

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

- | | |
|-------------------------------------|--|
| n/a | Confirmed |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> The statistical test(s) used AND whether they are one- or two-sided
<i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i> |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> A description of all covariates tested |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
<i>Give P values as exact values whenever suitable.</i> |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated |

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection: Data extraction for epidemiological research (DExTER)

Data analysis: Stata IC version 16

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The data that support the findings of this study are available from CPRD but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the CPRD, subject to Research Data Governance (RDG) approvals.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	A total of 486,149 non-hospitalised individuals had a coded record of confirmed SARS CoV-2 infection (infected cohort), and 8,030,224 had no records of either suspected or confirmed COVID-19 during the study period, which formed the pool of uninfected patients. From the uninfected pool, 1,944,580 individuals were propensity score matched to patients in the infected cohort. Of these patients, 384,137 and 1,501,689 infected and uninfected patients had a minimum follow-up of 12 months. This resulted in over 90% statistical power to detect the observed difference in the hazard of composite symptom outcome (aHR: 1.26, 1.25-1.28).
Data exclusions	(1) Patients aged 18 years or below on index date were excluded from the study. (2) Patients who were registered with a general practice for less than 12 months were excluded from the study. (3) Practices contributing data that were not of research quality were excluded from the study (4) Patients with a record of hospitalisation 14 days before or 42 days after infection (within 28 days of infection with a ± 14 -day grace period for clinical coding delays) were excluded from the study, since our study focused on non-hospitalised individuals. (5) Patients without a minimum follow-up period of four and twelve weeks were excluded from the symptom outcome analyses during the four to twelve weeks period after infection ("ongoing symptomatic COVID-19") and twelve weeks after infection (period of "post-COVID-19 condition" or "Long COVID"), respectively.
Replication	Data for this study was extracted in an analysable format from CPRD AURUM database using the in-house DEXTER software, which provides a non-invasive solution to generate quality datasets through a process that can be verified and reproducible. STATA codes used to clean and analyse the data can be made available upon request to authors to assess reproducibility of results.
Randomization	This was not a randomized controlled study. However, in this retrospective cohort study, we propensity score matched the infected and uninfected patients using a logistic regression model with a calliper width of 0.2, including a comprehensive range of covariates such as age, sex, body mass index, smoking status, socioeconomic deprivation, ethnicity, symptoms and comorbidities recorded at baseline. The result of this was well balanced cohorts in terms of baseline covariates, evidenced by the kernel density plots of propensity scores for the two cohorts before and after matching, and the estimated standardised mean differences for each of the baseline covariates, which were all < 0.1 .
Blinding	Blinding was not possible in this study as the data were collected retrospectively from primary care records.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology and archaeology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input type="checkbox"/>	<input checked="" type="checkbox"/> Human research participants
<input type="checkbox"/>	<input checked="" type="checkbox"/> Clinical data
<input checked="" type="checkbox"/>	<input type="checkbox"/> Dual use research of concern

Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

Human research participants

Policy information about [studies involving human research participants](#)

Population characteristics	The mean age of all eligible patients included in our study was 43.8 years (SD 16.9) and 55.3% were female. 64.7% were white, 12.2% were Asian and 4.0% were black, and ethnicity data were missing for 16.2%. 53.8% were overweight or obese (with BMI data missing for 13.0%) and 22.5% were current smokers (with smoking data missing for 4.3%). The most common comorbidities were depression (22.1%), anxiety (20.3%), asthma (20.1%), eczema (19.5%), and hay fever (18.1%).
Recruitment	We conducted a population-based retrospective matched cohort study using data from the Medicines and Healthcare

Recruitment

products Regulatory Agency (MHRA) Clinical Practice Research Datalink (CPRD) Aurum. CPRD Aurum is an anonymised database of primary care medical records of over seven million actively registered patients in general practices that use the EMIS clinical information system. A total of 486,149 non-hospitalised individuals had a coded record of confirmed SARS CoV-2 infection (infected cohort), and 8,030,224 had no records of either suspected or confirmed COVID-19 during the study period, which formed the pool of uninfected patients. A limitation of our study is potential misclassification bias due to retrospective recruitment of the infected and the uninfected cohort from primary care records using SNOMED CT codes. Community testing for SARS CoV-2 was very limited during the first wave of the pandemic, and many non-hospitalised individuals with COVID-19 were not tested. It is therefore possible that some members of our uninfected cohort had been infected with SARS CoV-2 but had simply not been tested. We attempted to partially account for this bias by excluding individuals from the uninfected pool if they had a coded diagnosis of suspected COVID-19. However, this is unlikely to be 100% sensitive in identifying individuals with unverified COVID-19 from the uninfected pool, which would potentially have the effect of attenuating the observed effect sizes.

Ethics oversight

CPRD obtains annual research ethics approval from the UK's Health Research Authority (HRA) Research Ethics Committee (REC) (East Midlands – Derby, REC reference number 05/MRE04/87) to receive and supply patient data for public health research. Therefore, no additional ethics approval is required for observational studies using CPRD Aurum data for public health research, subject to individual research protocols meeting CPRD data governance requirements. The use of CPRD Aurum data for the study has been approved by the CPRD Independent Scientific Advisory Committee (reference: 21_000423).

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Clinical data

Policy information about [clinical studies](#)

All manuscripts should comply with the ICMJE [guidelines for publication of clinical research](#) and a completed [CONSORT checklist](#) must be included with all submissions.

Clinical trial registration

ISRCTN15674970

Study protocol

<https://doi.org/10.1101/2021.12.20.21268098>

Data collection

We conducted a population-based retrospective matched cohort study between 31st January 2020 and 15th April 2021 using data from the Medicines and Healthcare products Regulatory Agency (MHRA) Clinical Practice Research Datalink (CPRD) Aurum. CPRD Aurum is an anonymised database of primary care medical records of over seven million actively registered patients in general practices that use the EMIS clinical information system. It captures data on patient demographics, diagnoses, symptoms, prescriptions, referrals, and tests. Structured data on diagnoses, symptoms and referrals are recorded using SNOMED CT coding terminology

Outcomes

We identified 115 relevant symptoms coded within primary care records through a systematic review and meta-analysis of Long COVID symptoms, a scoping search of Long COVID clinical assessment questionnaires, qualitative interviews with patients, a clinician survey, and refinement of the symptom list using psychometric methods. These were grouped into 15 domains: (1) breathing, (2) pain, (3) circulation, (4) fatigue, (5) cognitive health, (6) movement, (7) sleep, (8) ear, nose, and throat, (9) stomach and digestion, (10) muscles and joints, (11) mental health, (12) hair, skin and nails, (13) eyes, (14) reproductive health, and (15) other symptoms. Our primary outcome definition of Long COVID was predefined as the presence of at least one symptom included in the WHO case definition at ≥ 12 weeks post-infection. Our secondary outcome definition of Long COVID was derived post-hoc as the presence of at least one symptom that was statistically associated with SARS CoV-2 infection at ≥ 12 weeks post-infection within this study.