7·8%)	34 237 (37·2%)	79 719 (34·7%)	29 631 (38·4%)	67 886 (35·4%)	2731 (27·7%)	6527 (27·5%)	1875 (37·0%)	5306 (36·0%)
Circulatory (35)	1 116 173 (15·4%)	61 077 (66·3%)	141 864 (61·7%)	52 220 (67·6%)	119 092 (62·2%)	5248 (53·3%)	12 963 (54·6%)	3609 (71·3%)	9809 (66·5%)
Respiratory (16)	1 407 172 (19·4%)	31 867 (34·6%)	74 399 (32·3%)	27 018 (35·0%)	61 272 (32·0%)	2747 (27·9%)	7198 (30·3%)	2102 (41·5%)	5929 (40·2%)
Genitourinary (17)	530 319 (7·3%)	27 005 (29·3%)	54 127 (23·5%)	23 745 (30·8%)	46 162 (24·1%)	1790 (18·2%)	4430 (18·7%)	1470 (29·0%)	3535 (24·0%)
Digestive (25)	664 291 (9·2%)	22 300 (24·2%)	51 743 (22·5%)	19 093 (24·7%)	43 254 (22·6%)	1935 (19·7%)	4930 (20·8%)	1272 (25·1%)	3559 (24·1%)
Endocrine (7)	941 512 (13·0%)	39 260 (42·6%)	93 643 (40·7%)	32 595 (42·2%)	76 257 (39·8%)	4191 (42·6%)	10 360 (43·6%)	2474 (48·9%)	7026 (47·6%)
Haematological or immunological (18)	93 402 (1·3%)	4590 (5·0%)	8965 (3·9%)	3856 (5·0%)	7439 (3·9%)	430 (4·4%)	901 (3·8%)	304 (6·0%)	625 (4·2%)
Infectious diseases (26)	1 328 220 (18·3%)	37 574 (40·8%)	76 416 (33·2%)	33 442 (43·3%)	66 304 (34·6%)	2169 (22·0%)	5356 (22·5%)	1963 (38·8%)	4756 (32·2%)
Neurological (12)	235 511 (3·3%)	11 977 (13·0%)	26 157 (11·4%)	10 479 (13·6%)	22 273 (11·6%)	846 (8·6%)	2162 (9·1%)	652 (12·9%)	1722 (11·7%)
Cancers (41)	722 801 (10·0%)	16 998 (18·5%)	38 913 (16·9%)	14 740 (19·1%)	33 132 (17·3%)	1318 (13·4%)	3301 (13·9%)	940 (18·6%)	2480 (16·8%)
Benign neoplasms (7)	328 575 (4·5%)	11 383 (12·4%)	26 920 (11·7%)	9751 (12·6%)	22 587 (11·8%)	972 (9·9%)	2396 (10·1%)	660 (13·0%)	1937 (13·1%)
Mental Health (13)	1 566 420 (21·6%)	40 847 (44·3%)	93 979 (40·9%)	35 515 (46·0%)	79 581 (41·6%)	3315 (33·7%)	8550 (36·0%)	2017 (39·8%)	5848 (39·6%)
Musculoskeletal (19)	504 216 (7·0%)	25 607 (27·8%)	57 099 (24·8%)	22 853 (29·6%)	49 527 (25·9%)	1430 (14·5%)	3980 (16·8%)	1324 (26·2%)	3592 (24·3%)
Skin (9)	216 893 (3·0%)	4436 (4·8%)	10 065 (4·4%)	3821 (4·9%)	8549 (4·5%)	366 (3·7%)	857 (3·6%)	249 (4·9%)	659 (4·5%)
Eye (12)	509 135 (7·0%)	36 677 (39·8%)	83 328 (36·2%)	31 651 (41·0%)	70 653 (36·9%)	2895 (29·4%)	7091 (29·9%)	2131 (42·1%)	5584 (37·9%)
Ear (3)	21 340 (0·3%)	1463 (1·6%)	3265 (1·4%)	1329 (1·7%)	2935 (1·5%)	66 (0·7%)	152 (0·6%)	68 (1·3%)	178 (1·2%)
Perinatal (9)	204 830 (2·8%)	780 (0·8%)	1905 (0·8%)	662 (0·9%)	1584 (0·8%)	79 (0·8%)	208 (0·9%)	39 (0·8%)	113 (0·8%)
Median number of comorbidities (IQR)	1·00 (0·00–3·00)	6·00 (3·00–10·00)	5·00 (2·00–9·00)	7·00 (3·00–11·00)	5·00 (2·00–9·00)	4·00 (2·00–7·00)	4·00 (2·0–7·00)	6·00 (3·00–11·00)	6·00 (3·00–10·00)

