Prevalence and risk factors for long COVID among adults in Scotland using electronic health records: a national, retrospective, observational cohort study

Karen Jeffrey,[a] Lana Woolford,[a] Rishma Maini,[b] Siddharth Basetti,[c] Ashleigh Batchelor,[d] David Weatherill,[d] Chris White,[d] Vicky Hammersley,[a] Tristan Millington,[a] Calum Macdonald,[a] Jennifer K. Quint,[e] Robin Kerr,[f,g] Steven Kerr,[a] Syed Ahmar Shah,[a] Igor Rudan,[a] Adeniyi Francis Fagbamigbe,[l] Colin R. Simpson,[a,h] Srinivasa Vittal Katikireddi,[b,i] Chris Robertson,[b,j] Lewis Ritchie,[k,l] Aziz Sheikh,[a,m] and Luke Daines[a,m,*]

[a]Usher Institute, University of Edinburgh, Edinburgh, UK
[b]Public Health Scotland, Glasgow and Edinburgh, UK
[c]NHS Highland, Inverness, UK
[d]Patient and Public Contributors, Usher Institute, University of Edinburgh, Edinburgh, UK
[e]National Heart and Lung Institute, Imperial College London, London, UK
[f]NHS Borders, Melrose, UK
[g]NHS Dumfries & Galloway, Dumfries, UK
[h]School of Health, Wellington Faculty of Health, Victoria University of Wellington, Wellington, NZ
[i]MRC/CSO Social & Public Health Sciences Unit, University of Glasgow, Glasgow, UK
[j]Department of Mathematics and Statistics, University of Strathclyde, Glasgow, UK
[k]Academic Primary Care, University of Aberdeen, Aberdeen, UK
[l]Institute of Applied Health Sciences, University of Aberdeen, UK

eClinicalMedicine
2024;71: 102590
Published Online xxx
https://doi.org/10.1016/j.eclinm.2024.102590

Summary

Background Long COVID is a debilitating multisystem condition. The objective of this study was to estimate the prevalence of long COVID in the adult population of Scotland, and to identify risk factors associated with its development.

Methods In this national, retrospective, observational cohort study, we analysed electronic health records (EHRs) for all adults (≥18 years) registered with a general medical practice and resident in Scotland between March 1, 2020, and October 26, 2022 (98–99% of the population). We linked data from primary care, secondary care, laboratory testing and prescribing. Four outcome measures were used to identify long COVID: clinical codes, free text in primary care records, free text on sick notes, and a novel operational definition. The operational definition was developed using Poisson regression to identify clinical encounters indicative of long COVID from a sample of negative and positive COVID-19 cases matched on time-varying propensity to test positive for SARS-CoV-2. Possible risk factors for long COVID were identified by stratifying descriptive statistics by long COVID status.

Findings Of 4,676,390 participants, 81,219 (1.7%) were identified as having long COVID. Clinical codes identified the fewest cases (n = 1,092, 0.02%), followed by free text (n = 8,368, 0.2%), sick notes (n = 14,469, 0.3%), and the operational definition (n = 64,193, 1.4%). There was limited overlap in cases identified by the measures; however, temporal trends and patient characteristics were consistent across measures. Compared with the general population, a higher proportion of people with long COVID were female (65.1% versus 50.4%), aged 38–67 (63.7% versus 48.9%), overweight or obese (45.7% versus 29.4%), had one or more comorbidities (52.7% versus 36.0%), were immunosuppressed (6.9% versus 3.2%), shielding (7.9% versus 3.4%), or hospitalised within 28 days of testing positive (8.8% versus 3.3%%), and had tested positive before Omicron became the dominant variant (44.9% versus 35.9%). The operational definition identified long COVID cases with combinations of clinical encounters (from four symptoms, six investigation types, and seven management strategies) recorded in EHRs within 4–26 weeks of a positive SARS-CoV-2 test. These combinations were significantly ($p < 0.0001$) more prevalent in positive COVID-19 patients than in matched negative controls. In a case-crossover analysis, 16.4% of those identified by the operational definition had similar healthcare patterns recorded before testing positive.

Interpretation The prevalence of long COVID presenting in general practice was estimated to be 0.02–1.7%, depending on the measure used. Due to challenges in diagnosing long COVID and inconsistent recording of

*Corresponding author.
E-mail address: luke.daines@ed.ac.uk (L. Daines).
[m]Joint senior authors.

information in EHRs, the true prevalence of long COVID is likely to be higher. The operational definition provided a novel approach but relied on a restricted set of symptoms and may misclassify individuals with pre-existing health conditions. Further research is needed to refine and validate this approach.

Funding Chief Scientist Office (Scotland), Medical Research Council, and BREATHE.

Keywords: Long COVID; Population surveillance; Primary health care; Clinical coding; Matched-pair analysis

Research in context

Evidence before this study

We searched PubMed for studies that investigated the prevalence and risk factors associated with long COVID using data from electronic health records (EHRs), published up to January 31, 2023. We used the search terms ((“electronic health record*” OR “clinical cod*” OR “electronic patient record*” OR “electronic clinical record*” OR EHR OR EPR OR ECR) AND (“long covid” OR “post-COVID*” OR (“sequela*” AND “COVID*”) OR “post-acute COVID*” OR (“consequence*” AND COVID*) OR PASC OR “post-coronavirus” OR “ongoing coronavirus” OR “ongoing COVID*”)) and excluded studies that: considered only single-organ or single system effects; focussed on very specific sub-populations (such as pregnant women); primarily analysed survey data; or investigated acute SARS-CoV-2 infection. 23 studies analysed EHRs to identify risk factors and sequelae associated with long COVID. A subset used matched analyses to isolate the effects of SARS-CoV-2 infection on sequelae or used machine learning methods to identify symptom clusters or predict risk of long COVID. One study used natural language processing to identify symptoms of long COVID. However, none of the studies analysed population-level data to estimate the general prevalence of long COVID.

Added value of this study

The Early Pandemic Evaluation and Enhanced Surveillance of COVID-19 (EAVE II) platform allowed linkage of primary and secondary care data, with prescribing, vaccinations, SARS-CoV-2 testing and genomic sequencing records for over 4.6 million individuals (all adults registered with a general medical practice in Scotland). The opportunity to analyse free text entries, in addition to coded data contained in EHR, enabled a multi-measure approach to long-COVID identification. Using a multi-measure approach identified more patients with long COVID than a single measure would have. However, our results likely under-estimate the true prevalence of long COVID, due to incomplete recording of information in EHRs, and limitations of the operational definition itself. Specifically, the operational definition’s reliance on a limited set of symptoms and potential over-identification of individuals with pre-existing conditions must be considered when interpreting prevalence estimates.

Implications of all the available evidence

Accurately estimating the prevalence of long COVID is crucial for healthcare planning and service provision. Our study highlights the utility of combining different methods of identifying cases of long COVID using information recorded in EHRs. Although the withdrawal of widespread COVID-19 testing limits the viability of our operational definition, we consider analysis of free text recorded in primary care records and on sick notes, in conjunction with analysis of long COVID clinical codes, to be a promising (albeit conservative) approach for future surveillance of long COVID at a national level.

Introduction

Long COVID is a debilitating multisystem condition occurring after infection with severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2).[1] Individuals with long COVID describe wide ranging symptoms including fatigue, breathlessness, cognitive dysfunction and disturbances to taste or smell that can last for months or years, fluctuate over time and have a profound impact on daily life.[1–3] Long COVID is also associated with an increased incidence of type 2 diabetes,[4] myocardial infarction, stroke, venous thrombosis,[5] and dysregulation of the autonomic nervous system.[6] Given concerns about the burden of disease, accurate estimates of the number of people affected by long COVID are vital for policy makers and healthcare providers tasked with planning and providing healthcare services.

National surveys suggest that 2.9% of individuals in the UK and 5.0% in Scotland experienced symptoms for four or more weeks following a confirmed or suspected SARS-CoV-2 infection.[7,8] Analysis of patient questionnaires identified ongoing symptoms in 6.0% of symptomatic COVID-19 cases in Scotland,[9] 21.6–37.7% of positive cases in England,[10] and 7.8–17.0% of individuals with a history of COVID-19 in the UK.[11] The variation in these estimates reflects heterogeneous study designs, particularly in relation to the definition of long COVID used. Estimates derived from surveys or

questionnaires may also be subject to selection bias (for instance, if long COVID sufferers were over-represented among respondents) or information bias.

Others have analysed electronic health records (EHRs) to identify cases of long COVID. Findings show that long COVID is coded in EHRs considerably less frequently than is suggested by survey data. Analysis of English EHRs identified long COVID clinical codes in 0.04% of records (up to April 2021).[12] Analysis of American and German EHRs found explicit evidence of long COVID recorded in <0.01%[13] and 2.0%[14] of confirmed COVID-19 cases, respectively. This possible under-recording of long COVID in EHRs may reflect lack of presentation to clinical services, delays in clinical codes being made available at the beginning of the pandemic, clinicians' lack of familiarity with the codes, or hesitancy to code long COVID due to clinical uncertainty.[15]

In an effort to improve case identification, researchers have leveraged additional information recorded in EHRs to identify sequelae and phenotypes of long COVID. These studies analysed diagnostic codes[16–18] or free text entries[19] recorded in EHRs in the months following COVID-19 infection. To inform national policy and clinical deliberations, we combined these approaches, analysing clinical codes and free text data for the Scottish population to estimate the prevalence of long COVID. Others have used machine learning models trained on EHR data of individuals with explicit evidence of long COVID to predict potential cases of long COVID in the wider population.[14] Our analysis might complement this work by providing a method to identify a larger sample of patients with long COVID with which to train such models.

Methods

Study design, setting, and participants

The protocol describing this study was published in advance.[20] The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) checklist guided reporting.[21] It was not feasible to obtain consent from each of the 4,676,390 research participants; however, we were granted permission to access, within a secure trusted research environment, unconsented, whole-population, de-identified data from electronic health records for the purpose of surveillance during a public health emergency.

We estimated the prevalence of long COVID in the Scottish adult population by analysing data from EHRs hosted on the Early Pandemic Evaluation and Enhanced Surveillance of COVID-19 (EAVE II) platform. The EAVE II platform was established in response to the COVID-19 pandemic, to allow affiliated researchers access to pseudonymised, national-level data from primary care, secondary care, laboratory tests, and prescribing for all individuals registered with general medical practices (GPs) in Scotland (98–99% of the population[22]). The datasets hosted on the EAVE II platform (detailed in the Supplementary Materials) were curated by Scotland's national public health body, Public Health Scotland, which removed identifiable characteristics (such as date of birth and postcode) and provided pseudonymised identifiers to allow linkage of datasets.

In this study, we defined a cohort containing all adults (≥18 years) in the EAVE II platform who were resident in Scotland between March 1, 2020, and October 26, 2022 (n = 4,676,390) (Fig. 1). The study end-date was set at 26 weeks after the end of mass SARS-CoV-2 testing in Scotland to allow sufficient follow up of confirmed COVID-19 cases. We analysed data from the cohort to identify patterns in EHRs that were indicative of long COVID. This allowed us to create an operational definition to identify cases of long COVID, including where no long COVID clinical code had been recorded. We used the operational definition in conjunction with long COVID clinical codes and explicit references to long COVID in the free text of EHRs to estimate prevalence of long COVID within the cohort.

The EAVE II study obtained approvals from the West of Scotland Research Ethics Committee (reference: 22/WS/0071), and the Public Benefit and Privacy Panel for Health and Social Care (reference: 1920-0279).

Exposures

To allow us to identify potential indicators of long COVID for inclusion in our operational definition, we tested for significant differences in counts of several exposure variables recorded in the EHRs of patients with and without a confirmed SARS-CoV-2 infection in the 4–26 weeks following testing for SARS-CoV-2.

Primary care use

Clinical codes (in accordance with the coding system, Read Coded Clinical Terms version 2) were used to identify symptoms, clinical observations, and investigations recorded in primary care EHRs.[23] In accordance with data minimization requirements, we focused on a subset of 655 codes aggregated into 45 categories of indicators (Table S1). The selected codes represent a comprehensive range of potential long COVID indicators, extending previous work.[24]

Healthcare service use

We counted the number of times each individual in the cohort was seen by or had contact with seven areas of the healthcare system, as recorded in EHRs, including: GP visits where any of the 655 clinical codes were recorded, hospital admissions, outpatient attendances for respiratory conditions, A&E visits, out of hours encounters, admissions to Intensive Care Units (ICU), and NHS 24 telehealth interactions.

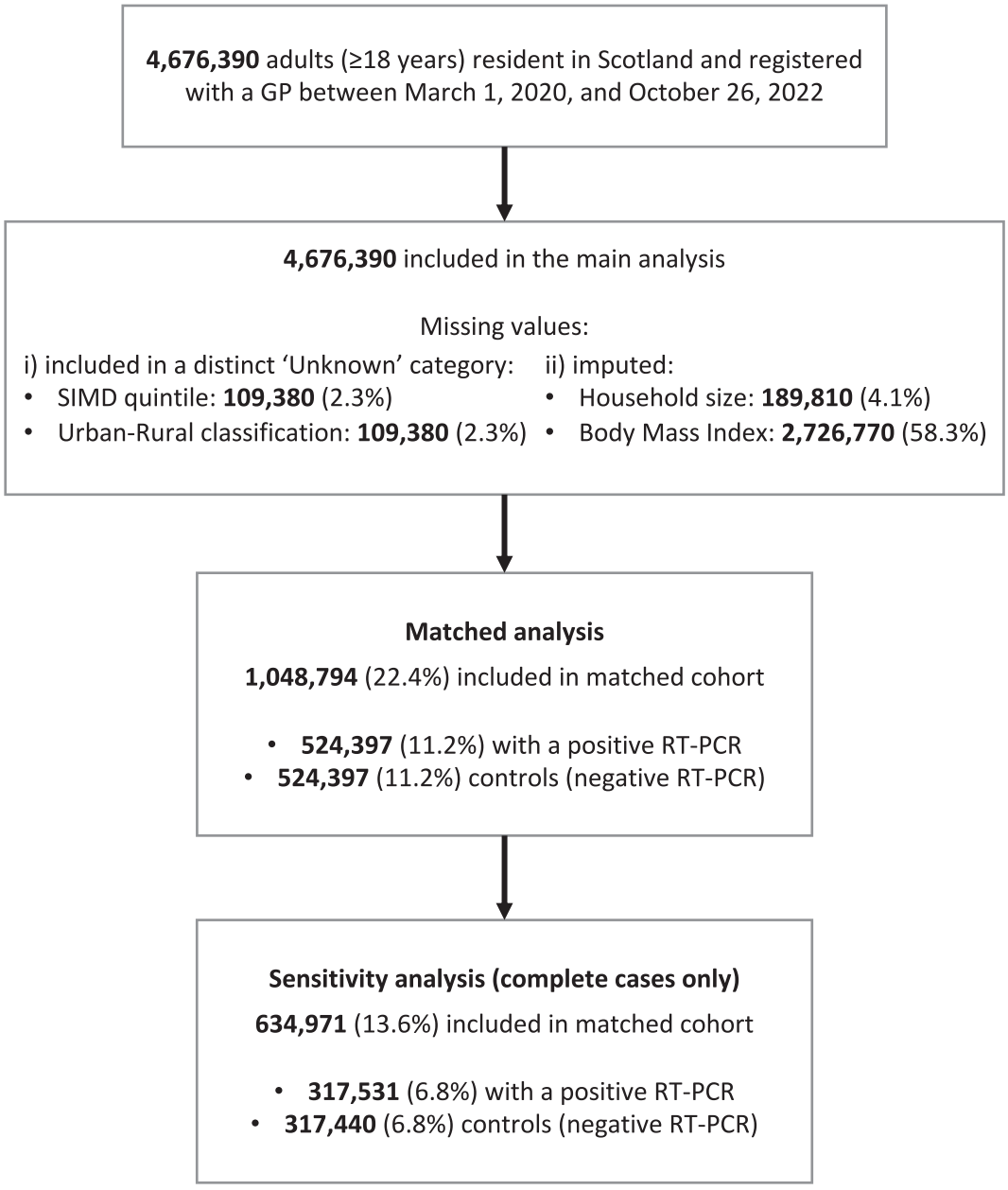

Fig. 1: **Participant inclusion**. The figure shows the number of participants included in the main analysis, the matched sample, and in sensitivity analysis that restricted the matched sample to respondents with no missing data. Percentages represent the share of the full cohort.

Prescribing data

We counted prescriptions dispensed in the community, which were automatically recorded in EHRs. To comply with data minimisation requirements and manage computational demands, we focused on 27 BNF sub-paragraphs (Table S2) representing 894 medicinal products deemed to be most relevant by the clinical and patient members of our research team. We restricted our analysis to new prescriptions that had not been dispensed during the 12 months prior to testing.

Patient characteristics

We report prevalence of long COVID within socio-demographic and clinical groups, including age, sex, Scottish Index of Multiple Deprivation (SIMD) quintiles, health boards (regional authorities with responsibility

for the delivery of health services), urban/rural residency, body mass index (BMI), vaccination status, and number of comorbidities.

Outcome measures

We used four outcome measures to identify cases of long COVID:

Operational definition

To allow us to identify cases of long COVID not explicitly recorded in EHRs, we developed an operational definition, which identified cases of long COVID based on patterns of clinical interactions recorded in EHRs.

Clinical codes

We used diagnostic codes for long COVID (Table S1), which were introduced in Scottish primary care on March 9, 2021, based on National Institute for health and Care Excellence (NICE) led working definitions of long COVID.[25,26]

Free text

We searched for terms indicative of long COVID recorded in the free text of primary care records.

Free text in sick notes

We searched for terms indicative of long COVID recorded in the free text of sick notes (also known as fitness to work certificates or "fit notes").

Potential sources of bias

Individuals who contracted COVID-19, but who did not receive a confirmatory reverse transcription polymerase chain reaction (RT-PCR) result would have been included in our control group. To minimise this potential bias, we required that controls had a negative RT-PCR test.

The operational definition, which identifies patients with long COVID based on patterns of clinical interactions recorded in EHRs, is contingent upon the systematic recording of clinical codes. However, substantial under-recording of clinical codes for long COVID symptoms[27] could lead to under-ascertainment of long COVID in our results. To mitigate this risk, we evaluated the operational definition in conjunction with the other measures of long COVID, as well as on its own.

The operational definition risks misclassifying individuals with pre-existing health conditions who test positive for SARS-CoV-2 as having long COVID. We conducted a series of analyses to evaluate the potential impact of such misclassification on our results.

Statistical analyses

Development of the operational definition of long COVID

The operational definition was derived using individual indicators of long COVID that were recorded at a significantly higher rate (adjusted-p <0.05) in the EHRs of individuals who had tested positive for COVID-19, relative to those who had tested negative, within 4–26 weeks of testing. Individuals with a specific combination of those indicators recorded in their EHRs within 4–26 weeks of receiving a positive RT-PCR or lateral flow test (LFT) result were identified by the operational definition as having long COVID, as described below. Additional details are in the Supplementary Methods (pp.S19-20).

To identify individual indicators of long COVID, we first matched individuals from the full cohort who had a positive RT-PCR test for COVID-19 (exposed group) to individuals with a negative RT-PCR test result (control group) in a 1:1 ratio based on estimates of their time-varying (by month) propensity to test positive (Equation S1). We used individuals with a negative RT-PCR test result as controls in order to minimise undocumented cases of COVID-19 in the control group, allowing us to better isolate the impact of COVID-19 on subsequent health outcomes. The resultant matched sample contained 54.3% (n = 524,397) of individuals from the full cohort who had a positive RT-PCR (all those for whom an appropriate match could be identified), and an equal number of controls (Fig. 1). Covariate balance plots (Figure S2) confirmed the adequacy of the matching.

We then used Poisson regression to estimate adjusted rate ratios (aRR) for each of our primary care use, healthcare service use, and prescribing indicators (shown in Figures S3 and S4) within the exposed group (relative to the control group) in the matched cohort. In each model, we included all predictors used in the propensity score estimation as covariates (Equation S1) and censored: controls who went on to test positive, positive cases who were reinfected, and all individuals who died. An offset term for the logarithm of follow-up days was included in each regression model. We adjusted p-values to reduce the false discovery rate, using the Benjamini-Hochberg correction.[28] Each of the primary care use, healthcare service use, or prescribing indicators that occurred at a significantly higher rate (adjusted-p <0.05) in the exposed group within either 4–12 weeks or >12–26 weeks following testing were taken to be indicative of long COVID (Figures S3–S10).

Next, we examined how the individual indicators clustered to form phenotypes (Figures S11 and S12 and Tables S3 and S4). However, we were unable to progress the cluster analysis as planned, due to sparse recording of clinical codes for symptoms and given that the clusters were strongly influenced by types of clinical codes; specifically, clinical codes indicating that blood tests had been requested in primary care clustered together. We therefore adopted a pragmatic alternative to clustering, informed by the clinical expertise of the project's steering group. We classified the indicators into three categories: symptoms, investigations, and management strategies (Fig. 2). Individuals in the full study cohort

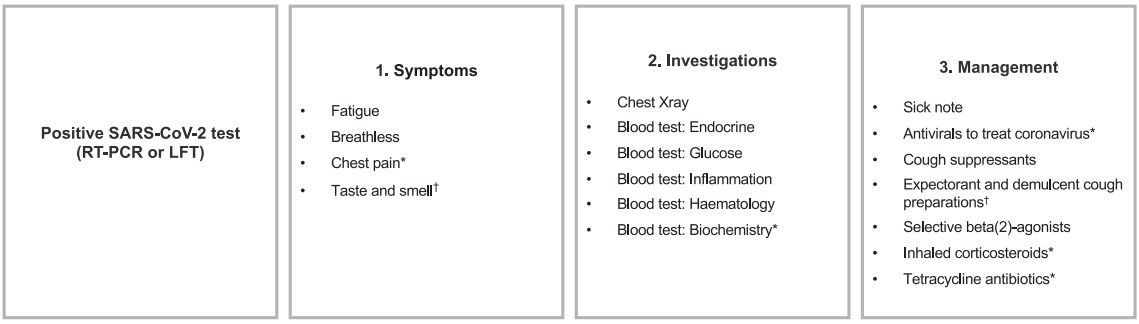

Fig. 2: **Operational definition of long COVID**. The operational definition classified individuals as having long COVID if they had both: (i) a positive RT-PCR or LFT result, and (ii) any outcome listed in two of the three categories (symptoms, investigations, management) recorded in their EHRs within 4–26 weeks of testing positive. * Observed at a significantly higher rate 4–12 weeks after testing only. † Observed at a significantly higher rate >12–26 weeks after testing only.

with a positive RT-PCR or LFT result who also had indicators in two or more of the three categories recorded in their EHRs during the 4–26 weeks following testing were identified as having long-COVID.

To assess the possibility that the operational definition was overfitted to individuals within the matched sample, we compared the proportion of patients identified by the operational definition and included in the matched sample to the proportion of patients identified by the operational definition, who had a positive RT-PCR test, but were not included in the matched sample.

Identification of long COVID in free text

We used natural language processing (NLP) to identify frequently occurring phrases in the free text fields of primary care records for the subset of individuals with long COVID clinical codes. We manually reviewed the 100 most frequently occurring phrases and identified the following as explicitly indicative of long COVID: "long covid", "post covid", "ongoing covid", "post coronavirus", "ongoing coronavirus" (including variations containing capitalisation or non-alphanumeric characters). We then used computer assisted coding to create a binary variable that indicated the presence or absence of any of those phrases in the free text of primary care records and on sick notes for each individual in the cohort. We examined the proportion of long COVID phrases that appeared within six words before or after a negation term and found this to be low (1.06%) (Table S5).

Descriptive statistics

Participant characteristics for the full sample, and stratified by outcome measures, were presented as counts and percentages, with continuous variables categorised in accordance with categories shown in Table 1.

Missing data

Missing values for descriptive statistics were reported using a distinct "Unknown" category. During the development of the operational definition (including propensity score estimation and in the Poisson regression models), missing BMI values were imputed using single imputation by chained equations. We considered using multiple imputation (which would have reduced the risk of bias); however, the approach would have been prohibitively time consuming, given the scale of our dataset and the complexity of our analysis. Household size, which was missing for a small minority of cases and likely missing at random, was mean imputed. To assess the impact of missing and imputed data, we re-ran our analyses omitting all cases with any imputed or "Unknown" variables.

Assessing the validity of the outcome measures

In the absence of a definitive 'ground truth' against which to assess the validity of each outcome measure, we described the congruence between measures, reporting the number of individuals identified as having long COVID by each outcome measure.

We conducted two further analyses to assess the possibility that the operational definition misclassified individuals with other health conditions as being patients with long COVID. Using a case-crossover design, we investigated the proportion of patients with long COVID that would have been identified by the operational definition (minus the requirement to have a positive SARS-CoV-2 test) based on information recorded in EHRs in the period leading up to each case's positive COVID-19 test result (we considered a lead up period equal to each individual's follow up period). Using logistic regression, we also estimated the odds ratio for the operational definition identifying long COVID in patients with a positive RT-PCR test, compared to matched negative controls (minus the requirement to have a positive RT-PCR or LFT result), while including an offset for the logarithm of days of follow up in the regression model.

Sensitivity analyses

During development of the operational definition, we repeated the matched analysis, stratified by time-periods

	Full sample		Any outcome		Long-COVID clinical code		Long-COVID in free text		Long-COVID on sick note		Operational definition	
	N	%	N	%	N	%	N	%	N	%	N	%
Total (% of sample)	4,676,390	100.0	81,219	1.7	1092	0.0	8368	0.2	14,469	0.3	64,193	1.4
Sex												
Female	2,359,166	50.4	52,851	65.1	698	63.9	5480	65.5	9947	68.7	41,597	64.8
Male	2,317,224	49.6	28,368	34.9	394	36.1	2888	34.5	4522	31.3	22,596	35.2
Age												
18–27	595,187	12.7	5991	7.4	64	5.9	532	6.4	708	4.9	4959	7.7
28–37	757,724	16.2	11,077	13.6	119	10.9	1080	12.9	2317	16.0	8432	13.1
38–47	729,889	15.6	15,060	18.5	246	22.5	1742	20.8	3498	24.2	11,196	17.4
48–57	801,896	17.1	19,333	23.8	310	28.4	2084	24.9	4532	31.3	14,573	22.7
58–67	757,380	16.2	17,349	21.4	222	20.3	1728	20.7	3228	22.3	13,834	21.6
68–77	577,749	12.4	7735	9.5	89	8.2	765	9.1	181	1.3	6931	10.8
78–87	334,459	7.2	3766	4.6	35	3.2	344	4.1	<5	–	3443	5.4
88–100	122,106	2.6	908	1.1	7	0.6	93	1.1	<5	–	825	1.3
Testing (RT-PCR or LFT)												
Positive	1,457,522	31.2	76,065	93.7	809	74.1	5700	68.1	12,141	83.9	64,193	100.0
Negative (and never positive)	1,591,959	34.0	3759	4.6	216	19.8	1817	21.7	1815	12.5	0.0	0.0
No tests recorded	1,626,909	34.8	1395	1.7	67	6.1	851	10.2	513	3.5	0.0	0.0
SIMD quintiles												
1–Most deprived	933,875	20.0	20,758	25.6	231	21.2	1645	19.7	4192	29.0	16,588	25.8
2	920,172	19.7	18,242	22.5	238	21.8	1836	21.9	3298	22.8	14,487	22.6
3	910,527	19.5	15,606	19.2	203	18.6	1788	21.4	2684	18.6	12,231	19.1
4	906,005	19.4	14,387	17.7	211	19.3	1818	21.7	2264	15.6	11,262	17.5
5–Least deprived	896,431	19.2	12,009	14.8	185	16.9	1232	14.7	1982	13.7	9510	14.8
Unknown	109,380	2.3	217	0.3	24	2.2	49	0.6	49	0.3	115	0.2
Urban-Rural												
Large urban areas	1,585,035	33.9	27,380	33.7	245	22.4	2001	23.9	5765	39.8	21,840	34.0
Other urban areas	1,694,708	36.2	33,047	40.7	564	51.6	3448	41.2	6109	42.2	25,767	40.1
Accessible small towns	424,007	9.1	6860	8.4	64	5.9	681	8.1	1192	8.2	5510	8.6
Remote small towns	217,586	4.7	3363	4.1	59	5.4	498	6.0	309	2.1	2727	4.2
Accessible rural	421,407	9.0	6332	7.8	93	8.5	769	9.2	800	5.5	5129	8.0
Remote rural	224,267	4.8	4020	4.9	43	3.9	922	11.0	245	1.7	3105	4.8
Unknown	109,380	2.3	217	0.3	24	2.2	49	0.6	49	0.3	115	0.2
Household size												
1	1,382,701	29.6	20,574	25.3	239	21.9	2115	25.3	3411	23.6	16,445	25.6
2	1,312,369	28.1	23,681	29.2	269	24.6	2473	29.6	3742	25.9	19,059	29.7
3–5	1,598,612	34.2	31,744	39.1	374	34.2	3283	39.2	6331	43.8	24,665	38.4
6–10	145,399	3.1	2404	3.0	35	3.2	241	2.9	408	2.8	1931	3.0
11+	47,499	1.0	388	0.5	<5	–	25	0.3	11	0.1	359	0.6
Unknown	189,810	4.1	2428	3.0	174	15.9	231	2.8	566	3.9	1734	2.7
BMI												
Underweight (BMI<18.5)	47,664	1.0	657	0.8	5	0.5	55	0.7	82	0.6	547	0.9
Normal weight (BMI 18.5 to <25.0)	526,232	11.3	8844	10.9	85	7.8	962	11.5	1220	8.4	7141	11.1
Overweight (BMI 25 to <30)	661,468	14.1	14,302	17.6	157	14.4	1410	16.8	1960	13.5	11,812	18.4
Obese (BMI ≥30)	714,256	15.3	22,832	28.1	248	22.7	2209	26.4	3259	22.5	19,037	29.7
Unknown	2,726,770	58.3	34,584	42.6	597	54.7	3732	44.6	7948	54.9	25,656	40.0
Comorbidities												
0	2,955,712	63.2	37,118	45.7	649	59.4	4305	51.4	8428	58.2	27,325	42.6
1	1,099,526	23.5	25,845	31.8	299	27.4	2590	31.0	4449	30.7	20,757	32.3
2	383,304	8.2	11,227	13.8	104	9.5	982	11.7	1271	8.8	9645	15.0
3+	201,000	4.3	5805	7.1	35	3.2	400	4.8	311	2.1	5314	8.3

(Table 1 continues on next page)

	Full sample		Any outcome		Long-COVID clinical code		Long-COVID in free text		Long-COVID on sick note		Operational definition	
	N	%	N	%	N	%	N	%	N	%	N	%
(Continued from previous page)												
Advised to shield												
Yes	159,495	3.4	6383	7.9	61	5.6	419	5.0	565	3.9	5679	8.8
No	4,516,895	96.6	74,836	92.1	1031	94.4	7949	95.0	13,904	96.1	58,514	91.2
Immunosuppressed												
Immunosuppressed	150,324	3.2	5579	6.9	54	4.9	350	4.2	580	4.0	4937	7.7
Not immunosuppressed	4,526,066	96.8	75,640	93.1	1038	95.1	8018	95.8	13,889	96.0	59,256	92.3
Vaccination doses												
0			19,763	24.3	275	25.2	2587	30.9	5259	36.3	14,720	22.9
1			4286	5.3	99	9.1	705	8.4	965	6.7	2937	4.6
2			22,290	27.4	271	24.8	1842	22.0	3761	26.0	18,278	28.5
3			33,139	40.8	412	37.7	2994	35.8	4375	30.2	26,859	41.8
4+			1741	2.1	35	3.2	240	2.9	109	0.8	1399	2.2
Variant period (positive cases only)												
Wild-type (01/03/2020–10/01/2021)	123,020	8.4	9733	12.8	160	19.8	1346	23.6	2978	24.5	7094	11.1
Alpha (11/01/2021–09/05/2021)	57,731	4.0	4641	6.1	86	10.6	505	8.9	1347	11.1	3477	5.4
Delta (24/05/2021–05/12/2021)	342,477	23.5	19,743	26.0	282	34.9	1578	27.7	3837	31.6	16,047	25.0
Omicron (27/12/2021–30/04/2022)	697,670	47.9	33,684	44.3	193	23.9	1717	30.1	3037	25.0	30,441	47.4
Severity of acute infection (positive cases only)												
Hospitalised within 28 days	47,930	3.3	6688	8.8	142	17.6	712	12.5	1122	9.2	5668	8.8
Not hospitalised within 28 days	1,409,592	96.7	69,377	91.2	667	82.4	4988	87.5	11,019	90.8	58,525	91.2

The table presents the number and percentage of individuals in each category indicated by the column headings. Percentages in the 'Total' row reflect the share of individuals in each category as a proportion of the total population. Statistics in the 'Variant period' and 'Severity of acute infection' rows relate to the subset of the population that has a positive LFT or RT-PCR test recorded in their EHRs. Population-level statistics for 'Vaccination doses (up to 14 days before positive test/outcome)' are intentionally omitted. Analysis of cases identified during the Omicron period was censored to align with the end of widespread RT-PCRT testing in Scotland on 30 April 2022. Cell counts <5 have been suppressed.

Table 1: **Patient characteristics stratified by outcome measures for long COVID.**

when different COVID-19 variants were dominant (we considered a variant to be dominant if it represented 60% or more of sequenced cases in a given week).[29] In an effort to ensure consistency in the share of individuals with the full 26 weeks of follow up data across variant periods, we required that individuals included in the Omicron period had tested positive by April 30, 2022, 26 weeks before the study end date. We also repeated the matched analysis using all individuals who had not yet taken an RT-PCR test up to the time of matching as controls.

R (version 3.6.1) was used for all analyses.

Patient and public involvement

Patient and public contributors were involved in the design and interpretation of this study (details in the Supplementary Materials, Tables S6 and S7).

Role of the funding source

The funder of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report. KJ, LD and AS verified all the data in the study and had final responsibility for the decision to submit the article for publication.

Results

Participants

Table 1 presents descriptive statistics. Of the 4,676,390 participants, 2,359,166 (50.4%) were female, the median age was 51 years (interquartile range: 35–65 years) and 1,457,522 (31.2%) had a positive LFT or RT-PCR test for SARS-CoV-2 during the study period (descriptive statistics stratified by testing status are presented in Table S8). 201,000 (4.3%) individuals had three or more comorbidities. 1,375,724 (70.6% of non-missing values) were categorised as overweight or obese. 150,324 (3.2%) were immunosuppressed and 159,495 (3.4%) had been advised to shield against COVID-19. Over the course of the study period, 4,007,666 (85.7%) individuals had any interaction with the healthcare system recorded in their EHRs (including primary care, secondary care, NHS 24 telehealth services, or RT-PCR or LFT testing), and 3,147,210 (67.3%) had one or more of the requested Read codes recorded in their EHRs. The proportion of missing data was most notable for BMI (58.3%), household composition (4.1%), urban-rural classification (2.3%), and SIMD (2.3%). Mean participant follow-up was 147.2 days, and the maximum possible was 154 days.

Main results

Prevalence of long COVID

Table 1 presents prevalence of long COVID by outcome measure. 81,219 (1.7%) individuals were identified as having long COVID using one or more outcome measure. Clinical codes identified the fewest patients with long COVID (n = 1,092, 0.02%). More individuals were identified by the long COVID terms in free text (n = 8,368, 0.2%) or on sick notes (n = 14,469, 0.3%). Most patients with long COVID were identified by the operational definition (n = 64,193, 1.7%) of which, 84.9% remained when the definition was restricted to indicators recorded >12–26 weeks after testing.

The incidence of long COVID over time

All outcome measures indicated a similar trend in the number of patients identified as having long COVID over time (Fig. 3). Panel A of Fig. 3 shows a general increase in new patients with long COVID between September 2020 and April 2022, interrupted by a temporary decline between February 2021 and May 2021. This was followed by a decline in newly recorded cases from April 2022 until the end of the study period, coinciding with the removal of mass testing in April 2022. Panel B of Fig. 3 shows that the share of people who tested positive for COVID-19 and went on to develop long COVID was greatest at the beginning of the study period, then oscillated but generally declined over time. The number of people identified by the operational definition as having long COVID, as a share of people who tested positive for COVID-19, declined sharply in the last six months of the study, reflecting increasingly truncated follow up times.

Characteristics of people with long COVID

Compared with the general population, each outcome measure identified higher prevalence of long COVID among females, those aged 38–67 years, obese individuals, and those with one or more comorbidity (Table 1). Individuals identified as having long COVID were also more likely to have a positive SARS-CoV-2 test, been advised to shield, be immunosuppressed, or to have been hospitalised or admitted to ICU within 28 days of testing positive, according to all measures. A disproportionately large percentage of patients who first tested positive for COVID-19 while the wild-type, Alpha, and Delta variants were dominant went on to develop long COVID, while a disproportionately small share of those who first tested positive during the Omicron period went on to develop long COVID. Compared with the general population, free text on sick notes identified fewer patients with long COVID aged over 68 years, likely reflecting lower employment levels among this group.

Relative to the general population, all outcome measures identified higher prevalence of long COVID among individuals with certain comorbidities, including asthma, severe mental illness, and rheumatoid arthritis or systemic lupus erythematous (SLE) (Table 2). The operational definition additionally identified higher prevalence of long COVID among individuals with type II diabetes mellitus, chronic obstructive pulmonary disease, and coronary heart disease. Relative to the other

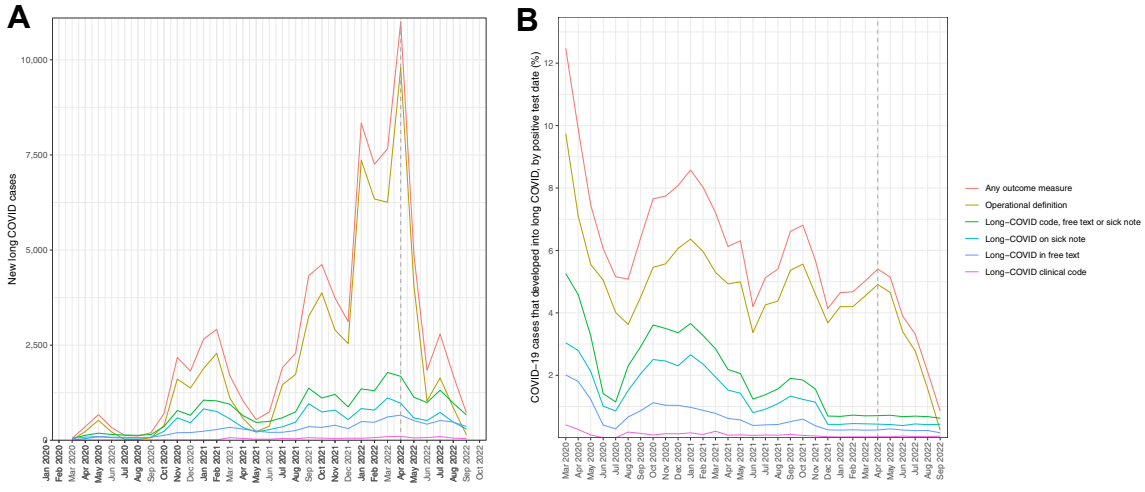

Fig. 3: **New patients with long COVID over time in absolute terms, and as a share of all patients with COVID-19**. Panel A presents the number of new cases of long COVID identified by each outcome measure in each month. Where an individual is identified by multiple outcome measures, only the first measure recorded is included in the aggregate measures. Cases of long COVID identified by the operational definition were dated from the date of the positive RT-PCR or LFT result, plus 28 days. All other indicators were counted according to the date they were recorded in EHRs. Panel B presents the percentage of COVID-19 cases that developed into long COVID, according to each outcome measure, by month of positive test date. Cases identified during October 2022 were omitted due to incomplete data for that month. The dashed line indicates the withdrawal of widespread LFT and RT-PCR testing in April 2022.

	Full sample		Any outcome		Long-COVID clinical code		Long-COVID in free text		Long-COVID on sick note		Operational definition	
	N	%	N	%	N	%	N	%	N	%	N	%
Total (% of population)	4,676,390	100.0	81,219	1.7	1092	0.0	8368	0.2	14,469	0.3	64,193	1.4
Atrial fibrillation	103,554	2.2	2152	2.6	7	0.6	165	2.0	90	0.6	1963	3.1
Asthma	559,662	12.0	17,896	22.0	172	15.8	1564	18.7	2540	17.6	14,958	23.3
Haematological cancer	21,075	0.5	559	0.7	<5	–	46	0.5	40	0.3	502	0.8
Heart failure	46,293	1.0	1117	1.4	<5	–	69	0.8	46	0.3	1034	1.6
Coronary heart disease	197,943	4.2	5043	6.2	36	3.3	379	4.5	262	1.8	4576	7.1
Chronic obstructive pulmonary disease (COPD)	130,527	2.8	4497	5.5	15	1.4	268	3.2	203	1.4	4169	6.5
Dementia	34,817	0.7	444	0.5	<5	–	25	0.3	5	0.0	418	0.7
Diabetes Type I	21,422	0.5	749	0.9	<5	–	43	0.5	92	0.6	669	1.0
Diabetes Type II	255,245	5.5	8328	10.3	63	5.8	552	6.6	691	4.8	7537	11.7
Epilepsy	63,718	1.4	1249	1.5	15	1.4	138	1.6	160	1.1	1026	1.6
Fracture	189,099	4.0	3610	4.4	36	3.3	327	3.9	545	3.8	2944	4.6
Neurological disorder	18,023	0.4	416	0.5	5	0.5	38	0.5	48	0.3	349	0.5
Parkinson's disease	9198	0.2	119	0.1	<5	–	8	0.1	<5	–	111	0.2
Pulmonary hypertension	8367	0.2	192	0.2	<5	–	14	0.2	<5	–	183	0.3
Rare pulmonary disease	22,445	0.5	726	0.9	5	0.5	64	0.8	39	0.3	650	1.0
Peripheral vascular disease	42,872	0.9	878	1.1	6	0.5	76	0.9	33	0.2	799	1.2
Rheumatoid arthritis or systemic lupus erythematous (SLE)	46,189	1.0	1577	1.9	13	1.2	100	1.2	146	1.0	1409	2.2
Respiratory cancer	10,465	0.2	213	0.3	<5	–	12	0.1	<5	–	200	0.3
Severe mental illness	532,744	11.4	15,383	18.9	188	17.2	1715	20.5	2415	16.7	12,444	19.4
Stroke/Transient Ischaemic Attack (TIA)	118,177	2.5	2469	3.0	16	1.5	194	2.3	132	0.9	2230	3.5
Thrombosis or pulmonary embolus	74,769	1.6	1782	2.2	19	1.7	173	2.1	184	1.3	1526	2.4
Chronic Kidney disease (level 3+)	155,810	3.3	3108	3.8	16	1.5	242	2.9	147	1.0	2823	4.4

The table presents the number and percentage of individuals in each category indicated by the column headings. Percentages in the 'Total' row reflect the share of individuals in each category as a proportion of the total population. Neurological disorder includes motor neurone disease, multiple sclerosis, myasthenia gravis and Huntington's chorea. Rare pulmonary disease includes cystic fibrosis, bronchiectasis or alveolitis. Severe mental illness includes bipolar affective disorder, psychosis, schizophrenia or schizoaffective disorder, and severe depression. Cell counts <5 have been suppressed.

Table 2: **Existing conditions stratified by outcome measures for long COVID.**

outcome measures, the operational definition generally identified a larger proportion of individuals with existing health conditions as having long COVID.

Geographic variation in prevalence of long COVID

To identify any geographic variation, we estimated prevalence of long COVID in each of Scotland's 14 health boards (regional authorities with responsibility for the delivery of health services). In each health board, the share of people with long COVID was proportionate to population shares, with the exception of a disproportionately large share in NHS Greater Glasgow and Clyde (27.5% compared with 21.9% of the general population) and a disproportionately small share in NHS Lothian (10.3% compared with 16.2% of the general population) (Table S9).

Congruence of the long COVID outcome measures

Table 3 shows the extent to which the outcome measures for long COVID co-occurred. The percentage of individuals with any outcome measure who also had at least one other outcome measure ranged from 9.8–42.8%. Of those with a long COVID clinical code, 34.1% were also identified by long COVID free text terms. However, there was no overlap between clinical codes and free text mentions of long COVID on sick notes.

Validity of the operational definition

Individual indicators included in the operational definition

The analysis underpinning the operational definition identified 17 indicators (four symptoms, six types of investigations, and seven types of management strategy) that occurred at a significantly higher rate among individuals with a positive RT-PCR test (relative to negative controls) in the 4–12 or >12–26 weeks after testing, shown in Fig. 2. Results from individual regressions are presented in Figures S3 and S4.

Each of the 17 indicators were observed at a higher rate among people identified by the other outcome measures as having long COVID, relative to all confirmed patients with COVID-19 (Table S10).

Case-crossover analysis

Of all patients identified by the operational definition as having long COVID, 98.4% (n = 63,189) had sufficient history to allow for a case-crossover analysis. Of those patients, 16.4% (n = 10,336) had information recorded in their EHRs in the lead-up to testing positive that was consistent with the criteria of the operational definition

	Full cohort	Long COVID clinical code	Long COVID in free text	Long COVID on sick note	Operational definition
N	4,676,390	1092	8368	14,469	64,193
Long-COVID clinical code (%)	0.0	100.0	4.5	0.0	0.3
Long-COVID in free text (%)	0.2	34.2	100.0	1.5	2.8
Long-COVID on sick note (%)	0.3	0.0	2.6	100.0	7.1
Operational definition (%)	1.4	19.3	21.6	31.3	100.0
Any of all other outcomes (%)	1.7	42.8	25.8	32.0	9.8

The table presents the percentage of individuals who meet the criteria for each outcome measure indicated by the row headings, as a share of the total number of individuals who meet the criteria for the outcome measure indicated by the column headings. The information presented is visualised in Figure S16.

Table 3: **Overlap in identification of long COVID across outcome measures.**

(minus the requirement to have a positive COVID-19 test result). Within this subset, 70.8% (n = 7319) had one or more existing health conditions, compared with 55.6% of all patients identified by the operational definition, and 36.0% of the full cohort. 7.8% (n = 818) of those identified based on data recorded in the lead up to testing positive were separately identified as having long COVID by one or more of the other outcome measures.

Identification of long COVID among matched controls

Within the matched sample, patients who had tested positive for COVID-19 were significantly ($p < 0.0001$) more likely to be identified as having long COVID by the operational definition, relative to negative controls (minus the operational definition's requirement to have a positive COVID-19 test result) (OR: 1.15, 95% CI 1.13–1.17).

Assessing the possibility of overfitting

The proportion of patients identified by the operational definition and included in the matched sample was similar to the proportion of patients identified by the operational definition, who had a positive RT-PCR test, but were not included in the matched sample (4.5% and 4.8%, respectively). This provides reassurance that the operational definition is unlikely to be overly influenced by the subset of patients with a positive RT-PCR test included in the matched sample.

Sensitivity analyses

Sensitivity analyses using controls who had neither a positive nor a negative RT-PCR test by the date of their positive match's test identified a larger set of individual indicators of long COVID, which included all 17 of those identified in the main analysis (Figures S5 and S6).

The individual indicators of long COVID identified were generally consistent during the wild-type, Alpha, and Delta periods, and fewer indicators were identified during the Omicron period (Figures S7–S10).

Excluding patients with imputed or "Unknown" data for any variable identified 11 of the 17 indicators of long COVID found in the main analysis (Figures S13–S15). However, this analysis was conducted using a smaller sample (n = 413,823 compared to n = 1,048,794 in the main analysis), inherently reducing statistical power. Odds ratios for the remaining six indicators were not significantly different from those estimated in the main analysis, suggesting general consistency of findings.

Discussion

Understanding the prevalence of long COVID is crucial for informing policy decisions on healthcare resource allocation, rehabilitation services, and research priorities. However, the multifaceted nature of long COVID and the lack of clearly defined diagnostic criteria make accurate estimation of long COVID prevalence challenging. To make progress on this important question, we analysed EHRs of almost the entire adult population of Scotland (n = 4,676,390) to estimate long COVID prevalence using four measures: long COVID clinical codes, free text recorded in GP records, free text on sick notes, and a novel operational definition.

Our analysis identified a prevalence of 0.02%–1.4%, depending on which of the four measures was used. The share of individuals identified by any one of the four measures was 1.7%. Long COVID clinical codes indicated a prevalence of just 0.02%, echoing findings from England[12] and confirming under-utilisation of the codes. Free text entries suggested higher prevalence, of 0.2% and 0.3% based on text recorded in primary care records and on sick notes (respectively), while the operational definition indicated prevalence of 1.4%.

Notably, only 42.8% of individuals identified by the long COVID clinical codes were also identified as having long COVID by any of the other measures. This likely reflects inconsistencies in the way that information was recorded in primary care. For instance, where a long COVID diagnosis was coded but symptoms were not, or where clinical codes or free text were recorded in isolation from one another. Despite this incongruence, we observed consistency in trends in new patients with long COVID over time and in patient characteristics across all four measures. The patient characteristics associated with long COVID also align with those identified in the literature.[9–12] This suggests the complementary nature of the four measures, underscoring the value of a multi-measure approach.

A key contribution of this study is the operational definition, which offers a novel method for improving the identification of conditions where diagnostic codes are unreliably recorded in EHRs. While lacking formal validation, the alignment of the operational definition with the other measures in terms of temporal trends and patient characteristics is encouraging. Notably, even when excluding the strict requirement for a positive SARS-CoV-2 test, the operational definition identified significantly more patients with long COVID among those with a positive COVID-19 test result relative to negative controls. However, we note that this was by a relatively modest margin, which may reflect bias from undocumented COVID-19 cases among controls.

The operational definition also had limitations. On one hand, it was a conservative measure, requiring both a documented positive SARS-CoV-2 test and presentation in general practice. Variation in patients' health-seeking behaviours and under-recording of symptoms[28] could cause conservative or biased identification of patients with long COVID. This possibility is indicated by the limited convergence between patients identified by long COVID clinical codes and the operational definition. Additionally, although the symptoms included in the operational definition were among those most prominently identified in the literature,[17–20] they do not reflect the full spectrum of COVID-19 sequelae. Symptoms with high baseline prevalence were less likely to be detected using our methodology due to lower discriminatory power between patients with a positive RT-PCR test result and controls where baseline prevalence was high. Moreover, the requirement that controls had a negative RT-PCR test may have caused over-representation of less healthy individuals within the control group. This likely reduced the number of indicators of long COVID identified, reducing the number of false positives at the cost of increasing the number of false negatives. As the operational definition is likely to under-estimate the number of patients with long COVID, estimates of prevalence should be interpreted carefully, and ideally evaluated alongside other measures.

Conversely, the operational definition may overestimate long COVID in individuals with existing health conditions, as suggested by higher prevalence of comorbidities among patients identified by the operational definition and by case crossover analysis. This limitation warrants careful consideration when interpreting our results and before applying the method to study other conditions. Despite this, we view the operational definition as a potentially valuable tool for long COVID identification. It complements existing identification approaches that may be subject to other biases, such as response bias in patient surveys. Future research should link data on patients' self-reported experiences of long COVID with EHR data to investigate overlap and discrepancies between approaches.

The scale and complexity of our dataset precluded the use of multiple imputation during development of the operational definition; however, sensitivity analysis suggested that this was unlikely to have biased our results.

With the removal of widespread testing, analysis of long COVID clinical codes and free text recorded in EHRs could be a promising (albeit conservative) approach for ongoing long COVID surveillance. Case detection using this approach could be improved by encouraging and enabling clinicians to accurately code long COVID symptoms and diagnoses. This might be achieved by communicating the value of coded data and by enhancing clinical software. In addition, enabling researchers to securely access free text data could improve long COVID detection by facilitating automated symptom and diagnosis extraction. Future work should explore the possibility of using such a multi-measure approach to enhance state-of-the-art machine learning models used for long COVID prediction,[14] by expanding the number of patients with long COVID available for model training.

Using data from virtually the entire adult population of Scotland, we estimated the prevalence of long COVID presenting in general practice between 0.02 and 1.7%. This likely underestimates the true prevalence, given incomplete health-seeking behaviour by patients, and under-recording of symptoms and long COVID diagnoses in primary care. In addition, among individuals with existing health conditions, it is possible that the operational definition overestimated the number of patients with long COVID. Using different measures to identify long COVID in EHR identified more patients with long COVID than a single measure would have. The multi-measure approach also highlighted inconsistencies in the way that routinely collected data is recorded in EHR and shows the challenges of using EHR data to estimate the prevalence of long COVID. We intend to utilise the measures of long COVID identification presented in this paper to develop a long COVID risk prediction model.[21]

Contributors

AS was the PI of this study. The manuscript was initially drafted by KJ and LD and further developed by the writing group. VH, CR, CS, LR, AS made substantial contributions to the acquisition of data. CR, LD, KJ, AS made substantial contributions to the conception and design of the work. AB, DW, CW, VH, LW carried out Patient and Public Involvement for the study. The analysis was carried out by KJ and code was checked by TM. KJ and TM have access to and verify the underlying study data. All authors contributed to data interpretation and critical review and revision of the manuscript.

