BJGP OPEN

Patient characteristics associated with clinically coded long COVID: an OpenSAFELY study using electronic health records

Wei, Yinghui; Horne, Elsie MF; Knight, Rochelle; Cezard, Genevieve; Walker, Alex J; Fisher, Louis; Denholm, Rachel; Taylor, Kurt; Walker, Venexia; Riley, Stephanie; Williams, Dylan M; Willans, Robert; Davy, Simon; Bacon, Sebastian; Goldacre, Ben; Mehrkar, Amir; Denaxas, Spiros; Greaves, Felix; Silverwood, Richard J; Sheikh, Aziz; Chaturvedi, Nish; Wood, Angela M; Macleod, John; Steves, Claire; Sterne, Jonathan; on behalf of the UK COVID-19 Longitudinal Health and Wellbeing National Core Study and CONVALESCENCE study

Author Accepted Manuscript

MANUSCRIPT TITLE

Patient characteristics associated with clinically coded long COVID: an OpenSAFELY study using electronic health records

AUTHOR

Yinghui Wei[1,2,*], Elsie MF Horne[2,3], Rochelle Knight[2,3,4,5], Genevieve Cezard[6], Alex Walker[7], Louis Fisher[7], Rachel Denholm[2,3,8], Kurt Taylor[2,5], Venexia Walker[2,4,9], Stephanie Riley[1], Dylan M Williams[10], Robert Willans[11], Simon Davy[7], Sebastian Bacon[7], Ben Goldacre[7], Amir Mehrkar[7], Spiros Denaxas[12,13,14,15], Felix Greaves[11,16], Richard J Silverwood[17], Aziz Sheikh[18], Nish Chaturvedi[10], Angela M Wood[6], John Macleod[2,5], Claire Steves[19], Jonathan AC Sterne[2,3,8] on behalf of the UK COVID-19 Longitudinal Health and Wellbeing National Core Study and CONVALESCENCE study

[1] Centre for Mathematical Sciences, School of Engineering, Computing and Mathematics, University of Plymouth, Plymouth, UK

[2] Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, UK

[3] NIHR Bristol Biomedical Research Centre, Bristol, UK

[4] MRC Integrative Epidemiology Unit, University of Bristol, Bristol, UK

[5] The National Institute for Health and Care Research Applied Research Collaboration West (NIHR ARC West) at University Hospitals Bristol and Weston NHS Foundation Trust, UK

[6] Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK

[7] The Bennett Institute for Applied Data Science, Nuffield Department of Primary Care Health Sciences, University of Oxford, Oxford, UK

[8] Health Data Research UK South-West, Bristol, UK

[9] Department of Surgery, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA

[10] MRC Unit for Lifelong Health and Ageing, University College London, London, UK

[11] National Institute for Health and Care Excellence, UK

[12] Institute of Health Informatics, University College London, London, UK

[13] British Heart Foundation Data Science Centre, Health Data Research UK

[14] NIHR University College London Hospitals Biomedical Research Centre (UCLH BRC), London, UK

[15] UCL BHF Research Accelerator, University College London, London, UK

[16] Department of Primary Care and Public Health, Imperial College London, UK

[17] Centre for Longitudinal Studies, UCL Social Research Institute, University College of London, UK

[18] Usher Institute, The University of Edinburgh, UK

[19] Department of Twin Research and Genetic Epidemiology, School of Life Course Sciences, King's College London, London, UK

*CORRESPONDING AUTHOR

Yinghui Wei, Centre for Mathematical Sciences, School of Engineering, Computing and Mathematics, University of Plymouth, Plymouth, PL4 8AA, UK; Email: yinghui.wei@plymouth.ac.uk

ABSTRACT

Background
Clinically coded long COVID cases in electronic health records are incomplete, despite reports of rising cases of long COVID.

Aim
To determine patient characteristics associated with clinically coded long COVID.

Design and Setting
With the approval of NHS England, we conducted a cohort study using electronic health records within the OpenSAFELY-TPP platform in England, to study patient characteristics associated with clinically coded long COVID from 29 January 2020 to 31 March 2022.

Methods
We summarised the distribution of characteristics for people with clinically coded long COVID. We estimated age-sex adjusted hazard ratios and fully adjusted hazard ratios for coded long COVID. Patient characteristics included demographic factors, and health behavioural and clinical factors.

Results
Among 17,986,419 adults, 36,886 (0.21%) were clinically coded with long COVID. Patient characteristics associated with coded long COVID included female sex, younger age (under 60 years), obesity, living in less deprived areas, ever smoking, greater consultation frequency, and history of diagnosed asthma, mental health conditions, pre-pandemic post-viral fatigue, or psoriasis. These associations were attenuated following two-doses of COVID-19 vaccines compared to before vaccination. Differences in the predictors of coded long COVID between the pre-vaccination and post-vaccination cohorts may reflect the different patient characteristics in these two cohorts rather than the vaccination status. Incidence of coded long COVID was higher in those with hospitalised COVID than with those non-hospitalised COVID-19.

Conclusions
We identified variation in coded long COVID by patient characteristic. Results should be interpreted with caution as long COVID was likely under-recorded in electronic health records.

How this fits in

Electronic health records for long COVID are incomplete. It is important to understand the characteristics of people who have had their long COVID coded in electronic health records. This study identified a set of patient characteristics associated with clinically coded long COVID. This includes the frequency of prior GP-patient interaction, socio-demographical variables, history of diagnosed diseases, and SARS-COV-2 severity.

KEYWORDS

Number of display items: 2 tables and 4 figures

Word count: 3,999

3

INTRODUCTION

Long COVID[1], also known as post-acute sequelae of SARS-CoV-2 (PASC)[2], or post-COVID-19 syndrome[3], is an overarching term for the persistent symptoms for weeks, months[4], or years, following the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection. National Institute for Health and Care Excellence guidance on supporting patients with long COVID includes assessing people with symptoms after acute SARS-CoV-2, investigations and referrals[5].

Understanding risk factors for long COVID is a public health priority. Counts and rates of people having a long COVID code in English primary care varied by demographic factors but also considerably by the practice clinical software system[6]. UK longitudinal cohort studies reported that risk factors for having a long COVID included increasing age, female sex, obesity, poor pre-pandemic general and mental health, and asthma[7][8].

Previous electronic health records (EHR) analyses were based on the study period from 1 February 2020 to 9 May 2021[7], during which 4,189 long COVID cases were clinically coded. This represents considerable under-reporting, compared with the Office for National Statistics (ONS)'s estimate of 1.0 million people with self-reported long COVID[9] in the UK in May 2021. The usage of long COVID codes has improved with time[10]. General practice (GP) services were encouraged to enhance their knowledge on assessing and referring patients with long COVID as set out in NHS actions on long COVID for 2021/22[11].

We conducted a cohort study within the OpenSAFELY-TPP database (https://www.opensafely.org/), which includes detailed linked data on around 24 million people registered with an English GP using TPP SystmOne EHR software (see 'Data source'). We aimed to quantify associations of patient characteristics, including vaccination status, COVID-19 severity, and history of a range of disease diagnoses, with coded long COVID in English primary care.

METHODS

Data source

We used patient data from primary care records managed by the GP software provider, TPP SystmOne, covering around 40% of the population in England. These data include clinically coded long COVID, information on socio-demographics, pre-existing health conditions, and frequencies of GP-patient interactions, which may be consultations or any practice contacts. Data were linked to national SARS-CoV-2 testing records (Second Generation Surveillance System), vaccination data (National Immunisation Management Service), Index of Multiple Deprivation (IMD), and the ONS death registry. Admitted Patient Care Spells (APCS) is part of Hospital Episode Statistics (HES) and is provided to OpenSAFELY via NHS Digital's Secondary Use Service (SUS). OpenSAFELY includes pseudonymized data such as coded diagnoses, medications, and physiological parameters, but does not include free text data.

Study population and cohort definitions

Our study population consisted of adults aged between 18 and 105 years, with known sex and region, who were registered as active patients in a TPP GP on 29 January 2020 (the date when the first two SARS-CoV-2 cases were reported in the UK) and had at least one year of prior follow-up in a GP, to ensure that baseline characteristics could be adequately captured.

We constructed four cohorts (Supplement, Figure S1, Table S1): (1) a primary general population cohort, with follow-up start date 29 January 2020 and end date the earliest of first record of any long COVID code, death date, or 31 March 2022 (the day before free SARS-CoV-2 testing in England ended[12]); (2) a post-COVID diagnosis cohort, defined regardless of vaccination status, with follow-up start date the first recorded COVID-19 diagnosis and end date the earliest of first record of any long COVID code, death date, or 31 March 2022; (3) a pre-vaccination cohort with follow-up start date 29 January 2020 and end date the earliest of first record of any long COVID code, date of receipt of first COVID-19 vaccine dose, death date, or 31 March 2022; (4) a post-vaccination cohort, with follow-up start date 14 days after receipt of second COVID-19 vaccine dose and end date the earliest of first record of any long COVID code, death date, or 31 March 2022. In each cohort, people with a history of SARS-CoV-2 infection and/or long COVID code prior to their follow-up start date, were excluded.

Outcomes

The outcome was clinically coded long COVID, constructed from the date of the first record of any of the 15 UK SNOMED-CT codes for long COVID[6] in English primary care records, consisting of two diagnostic codes, three referral codes and 10 assessment codes (Supplement, Table S2). Time to the outcome event was defined as days from participant specific follow-up start date (Supplement, Table S1).

COVID-19 diagnosis

Date of COVID-19 diagnosis was defined as the earliest of: record of a positive SARS-CoV-2 polymerase chain reaction or antigen test; confirmed COVID-19 diagnosis in primary care or secondary care hospital admission records; or death certificate with SARS-CoV-19 infection listed as primary or underlying cause.

Patient characteristics

Patient characteristics included demographic variables, and health behavioural and clinical factors that may be associated with coded long COVID[6 7], and the frequency of GP-patient interactions, which could be an indicator of patient access to care and ability to interact with GP. There is only one entry for sex in the EHR for each patient. All other coded values were the latest record on or before the cohort and participant specific follow-up start date. A full description of patient characteristics is in the supplement, Table S2.

