Risk factors for long COVID: analyses of 10 longitudinal studies and electronic health records in the UK

Ellen J. Thompson[1]†^*, Dylan M. Williams[2,3] †^*, Alex J. Walker [4]†^, Ruth E. Mitchell[5,6]^, Claire L. Niedzwiedz[7]^, Tiffany C. Yang[8]^, Charlotte F. Huggins[9]^, Alex S. F. Kwong[5,10]^, Richard J. Silverwood[11], Giorgio Di Gessa[12], Ruth C.E. Bowyer[1], Kate Northstone[6], Bo Hou[8], Michael J. Green[13], Brian Dodgeon[11], Katie J. Doores[14], Emma L. Duncan[1], Frances Williams[1], OpenSAFELY Collaborative, Andrew Steptoe[12], David J. Porteous[9], Rosemary R. C. McEachan[8], Laurie Tomlinson[15], Ben Goldacre[4], Praveetha Patalay[2,11], George B. Ploubidis[11], Srinivasa Vittal Katikireddi[13], Kate Tilling[5], Christopher T. Rentsch[15,16], Nicholas J Timpson[5,6], Nishi Chaturvedi[2], Claire J. Steves[1]*

† Joint first

^ Lead analyst team

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

[2] MRC Unit for Lifelong Health and Ageing at UCL, University College London

[3] Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden

[4] The DataLab, Nuffield Department of Primary Care Health Sciences, University of Oxford

[5]MRC Integrative Epidemiology Unit at the University of Bristol, United Kingdom

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

[7]Institute of Health & Wellbeing, University of Glasgow

[8]Bradford Institute for Health Research, Bradford Teaching Hospitals NHS Foundation Trust, Bradford BD9 6RJ, UK

[9]Centre for Genomic and Experimental Medicine, Institute of Genetics and Cancer, University of Edinburgh

[10]Division of Psychiatry, University of Edinburgh, UK

[11]Centre for Longitudinal Studies, UCL Social Research Institute, University College London

[12]Department of Epidemiology and Public Health, University College London

[13]MRC/CSO Social & Public Health Sciences Unit, University of Glasgow

[14]School of Immunology & Microbial Sciences, King's College London

[15]London School of Hygiene & Tropical Medicine, London, UK.

[16]VA Connecticut Healthcare System, West Haven, Connecticut, USA.

*Correspondence to: Claire J. Steves (claire.j.steves@kcl.ac.uk), Ellen J. Thompson (ellen.thompson@kcl.ac.uk), Dylan M. