BMJ Open

Persons diagnosed with COVID-19 in England in the Clinical Practice Research Datalink (CPRD): a cohort description

Kathleen M Andersen ,[1] Leah J McGrath,[1] Maya Reimbaeva,[1] Diana Mendes,[2] Jennifer L Nguyen,[1] Kiran K Rai,[3] Theo Tritton,[3] Carmen Tsang,[2] Deepa Malhotra,[1] Jingyan Yang[1,4]

To cite: Andersen KM, McGrath LJ, Reimbaeva M, *et al*. Persons diagnosed with COVID-19 in England in the Clinical Practice Research Datalink (CPRD): a cohort description. *BMJ Open* 2024;**14**:e073866. doi:10.1136/bmjopen-2023-073866

► Prepublication history and additional supplemental material for this paper are available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2023-073866).

Received 22 March 2023
Accepted 06 December 2023

[1]Pfizer Inc, New York, New York, USA
[2]Pfizer Ltd, Tadworth, UK
[3]Adelphi Real World, Bollington, UK
[4]Institute for Social and Economic Research and Policy, Columbia University, New York, New York, USA

Correspondence to
Dr Kathleen M Andersen; kathleen.andersen@pfizer.com

ABSTRACT

Objective To create case definitions for confirmed COVID-19 diagnoses, COVID-19 vaccination status and three separate definitions of high risk of severe COVID-19, as well as to assess whether the implementation of these definitions in a cohort reflected the sociodemographic and clinical characteristics of COVID-19 epidemiology in England.

Design Retrospective cohort study.

Setting Electronic healthcare records from primary care (Clinical Practice Research Datalink, CPRD) linked to secondary care data (Hospital Episode Statistics) data covering 24% of the population in England.

Participants 2 271 072 persons aged 1 year and older diagnosed with COVID-19 in CPRD Aurum between 1 August 2020 and 31 January 2022.

Main outcome measures Age, sex and regional distribution of COVID-19 cases and COVID-19 vaccine doses received prior to diagnosis were assessed separately for the cohorts of cases identified in primary care and those hospitalised for COVID-19 (primary diagnosis code of ICD-10 U07.1 'COVID-19'). Smoking status, body mass index and Charlson Comorbidity Index were compared for the two cohorts, as well as for three separate definitions of high risk of severe disease used in the UK (National Health Service Highest Risk, PANORAMIC trial eligibility, UK Health Security Agency Clinical Risk prioritisation for vaccination).

Results Compared with national estimates, CPRD case estimates under-represented older adults in both the primary care (age 65–84: 6% in CPRD vs 9% nationally) and hospitalised (31% vs 40%) cohorts, and over-represented people living in regions with the highest median wealth areas of England (20% primary care and 20% hospital admitted cases in South East vs 15% nationally). The majority of non-hospitalised cases and all hospitalised cases had not completed primary series vaccination. In primary care, persons meeting high-risk definitions were older, more often smokers, overweight or obese, and had higher Charlson Comorbidity Index score.

Conclusions CPRD primary care data are a robust real-world data source and can be used for some COVID-19 research questions, however, limitations of the data availability should be carefully considered. Included in this publication are supplemental files for a total of over 28 000 codes to define each of three definitions of high risk of severe disease.

STRENGTHS AND LIMITATIONS OF THIS STUDY

⇒ This study used the Clinical Practice Research Datalink, a longitudinal and anonymised electronic health record database of primary healthcare interactions in England.
⇒ Definitions were built using reproducible methods with independent local clinician review.
⇒ Results were compared with government published COVID-19 case and vaccination counts to evaluate external validity of this data source.
⇒ Limitations of the data include data recency lag, with primary care data ending in April 2022 and hospitalisation data ending in March 2021.
⇒ Owing to the primary care setting of the data, this study does not represent people without a general practitioner, or those who presented directly to hospital without a prior COVID-19-related primary care encounter.

INTRODUCTION

As of 3 February 2023, there have been over 20 million confirmed COVID-19 cases and more than 183 000 related deaths in England.[1] COVID-19 severity ranges from asymptomatic cases to severe disease requiring hospital admission and sometimes death, with older adults and people with chronic health conditions at disproportionate risk.[2 3] Therefore, electronic health records (EHR) with long-standing capture of patients' medical history are uniquely positioned to answer population-based questions related to healthcare resource utilisation, economic impact and pharmaceutical intervention associated with COVID-19.

The Clinical Practice Research Datalink (CPRD) is a longitudinal and anonymised EHR database of primary healthcare

interactions in England.[4] Primary care is the cornerstone of healthcare in England, with over 98% of the population registered with a general practitioner (GP). From August 2020 to March 2022, all tests booked via the National Health Service (NHS) website for PCR tests for SARS-CoV-2, regardless of result, were reported to GP offices that use the EMIS EHR software.[5] Similarly, COVID-19 vaccines administered at any location in the country were also mandated to be reported to patient's GP.[6] Thus, during this period, CPRD can be considered a closed network of confirmed COVID-19 cases for persons in the network, with accurate information on COVID-19 vaccination status.

While the CPRD has been used in over 3000 peer-reviewed publications, it is not known whether interruptions in healthcare, particularly in-person primary care visits, during the first years of the COVID-19 pandemic affected the previously described characteristics of the CPRD population. The objective of this study was to create case definitions for COVID-19 diagnoses, COVID-19 vaccination and three separate definitions of high risk of severe COVID-19. Second, we aimed to evaluate these definitions in a sample of persons with COVID-19, using the CPRD, to assess whether this cohort's sociodemographic and clinical characteristics were generalisable to population-level COVID-19 epidemiology in England. The methodology developed in the study can be leveraged in future COVID-19 research in CPRD data.

METHODS

Study setting and population

CPRD Aurum contains data which are routinely collected from primary care practices that use the EMIS web digital clinical system that includes electronic patient records.[7] The May 2022 release of CPRD Aurum contained data from approximately 24% of persons in England.[8 9] Data captured include age, sex, body weight, medical diagnoses, referrals to specialists and/or secondary care, prescriptions issued in primary care, laboratory tests, vaccinations administered, smoking and alcohol consumption status, and all other types of care delivered as part of routine primary care practice.

CPRD Aurum was linked to Hospital Episode Statistics Admitted Patient Care (HES APC) records, using deterministic patient-level linkage with an eight-stage algorithm carried out by NHS Digital. Over 99% of practices contributing to CPRD Aurum participate in HES linkage. The HES APC dataset includes patient demographics, date and method of hospital admission and discharge, diagnoses, specialty care, and procedures. HES APC data from April 1997 to March 2021 were available in this study.

Inclusion and exclusion criteria

We included persons of any age diagnosed with COVID-19 (described below) from 1 August 2020 to 31 January 2022. First, we required records to be of acceptable research quality, as defined by CPRD. Second, we required people to be continuously registered with their GP practice for at least 365 days prior to COVID-19 diagnosis, to establish pre-COVID-19 health history. Third, we required persons to be HES APC linkage eligible to ensure the exclusion of patients where confirmed hospital admission status (via HES APC, during the time period available) could not be known. Fourth, we excluded persons who were admitted to the hospital with a primary diagnosis of U07.1 on or before their primary care recorded date of COVID-19 diagnosis. Due to mandatory reporting guidelines, the GP's date of notification from the hospital may be delayed from the date of true test collection, and therefore, may not accurately reflect date of diagnosis. Lastly, we excluded persons with a registration end date, practice last collection date or death date that was prior to their COVID-19 diagnosis.

COVID-19 case definition

With each monthly data release, CPRD publishes feasibility counts for SARS-CoV-2-related codes in CPRD primary care data with corresponding code lists.[8 10] The code types include vaccination, tests (including PCR, antibody and antigen tests), diagnosis, advice, possible cases and post-COVID-19 clinic referral codes. Three reviewers (KMA, QM and AS) independently screened the CPRD code list to determine which of the codes represented a confirmed and current infection. Discrepancies were adjudicated by a fourth reviewer (LJM), and the final definition for the COVID-19 case definition was reviewed by UK and non-UK clinicians (online supplemental table 1). We defined a current and confirmed COVID-19 episode as a diagnosis code, positive PCR or antigen test. We did not include COVID-19 vaccination, antibody tests, possible cases, exposure to COVID-19 or post-COVID-19 clinic referral codes in the COVID-19 case definition.

Hospitalisations for COVID-19 were defined as persons admitted with a primary diagnosis of COVID-19 (ICD-10 U07.1 'COVID-19') within 12 weeks of the initial diagnosis recorded in primary care. Non-hospitalised COVID-19 cases were defined as the subset of persons for whom secondary care data were available but had no record of hospital admission. Primary care COVID-19 cases were defined as persons who were not admitted to the hospital within 12 weeks of diagnosis, and those after 1 April 2021 for which hospital admission data were not available.

COVID-19 vaccination definition

The first COVID-19 vaccine dose in England was administered on 8 December 2020, with the initially limited supply prioritised for groups as outlined by the Joint Committee for Vaccinations and Immunisations (JCVI).[11] In England, COVID-19 vaccines produced by Pfizer-BioNTech and Moderna have been available since December 2020. In April 2021, JCVI announced that persons who received a first dose of an AstraZeneca vaccine would receive a second dose of the same brand but persons who had not yet received a vaccine dose would be 'preferentially offered an alternative'.[12] This study did not consider

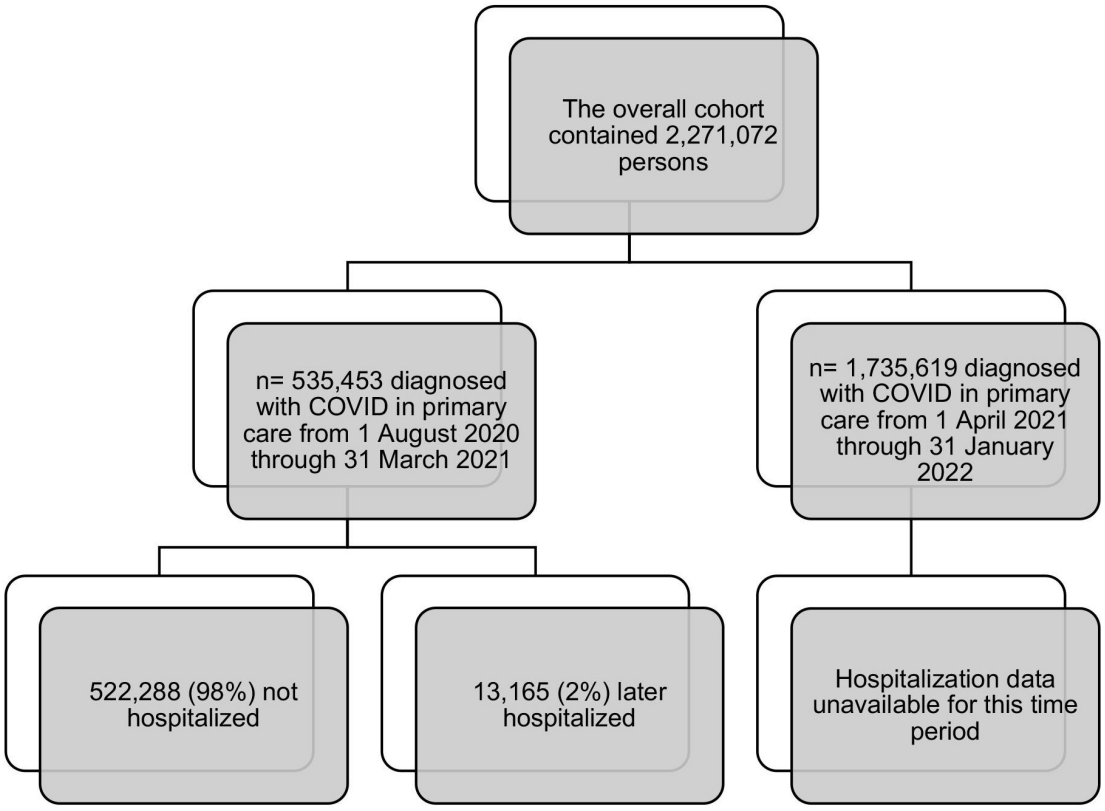

Figure 1 Cohort description. Attrition diagram of final cohort derivation. For further detail, see online supplemental table 6.

Novavax vaccinations, as it was approved for use after our index date, or the Janssen or Valneva vaccines, which have not been used in England as of February 2023.[13]

In phase 1 (December 2020–March 2021), people aged 50 and older, as well as front-line health and social care workers, clinically extremely vulnerable persons and persons aged 16 and older with underlying health conditions were eligible to receive two doses. In phase 2 (beginning April 2021), access broadened for persons aged 18–49 to receive two doses. Additionally, immunocompromised persons who received two doses in phase 1 were recommended to receive a third dose at least 8 weeks after their second dose in order to complete their primary series. Phase 3 included a single dose for 16–17 years in August 2021. In September 2021, 12–15 years could receive a single dose and a booster dose for adults was recommended at least 6 months after the completion of a primary series. In November 2021, 12–17 years could receive a second dose, and the recommendation for booster dose was shortened to at least 3 months after the completion of a primary series.[11]

We used a combination of product codes, which specify brand and dose of vaccine, as well as medical codes which indicated administration of a non-specific COVID-19 vaccine (online supplemental table 2). The code lists were derived from the published set by CPRD.[10] While patients may have multiple vaccination codes in their record on a given day, such as a code to indicate both the administration as well as a separate code for the specific product given, persons were not counted as receiving more than one vaccine dose per day.

Vaccination status at COVID-19 diagnosis date was reported, regardless of brand and separately for immunocompromised versus non-immunocompromised persons. Persons were considered unvaccinated if there was no record of COVID-19 vaccination, or for up to 13 days after first dose to account for time needed to build immunity following immunisation. Dose 1 was defined starting 14 days after the record of the first COVID-19 vaccine administration until 13 days after the receipt of second dose, or until end of follow-up. Dose 2 was defined starting 14 days after the record of the second COVID-19 vaccine administration and administered at least 21 days after the first dose. Persons were considered to have had two doses until 13 days before their next vaccine. For non-immunocompromised persons, the primary vaccination series was completed 14 days after second dose. For immunocompromised persons only, dose 3 was part of their primary series at least 21 days after their second dose. First booster doses (winter 2021) were defined 21 or more days after the receipt of the last dose in primary series. For all vaccination definitions, no upper limit on time between doses was applied.

COVID-19 risk status

We examined three separate definitions of COVID-19 risk status, each of which are set forth by different groups. While similarities exist, there are differences in the

Table 1 Characteristics of COVID-19 cases in CPRD Aurum, 1 August 2020–31 January 2022

	Primary care COVID-19 cases (n=2 257 907)	Hospitalised COVID-19 cases (n=13 165)
Age at COVID-19 diagnosis, years		
1–4	40 658 (2%)	15 (<1%)
5–11	260 186 (11%)	21 (<1%)
12–17	263 800 (11%)	24 (<1%)
18–49	1 161 843 (52%)	3127 (24%)
50–64	379 528 (17%)	4844 (37%)
65–74	92 573 (4%)	2386 (18%)
75–84	40 481 (2%)	1690 (13%)
85+	18 838 (1%)	1058 (8%)
Sex		
Male	1 046 275 (46%)	7537 (57%)
Female	1 211 592 (54%)	5628 (43%)
Unknown	40 (<1%)	0
General practitioner practice region		
North East	83 834 (4%)	467 (4%)
North West	471 195 (21%)	2886 (22%)
Yorkshire and The Humber	71 429 (3%)	340 (3%)
East Midlands	49 651 (2%)	214 (2%)
West Midlands	377 125 (17%)	2065 (16%)
East of England	95 139 (4%)	414 (3%)
London	424 908 (19%)	2984 (23%)
South East	448 536 (20%)	2682 (20%)
South West	235 785 (10%)	1113 (8%)
Unknown	305 (<1%)	0

CPRD, Clinical Practice Research Datalink.

populations. First, the NHS highest risk group, which was a list of conditions set forth by an advisory group commissioned by England's Deputy Chief Medical Officer to identify persons at the very highest risk of COVID-19 hospital admission and death.[14] Second, eligibility for the PANORAMIC (Platform Adaptive trial of NOvel antiviRals for eArly treatMent of COVID-19 In the Community) study, which began in December 2021 and is a platform randomised trial of antiviral therapeutic agents.[15] The persons who qualify for antiviral treatment in the trial are those at a higher risk of hospital admission and death. Third, UK Health Security Agency (UKHSA) clinical risk groups as outlined in 'The Green Book' chapter 14a, which is COVID-19 vaccination prioritisation from the JCVI.[16] For each of these three risk definitions, we operationalised the clinical conditions lists into SNOMED and ICD-10 (International Statistical Classification of Diseases and Related Health Problems, 10th revision) code lists. Where available, we used published code lists.[17–23] For concepts not found in the literature, we developed wildcard-based search terms in the CPRD code browser (online supplemental tables 3–5). At least one practising clinician from the UK, who was independent and external to Pfizer, reviewed each of the search term strategies, as well as ensuing code lists.

Comparison to publicly available data

We compared our results to estimates published on the UK Government's Coronavirus dashboard, and restricted results to those specific to England, as the CPRD Aurum population that is HES-linkage eligible contains data from England only.[24 25] The data in this manuscript reflect figures from the dashboard as accessed on 31 January 2023. Where possible, we restricted the dashboard to cases from 1 August 2020 to 31 January 2022 to reflect the study period, although age, sex and region-specific estimates were reported 'since the start of the pandemic'.[24 25] For whole country (non-COVID-19-related) comparisons, we used the Office of National Statistics' 2021 Census.[9]

Statistical analyses

All results are presented separately for the cohorts of persons with primary care versus hospitalised COVID-19 cases. Continuous variables are presented as means (SD) or medians (IQRs) and categorical variables as counts (percentages). Missing data for sex, region, smoking status and body mass index are shown in tables. The absence of codes for comorbidities was assumed to be the absence of the comorbidity rather than missing data. Standardised mean differences (SMDs) were used to compare groups, with SMD>10% indicating a significant difference. As per CPRD privacy rules, any cell with 10 persons or fewer (but not zero) was suppressed and other cells related to the small cell count were redacted to ensure no back-calculation could populate the count.

Data management and analyses for this study used SAS, V.9.4 (SAS Institute).

Patient and public involvement

There were no directly involved patients or public involvement in this study.

RESULTS

From 1 August 2020 to 31 January 2022, the UK Government's Coronavirus dashboard reported 14 744 991 COVID-19 cases in England.[24] The final CPRD cohort contained 2 271 072 persons diagnosed with COVID-19, regardless of care setting, in England in the same time period (online supplemental table 6). The case trends followed national estimates, with a series of peaks and troughs with a large peak in winter 2021 (online supplemental figures 1 and 2). Hospital admissions increased in fall 2020 and winter 2020 (online supplemental figure 3).

There were 2 257 907 persons with COVID-19 cases observed in primary care, which included persons who were not hospitalised during the period of time for which

hospital admission data was available (n=522 288) as well as all COVID-19 diagnoses recorded in the period where hospital admission data ended (n=1 735 619). Separately, there were 13 165 persons (2% of COVID-19 cases from August 2020 to March 2021) who were hospitalised within 12 weeks of initial COVID-19 diagnosis with a primary diagnosis of COVID-19 (figure 1).

Characteristics of COVID-19 cases

For both primary care and hospitalised cases, the majority of persons with COVID-19 in CPRD were adults aged 18–64 (table 1). The age distribution of CPRD primary care cases generally followed the national case count distribution, with young and middle-aged adults representing the largest groups, as well as the largest fraction of the English population (online supplemental table 7). Older adults were under-represented in both cohorts as compared with national population estimates. Within the primary care cohort, 6% of patients were aged 65–84 years, whereas 9% of national COVID-19 cases occurred in this age group. Notable differences were observed in the hospitalised cohort when compared with the national estimates; 31% vs 40% were aged 65–84 years and 8% vs 21% were aged >85 years, respectively.

The sex distribution of COVID-19 cases in the primary care cohort was comparable to national estimates, with 54% vs 55% female in each (table 1 and online supplemental table 7). There were more males (57%) than females in the hospitalised cohort; there are no national sex-specific hospital admission counts available.

The regional distributions of the CPRD primary care and hospitalised cohorts were similar to each other but not the overall country. Both of the study cohorts had larger proportions of persons with GP practices in higher-income regions of England as compared with national case counts.[26] For example, 20% of persons in the primary care cohort and 20% in the hospitalised cohort lived in South East region of England, the region with the highest median total household wealth as reported in March 2020. This is compared with 20% of CPRD contributing practices, 15% of all English cases and 16% of the population lived in London (online supplemental table 8). Similar trends are seen with lower median total household wealth regions being proportionally under-represented in the COVID-19 CPRD cohort as compared with national case estimates or population distribution.

COVID-19 vaccination status

At the time of COVID-19 diagnosis, 27% of non-immunocompromised persons and 36% of immunocompromised persons had completed a primary series of COVID-19 vaccination (table 2). Among the non-immunocompromised hospitalised COVID-19 cases, 98% were unvaccinated and none had completed a primary series. For immunocompromised hospitalised COVID-19 cases, 56% were unvaccinated, 44% had received one dose and none had completed a primary series. As of 1 December 2021, 86% of adults had ≥1 and 80% had ≥2 COVID-19 vaccine doses (2% and 0.2% lower than official reports, respectively).

Table 2 Vaccination status at COVID-19 diagnosis in CPRD Aurum, 1 August 2020–31 January 2022

	Primary care COVID-19 cases (n=2 257 907)	Hospitalised COVID-19 cases (n=13 165)
Non-immunocompromised	2 193 241 (100%)	12 984 (100%)
Unvaccinated	1 170 509 (53%)	12 744 (98%)
1 COVID-19 vaccine dose	173 853 (8%)	240 (2%)
2 COVID-19 vaccine doses	598 367 (27%)	0
First booster	250 512 (12%)	0
Immunocompromised	64 666 (100%)	181 (100%)
Unvaccinated	1306 (2%)	102 (56%)
1 COVID-19 vaccine dose	6033 (9%)	79 (44%)
2 COVID-19 vaccine doses	33 953 (53%)	0
3 COVID-19 vaccine doses	23 108 (36%)	0
First booster	266 (<1%)	0

CPRD, Clinical Practice Research Datalink.

Populations at risk of severe disease

After wildcard searches, clinicians reviewed nearly 50 000 codes to set the definitions for high-risk lists. The final lists contained 12 390 codes for NHS Highest Risk, 9132 codes for PANORAMIC trial criteria and 7343 codes for UKHSA Clinical Risk (code lists available on reasonable request). In the primary care cohort, 11% met NHS Highest Risk, 31% PANORAMIC and 10% UKHSA Clinical Risk criteria (table 3). With each definition, primary care cohort cases at high risk were more often current smokers (ranging 17%–21% vs 13% in full cohort) or former smokers (24%–28% vs 13%), overweight (16%–19% vs 9%) or obese (20%–28% vs 10%) and a larger proportion had at least 1 comorbidity in the Charlson Index (29%–54% vs 11%), where this was recorded. Primary care cases at high risk of severe disease were also older (mean age 49–55 years vs 34 years), more often female (55%–63% vs 54%) and had more recorded vaccine doses prior to COVID-19 diagnosis (online supplemental table 9).

Among hospitalised COVID-19 cases, 33% met NHS Highest Risk, 84% PANORAMIC and 41% UKHSA Clinical Risk criteria (table 4). The high-risk hospitalised groups were similar to the overall hospitalised cohort. Nearly half of persons were current smokers (13%–16% vs 13%) or former smokers (32%–39% vs 29%) and more than half were overweight (20%–23% vs 18%) or obese (37%–42% vs 35%). Among the hospital admitted cohort, people meeting each high-risk definition were older (mean age 65–66 years vs 60 years) (online supplemental table 10). Subgroups of patients with the NHS and UKHSA definitions had higher mean Charlson Comorbidity Indexes, and more females, than high-risk patients identified with the PANORAMIC criteria or the entire hospitalised population.

Table 3 Clinical characteristics of primary care COVID-19 cases in CPRD Aurum, by high-risk definitions

	All (n=2 257 907)	Meeting NHS highest risk conditions (n=249 972)	Meeting PANORAMIC criteria (n=691 593)	Meeting UKHSA clinical risk (n=225 051)
Primary care cases meeting definition	--	11%	31%	10%
Smoking status				
Current smoker	298 735 (13%)	44 365 (17%)	121 019 (18%)	46 436 (21%)
Former smoker	285 722 (13%)	67 637 (26%)	168 343 (24%)	63 124 (28%)
Never smoked	580 239 (26%)	87 217 (34%)	228 579 (33%)	77 157 (34%)
Unknown	1 093 211 (48%)	60 753 (23%)	173 652 (25%)	38 334 (17%)
Body mass index (kg/m^2)				
Mean (SD)	27.6 (7.1)	29.1 (6.9)	29.5 (6.7)	29.5 (7.5)
Underweight (< 18.5)	45 194 (2%)	3720 (1%)	5305 (1%)	5926 (3%)
Not overweight (18.5–24.9)	226 272 (10%)	35 766 (14%)	81 363 (12%)	36 716 (16%)
Overweight (25.0–29.9)	204 360 (9%)	45 261 (17%)	113 863 (16%)	43 758 (19%)
Obese (≥30.0)	219 253 (10%)	53 814 (21%)	138 035 (20%)	62 494 (28%)
Unknown	1 562 828 (69%)	121 411 (47%)	353 027 (51%)	76 157 (34%)
Charlson comorbidity index				
Mean (SD)	0.2 (0.6)	0.7 (1.3)	0.5 (1.0)	1.0 (1.3)
0	2 003 257 (89%)	165 823 (64%)	487 550 (71%)	102 790 (46%)
1 or 2	221 370 (10%)	74 805 (29%)	174 287 (25%)	99 830 (44%)
3 or 4	24 554 (1%)	12 659 (5%)	21 574 (3%)	15 961 (7%)
5+	8726 (<1%)	6685 (3%)	8182 (1%)	6470 (3%)

CPRD, Clinical Practice Research Datalink; NHS, National Health Service; PANORAMIC, Platform Adaptive trial of NOvel antiviRals for eArly treatMent of COVID-19 In the Community; UKHSA, UK Health Security Agency.

DISCUSSION

Key results

In this work, we defined and benchmark results from three key variables related to COVID-19 research using CPRD: index COVID-19 diagnoses, COVID-19 vaccinations and persons at high risk of severe disease.

We identified 2 271 072 COVID-19 cases in CPRD Aurum between 1 August 2020 and 31 January 2022. Younger age and lower socioeconomic deprivation have been consistently associated with reductions in COVID-19 incidence and severity.[27 28] These factors may explain why this CPRD cohort, which proportionally under-represented persons age 65 and older and over-represented persons living in the regions with higher median total household wealth, captured 15% of COVID-19 cases in a database that covers 24% of persons in England. The requirement for an NHS number in order for results to be shared may explain some of this attrition as well. Future work using CPRD for COVID-19 research will need to consider these limitations of under ascertainment of cases. Moving beyond this study's time period, the transition to at home testing, as well as the end of free PCR testing for the general public on 1 April 2022, will need to be additionally considered.

This manuscript reports results from a case definition of confirmed and current infection. We did not include codes for immunoglobulin titres, as measurable antibodies indicate a resolved infection rather than date of onset. We did not include codes indicating a sequela of prior infection, as these most often occur on a later date than index diagnosis. We did not include codes indicating a test without a result, as people with a negative test result should not be included in a COVID-19 case definition. Our results, therefore, identified fewer cases, although with greater specificity, than other studies in published literature that allow for such heterogeneity.[29]

COVID-19 vaccination events were well captured in the CPRD. This stands in stark contrast to most administrative claims and EHR databases in the USA, where less than 50% of COVID-19 vaccines are recorded in comparison to estimates provided by the Centers for Disease Control and Prevention.[30 31] England's national healthcare system, as well as the NHS data infrastructure to facilitate capture of COVID-19-related events and long-standing structure of GPs as the central node in a person's healthcare coordination, enabled the high coverage of COVID-19 vaccination records. Researchers can be more confident with CPRD data that the absence of a vaccination record indicates unvaccinated status than they otherwise would be with most other real-world datasets, which is a critical consideration for studies related to COVID-19 disease burden, vaccination or treatments.

The proportion of persons who had completed primary series vaccination prior to infection was low among primary care cases. Notably, among hospitalised cases, no

Table 4 Clinical characteristics of hospitalised COVID-19 cases in CPRD Aurum-HES linked data, by high-risk definitions

	All (n=13 165)	Meeting NHS highest risk conditions (n=4333)	Meeting PANORAMIC criteria (n=11 011)	Meeting UKHSA clinical risk (n=5353)
Hospitalised cases meeting definition	--	33%	84%	41%
Smoking status				
Current smoker	1690 (13%)	655 (15%)	1476 (13%)	840 (16%)
Former smoker	3810 (29%)	1574 (36%)	3544 (32%)	2070 (39%)
Never smoked	4205 (32%)	1362 (31%)	3513 (32%)	1746 (33%)
Unknown	3460 (26%)	742 (17%)	2478 (23%)	697 (13%)
Body mass index (kg/m^2)				
Mean (SD)	31.9 (7.5)	31.4 (7.5)	31.7 (7.4)	31.4 (7.7)
Underweight (<18.5)	96 (1%)	51 (1%)	90 (1%)	70 (1%)
Not overweight (18.5–24.9)	1162 (9%)	509 (12%)	1060 (10%)	729 (14%)
Overweight (25.0–29.9)	2390 (18%)	949 (22%)	2161 (20%)	1239 (23%)
Obese (≥30.0)	4586 (35%)	1662 (38%)	4053 (37%)	2231 (42%)
Unknown	4931 (37%)	1162 (27%)	3647 (33%)	1084 (20%)
Charlson Comorbidity Index				
Mean (SD)	1.0 (1.6)	1.9 (2.1)	1.2 (1.7)	2.0 (2.0)
0	7128 (54%)	1283 (30%)	5158 (47%)	894 (17%)
1 or 2	4241 (32%)	1833 (43%)	4094 (37%)	2.927 (55%)
3 or 4	1161 (9%)	711 (16%)	1129 (10%)	963 (18%)
5+	635 (5%)	506 (12%)	630 (6%)	569 (11%)

CPRD, Clinical Practice Research Datalink; HES, Hospital Episode Statistics; NHS, National Health Service; PANORAMIC, Platform Adaptive trial of NOvel antiviRals for eArly treatMent of COVID-19 In the Community; UKHSA, UK Health Security Agency.

patients had completed a primary COVID-19 vaccination series. These findings may be explained by several factors. First, the COVID-19 vaccine was first made available in England on 8 December 2020, and initially, second doses were given up to 12 weeks later to maximise limited supply for as many people as possible. Therefore, the calendar period under study allowed for most persons to have had a COVID-19 diagnosis in periods at which 'full vaccination' was not achievable. Second, it is possible that one vaccine offered protection against severe illness.[32–34]

We operationalised a total of over 28 000 codes, from an initial set of nearly 50 000, for three definitions of persons at risk of severe disease. We have made the search terms available, for reproducibility, as well as the resulting code lists, for other research groups to implement in their work. After the completion of this work, NHS Digital published a code list for 'Targeted Conditions', which includes each element in the NHS Highest Risk category. Among these, the NHS code list can be repurposed for 4 of the 14 conditions in PANORAMIC criteria and 3 of the 11 conditions in UKHSA Clinical Risk. To our knowledge, we offer the first publication of code lists for capture of all elements in the PANORAMIC criteria as well as UKHSA Clinical Risk criteria for these high-risk definitions, which can now be readily used in datasets that contain CPRD medical and product, ICD-10 and OPCS Classification of Interventions and Procedures codes.

While there are similarities between the three definitions, differences do exist. In the example of renal disease, NHS Highest Risk is defined as chronic kidney disease stage 4 or 5, PANORAMIC trial criteria stipulate stage 2 or 3 and UKHSA Clinical Risk are for stages 3–5. PANORAMIC trial eligibility capture persons with mild renal disease, as some antiviral treatments are not approved for use in persons with severe renal disease. However, UKHSA Clinical Risk prioritised vaccination access for persons at highest risk of disease, which would include persons with more advanced renal disease. Notably, persons who have renal disease as defined by PANORAMIC trial criteria by definition do not have renal disease by NHS definition, and people in each of these may (or may not) meet UKHSA prioritisation. The choice of which high-risk definition to implement in future studies will need to be guided by the study population and research question.