Demographic variables included age, sex, obesity, ethnicity, region and deprivation. Where categorised, age groups were: 18–39, 40–59, 60–79, 80–105 years. Obesity was grouped based on body mass index (BMI kg/m^2) using categories derived from the World health organization (WHO)[13]: no evidence of obesity BMI<30; obese class I, BMI 30–34.9; obese class II, BMI 35–39.9; and obese class III, BMI ≥ 40. Ethnic groups were White, Mixed, Asian or Asian British, Black or Black British and Chinese or other ethnic groups. All nine regions in England were included (East, London, East Midlands, North East, North West, West Midlands, Yorkshire and the Humber)[14]. IMD was determined based on residential area categorised into five quintiles based on relative disadvantage, with quintile 1 (Q1) being the most deprived, and quintile 5 (Q5) being the least deprived.

Health behavioural and clinical factors included smoking status, frequency of GP-patient interaction and history of disease diagnoses. Smoking status was grouped into current-, ever-, and never-smokers. Frequency of GP-patient interaction was defined during the 12 months prior to participants' follow-up start date, and categorised as: without any interaction; 1–3; 4–8; 9–12 and 13+ interactions. History of the disease diagnoses, chosen based on previous literature on risk factors for long COVID[7] and defined on or before the cohort and participant specific follow-up start date, was coded as separate indicator variables: asthma, cancer, chronic cardiac disease, chronic kidney disease, chronic liver disease, chronic obstructive pulmonary disease, chronic respiratory disease, dementia, diabetes, dysplenia (dysfunctional-spleen), haematological cancer, heart failure, hypertension, mental health condition, organ transplant, other immunosuppressive condition, other neurological condition, post-viral fatigue, psoriasis, rheumatoid arthritis, systematic lupus erythematosus (SLE), and stroke. History of diagnosed post-viral fatigue was defined prior to 29 January 2020 due to the potential use of the corresponding codes as a proxy for long COVID before the introduction of long COVID clinical codes in December 2020.

Hospitalisation for COVID-19 was defined as a hospital admission record with confirmed COVID-19 diagnosis in primary position within 28 days of the first COVID-19 diagnosis and COVID-19 without hospitalisation as a COVID-19 diagnosis that was not followed by hospitalisation within 28 days[15].

Statistical analyses

Rates of coded long COVID were quantified as the number of first long COVID events per 1000 person-years. The cumulative probability of coded long COVID was estimated, using the Kaplan-Meier approach, by age group and sex. In each cohort, hazard ratios with 95% confidence intervals (CI) for each patient characteristic were estimated from age-and-sex adjusted Cox proportional hazards (PH) models, and then all patient characteristics were included in a multivariable Cox PH model. Age was modelled using a restricted cubic spline, and estimated log hazard ratios against continuous age were plotted. In the post-COVID diagnosis cohort, we included COVID-19 severity (hospitalised vs non-hospitalised COVID-19) as an additional factor. Hazard ratios by age group (40 to 59, 60 to 79 and 80 to 105 years compared with 18 to 39 years(reference)), were estimated from models including age as a categorical variable, instead of a cubic spline.

For computational efficiency, we used the full population with coded long COVID and a randomly sampled population without coded long COVID with a ratio of 1:20. We used inverse probability weighting and robust standard errors to account for the sampling approach. The discriminative ability of the fitted model was quantified using C-statistics[16].

We included a missing category for ethnicity, smoking status and IMD. All other covariates were defined using the presence versus absence of specific codes, and thus have no identifiable missing values.

Data management and analysis were conducted using Python 3.8 and R version 4.2.1 according to a prespecified protocol. Our protocol, analysis code and code lists are available[17].

RESULTS

Study population

In total, 17,986,419 adults were included in the primary and pre-vaccination cohorts, 13,401,208 in the post-vaccination cohort and 3,507,738 in the post-COVID diagnosis cohort (Table 1). In the primary cohort, there were missing data for ethnicity (4,809,699, 26.74%), smoking status (744,851, 4.14%) and IMD (298,586, 1.66%). There were 1,855,613 (10.32%) people with ethnicity recorded as from minority groups, including Asian or Asian British, Black or Black British, Chinese or other ethnic groups, or Mixed. People in the post-vaccination and post-COVID diagnosis cohorts were more likely to have had at least one GP interaction 12 months before follow-up than those in the primary cohort. In each cohort, the most prevalent previous diagnoses were of asthma, chronic cardiac disease, diabetes, hypertension, and mental health conditions. People in the post-vaccination cohort were older, less likely to be recorded as from a minority ethnic group, and more likely to have a history of prior disease diagnoses than those in the pre-vaccination cohort. People in the post-COVID diagnosis cohort were younger, more likely to be male, and more likely to be recorded as from a minority ethnic group than those in the primary cohort. This motivates future research on increasing vaccine uptake in minority ethnic groups.

The numbers of people with coded long COVID were 36,886 (0.2%), 7,155 (0.04%), 17,376 (0.1%) and 29,268 (0.8%) in the primary, pre-vaccination, post-vaccination and post-COVID diagnosis cohorts, respectively (Table 2). The corresponding incidence rates of coded long COVID were 1.0, 0.3, 1.6 and 12.8 per 1000 person-years respectively. In the primary cohort, rate was highest in people aged 40-59 years (1.4), females (1.2) and people with BMI greater than 40 kg/m^2 (1.8). In the post-COVID diagnosis cohort, the incidence rate was highest in people aged 40–59 years (17.0), females (14.8), and people with BMI greater than 40 kg/m^2 (20.2), of white ethnicity (14.0), and living in less deprived areas (IMD Q4:14.7).

In the primary cohort, the overall cumulative probability of coded long COVID was less than 0.1% in people aged 80 years or over, rising to around 0.4% and 0.2% respectively in women and men aged 40-59 years (Supplement, Figure S2). In the post-COVID diagnosis cohort, the overall cumulative

probability of coded long COVID was less than 0.5% in people aged 80 years or over, rising to around 1.3% and 0.9% respectively in women and men aged 40-59 years (Supplement, Figure S3). The low cumulative probability of coded long COVID for people aged 80 years or over may have been due to higher risk of mortality, which can censor diagnosis of long COVID.

Demographic factors – Primary and post-COVID diagnosis cohorts

Fully adjusted hazard ratios (aHRs) for sex, obesity and ethnicity were generally attenuated towards 1, compared with age-sex adjusted hazard ratios (Figure 1). The incidence of coded long COVID declined markedly with age in the primary cohort (aHRs 0.51 (95% CI 0.43-0.60) and 0.19 (0.15-0.24) for age groups 60-79 and 80-105 years respectively, compared with age group 18-39 years). This decline was less marked in the post-COVID diagnosis cohort. The aHRs comparing age groups were consistent with those when age was modelled by restricted cubic spline (supplement Figure S4). The incidence of coded long COVID was higher in females than males in (aHRs 1.33 (1.27-1.39) and 1.20 (1.14-1.27) in the primary and post-COVID diagnosis cohorts respectively). In the primary cohort, the incidence of coded long COVID was lower in people from Black or Black British ethnicity (aHR 0.84 (0.74-0.96)) and Chinese or other ethnic groups (aHR 0.66 (0.56-0.77)), compared with those of white ethnicity. These differences were attenuated towards 1 in the post-COVID diagnosis cohort. In each cohort, the incidence of coded long COVID was higher in North East, and increased with increasing obesity and decreasing deprivation.

Demographic factors – Pre-vaccination and post-vaccination cohorts

Fully adjusted hazard ratios (aHRs) for sex and BMI were generally attenuated towards 1, compared with age-sex adjusted hazard ratios (Figure 2) in both pre-vaccination and post-vaccination cohorts. The incidence of coded long COVID declined in older adults in the post-vaccination cohort (aHRs 0.36 (95% CI 0.30-0.44) and 0.12 (0.09-0.16) for age groups 60-79 and 80-105 years respectively, compared with younger adults aged 18-39 years). This decline was less marked in the pre-vaccination cohort. The incidence of coded long COVID was higher in females than males (aHRs 1.31 (1.22-1.41) and 1.23 (1.16-1.30) in the pre-vaccination and post-vaccination cohorts respectively). In the pre-vaccination cohort, the incidence of coded long COVID was increased with increasing obesity. This pattern was less clear in the post-vaccination cohort. In both cohorts, the incidence of coded long COVID was lower in people of Chinese or other ethnic groups (aHRs 0.63 (0.50-0.81) and 0.72 (0.56-0.92) in the pre-vaccination and post-vaccination cohorts, respectively), compared with those of white ethnicity. The incidence of coded long COVID was lower in people of Black or Black British ethnicity compared to white ethnicity in the post-vaccination cohort (aHR 0.67 (0.56-0.0.81)), but not in the pre-vaccination cohort (aHR 1.10 (0.93-1.30)). In the pre-vaccination cohort, the incidence of coded long COVID was slightly higher in North East. In the post-vaccination cohort, it was slightly higher in North West. In each cohort, the incidence of long COVID increased with decreasing deprivation.

Health behavioural and clinical factors – Primary and post-COVID diagnosis cohorts

In the primary cohort, the incidence of coded long COVID was lower in current smokers and people with a missing smoking status, compared with people who never smoked (Figure 3). These differences were attenuated towards 1 in the post-COVID diagnosis cohort. In each cohort, the incidence of coded long COVID increased with increasing frequency of GP-patient interactions, during 12 months prior to the follow-up start date. The aHRs for GP-patient interaction were generally attenuated, compared with age-sex adjusted hazard ratios.

In the primary cohort, the incidence of coded long COVID was higher in people with than without a history of diagnosed asthma, chronic cardiac disease, chronic respiratory disease, haematological cancer, mental health conditions, pre-pandemic post-viral fatigue, psoriasis, or rheumatoid arthritis. These differences were generally attenuated in the post-COVID diagnosis cohort. In both cohorts, aHRs for these diseases were attenuated towards 1, compared with age-sex adjusted hazard ratios. The largest aHRs were for pre-pandemic post-viral fatigue (pre-vaccination cohort 2.01, 95% CI 1.72-2.35; post-vaccination cohort 1.96, 95% CI 1.63-2.35). In the primary cohort, the incidence of coded long COVID was lower in people with than without a history of diagnosed cancer, COPD, diabetes, heart failure, hypertension, or other neurological disorders. In the post-COVID diagnosis cohort, incidence of coded long COVID was similar in people with and without a history of diagnosed hypertension (aHR 1.00 (0.97-1.04). In the post-COVID diagnosis cohort, people with hospitalised COVID-19 had higher incidence of coded long COVID (aHR 1.37 (1.21-1.55)) than those with non-hospitalised COVID.

Health behavioural and clinical factors – Pre-vaccination and post-vaccination cohorts

In the pre-vaccination cohort, the incidence of coded long COVID was lowest in current smokers and people with a missing smoking status, and highest in ever smokers, compared with people who never smoked (Figure 4). The aHRs for smoking status were attenuated towards 1 in the post-vaccination cohort, compared with the pre-vaccination cohort. The incidence of coded long COVID increased with increasing frequency of GP-patient interaction, although aHRs were attenuated towards 1 in the post-vaccination cohort, compared with the pre-vaccination cohort. The aHRs for GP-patient interaction were generally attenuated, compared with age-sex adjusted hazard ratios.

In the pre-vaccination cohort, the incidence of coded long COVID was higher in people with than without a history of diagnosed asthma, mental health conditions, pre-pandemic post-viral fatigue, and psoriasis. These differences were attenuated in the post-vaccination cohort, compared with the pre-vaccination cohort. The aHRs for these diseases were attenuated, compared with age-sex adjusted hazard ratios. In the post-vaccination cohort, but not the pre-vaccination cohort, the incidence of coded long COVID was higher in people with than without a history of organ transplant. The incidence of coded long COVID was higher in people with than without a history of diagnosed pre-pandemic post-viral fatigue, in both the pre-vaccination and post-vaccination cohorts.

DISCUSSION

Summary

9

Despite an estimated 2.8% of the UK population having self-reported symptoms of long COVID[18] as of 3 April 2022, only 36,886 (0.2%) of the eligible general adult population in this study of up to 18 million adults had a diagnosis of long COVID recorded in their primary care record.

Patient characteristics associated with higher incidence of coded long COVID included female sex, younger age (below 60 years), greater BMI, ever having smoked, and a history of diagnosed asthma, mental health conditions, and psoriasis. The incidence of coded long COVID was higher with increasing GP-patient interaction. Coded long COVID was more than twice as likely in people with than without a diagnosis of post-viral fatigue before the pandemic. The incidence of coded long COVID was higher after hospitalised than non-hospitalised COVID-19.

Differences between factors associated with coded long COVID in the four cohorts studied may reflect differences between risk factors for infection with SARS-CoV-2, developing severe COVID-19, and developing long COVID having been infected with SARS-CoV-2. They may also reflects the influence of vaccination on developing long COVID, and changes in primary care coding practice and health care seeking behaviours during the pandemic. There were only minor differences between the cohorts in associations of demographic factors with coded long COVID (for example, lower incidence compared with White ethnicity for Chinese or other ethnic groups apart from the post-COVID diagnosis cohort, and for Asian or Asian British only in the post-vaccination cohort). Similarly, there were inverse associations with coded long COVID of current smoking compared with never smoking, and positive associations with number of previous GP-patient interactions, across the four cohorts, although the magnitude of this association was lower in the post-COVID diagnosis and post-vaccination cohorts than in the primary and pre-vaccination cohorts. Associations with previous disease diagnoses were also broadly consistent across the four cohorts. Further, COVID-19 vaccination did not substantially modify associations of factors with coded long COVID-19, although it is likely to have substantially attenuated the overall incidence of COVID-19[19].

Strengths and limitations

A key strength of this study is its use of the data from the OpenSAFELY-TPP platform, which includes over 40% of the English population[20]. We analysed data from all eligible adults with follow-up of up to 26 months. The prevalence of coded long COVID was higher in people registered in an NHS primary care GP using EMIS EHR software than in practices using TPP software[6]. The type of EHR software used is geographically clustered[21]. However, we were not able to access data from practices using EMIS software.

The prevalence of coded long COVID in English primary care records was substantially lower than that found in population surveys. There is likely to be considerable under-ascertainment of long COVID in these records due to difficulties in accessing care during the pandemic. Future research can investigate if access to free text records might help decrease under-ascertainment[22]. However, free text was not available for our analyses.

Fully adjusted hazard ratios quantify the contribution of each patient characteristic to predicting the outcome, having accounted for the value of each other patient characteristic. However, they do not have causal interpretations, because they do not distinguish between adjustment for confounders and mediators. Such misinterpretation of multiple adjusted effect estimates presented in a single table has been referred to as the 'Table 2 Fallacy"[23].

As described in the "COVID-19 diagnosis" section, we used data from multiple sources to capture COVID-19 diagnosis as accurately as possible. Patient characteristics were determined by using primary care records, and additional data from secondary care can improve the completeness and updated analysis can be conducted.

In the pre-vaccination cohort, follow-up was censored at the time of vaccination. Such censoring could lead to bias if it was informative, which would be the case if the incidence of coded long COVID differed systematically between people who were and were not vaccinated, having accounted for baseline covariates. We believe that our analyses adjusted for the major predictors of COVID-vaccination (eg age, sex, ethnicity, IMD), which should have limited informative censoring, but cannot exclude the possibility that estimated associations were biased because of the censoring.

Comparison with existing literature

Similar to other studies[7 24 25], we found positive associations of coded long COVID with female sex, obesity, mental health conditions and living in less deprived areas. The latter association contrasts with the increased risk of SARS-CoV-2 infection with increasing deprivation, and illustrates the distinction between long COVID and coded long COVID, which depends on the ability of people with long COVID to access health care for their condition at a time of extreme pressure on health services. A previous EHR analysis also found that people living in less deprived areas had higher incidence of coded long COVID. However, in the same study the analysis of longitudinal cohort studies found no association between IMD and self-reported long COVID[7].

Among the general population, the incidence for coded long COVID was lower in people of Black ethnicity, similar to a previous study[7]. We found similar incidence of coded long COVID in Asian and Asian British people and people of White ethnicity. Additionally, the incidence of coded long COVID was lower in Chinese or other ethnic groups, compared to people of White ethnicity. In general, the incidence of coded long COVID was higher in ever smokers but lower in current smokers, compared to never smokers. A previous study[7] included only two categories for smoking status, and found no difference in the incidence of coded long COVID between current smokers and non-smokers. Smoking status in EHR may not be up-to-date, especially for people who had less frequent interaction with their GP.

A study in Moscow identified that pre-existing hypertension was associated with higher risk of long COVID 12 months since discharge from hospitalisation[26]. Our fully adjusted model in the primary cohort showed that the incidence of coded long COVID was lower for people with a history of diagnosed hypertension, although the incidence was higher when only adjusted for age and sex. In

11

other three cohorts, no association with hypertension was observed from the fully adjusted models. In the Moscow study, long COVID was assessed by clinicians after hospitalised COVID, whilst our study relied on people getting access to their GP and the diagnosis then being recorded.

A previous report to the UK Government's Scientific Advisory Group for Emergencies found that the risk of coded long COVID was higher in adults with hospitalised than non-hospitalised COVID-19[27]. Our study was restricted to adults. Other studies report that hospitalised COVID was also associated with higher risk of long COVID in children[25 28 29]. A systematic review of 20 studies, identified higher risk of long COVID with female sex, mental health conditions, fatigue and acute disease severity with respiratory symptoms[30].

Implication for practice

A potentially large proportion of people with long COVID did not have a long COVID code in their health records. The incidence of coded long COVID was influenced by the frequency of previous GP-patient interaction. Despite controlling for this factor, we identified a set of patient characteristics associated with coded long COVID, including socio-demographical variables, history of diagnosed diseases, and SARS-COV-2 severity.

CONFLICTS OF INTEREST

Over the past five years BG has received research funding from the Laura and John Arnold Foundation, the NHS National Institute for Health Research (NIHR), the NIHR School of Primary Care Research, the NIHR Oxford Biomedical Research Centre, the Mohn-Westlake Foundation, NIHR Applied Research Collaboration Oxford and Thames Valley, the Wellcome Trust, the Good Thinking Foundation, Health Data Research UK (HDRUK), the Health Foundation, and the World Health Organization; he also receives personal income from speaking and writing for lay audiences on the misuse of science.

ACKNOWLEDGEMENT

FUNDING

INFORMATION GOVERNANCE AND ETHICAL APPROVAL

NHS England is the data controller for OpenSAFELY-TPP; TPP is the data processor; all study authors using OpenSAFELY have the approval of NHS England[31]. This implementation of OpenSAFELY is hosted within the TPP environment which is accredited to the ISO 27001 information security standard and is NHS IG Toolkit compliant[32].

Patient data has been pseudonymised for analysis and linkage using industry standard cryptographic hashing techniques; all pseudonymised datasets transmitted for linkage onto OpenSAFELY are encrypted; access to the platform is via a virtual private network (VPN) connection, restricted to a small group of researchers; the researchers hold contracts with NHS England and only access the platform to initiate database queries and statistical models; all database activity is logged; only aggregate statistical outputs leave the platform environment following best practice for anonymisation of results such as statistical disclosure control for low cell counts[33].

The service adheres to the obligations of the UK General Data Protection Regulation (UK GDPR) and the Data Protection Act 2018. The service previously operated under notices initially issued in February 2020 by the Secretary of State under Regulation 3(4) of the Health Service (Control of Patient Information) Regulations 2002 (COPI Regulations), which required organisations to process confidential patient information for COVID-19 purposes; this set aside the requirement for patient consent[34]. As of 1 July 2023, the Secretary of State has requested that NHS England continue to operate the Service under the COVID-19 Directions 2020[35]. In some cases of data sharing, the common law duty of confidence is met using, for example, patient consent or support from the Health Research Authority Confidentiality Advisory Group[36].

Taken together, these provide the legal bases to link patient datasets on the OpenSAFELY platform. GP practices, from which the primary care data are obtained, are required to share relevant health information to support the public health response to the pandemic and have been informed of the OpenSAFELY analytics platform.

This study was approved by NHS London - Harrow Research Ethics Committee (IRAS reference: 310808, NHS REC reference: 22/LO/0105); and by the University of Plymouth Research Ethics and Integrity Panel (reference: 3193).

TABLES

Table 1. Patient characteristics. Summary statistics are number (percentage) except where indicated.