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). Counts of less than 5 are presented as <5 to maintain anonymity.

Table 2: **Comparison of COVID-19 hospital admissions between pandemic waves 1 and 2**

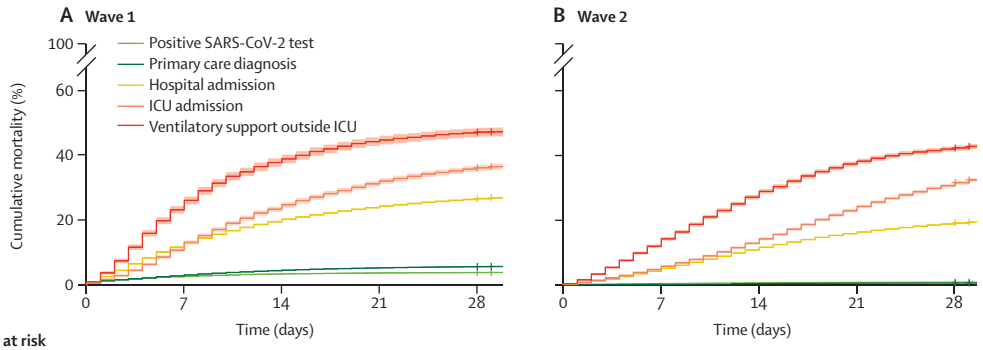
230 033 individuals]). Despite these changes, no significant difference in mortality was observed among patients admitted to the ICU between waves 1 and 2 (40·4% [3980 of 9843 individuals] *vs* 41·2% [9781 of 23 753 individuals]; p=0·21). By contrast, mortality decreased between waves 1 and 2 among inpatients 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% [2569 of 5063 individuals] *vs* 46·5% [6854 of 14 752 individuals]; p<0·0001).

Kaplan-Meier curves for 28-day mortality corroborated the finding that no differences in overall mortality were observed among patients admitted to the ICU, and indicated a prolongation of survival time in wave 2, indicated by the reduced slope (figure 2). This increase in survival time was 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 19–23).

No single source captured all COVID-19 cases: 2 641 682 (39·0%) of 6 778 342 individuals with a positive SARS-CoV-2 test also received a primary care diagnosis, whereas 3 936 053 (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 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



Number at risk										
Positive SARS-CoV-2 test	62773	61144	60618	60409	60327	797642	794758	792704	791685	791236
Primary care diagnosis	94397	91680	90170	89403	89020	1855371	1852171	1848033	1845750	1844761
Hospital admission	83436	73671	67141	63520	61423	194587	186321	173652	163855	158063
ICU admission	10690	9538	8189	7355	6869	23861	22733	20803	18406	16559
Ventilatory support outside ICU	5285	4059	3290	2950	2808	15022	13231	10959	9421	8731
Cumulative number of events										
Positive SARS-CoV-2 test	627	1740	2199	2384	2456	1162	3218	5160	6049	6440
Primary care diagnosis	976	2994	4379	5068	5412	390	3805	7794	9820	10701
Hospital admission	640	11004	16939	20329	22232	280	10080	22618	31754	37131
ICU admission	63	1403	2646	3429	3863	54	1369	3388	5774	7524
Ventilatory support outside ICU	49	1380	2054	2362	2491	52	2133	4332	5734	6350

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.

in the vaccinated group than the unvaccinated group (log-rank p<0·0001 for all phenotypes; figure 5).

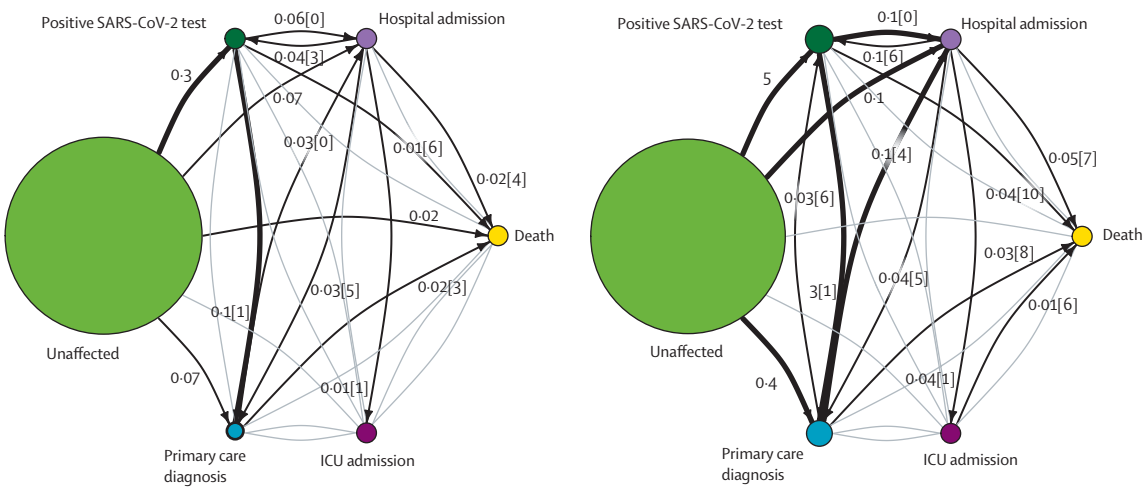
Discussion

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 major venous, arterial, and thrombocytopenic events,[15] assessing the use of antithrombotic medication on COVID-19 outcomes,[16] and studying the incidence of vascular diseases after COVID-19 infection.[17]

Our study expands the literature in several key ways. To our knowledge, this is the largest study in terms of sample size and data fidelity to create and evaluate computable COVID-19 phenotypes by leveraging multiple sources of linked data at the national level spanning electronic health records, administrative health-care billing data, disease audits, and national registers. The use of multiple sources enabled the ascertainment of events that would have previously been missed (eg, 10% of COVID-19 deaths occurred without COVID-19 being listed as a cause of death) or incorrectly aggregated (eg, mortality was highest among patients who received ventilation outside of the ICU). The framework for defining disease severity across multiple settings can be adapted and applied in other countries with similar national or regional EHR sources (eg, Denmark, Korea, Canada, and the USA). Although each country will have different data recording patterns, and the algorithms might require adaptation, the overall 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[9] and performed methods of validation including replication of findings consistent with existing literature. We internally validated our COVID-19 phenotypes through demonstrating 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% (25 569 of 35 377) and 77·5% (17 380 of 22 431) 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



Transition	Proportion in wave 1 (%)	Median time to transition between phenotypes in wave 1, days	Difference in proportion between waves (wave 2–wave 1)	Difference in median time to transition between phenotypes (wave 2–wave 1), days
Unaffected only	99·56%	..	−4·59	..
Unaffected → Positive SARS-CoV-2 test	0·28%	..	4·27	..
Positive SARS-CoV-2 test only	0·10%	..	1·29	..
Positive SARS-CoV-2 test → Primary care diagnosis	0·14%	1	2·90	0
Positive SARS-CoV-2 test → Hospital admission	0·06%	0	0·09	0
Positive SARS-CoV-2 test → Death	0·01%	6	0·03	4
Unaffected → Primary care diagnosis	0·07%	..	0·29	..
Primary care diagnosis only	0·06%	..	0·27	..
Primary care diagnosis → Positive SARS-CoV-2 test	<0·01%	0	0·02	4
Primary care diagnosis → Hospital admission	0·03%	3	0·11	5
Primary care diagnosis → Death	0·02%	..	0·01	..
Unaffected → Hospital admission	0·07%	..	0·05	..
Hospital admission only	0·02%	..	−0·01	..
Hospital admission → Positive SARS-CoV-2 test	0·04%	3	0·06	3
Hospital admission → Primary care diagnosis	0·03%	5	0·01	0
Hospital admission → ICU admission	0·01%	1	0·02	0
Hospital admission → Death	0·02%	4	0·02	3
Unaffected → ICU admission	<0·01%	..	0	..
ICU admission only	0%	..	0	..
ICU admission → Death	<0·01%	..	0·01	..
Unaffected → Death	0·02%	..	−0·01	..
Death only	0·01%	..	−0·01	..

Figure 3: **COVID-19 trajectory networks**
The size of the circles represent the number of individuals with that event relative to the total study population. Numbers on arrows show the proportion of individuals who transitioned to each phenotype (relative to the number of individuals in that COVID-19 wave). Numbers in square brackets show median number of days between events across all individuals with that transition. Median days between individuals who were unaffected (ie, no recorded COVID-19 phenotype) and other severity phenotypes are not shown since they were not directly comparable between waves, due to difference in length of the two periods. Thick arrows represent transitions that occurred in 0·1% of individuals or more. Thin black arrows represent transitions that occurred in 0·01% individuals or more. Any transitions that occurred in fewer than 0·01% of individuals are not shown. All included individuals were alive and had no previous COVID-19 events recorded before the start date of the specified waves.

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,[3] secondary care,[18] and the ICU.[19] Additionally, the ascertained infection rate of 12·7%, was comparable with official estimates from Public Health

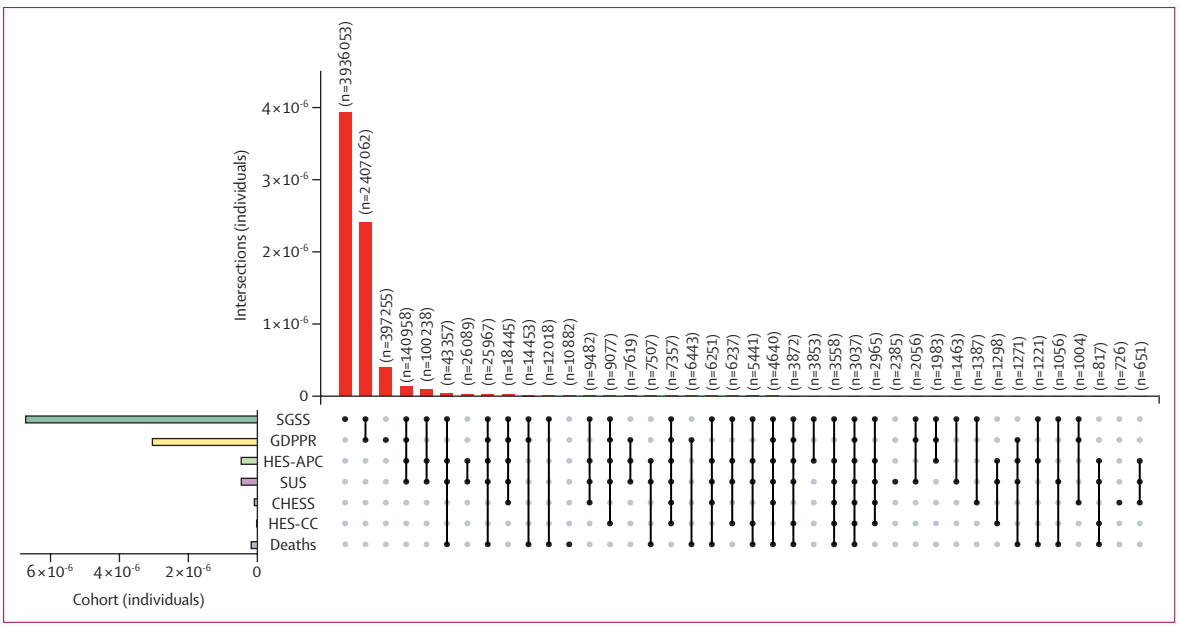


Figure 4: **UpSet plots of individuals with one or more COVID-19 event phenotypes across seven datasets**
Vertical bars report unique individuals in the intersection denoted by the intersection matrix below. Empty intersections are not shown. Horizontal bars report unique individuals identified from 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.

England (ascertained from the UK Government COVID-19 Dashboard).[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).[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.[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.[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.[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

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

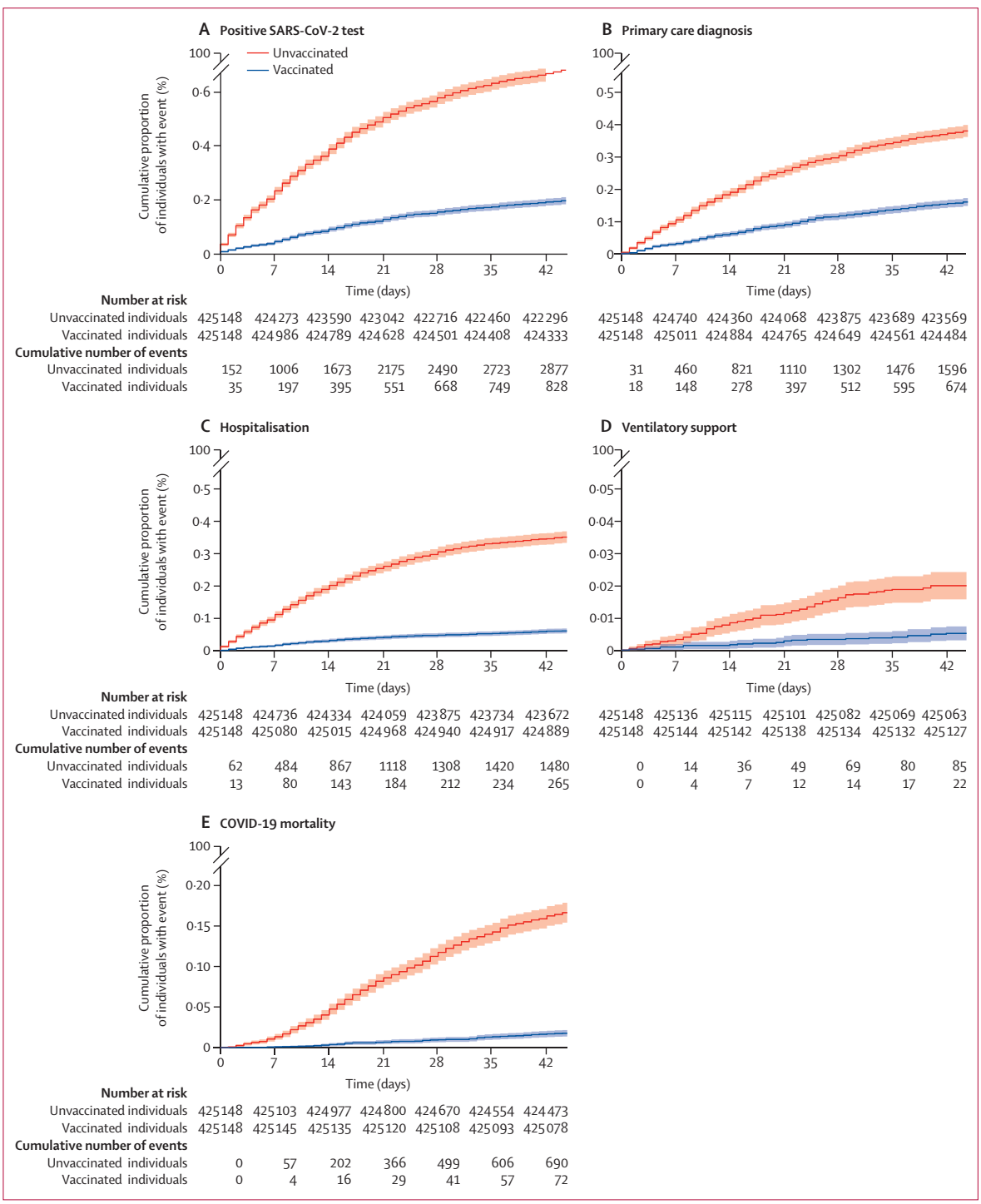


Figure 5: **Kaplan-Meier curve of cumulative COVID-19 events in vaccinated and unvaccinated individuals**
Cumulative proportion of individuals with a positive SARS-CoV-2 test (A), primary care diagnosis (B), hospital admission (C), ventilatory support (D), and COVID-19 mortality (E) in matched groups 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<0·0001, for all subplots.

prioritise care pathways associated with adverse outcomes and identify so-called patient touch points with the heath-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 Scale[24]), and monitor and assess the ongoing impact of the pandemic on health-care systems.

The use of EHR and administrative data for research is associated with known challenges in terms of variable data quality, missingness, and consistency.[25] 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.[9]

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 Scale[24] 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 Heath 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 External advisory board, 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 Afimmune, Amgen, AstraZeneca, Boehringer Ingelheim, Eli Lilly, Hanmi 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, Astra Zeneca, 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/BHFDSC/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/7e5f0247-f033-4f98-aed3-3d7422b9dc6d. 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-Y9C5K). 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 de-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, biostatisticians, 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 CONVALESCENCE 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 (HDR-9006) funded by the UK Medical Research Council, Engineering and Physical Sciences Research Council, Economic and Social Research Council, Department of Health and Social Care (England), Chief Scientist Office of the Scottish Government Health and Social Care Directorates, Health and Social Care Research and Development Division (Welsh Government), Public Health Agency (Northern Ireland), BHF, and the Wellcome Trust. AA is supported by HDR UK (HDR-9006); 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 (204841/Z/16/Z), the NIHR University College London Hospitals Biomedical Research Centre (BRC714/HI/RW/101440), NIHR Great Ormond Street Hospital Biomedical Research Centre (19RX02), and the Academy of Medical Sciences (SBF006\1084). 