TITLE

Study protocol: Risk factors for clinically diagnosed long COVID: an analysis of linked electronic health records using the OpenSAFELY platform

AUTHORS

Yinghui Wei[1,2,#], Jonathan AC Sterne[2] on behalf of the Longitudinal Health and Wellbeing UK COVID-19 National Core Study and CONVALESCENCE study

[1]University of Plymouth

[2]University of Bristol

[#] Correspondence to Yinghui Wei, Email: yinghui.wei@plymouth.ac.uk

RESARCH QUESTIONS

- Among the general population, what factors are associated with long COVID?
- Among the post-COVID population, who are more likely to have long COVID?
- Can vaccination modify the risk of long COVID?

DATA SOURCES

This research will be conducted using OpenSAFELY and requires the following data sources:

- Primary care data (TPP)
- Second Generation Surveillance System (SGSS) for Pillar 1 and Pillar 2 SARS-COV-2 infection laboratory testing data
- Secondary Uses Service (SUS)
- Index of Multiple Deprivation (IMD)
- Office of National Statistics (ONS) death registry

STUDY DESIGN

This is a population-based cohort study based on electronic health records using the OpenSAFELY platform. The study period is from 29 January 2020 to 31 March 2022.

STUDY POPULATION

The study population is adults with age between 18 and 105 years, alive and registered in an English general practice which employs the TPP system at the study start date 29 January 2020. The inclusion and exclusion criteria are listed in Table 1.

Table 1. Overview of the inclusion and exclusion criteria of the study population

Inclusion	Exclusion
• Adult, aged between 18 and 105 years on study start date • Known sex on study start date • Known age on study start date • Know region on study start date • Registered on study start date in an English general practice which employs the TPP system.	• COVID infection prior to the start date of individual follow-up • Long COVID record prior to the start date of individual follow-up • Less than 12 months of follow-up prior to cohort start date

To answer the research questions, we consider four cohorts (Table 2). The primary cohort includes all eligible general population. In the pre-vaccination cohort, individuals are additionally censored by the first vaccination date. In the post-vaccination cohort, the second vaccination date plus 14 days is the individual specific follow-up start date. In the post-COVID cohort, the first COVID date during the study period is the individual specific follow-up start date.

Table 2. Description of study cohorts

Cohorts	**Study period**	**Participant follow-up start date (Index date)**	**Participant follow-up end date**	**Population**	**Outcome**
Primary cohort	29/01/2020 to 31/03/2022	29/01/2020	Earliest of first long COVID record, death or end date of study period	All eligible adults	Time from 29 January 2020 to first long COVID record, individuals censored at first long COVID record, death or end date of study period
Pre-vaccination cohort	29/01/2020 to 31/03/2022	29/01/2020	Earliest of first long COVID record, first COVID vaccine, death, or end date of study period	Eligible unvaccinated adults	Time from 29 January 2020 to first long COVID record, individuals censored at first vaccination date, first long COVID record, death or end date of study period.
Post-vaccination cohort	29/01/2022 to 31/03/2022	The second vaccination + 14 days	Earliest of first long COVID record, death or end date of study period	Eligible vaccinated adults	Time from the second vaccination +14 days to first long COVID record, individuals censored at first long COVID record, death or end date of study period.
Post-infection cohort	29/01/2020 to 31/03/2022	First COVID date during the study period	Earliest of first long COVID record, death or end date of study period	All eligible adults with COVID infection	Time from COVID infection to first long COVID record, individuals censored at first long COVID, death or end date of study period.

OUTCOMES

The outcome of interest is any record of long COVID in the primary care record, and is defined using a list of 15 UK SNOMED-CT codes[1]. The outcome is measured between the study start date (29 January 2020) and the end date (31 March 2022). Timing of outcomes was determined by the first record of a SNOMED code for each person, as determined by the date recorded by clinician. The outcome is measured by the time (in days) from index date to the first record of long COVID code.

PATIENT CHARACTERISTICS

Patient characteristics for consideration are based upon medical knowledge and informed by previous work[1 2]. We will include demographic variables, clinical variables, vaccination status and variables indicating health-related behaviours as covariates. The measurement time is on the index date for demographic variable, on or before index date for clinical variables, 12 months prior to index date for GP consultation rate, and on or before pandemic start date (29 January 2020) for post-viral fatigue. The rationale is the potential use of post-viral fatigue as a proxy for long COVID, particularly in the early stages of the pandemic.