Characteristic		Cohort		
		Primary / Pre-vaccination	Post-vaccination	Post-COVID diagnosis
All		17,986,419	13,401,208	3,507,738
Mean (SD) age in years		49.72 (18.69)	53.45 (18.40)	44.55 (17.26)
Age	18 – 39	6,163,161 (34.27)	3,433,136 (25.62)	1,508,578 (43.01)
	40 – 59	6,143,985 (34.16)	4,732,533 (35.31)	1,340,894 (38.23)
	60 – 79	4,513,280 (25.09)	4,162,621 (31.06)	515,422 (14.69)
	80 – 105	1,165,993 (6.48)	1,072,918 (8.01)	142,844 (4.07)
Sex	Female	8,971,008 (49.88)	6,451,356 (48.14)	1,601,255 (45.65)
	Male	9,015,411 (50.12)	6,949,852 (51.86)	1,906,483 (54.35)
BMI	Not obese	13,826,227 (76.87)	9,964,252 (74.35)	2,617,431 (74.62)
	Obese i (30 – 34.9)	2,602,204 (14.47)	2,128,044 (15.88)	528,931 (15.08)
	Obese ii (35 – 39.9)	988,672 (5.50)	819,465 (6.11)	221,223 (6.31)
	Obese iii (40 +)	569,316 (3.17)	489,447 (3.65)	140,153 (4.00)
Ethnicity	White	11,321,107 (62.94)	9,100,727 (67.91)	2,290,746 (65.31)
	Asian or Asian British	1,056,550 (5.87)	678,284 (5.06)	228,256 (6.51)
	Black or Black British	345,940 (1.92)	183,283 (1.37)	67,597 (1.93)
	Chinese or other ethnic group	277,598 (1.54)	135,311 (1.01)	33,780 (0.96)
	Mixed	175,525 (0.98)	98,685 (0.74)	37,618 (1.07)
	Missing	4,809,699 (26.74)	3,204,918 (23.92)	849,741 (24.22)
Region	East	4,152,253 (23.09)	3,137,907 (23.42)	779,917 (22.23)
	East Midlands	3,116,231 (17.33)	2,358,504 (17.60)	641,785 (18.30)
	London	1,190,596 (6.62)	643,154 (4.80)	195,859 (5.58)
	North East	864,297 (4.81)	647,350 (4.83)	202,114 (5.76)
	North West	1,608,661 (8.94)	1,235,875 (9.22)	374,128 (10.67)
	South East	1,214,019 (6.75)	920,943 (6.87)	201,174 (5.74)
	South West	2,502,709 (13.91)	2,047,734 (15.28)	398,249 (11.35)
	West Midlands	728,973 (4.05)	475,426 (3.55)	153,392 (4.37)
	Yorkshire and the Humber	2,608,680 (14.50)	1,934,315 (14.43)	561,120 (16.00)
Index of multiple deprivation	1 (most deprived)	3,419,935 (19.01)	2,175,145 (16.23)	721,264 (20.56)
	2	3,504,610 (19.48)	2,450,814 (18.29)	695,034 (19.81)
	3	3,761,299 (20.91)	2,843,309 (21.22)	699,051 (19.93)
	4	3,666,169 (20.38)	2,900,059 (21.64)	683,882 (19.50)
	5 (least deprived)	3,335,820 (18.55)	2,755,089 (20.56)	624,221 (17.80)
	0 (missing)	298,586 (1.66)	276,792 (2.07)	84,286 (2.40)
Smoking status	Never smoker	8,257,661 (45.91)	6,221,555 (46.43)	1,681,507 (47.94)
	Current smoker	3,043,874 (16.92)	1,945,883 (14.52)	508,586 (14.5)
	Ever smoker	5,940,033 (33.03)	4,753,679 (35.47)	1,125,675 (32.09)
	Missing	744,851 (4.14)	480,091 (3.58)	191,970 (5.47)
GP-Patient interaction	0	4,887,100 (27.17)	2,874,697 (21.45)	787,607 (22.45)
	1 to 3	4,481,433 (24.92)	3,535,283 (26.38)	970,395 (27.66)
	4 to 8	2,364,870 (13.15)	1,887,389 (14.08)	467,316 (13.32)
	9 to 12	4,517,010 (25.11)	3,682,594 (27.48)	935,897 (26.68)
	13 or more	1,736,006 (9.65)	1,421,245 (10.61)	346,523 (9.88)
History of disease diagnosis	Asthma	3,039,981 (16.90)	2,366,051 (17.66)	693,139 (19.76)
	Cancer	898,197 (4.99)	830,573 (6.20)	131,263 (3.74)
	Chronic cardiac disease	1,216,263 (6.76)	1,083,971 (8.09)	188,628 (5.38)
	Chronic kidney disease	25,440 (0.14)	21,285 (0.16)	8,085 (0.23)
	Chronic liver disease	104,376 (0.58)	90,055 (0.67)	18,422 (0.53)
	Chronic obstructive pulmonary disease	597,757 (3.32)	526,800 (3.93)	94,218 (2.69)
	Chronic respiratory disease*	734,352 (4.08)	643,188 (4.80)	111,337 (3.17)
	Dementia	42,978 (0.24)	31,376 (0.23)	12,182 (0.35)
	Diabetes	1,827,304 (10.16)	1,678,045 (12.52)	322,903 (9.21)
	Dysplenia	25,815 (0.14)	22,137 (0.17)	4,316 (0.12)
	Hematological cancer	104,655 (0.58)	95,011 (0.71)	19,085 (0.54)
	Heart failure	315,575 (1.75)	292,006 (2.18)	56,106 (1.60)
	Hypertension	3,845,579 (21.38)	3,409,919 (25.44)	563,055 (16.05)
	Mental health	3,677,686 (20.45)	2,946,356 (21.99)	784,619 (22.37)
	Organ transplant	20,848 (0.12)	17,921 (0.13)	5,732 (0.16)
	Other immunosuppressive condition	90,212 (0.50)	78,650 (0.59)	20,283 (0.58)
	Other neurological disease	178,921 (0.99)	153,470 (1.15)	32,107 (0.92)
	Pre-pandemic post-viral fatigue	33,616 (0.19)	28,482 (0.21)	5,912 (0.17)
	Psoriasis	697,571 (3.88)	575,811 (4.30)	143,676 (4.10)
	Rheumatoid arthritis	183,330 (1.02)	165,080 (1.23)	32,030 (0.91)
	Systematic lupus erythematosus	30,177 (0.17)	25,819 (0.19)	5,495 (0.16)
	Stroke	381,038 (2.12)	339,592 (2.53)	62,944 (1.79)

Table 2. Event count /1000 person years (pyrs) and incidence rate (IR) per 1000 person years for long COVID.

		Primary		Pre-vaccination		Post-vaccination		Post-COVID diagnosis	
Characteristic		Count/1000 pyrs	IR	Count/1000 pyrs	IR	Count	IR	Count / 1000 pyrs	IR
All		36,886/38,520.8	1.0	7,155/23,609.2	0.3	17,376/11,142.9	1.6	29,268/2,293.2	12.8
Age	18 – 39	12,031/13,368.9	0.9	3,405/ 9,659.6	0.4	4,446/ 2,366.3	1.9	9,181/1,000.5	9.2
	40 – 59	18,661/13,287.1	1.4	3,092/ 7,882.5	0.4	9,218/ 3,917.5	2.4	14,886/ 876.1	17.0
	60 – 79	5,713/ 9,611.4	0.6	616/ 4,946.9	0.1	3,371/ 3,805.5	0.9	4,794/ 334.4	14.3
	80 – 105	481/ 2,253.5	0.2	42/ 1,120.2	0.0	341/ 1,053.6	0.3	407/ 82.1	5.0
Sex	Male	13,569/19,209.6	0.7	2,803/12,151.4	0.2	6,440/ 5,263.6	1.2	10,836/1,047.4	10.3
	Female	23,317/19,311.3	1.2	4,352/11,457.8	0.4	10,936/ 5,879.4	1.9	18,432/1,245.8	14.8
BMI	Not obese	24,736/29,616.8	0.8	5,203/18,604.8	0.3	11,685/ 8,179.8	1.4	19,045/1,694.2	11.2
	Obese i (30 – 34.9)	6,881/ 5,570.8	1.2	1,138/ 3,146.3	0.4	3,235/ 1,832.8	1.8	5,679/ 354.3	16.0
	Obese ii (35 – 39.9)	3,083/ 2,116.4	1.5	497/ 1,187.6	0.4	1,437/ 702.7	2.0	2,627/ 149.9	17.5
	Obese iii (40 +)	2,186/ 1,216.8	1.8	317/ 670.6	0.5	1,019/ 427.7	2.4	1,917/ 94.7	20.2
Ethnicity	White	24,553/24,226.3	1.0	4,363/14,430.7	0.3	12,385/ 7,682.1	1.6	20,045/1,436.2	14.0
	Asian or Asian British	2,172/ 2,279.4	1.0	598/ 1,519.2	0.4	612/ 525.8	1.2	1,795/ 190.3	9.4
	Black or Black British	544/ 746.1	0.7	244/ 545.4	0.4	138/ 140.7	1.0	428/ 48.2	8.9
	Chinese or other ethnic group	289/ 600.6	0.5	87/ 461.8	0.2	107/ 104.1	1.0	222/ 23.4	9.5
	Mixed	345/ 379.4	0.9	133/ 274.7	0.5	105/ 75.6	1.4	276/ 25.6	10.8
	Missing	8,983/10,289.0	0.9	1,730/ 6,377.4	0.3	4,029/ 2,614.6	1.5	6,502/ 569.4	11.4
Region	East	6,476/ 8,897.2	0.7	1,225/ 5,428.5	0.2	3,061/ 2,600.3	1.2	4,744/ 499.2	9.5
	East Midlands	5,100/ 6,671.6	0.8	1,048/ 4,035.3	0.3	1,928/ 1,959.6	1.0	4,066/ 425.2	9.6
	London	1,500/ 2,566.1	0.6	476/ 1,854.5	0.3	486/ 512.7	0.9	974/ 131.9	7.4
	North East	3,484/ 1,848.2	1.9	495/ 1,111.4	0.4	1,858/ 542.7	3.4	3,067/ 137.0	22.4
	North West	4,329/ 3,438.5	1.3	704/ 2,037.2	0.3	2,269/ 1,035.1	2.2	3,451/ 251.9	13.7
	South East	2,824/ 2,598.3	1.1	763/ 1,581.8	0.5	1,051/ 772.5	1.4	2,308/ 123.1	18.7
	South West	5,329/ 5,355.4	1.0	879/ 3,138.3	0.3	3,150/ 1,718.2	1.8	4,244/ 227.8	18.6
	West Midlands	1,223/ 1,561.0	0.8	329/ 1,011.7	0.3	454/ 392.3	1.2	909/ 110.9	8.2
	Yorkshire and The Humber	6,621/ 5,584.6	1.2	1,236/ 3,410.5	0.4	3,119/ 1,609.5	1.9	5,505/ 386.3	14.3
Index of multiple deprivation	1 (most deprived)	6,529/ 7,318.8	0.9	1,656/ 4,828.4	0.3	2,392/ 1,753.7	1.4	5,254/ 509.3	10.3
	2	6,964/ 7,503.8	0.9	1,450/ 4,745.7	0.3	3,011/ 2,013.6	1.5	5,522/ 466.5	11.8
	3	7,430/ 8,054.1	0.9	1,449/ 4,890.9	0.3	3,429/ 2,375.3	1.4	5,877/ 448.6	13.1
	4	8,185/ 7,853.4	1.0	1,371/ 4,641.4	0.3	4,255/ 2,442.7	1.7	6,357/ 431.9	14.7
	5 (least deprived)	7,135/ 7,150.1	1.0	1,108/ 4,112.1	0.3	3,907/ 2,333.9	1.7	5,629/ 385.5	14.6
	Missing	643/ 640.6	1.0	121/ 390.7	0.3	382/ 223.7	1.7	629/ 51.4	12.2