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 (BCDSA\100005) and by core funding from UK MRC (MR/L003120/1), BHF (RG/13/13/30194; RG/18/13/33946), and NIHR Cambridge Biomedical Research Centre (BRC-1215–20014). 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 CONVALESCENCE study. HH is an NIHR Senior Investigator. SD and HH are supported by the BHF Accelerator Award (AA/18/6/24223). CT is supported by a UCL UKRI Centre for Doctoral Training in AI-enabled Healthcare studentship (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_AWARD01786), the NIHR University College London Hospitals NHS Foundation Trust Bioemedical Research Centre (BRC713/HI/RW/101440), and the UCL Higher Education Innovation Fund (KEI2021–03–16).

References

1 The OpenSAFELY Collaborative, Walker AJ, MacKenna B, et al. Clinical coding of long COVID in English primary care: a federated analysis of 58 million patient records in situ using OpenSAFELY. *medRxiv* 2021; published online May 13. https://doi.org/10.1101/2021.05.06.21256755 (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 Mathur 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/10.1101/2021.03.16.21253770 (preprint).

5 Klann JG, 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-Izquierdo 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/BHFDSC/CCU013_01_ENG-COVID-19_event_phenotyping (accessed Aug 11, 2021).

11 UK Government. English 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-Izquierdo A, et al. A chronological map of 308 physical and mental health conditions from 4 million individuals in the English National Health Service. *Lancet Digit Health* 2019; **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 Siggaard T, Reguant R, Jørgensen IF, et al. Disease trajectory browser for exploring temporal, population-wide disease progression patterns in 7·2 million Danish patients. *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-2021-320325.

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/10.1101/2021.11.22.21266512 (preprint).

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

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 (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.