Data sharing statement

All code, metadata and documentation for this project is publicly available at https://github.com/EAVE-II/Long-COVID. Most of the data used in this study are highly sensitive and will not be made available publicly.

Declaration of interests

AS reports grants from HDRUK, NIHR, MRC, ICSF, and CSO during the conduct of the study; and being a Member of the Scottish Government's CMO COVID-19 Advisory Group and Standing Committee on Pandemics. CR reports support from PHS, CSO and MRC; and being a Member of SPI-M, Scottish Government Scientific Advisory Committee, MHRA Covid vaccine benefit and risk expert working group. CS reports grants from MBIE (New Zealand), Ministry of Health (New Zealand), and HRC (New Zealand). JKQ reports grants from MRC, HDR UK, GlaxoSmithKline, BI, Asthma + Lung UK, and AstraZeneca and consulting fees from GlaxoSmithKline, Evidera, AstraZeneca, Insmed. SVK reports grants from CSO and MRC. All other authors declare no competing interests.

Acknowledgements

This work was supported by the Chief Scientist Office, grant number COV/LTE/20/15. EAVE II is supported by a grant (MC_PC_19075) from the Medical Research Council; and a grant (MC_PC_19004) from BREATHE–The Health Data Research Hub for Respiratory Health, funded through the UK Research and Innovation Industrial Strategy Challenge Fund. LD was supported by a post-doctoral clinical fellowship from the Asthma UK Centre for Applied Research. SVK acknowledges funding from a NRS Senior Clinical Fellowship (SCAF/15/02), the Medical Research Council (MC_UU_00022/2) and the Scottish Government Chief Scientist Office (SPHSU17). The authors would like to acknowledge the support of Dave Kelly and Lamorna Brown of Albasoft Ltd., and Sharon Kennedy, Mike Birnie, Safraj Shahul Hameed and Elliott Hall of Public Health Scotland for their involvement in obtaining approvals, provisioning, and linking data and the use of the secure analytical platform within the National Safe Haven.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.eclinm.2024.102590.

References

1 Davis HE, McCorkell L, Vogel JM, Topol EJ. Long COVID: major findings, mechanisms and recommendations. *Nat Rev Microbiol*. 2023;21:133–146.
2 Soriano JB, Murthy S, Marshall JC, Relan P, Diaz JV, Group WC. A clinical case definition of post-COVID-19 condition by a Delphi consensus. *Lancet Infect Dis*. 2022;22(4):e102–e107.
3 Han Q, Zheng B, Daines L, Sheikh A. Long-Term sequelae of COVID-19: a systematic review and meta-analysis of one-year follow-up studies on post-COVID symptoms. *Pathogens*. 2022;11(2):269.
4 Xie Y, Al-Aly Z. Risks and burdens of incident diabetes in long COVID: a cohort study. *Lancet Diabetes Endocrinol*. 2022;10(5):311–321.
5 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.
6 Larsen NW, Stiles LE, Miglis MG. Preparing for the long-haul: autonomic complications of COVID-19. *Auton Neurosci*. 2021;235:102841.
7 Office for National Statistics. Prevalence of ongoing symptoms following coronavirus (COVID-19) infection in the UK: 30 March 2023. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/bulletins/prevalenceofongoingsymptomsfollowingcoronaviruscovid19infectionintheuk/2february2023. Accessed January 17, 2024.
8 Birtwistle S, Deakin E, Whitford R, Hinchliffe S, Daniels-Creasey A, Rule S. *The scottish health survey*. 2021 edition. vol. 1. Main Report. Scottish Government; 2022.
9 Hastie CE, Lowe DJ, McAuley A, et al. Outcomes among confirmed cases and a matched comparison group in the Long-COVID in Scotland study. *Nat Commun*. 2022;13(1):5663.
10 Whitaker M, Elliott J, Chadeau-Hyam M, et al. Persistent COVID-19 symptoms in a community study of 606,434 people in England. *Nat Commun*. 2022;13(1):1957.
11 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.
12 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–e814.
13 Pfaff ER, Girvin AT, Bennett TD, et al. Identifying who has long COVID in the USA: a machine learning approach using N3C data. *Lancet Digit Health*. 2022;4(7):e532–e541.
14 Kessler R, Philipp J, Wilfer J, Kostev K. Predictive attributes for developing long COVID—a study using machine learning and real-world data from primary care physicians in Germany. *J Clin Med*. 2023;12(10):3511.
15 Kingstone T, Taylor AK, O'Donnell CA, et al. Finding the 'right' GP: a qualitative study of the experiences of people with long-COVID. *BJGP Open*. 2020;4(5):bjgpopen20X101143. https://doi.org/10.3399/bjgpopen20X101143.
16 Taquet M, Dercon Q, Luciano S, et al. Incidence, co-occurrence, and evolution of long-COVID features: a 6-month retrospective cohort study of 273,618 survivors of COVID-19. *PLoS Med*. 2021;18(9):e1003773.
17 Xie Y, Bowe B, Al-Aly Z. Burdens of post-acute sequelae of COVID-19 by severity of acute infection, demographics and health status. *Nat Commun*. 2021;12(1):6571.
18 Zhang H, Zang C, Xu Z, et al. Data-driven identification of post-acute SARS-CoV-2 infection subphenotypes. *Nat Med*. 2023;29:226–235. https://doi.org/10.1038/s41591-022-02116-3.
19 Wang L, Foer D, MacPhaul E, et al. PASCLex: a comprehensive post-acute sequelae of COVID-19 (PASC) symptom lexicon derived from electronic health record clinical notes. *J Biomed Inf*. 2022;125:103951.
20 Daines L, Mulholland RH, Vasileiou E, et al. Deriving and validating a risk prediction model for long COVID-19: protocol for an observational cohort study using linked Scottish data. *BMJ Open*. 2022;12(7):e059385.
21 Von Elm E, Altman DG, Egger M, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *Ann Intern Med*. 2007;147(8):573–577.
22 Mulholland RH, Vasileiou E, Simpson CR, et al. Cohort profile: early pandemic evaluation and enhanced surveillance of COVID-19 (EAVE II) database. *Int J Epidemiol*. 2021;50(4):1064–1074.
23 Public Health Scotland. Grouping of codes for conditions (RCGs). https://www.isdscotland.org/Health-Topics/General-Practice/GP-Consultations/Grouping-clinical-codes.asp. Accessed February 28, 2023.
24 Whittaker HR, Gulea C, Koteci A, et al. GP consultation rates for sequelae after acute covid-19 in patients managed in the community or hospital in the UK: population based study. *BMJ*. 2021;375:e065834.
25 Scottish Government. *Management and recording of the long-term effects of COVID-19 (Long COVID)*; 2020. Available from: https://www.scimp.scot.nhs.uk/wp-content/uploads/CMO-Letter-09.03.21.pdf. Accessed February 28, 2023.
26 National Institute for Health and Care Excellence. *COVID-19 rapid guideline: managing the long-term effects of COVID-19*; 2020. Last updated: 11 November 2021. Available from: https://www.nice.org.uk/guidance/ng188. Accessed February 28, 2023.
27 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. https://doi.org/10.1371/journal.pone.0290583.
28 Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Roy Stat Soc B*. 1995;57(1):289–300.
29 Magnusson K, Kristoffersen DT, Dell'Isola A, et al. Post-covid medical complaints following infection with SARS-CoV-2 Omicron vs Delta variants. *Nat Commun*. 2022;13(1):1–9.