Figure 1. Primary and post-COVID diagnosis cohorts: age-and-sex adjusted and fully adjusted hazard ratios of coded long COVID for demographic variables.

Characteristic	Primary	Age-sex aHR	Fully aHR	Post-COVID diagnosis	Age-sex aHR	Fully aHR
Age group						
18 to 39 (reference)						
40 to 59		1.56 (1.38, 1.76)	1.34 (1.18, 1.52)		1.85 (1.61, 2.13)	1.62 (1.41, 1.85)
60 to 79		0.66 (0.56, 0.76)	0.51 (0.43, 0.60)		1.59 (1.34, 1.88)	1.27 (1.09, 1.48)
80 to 105		0.24 (0.20, 0.30)	0.19 (0.15, 0.24)		0.56 (0.45, 0.70)	0.45 (0.37, 0.55)
Sex						
Male (reference)						
Female		1.76 (1.66, 1.87)	1.33 (1.27, 1.39)		1.44 (1.33, 1.55)	1.20 (1.14, 1.27)
BMI						
Not obese (reference)						
Obese i (30-34.9)		1.44 (1.37, 1.51)	1.24 (1.19, 1.28)		1.27 (1.19, 1.35)	1.18 (1.12, 1.23)
Obese ii (35-39.9)		1.56 (1.46, 1.66)	1.27 (1.22, 1.33)		1.35 (1.24, 1.46)	1.21 (1.15, 1.27)
Obese iii (40+)		1.77 (1.62, 1.94)	1.36 (1.28, 1.44)		1.54 (1.38, 1.72)	1.33 (1.25, 1.42)
Ethnicity						
White (reference)						
Asian or Asian British		0.85 (0.66, 1.10)	1.03 (0.81, 1.32)		0.74 (0.55, 0.99)	0.92 (0.70, 1.23)
Black or Black British		0.64 (0.54, 0.76)	0.84 (0.74, 0.96)		0.65 (0.54, 0.79)	0.87 (0.76, 1.00)
Chinese or other ethnic group		0.45 (0.38, 0.54)	0.66 (0.56, 0.77)		0.70 (0.57, 0.85)	0.90 (0.75, 1.07)
Missing		0.94 (0.83, 1.07)	1.05 (0.93, 1.19)		0.97 (0.84, 1.11)	1.00 (0.89, 1.13)
Mixed		0.80 (0.71, 0.90)	0.97 (0.86, 1.08)		0.83 (0.73, 0.94)	0.98 (0.86, 1.11)
Region						
East (reference)						
London		0.76 (0.57, 1.00)	0.93 (0.70, 1.23)		0.80 (0.61, 1.04)	0.91 (0.69, 1.19)
East Midlands		1.06 (0.79, 1.41)	1.04 (0.78, 1.39)		1.02 (0.76, 1.38)	1.02 (0.76, 1.37)
North East		2.60 (1.23, 5.53)	2.72 (1.25, 5.90)		2.43 (1.09, 5.41)	2.58 (1.13, 5.88)
North West		1.75 (1.12, 2.75)	1.78 (1.14, 2.77)		1.47 (0.90, 2.39)	1.52 (0.94, 2.45)
South East		1.53 (0.70, 3.31)	1.59 (0.74, 3.43)		1.95 (0.78, 4.88)	2.01 (0.81, 5.00)
South West		1.42 (0.75, 2.66)	1.40 (0.74, 2.63)		1.90 (0.89, 4.04)	1.89 (0.89, 4.02)
West Midlands		1.08 (0.74, 1.58)	1.25 (0.84, 1.84)		0.91 (0.61, 1.34)	1.05 (0.70, 1.56)
Yorkshire and the Humber		1.64 (0.94, 2.85)	1.68 (0.93, 3.01)		1.54 (0.82, 2.88)	1.60 (0.83, 3.10)
Deprivation (IMD) quintile						
1 (most deprived, reference)						
2		1.06 (0.92, 1.21)	1.19 (0.99, 1.43)		1.12 (0.97, 1.31)	1.22 (1.00, 1.49)
3		1.08 (0.90, 1.29)	1.24 (1.01, 1.52)		1.22 (1.00, 1.49)	1.33 (1.07, 1.65)
4		1.22 (0.99, 1.49)	1.38 (1.11, 1.71)		1.35 (1.09, 1.67)	1.46 (1.16, 1.83)
5 (least deprived)		1.17 (0.92, 1.49)	1.35 (1.02, 1.78)		1.32 (1.02, 1.71)	1.50 (1.10, 2.03)
0 (missing)		1.13 (0.97, 1.31)	1.11 (0.95, 1.30)		1.15 (1.01, 1.32)	1.17 (1.01, 1.36)

Hazard ratio: 0.1 0.25 0.5 1 2 4 10

Cox Model   ◇ Age-and-sex adjusted   ■ Fully adjusted

Figure 2. Pre-vaccination and post-vaccination cohorts: age-and-sex adjusted and fully adjusted hazard ratios of coded long COVID for demographic variables.

Characteristic	Pre-vaccination	Age-sex aHR	Fully aHR	Post-vaccination	Age-sex aHR	Fully aHR
Age group						
18 to 39 (reference)						
40 to 59		1.73 (1.51, 1.99)	1.54 (1.33, 1.77)		1.13 (0.97, 1.32)	1.08 (0.92, 1.27)
60 to 79		0.98 (0.81, 1.18)	0.77 (0.62, 0.95)		0.39 (0.33, 0.48)	0.36 (0.30, 0.44)
80 to 105		0.45 (0.32, 0.64)	0.36 (0.24, 0.51)		0.13 (0.10, 0.17)	0.12 (0.09, 0.16)
Sex						
Male (reference)						
Female		2.01 (1.85, 2.19)	1.31 (1.22, 1.41)		1.48 (1.36, 1.60)	1.23 (1.16, 1.30)
BMI						
Not obese (reference)						
Obese i (30-34.9)		1.59 (1.49, 1.71)	1.25 (1.17, 1.34)		1.24 (1.16, 1.32)	1.13 (1.08, 1.19)
Obese ii (35-39.9)		1.77 (1.61, 1.94)	1.32 (1.20, 1.46)		1.29 (1.17, 1.43)	1.13 (1.06, 1.21)
Obese iii (40+)		2.13 (1.88, 2.41)	1.50 (1.32, 1.70)		1.28 (1.12, 1.47)	1.08 (0.99, 1.17)
Ethnicity						
White (reference)						
Asian or Asian British		1.02 (0.84, 1.22)	1.23 (1.02, 1.47)		0.60 (0.47, 0.75)	0.78 (0.67, 0.92)
Black or Black British		0.92 (0.75, 1.12)	1.10 (0.93, 1.30)		0.49 (0.38, 0.63)	0.67 (0.56, 0.81)
Chinese or other ethnic group		0.42 (0.33, 0.55)	0.63 (0.50, 0.81)		0.54 (0.42, 0.70)	0.72 (0.56, 0.92)
Missing		0.92 (0.79, 1.07)	1.05 (0.92, 1.21)		0.99 (0.84, 1.17)	1.07 (0.91, 1.26)
Mixed		1.15 (0.96, 1.38)	1.33 (1.11, 1.60)		0.67 (0.56, 0.81)	0.81 (0.66, 1.00)
Region						
East (reference)						
London		0.76 (0.58, 1.00)	0.93 (0.71, 1.22)		0.73 (0.48, 1.12)	0.91 (0.60, 1.36)
East Midlands		1.23 (0.92, 1.64)	1.23 (0.92, 1.64)		0.84 (0.56, 1.24)	0.84 (0.56, 1.24)
North East		2.10 (1.27, 3.50)	2.27 (1.35, 3.81)		2.87 (0.98, 8.42)	3.09 (1.02, 9.32)
North West		1.71 (1.01, 2.91)	1.69 (1.00, 2.84)		1.88 (1.01, 3.48)	1.97 (1.07, 3.60)
South East		2.22 (1.03, 4.78)	2.35 (1.09, 5.03)		1.19 (0.51, 2.80)	1.21 (0.52, 2.85)
South West		1.42 (0.93, 2.15)	1.40 (0.92, 2.12)		1.61 (0.65, 3.96)	1.58 (0.64, 3.93)
West Midlands		1.24 (0.88, 1.74)	1.34 (0.94, 1.91)		0.97 (0.54, 1.74)	1.18 (0.66, 2.12)
Yorkshire and the Humber		1.63 (0.99, 2.67)	1.65 (0.96, 2.83)		1.63 (0.72, 3.69)	1.71 (0.74, 3.96)
Deprivation (IMD) quintile						
1 (most deprived, reference)						
2		0.98 (0.85, 1.14)	1.11 (0.94, 1.32)		1.13 (0.93, 1.38)	1.27 (0.96, 1.70)
3		1.05 (0.84, 1.31)	1.22 (0.99, 1.50)		1.14 (0.92, 1.42)	1.31 (0.99, 1.74)
4		1.12 (0.94, 1.35)	1.30 (1.05, 1.60)		1.38 (1.04, 1.83)	1.58 (1.17, 2.12)
5 (least deprived)		1.13 (0.90, 1.42)	1.33 (1.01, 1.73)		1.34 (0.99, 1.82)	1.59 (1.10, 2.30)
0 (missing)		1.10 (0.87, 1.39)	1.16 (0.92, 1.45)		1.24 (1.04, 1.48)	1.30 (1.05, 1.61)