Table 3. Derivation of patient characteristics

Variable	**Type**	**Definition**	**Measurement time**	**Data sources**
Age	Continuous	Age in years	Index	Primary care
Age	Categorical	18 – 39 40 – 59 60 – 79 80+	Index date	Primary care
Sex*	Categorical	Male, Female	As per the record	Primary care
Obesity	Categorical	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+.	Index date	Primary care
Ethnicity	Categorical	White Mixed Asian or Asian British Black or Black British Chinese or Other Ethnic Groups	Index date	Primary care
Region	Categorical	North East North West Yorkshire and the Humber East Midlands West Midlands East London South East South West	Index date	Primary care
Deprivation	Categorical	Deprivation	Index date	Index of Multiple

		quintiles: 1 (most deprived) 2 3 4 5 (least deprived)		Deprivation
Smoking status	Categorical	Never smoker Ever smoker Current smoker	Index date	Primary care
GP-Patient Interaction	Categorical	Number of GP consultations in the following categories 0; 1-3; 4-8; 9-12; 13+	12 months prior to index date	Primary care
Asthma	Binary	Yes; No	On or before index date	Primary care
Cancer (exclude lung and haematological cancer)	Binary	Yes; No	On or before index date	Primary care
Chronic cardiac disease	Binary	Yes; No	On or before index date	Primary care
Chronic kidney disease	Binary	Yes; No	On or before index date	Primary care
Chronic liver disease	Binary	Yes; No	On or before index date	Primary care
Chronic obstructive pulmonary disease	Binary	Yes; No	On or before index date	Primary care
Chronic respiratory conditions	Binary	Yes; No	On or before index date	Primary care
Dementia	Binary	Yes; No	On or before index date	Primary care
Diabetes	Binary	Yes; No	On or before index date	Primary care
Dysplenia (Dysfunctional-Spleen)	Binary	Yes; No	On or before index date	Primary care
Haematological cancer	Binary	Yes; No	On or before index date	Primary care
Heart failure	Binary	Yes; No	On or before index date	Primary care
Hypertension	Binary	Yes; No	On or before index date	Primary care

Mental health conditions	Binary	Yes; No	On or before index date	Primary care
Organ transplant	Binary	Yes; No	On or before index date	Primary care
Other immunosuppressive condition	Binary	Yes; No	On or before index date	Primary care
Other neurological condition not including dementia or stroke	Binary	Yes; No	On or before index date	Primary care
Post-viral fatigue	Binary	Yes; No	Before pandemic start	Primary care
Psoriasis	Binary	Yes; No	On or before index date	Primary care
Rheumatoid arthritis	Binary	Yes; No	On or before index date	Primary care
Systemic lupus erythematosus	Binary	Yes; No	On or before index date	Primary care
Stroke	Binary	Yes; No	On or before index date	Primary care

*This variable is derived once per patient without a date specification so is an exception to 'most recent data prior to the study start date'.

STATISTICAL ANALYSES

Patient baseline characteristics will be presented using number and percentage for categorical variables, and mean ± standard deviation, median and interquartile range for continuous variables. For categorical variables, missing data are included as a missing category. Due to the large number of observations in electronic health records, it is anticipated the proportion of missing data is small and the incorporating missing data as a category allows to include all eligible individuals in the analysis.

We will include a missing category for ethnicity, smoking status and index of multiple deprivation. All other covariates will be defined using the presence versus absence of specific codes, so have no identifiable missing values.

Rates of clinically diagnosed long COVID will be quantified as number of patients with long COVID diagnosis per 1000 person-years. The cumulative probability of clinically diagnosed long COVID will be estimated by age group and sex. For each patient characteristic, an age-and-sex Cox proportional hazards model will be fitted. The time to the event outcome will be defined as days from participant specific follow-up start date to participant specific follow-up end date. We will model age by restricted cubic spline. All potential risk factors, including demographic and clinical factors listed in 'potential risk factors', will then be included in a single multivariable Cox proportional hazards model, with age modelled as restricted cubic spline. Both age-and-sex adjusted, and fully adjusted models will be implemented for each cohort. In the post-COVID cohort, we will include COVID-19 severity (hospitalised vs non-hospitalised infection) as an additional risk factor. Hazard ratios from the age-and-sex adjusted and fully adjusted models will be reported with 95% confidence intervals. The log hazard ratios against continuous age in years will be plotted. Models will be also refitted with age as a categorical variable to obtain hazard ratios by age group.

Since the large sample size available from electronic health records for analysis, overfitting was expected to be minimal, such that regularization of the predictor effects will not be considered.

For computational efficiency, if the number of patients without a long COVID code is 20 times higher than the number of patients with a long COVID code, we will sample the population without long COVID and use the whole population with long COVID. We will then use inverse probability weighting to reflect the original sample size.

The performance of the fitted model will be assessed by discrimination measure, C-statistic.

STUDY CONDUCT

Data curation will be conducted using Python. All data analysis will be performed in R or Stata. Analytical scripts will be shared on a GitHub repository for reproducible research.

PROPOSED OUTPUTS

Table 1. Patient characteristics. Summary statistics are number (percentage) for categorical variables and mean (standard deviation) for continuous variables.

Table 2. Event outcome / 1000 person years and incidence rates per 1000 person years for long COVID by demographics.

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

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

Figure 3. Primary and post-COVID cohorts: age and sex adjusted and fully adjusted hazard ratios for clinical variables

Figure 4. Pre-vaccination and post-vaccination cohorts: age and sex adjusted and fully adjusted hazard ratios for clinical variables

REFERENCES

1. 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. *British Journal of General Practice* 2021;71(721): e806–e14.
2. 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]