In the primary care cases, persons at high risk were older, more often smokers, had larger body sizes and higher Charlson Comorbidity Indices from the overall group of primary care cases. Among hospitalised cases, the high-risk groups were similar to the entire hospitalised

group. These findings are consistent with existing understanding of high-risk definitions, and perhaps provide reassurance that the code lists measure the purported phenomenon.

The study periods in this report represent the most recent data available from CPRD as of 3 February 2023. During the Autumn of 2022, the Aurum database experienced data quality issues related to the EMIS data flows from legacy systems, and no primary care data have been made available to researchers since the May 2022 release (data through March 2022, with some early view of April 2022). Separately, HES secondary care data have not been updated since March 2021, as NHS Digital has undergone a change in the way they process and link data. COVID-19 remains a serious disease for some people, and it is certain that some of the 1.7 million persons diagnosed with COVID-19 after 1 April 2021 would have been later admitted for COVID-19, but we do not have the hospital admission data to distinguish them from those whose cases were managed entirely in the community setting. Throughout, we have used the term 'primary care records' as the combined groups of those known to be non-hospitalised (cases where HES data were available, but the person was not hospitalised), as well as those whom we have GP encounters for but unknown eventual hospital admission status. It is difficult to approximate the number of hospital admissions that would be expected with full data availability. Carrying forward the 2% hospital admission incidence seen in the early pandemic period may not be appropriate, given 2021 introduced periods of increased (delta variant) and decreased (omicron variant) risk of hospital admission, as well the uptake of COVID-19 vaccinations and antiviral treatments. Finally, the population structure of this cohort outlined in this work further challenge the direct application of national estimates to CPRD cohorts.

This study does not capture persons not under GP care such as prisoners, some residential homes and persons without a place of residence. Additionally, CPRD Aurum, when linked with HES data, reduces the population to persons registered at eligible GP practices in England, and therefore, may not represent persons in other countries in the UK or countries outside the UK. This study does not include persons who presented directly to hospital without any prior GP interaction. In particular, persons with more severe disease such as older adults may require immediate hospital admission before seeking primary care, which could explain some of the gaps in representation we have reported. Given that CPRD is a primary care database, and the limited time period of hospital data availability, we decided to design our study as an initial cohort of persons with primary care records of COVID-19. Studies looking for complete capture of all hospitalised COVID-19 patients might consider other data sources.

CONCLUSION

In conclusion, we present a cohort of over 2 million COVID-19 cases in linked CPRD-HES data, using published definitions for COVID-19 cases, vaccinations and each of three UK-specific definitions of persons at increased risk of severe disease. CPRD primary care data are a robust real-world data source and can be used for COVID-19-related research questions; however, limitations of the data availability should be carefully considered.

Acknowledgements The authors gratefully acknowledge Bethany Backhouse, Poppy Payne, Elke Rottier and Robert Wood from Adelphi Real World (Bollington, UK), Tamuno Alfred, Darren Kailung Jeng, Tendai Mugwagwa, Qiao Mu and Chern Chuan Soo from Pfizer Inc. (New York, United States) Agnieszka Gajewska, Tomasz Mikołajczyk and Ewa Śleszyńska-Dopiera from Quanticate (Warsaw, Poland) and Andy Surinach from Genesis Research (Hoboken, United States).

Contributors KMA, LJM, MR, DM, JLN, CT, DM and JY of Pfizer, as well as KKR and TT of Adelphi Real World which has received consulting fees from Pfizer, were involved in the study design (collection, analysis and interpretation of the data), writing of the report and in the decision to submit the article for publication. All authors had full access to all statistical reports and tables in the study and take responsibility for the integrity of the data and accuracy of the data analysis. All results presented in all tables were quality control verified by a non-Pfizer non-Adelphi employee. KMA as guarantor accepts full responsibility for the work and/or conduct of the study, had access to the data, and controlled the decision to publish.

Funding This study was funded by Pfizer Inc (grant number N/A).

Competing interests KMA, LJM, MR, DM, JLN, CT, DM and JY are employees of Pfizer or Pfizer and may hold Pfizer stock or stock options. KKR and TT are employees of Adelphi Real World which has received consulting fees from Pfizer.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Patient consent for publication Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data may be obtained from a third party and are not publicly available. Electronic health records are considered sensitive data in the UK by the Data Protection Act and cannot be shared. The primary care data can be requested via application to the Clinical Practice Research Datalink, with secondary care data and mortality data through linkage on application. Information is available from https://www.cprd.com/research-applications.

ORCID iD

Kathleen M Andersen http://orcid.org/0000-0003-2670-800X

REFERENCES

1 UK Government. England Summary | Coronavirus (COVID-19) in the UK, Available: https://coronavirus.data.gov.uk
2 CDC. Risk for COVID-19 infection, hospitalization, and death by age group. *Cent Dis Control Prev* 2020. Available: https://www.cdc.gov/coronavirus/2019-ncov/covid-data/investigations-discovery/hospitalization-death-by-age.html

3 NHS. Who is at high risk from Coronavirus (COVID-19). 2021. Available: https://www.nhs.uk/conditions/coronavirus-covid-19/people-at-higher-risk/who-is-at-high-risk-from-coronavirus/
4 Herrett E, Gallagher AM, Bhaskaran K, *et al*. Data resource profile: clinical practice research Datalink (CPRD). *Int J Epidemiol* 2015;44:827–36.
5 NDRS. Coronavirus test results now visible to GPs, Available: https://digital.nhs.uk/news/2020/coronavirus-test-results-now-visible-to-gps
6 NHS Digit. COVID-19 vaccination record queries. n.d. Available: https://digital.nhs.uk/coronavirus/vaccinations/data-flows-and-resolving-data-queries/covid-19-vaccination-record-queries
7 Wolf A, Dedman D, Campbell J, *et al*. Data resource profile: clinical practice research Datalink (CPRD) Aurum. *Int J Epidemiol* 2019;48:1740.
8 CPRD. Data highlights, Available: https://www.cprd.com/data-highlights
9 Office for National Statistics. Population and household estimates, England and Wales: Census 2021, Available: https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/datasets/populationandhouseholdestimatesenglandandwalescensus2021
10 CPRD. Feasibility counts for SARS-Cov-2-related codes in CPRD primary care data. 2022. Available: https://cprd.com/sites/default/files/2022-05/SARS-CoV-2%20counts%20May2022.pdf
11 National Audit Office (NAO). The Rollout of the COVID-19 vaccination programme in England - national audit office (NAO) report. 2022. Available: https://www.nao.org.uk/reports/the-roll-out-of-the-covid-19-vaccine-in-england/
12 Coronavirus » JCVI announcement regarding AstraZeneca vaccine and next steps, Available: https://www.england.nhs.uk/coronavirus/documents/jcvi-announcement-regarding-astrazeneca-vaccine-and-next-steps/
13 Coronavirus (COVID-19) vaccine. Nhs.UK. 2022. Available: https://www.nhs.uk/conditions/coronavirus-covid-19/coronavirus-vaccination/coronavirus-vaccine/
14 Department of Health and Science. Defining the highest-risk clinical subgroups upon community infection with SARS-Cov-2 when considering the use of Neutralising Monoclonal antibodies (nMABs) and antiviral drugs: independent advisory group report. N.d Available: https://www.gov.uk/government/publications/higher-risk-patients-eligible-for-covid-19-treatments-independent-advisory-group-report/defining-the-highest-risk-clinical-subgroups-upon-community-infection-with-sars-cov-2-when-considering-the-use-of-neutralising-monoclonal-antibodies
15 Google My Maps. PANORAMIC Active GP Sites, Available: https://www.google.com/maps/d/viewer?mid=1fV1C91Aj4XtRUg1jwPL0oMlDUM8br6mJ
16 COVID-19: the green book, Chapter 14A. 2022. Available: https://www.gov.uk/government/publications/covid-19-the-green-book-chapter-14a
17 Davidson J, Warren-Gash C. Clinical Codelist - CPRD Aurum - chronic neurological disease. 2022. Available: https://datacompass.lshtm.ac.uk/id/eprint/2817/
18 Davidson J, Warren-Gash C, Mcdonald H. Clinical Codelist - CPRD Aurum - chronic respiratory disease. 2021. Available: https://datacompass.lshtm.ac.uk/id/eprint/2214/
19 Davidson J, Warren-Gash C, Mcdonald H, *et al*. Clinical Codelist - chronic kidney disease. 2021. Available: https://datacompass.lshtm.ac.uk/id/eprint/2406/
20 Davidson J, Warren-Gash C, Mcdonald H. Clinical Codelist - CPRD Aurum - Immuosuppressive conditions. 2021. Available: https://datacompass.lshtm.ac.uk/id/eprint/2234/
21 Dedman D, Carreira H, Strongman H. Clinical Codelist - cancer. 2021. Available: https://datacompass.lshtm.ac.uk/id/eprint/2408/
22 Muzambi R. Clinical Codelist - chronic liver disease ICD-10 codes. 2020. Available: https://datacompass.lshtm.ac.uk/id/eprint/2032/
23 Harriet Forbes, London School of Hygiene & Tropical Medicine, London, United Kingdom. *Clinical Code List - Moderate Immunosuppression OPCS Codes*.
24 Cases in England. Coronavirus in the UK. 2023. Available: https://coronavirus.data.gov.uk/details/cases?areaType=nation&areaName=England
25 Healthcare in England. Coronavirus in the UK. 2023. Available: https://coronavirus.data.gov.uk/details/healthcare?areaType=nation&areaName=England
26 Office for National Statistics. Household total wealth in Great Britain. Available: https://www.ons.gov.uk/peoplepopulationandcommunity/personalandhouseholdfinances/incomeandwealth/bulletins/totalwealthingreatbritain/april2018tomarch2020
27 Public Health England. Disparities in the risk and outcomes of COVID-19.
28 Office for National Statistics. Coronavirus (COVID-19) case rates by socio-demographic characteristics, England: 1 September 2020 to 10 December 2021. 2021.
29 Thygesen JH, Tomlinson C, Hollings S, *et al*. COVID-19 Trajectories among 57 million adults in England: a cohort study using electronic health records. *Lancet Digit Health* 2022;4:e542–57.
30 Center for Medicare & Medicaid Services. n.d. Assessing the completeness of Medicare claims data for measuring COVID-19 vaccine administration.
31 Wiemken TL, McGrath LJ, Andersen KM, *et al*. Coronavirus disease 2019 severity and risk of subsequent cardiovascular events. *Clin Infect Dis* 2023;76:e42–50.
32 Polack FP, Thomas SJ, Kitchin N, *et al*. Safety and efficacy of the BNT162b2 mRNA Covid-19 vaccine. *N Engl J Med* 2020;383:2603–15.
33 Baden LR, El Sahly HM, Essink B, *et al*. Efficacy and safety of the mRNA-1273 SARS-Cov-2 vaccine. *N Engl J Med* 2021;384:403–16.
34 Voysey M, Clemens SAC, Madhi SA, *et al*. Safety and efficacy of the Chadox1 nCoV-19 vaccine (Azd1222) against SARS-Cov-2: an interim analysis of four randomised controlled trials in Brazil, South Africa, and the UK. *Lancet* 2021;397:99–111.