Hazard ratio scale (left and right): 0.1 0.25 0.5 1 2 4 10

Hazard ratio

Cox Model   ◇ Age-and-sex adjusted   ■ Fully adjusted

Figure 3. Primary and post-COVID diagnosis cohorts: age-and-sex adjusted and fully adjusted hazard ratios of coded long COVID for health behavioural and clinical variables.
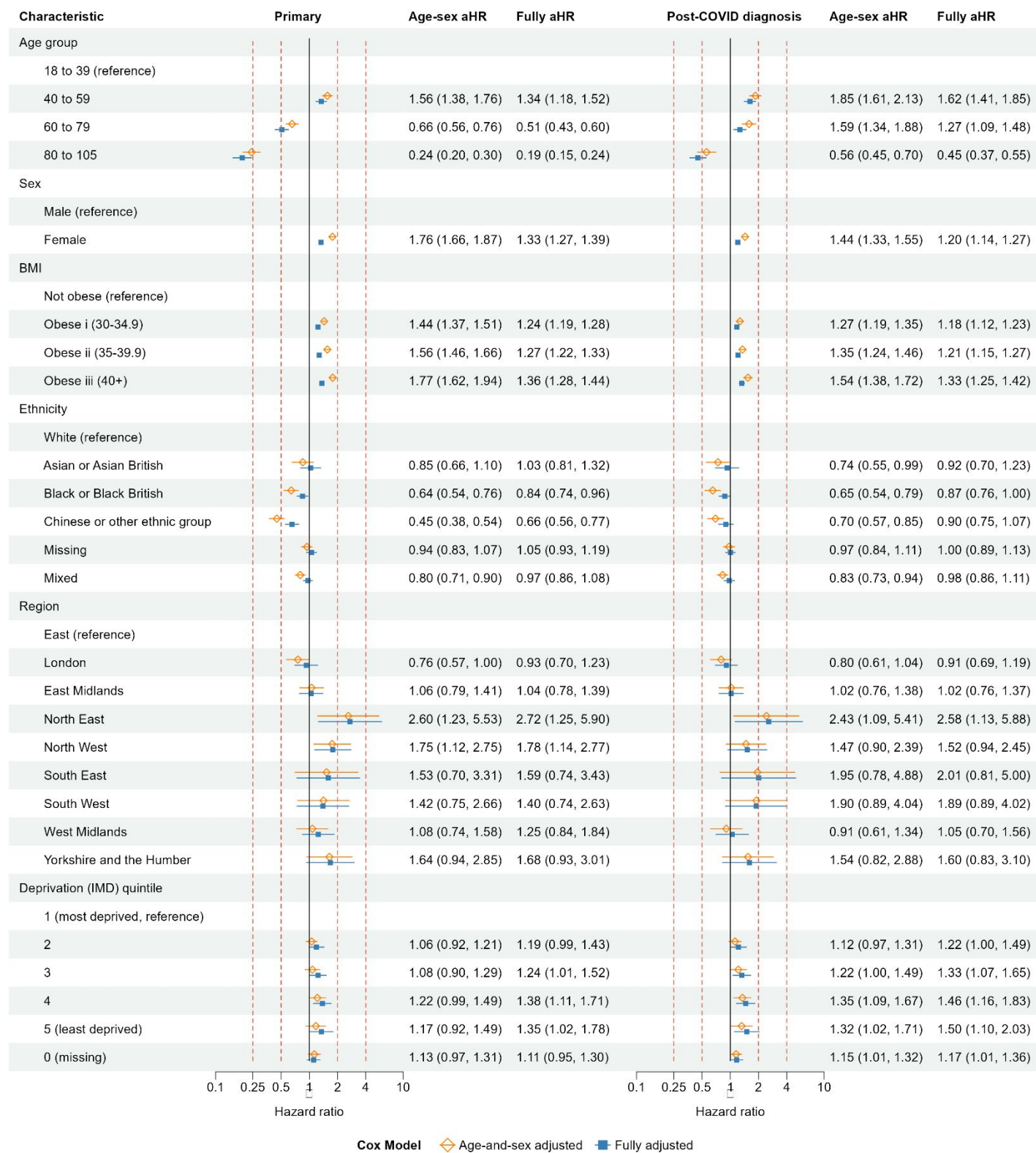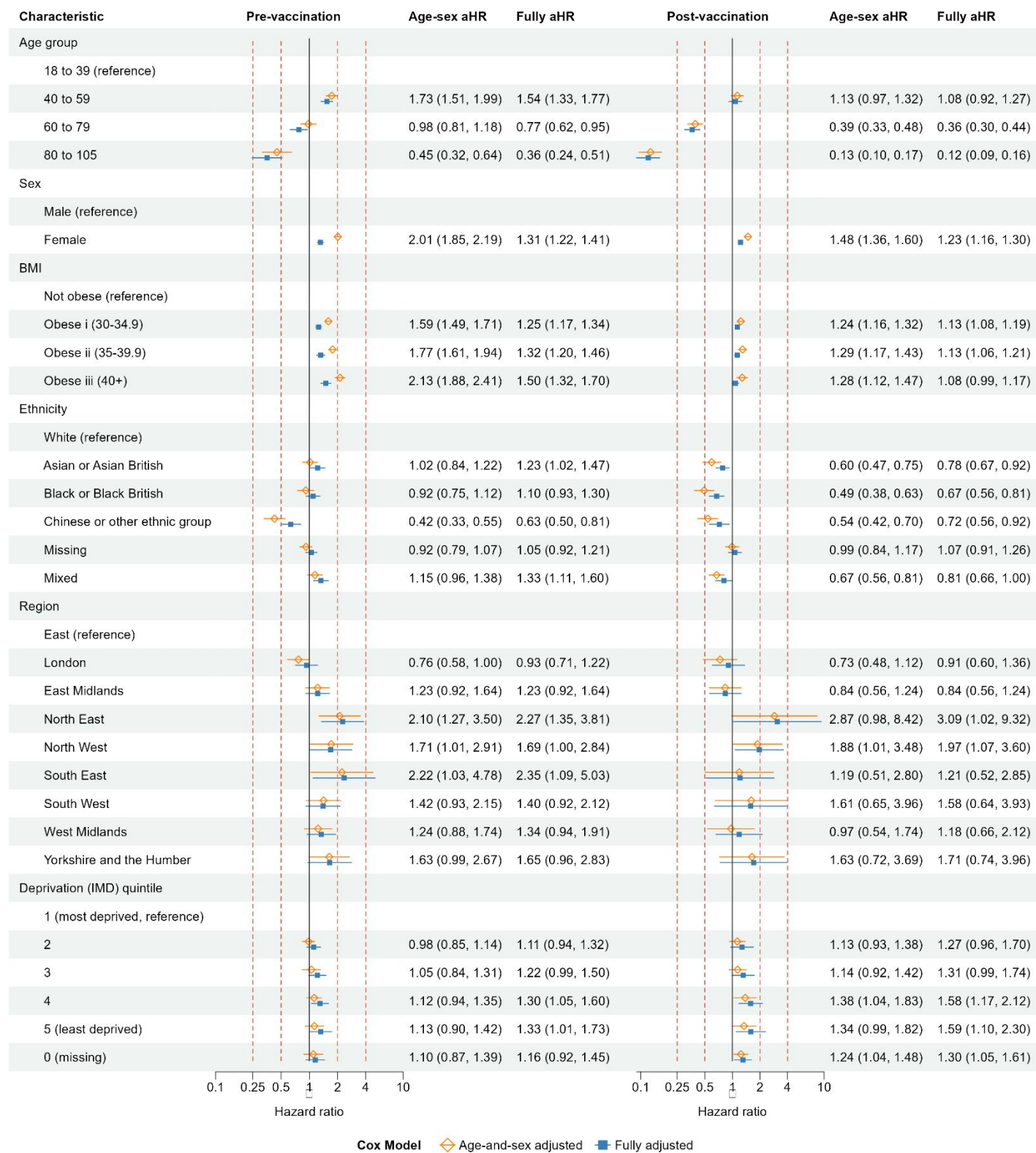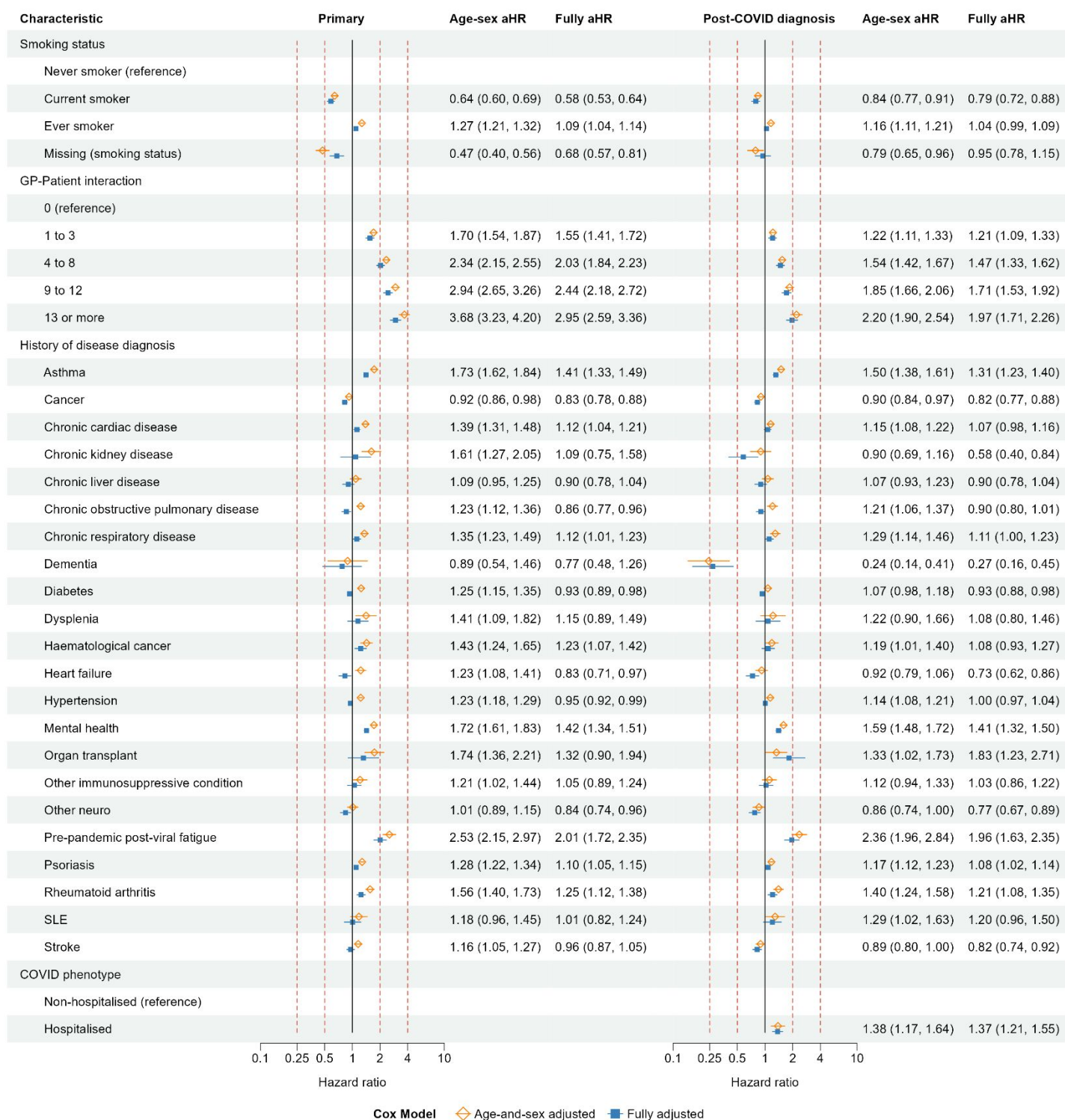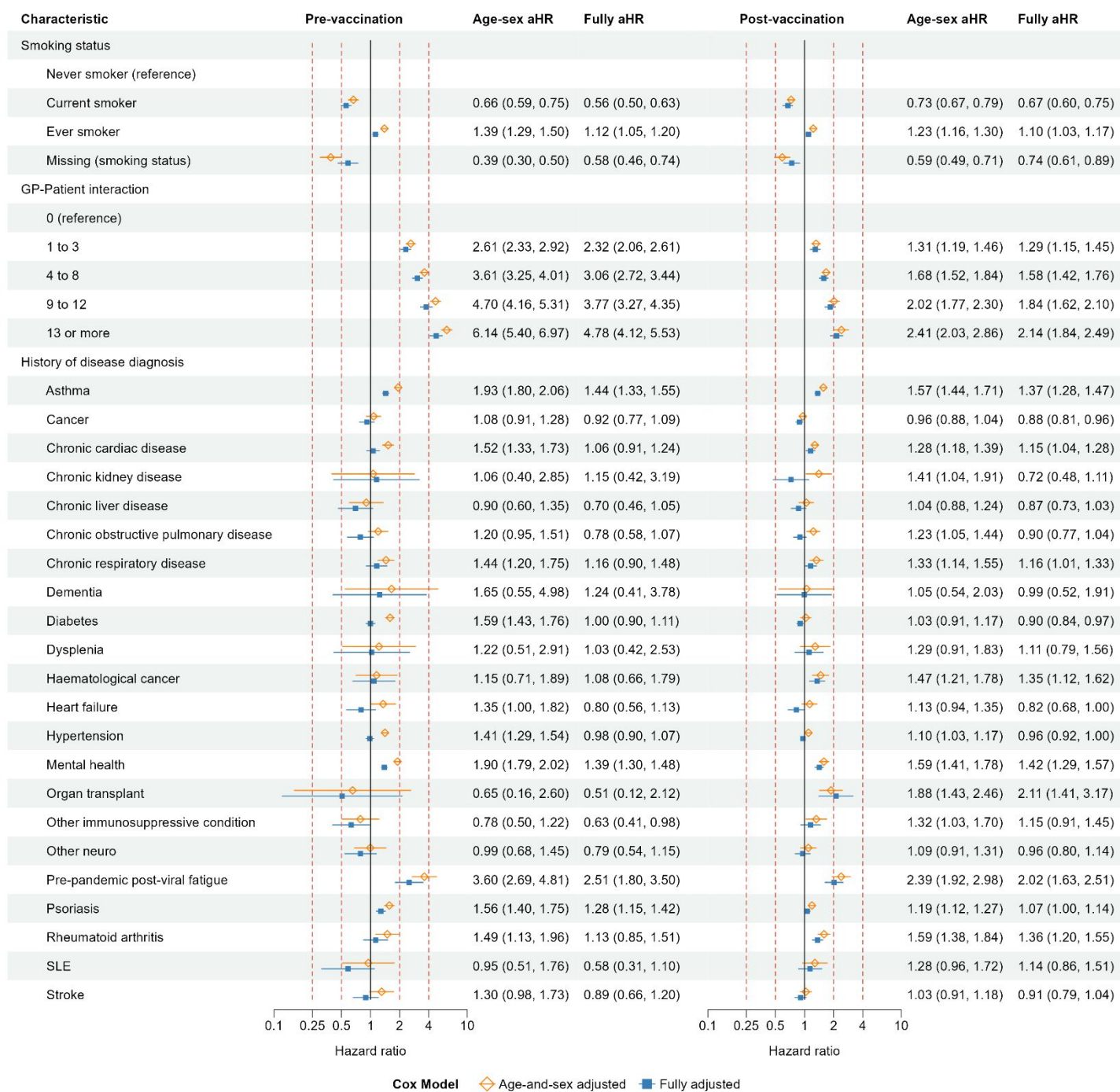
Figure 4. Pre-vaccination and post-vaccination cohorts: age-and-sex adjusted and fully adjusted hazard ratios of coded long COVID for health behavioural and clinical variables.

Characteristic	Pre-vaccination	Age-sex aHR	Fully aHR	Post-vaccination	Age-sex aHR	Fully aHR
Smoking status						
Never smoker (reference)						
Current smoker		0.66 (0.59, 0.75)	0.56 (0.50, 0.63)		0.73 (0.67, 0.79)	0.67 (0.60, 0.75)
Ever smoker		1.39 (1.29, 1.50)	1.12 (1.05, 1.20)		1.23 (1.16, 1.30)	1.10 (1.03, 1.17)
Missing (smoking status)		0.39 (0.30, 0.50)	0.58 (0.46, 0.74)		0.59 (0.49, 0.71)	0.74 (0.61, 0.89)
GP-Patient interaction						
0 (reference)						
1 to 3		2.61 (2.33, 2.92)	2.32 (2.06, 2.61)		1.31 (1.19, 1.46)	1.29 (1.15, 1.45)
4 to 8		3.61 (3.25, 4.01)	3.06 (2.72, 3.44)		1.68 (1.52, 1.84)	1.58 (1.42, 1.76)
9 to 12		4.70 (4.16, 5.31)	3.77 (3.27, 4.35)		2.02 (1.77, 2.30)	1.84 (1.62, 2.10)
13 or more		6.14 (5.40, 6.97)	4.78 (4.12, 5.53)		2.41 (2.03, 2.86)	2.14 (1.84, 2.49)
History of disease diagnosis						
Asthma		1.93 (1.80, 2.06)	1.44 (1.33, 1.55)		1.57 (1.44, 1.71)	1.37 (1.28, 1.47)
Cancer		1.08 (0.91, 1.28)	0.92 (0.77, 1.09)		0.96 (0.88, 1.04)	0.88 (0.81, 0.96)
Chronic cardiac disease		1.52 (1.33, 1.73)	1.06 (0.91, 1.24)		1.28 (1.18, 1.39)	1.15 (1.04, 1.28)
Chronic kidney disease		1.06 (0.40, 2.85)	1.15 (0.42, 3.19)		1.41 (1.04, 1.91)	0.72 (0.48, 1.11)
Chronic liver disease		0.90 (0.60, 1.35)	0.70 (0.46, 1.05)		1.04 (0.88, 1.24)	0.87 (0.73, 1.03)
Chronic obstructive pulmonary disease		1.20 (0.95, 1.51)	0.78 (0.58, 1.07)		1.23 (1.05, 1.44)	0.90 (0.77, 1.04)
Chronic respiratory disease		1.44 (1.20, 1.75)	1.16 (0.90, 1.48)		1.33 (1.14, 1.55)	1.16 (1.01, 1.33)
Dementia		1.65 (0.55, 4.98)	1.24 (0.41, 3.78)		1.05 (0.54, 2.03)	0.99 (0.52, 1.91)
Diabetes		1.59 (1.43, 1.76)	1.00 (0.90, 1.11)		1.03 (0.91, 1.17)	0.90 (0.84, 0.97)
Dysplenia		1.22 (0.51, 2.91)	1.03 (0.42, 2.53)		1.29 (0.91, 1.83)	1.11 (0.79, 1.56)
Haematological cancer		1.15 (0.71, 1.89)	1.08 (0.66, 1.79)		1.47 (1.21, 1.78)	1.35 (1.12, 1.62)
Heart failure		1.35 (1.00, 1.82)	0.80 (0.56, 1.13)		1.13 (0.94, 1.35)	0.82 (0.68, 1.00)
Hypertension		1.41 (1.29, 1.54)	0.98 (0.90, 1.07)		1.10 (1.03, 1.17)	0.96 (0.92, 1.00)
Mental health		1.90 (1.79, 2.02)	1.39 (1.30, 1.48)		1.59 (1.41, 1.78)	1.42 (1.29, 1.57)
Organ transplant		0.65 (0.16, 2.60)	0.51 (0.12, 2.12)		1.88 (1.43, 2.46)	2.11 (1.41, 3.17)
Other immunosuppressive condition		0.78 (0.50, 1.22)	0.63 (0.41, 0.98)		1.32 (1.03, 1.70)	1.15 (0.91, 1.45)
Other neuro		0.99 (0.68, 1.45)	0.79 (0.54, 1.15)		1.09 (0.91, 1.31)	0.96 (0.80, 1.14)
Pre-pandemic post-viral fatigue		3.60 (2.69, 4.81)	2.51 (1.80, 3.50)		2.39 (1.92, 2.98)	2.02 (1.63, 2.51)
Psoriasis		1.56 (1.40, 1.75)	1.28 (1.15, 1.42)		1.19 (1.12, 1.27)	1.07 (1.00, 1.14)
Rheumatoid arthritis		1.49 (1.13, 1.96)	1.13 (0.85, 1.51)		1.59 (1.38, 1.84)	1.36 (1.20, 1.55)
SLE		0.95 (0.51, 1.76)	0.58 (0.31, 1.10)		1.28 (0.96, 1.72)	1.14 (0.86, 1.51)
Stroke		1.30 (0.98, 1.73)	0.89 (0.66, 1.20)		1.03 (0.91, 1.18)	0.91 (0.79, 1.04)

Cox Model ◇ Age-and-sex adjusted ■ Fully adjusted

REFERENCES

1. Munblit D, O'Hara ME, Akrami A, et al. Long COVID: aiming for a consensus. *Lancet Respir Med* 2022;10(7):632-34. doi: 10.1016/s2213-2600(22)00135-7 [published Online First: 20220504]
2. Al-Aly Z, Xie Y, Bowe B. High-dimensional characterization of post-acute sequelae of COVID-19. *Nature* 2021;594(7862):259-64. doi: 10.1038/s41586-021-03553-9 [published Online First: 2021/04/23]
3. NHS. Post-COVID Syndrome (Long COVID) https://www.england.nhs.uk/coronavirus/post-covid-syndrome-long-covid/ [accessed 04 August 2022].
4. Soriano JB, Murthy S, Marshall JC, et al. A clinical case definition of post-COVID-19 condition by a Delphi consensus. *Lancet Infect Dis* 2022;22(4):e102-e07. doi: 10.1016/s1473-3099(21)00703-9 [published Online First: 20211221]
5. NICE. COVID-19 rapid guideline: managing the long-term effects of COVID-19 https://www.nice.org.uk/guidance/NG1882021 [accessed 31 July 2022].
6. Walker AJ, MacKenna B, Inglesby P, et al. Clinical coding of long COVID in English primary care: a federated analysis of 58 million patient records in situ using OpenSAFELY. *Br J Gen Pract* 2021;71(712):e806-e14. doi: 10.3399/bjgp.2021.0301
7. Thompson EJ, Williams DM, Walker AJ, et al. Long COVID burden and risk factors in 10 UK longitudinal studies and electronic health records. *Nat Commun* 2022;13(1):3528. doi: 10.1038/s41467-022-30836-0 [published Online First: 20220628]
8. Sudre CH, Murray B, Varsavsky T, et al. Attributes and predictors of long COVID. *Nat Med* 2021;27(4):626-31. doi: 10.1038/s41591-021-01292-y [published Online First: 20210310]
9. ONS. Prevalence of ongoing symptoms following coronavirus (COVID-19) infection in the UK: 4 June 2021 https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/bulletins/prevalenceofongoingsymptomsfollowingcoronaviruscovid19infectionintheuk/4june2021 [accessed 2 August 2022].
10. Willans R, Allsopp G, Jonsson P, et al. Primary Care Post-COVID syndrome Diagnosis and Referral Coding. *medRxiv* 2023:2023.05.23.23289798. doi: 10.1101/2023.05.23.23289798
11. NHS E. Long COVID: the NHS plan for 2021/22 2021 [Available from: https://www.england.nhs.uk/coronavirus/documents/long-covid-the-nhs-plan-for-2021-22/#action-summary accessed 02 June 2023 2023.
12. UKHSA. Changes to COVID-19 testing in England from 1 April 2022 [Available from: https://www.gov.uk/government/news/changes-to-covid-19-testing-in-england-from-1-april.
13. WHO. A healthy lifestyle - WHO recommendations https://www.who.int/europe/news-room/fact-sheets/item/a-healthy-lifestyle---who-recommendations2010 [accessed 19 May 2023.
14. ONS. Eurostat - An overview of the 3 NUTS and 2 LAU layers in the UK [Available from: https://www.ons.gov.uk/methodology/geography/ukgeographies/eurostat accessed 14 November 2021.
15. Knight R, Walker V, Ip S, et al. Association of COVID-19 With Major Arterial and Venous Thrombotic Diseases: A Population-Wide Cohort Study of 48 Million Adults in England and Wales. *Circulation* 2022;146(12):892-906. doi: 10.1161/circulationaha.122.060785 [published Online First: 20220919]
16. Uno H, Cai T, Pencina MJ, et al. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Stat Med* 2011;30(10):1105-17. doi: 10.1002/sim.4154 [published Online First: 2011/04/13]
17. GitHub. opensafely/long-covid-risk-factors-and-prediction https://github.com/opensafely/long-covid-risk-factors-and-prediction [accessed 18 December 2024].
18. ONS. Prevalence of ongoing symptoms following coronavirus (COVID-19) infection in the UK: 6 May 2022

https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/bulletins/prevalenceofongoingsymptomsfollowingcoronaviruscovid19infectionintheuk/6may20222022 [accessed 3 August 2022].

19. Hulme WJ, Horne EMF, Parker EPK, et al. Comparative effectiveness of BNT162b2 versus mRNA-1273 covid-19 vaccine boosting in England: matched cohort study in OpenSAFELY-TPP. *BMJ* 2023;380:e072808. doi: 10.1136/bmj-2022-072808

20. Andrews C, Schultze A, Curtis H, et al. OpenSAFELY: Representativeness of electronic health record platform OpenSAFELY-TPP data compared to the population of England. *Wellcome Open Res* 2022;7:191. doi: 10.12688/wellcomeopenres.18010.1

21. Kontopantelis E, Stevens RJ, Helms PJ, et al. Spatial distribution of clinical computer systems in primary care in England in 2016 and implications for primary care electronic medical record databases: a cross-sectional population study. *BMJ Open* 2018;8(2):e020738. doi: 10.1136/bmjopen-2017-020738 [published Online First: 20180228]

22. Shah AD, Subramanian A, Lewis J, et al. Long Covid symptoms and diagnosis in primary care: A cohort study using structured and unstructured data in The Health Improvement Network primary care database. *PLoS One* 2023;18(9):e0290583. doi: 10.1371/journal.pone.0290583

23. Westreich D, Greenland S. The table 2 fallacy: presenting and interpreting confounder and modifier coefficients. *Am J Epidemiol* 2013;177(4):292-8. doi: 10.1093/aje/kws412

24. Subramanian A, Nirantharakumar K, Hughes S, et al. Symptoms and risk factors for long COVID in non-hospitalized adults. *Nat Med* 2022 doi: 10.1038/s41591-022-01909-w [published Online First: 20220725]

25. Peters C, Dulon M, Westermann C, et al. Long-Term Effects of COVID-19 on Workers in Health and Social Services in Germany. *Int J Environ Res Public Health* 2022;19(12) doi: 10.3390/ijerph19126983 [published Online First: 20220607]

26. Pazukhina E, Andreeva M, Spiridonova E, et al. Prevalence and risk factors of post-COVID-19 condition in adults and children at 6 and 12 months after hospital discharge: a prospective, cohort study in Moscow (StopCOVID). *BMC Med* 2022;20(1):244. doi: 10.1186/s12916-022-02448-4 [published Online First: 20220706]

27. Park C, Chaturvedi N, Sterne J, et al. SAGE Report - Short report on Long COVID, 2021.

28. Osmanov IM, Spiridonova E, Bobkova P, et al. Risk factors for post-COVID-19 condition in previously hospitalised children using the ISARIC Global follow-up protocol: a prospective cohort study. *Eur Respir J* 2022;59(2) doi: 10.1183/13993003.01341-2021 [published Online First: 20220203]

29. Pérez-González A, Araújo-Ameijeiras A, Fernández-Villar A, et al. Long COVID in hospitalized and non-hospitalized patients in a large cohort in Northwest Spain, a prospective cohort study. *Sci Rep* 2022;12(1):3369. doi: 10.1038/s41598-022-07414-x [published Online First: 20220301]

30. Maglietta G, Diodati F, Puntoni M, et al. Prognostic Factors for Post-COVID-19 Syndrome: A Systematic Review and Meta-Analysis. *J Clin Med* 2022;11(6) doi: 10.3390/jcm11061541 [published Online First: 20220311]

31. The NHS England OpenSAFELY COVID-19 service - privacy notice. NHS Digital (Now NHS England). [Available from: https://digital.nhs.uk/coronavirus/coronavirus-covid-19-response-information-governance-hub/the-nhs-england-opensafely-covid-19-service-privacy-notice accessed 2 April 2025.

32. Data Security and Protection Toolkit - NHS Digital. NHS Digital (Now NHS England). [Available from: https://digital.nhs.uk/services/data-security-and-protection-toolkit/data-security-and-protection-toolkit accessed 2 April 2025.

33. ISB1523: Anonymisation Standard for Publishing Health and Social Care Data. NHS Digital (Now NHS England) [Available from: https://digital.nhs.uk/data-and-information/information-

standards/information-standards-and-data-collections-including-extractions/publications-and-notifications/standards-and-collections/isb1523-anonymisation-standard-for-publishing-health-and-social-care-data accessed 04 April 2025.

34. Coronavirus (COVID-19): notice under regulation 3(4) of the Health Service (Control of Patient Information) Regulations 2002 – general. 2022. [Available from: https://www.gov.uk/government/publications/coronavirus-covid-19-notification-of-data-controllers-to-share-information/coronavirus-covid-19-notice-under-regulation-34-of-the-health-service-control-of-patient-information-regulations-2002-general--2.

35. Secretary of State for Health and Social Care - UK Government. COVID-19 Public Health Directions 2020: notification to NHS Digital. [Available from: https://digital.nhs.uk/about-nhs-digital/corporate-information-and-documents/directions-and-data-provision-notices/secretary-of-state-directions/covid-19-public-health-directions-2020 accessed 2 April 2025.

36. Confidentiality Advisory Group. Health Research Authority. [Available from: https://www.hra.nhs.uk/about-us/committees-and-services/confidentiality-advisory-group/ accessed 2 April 2025.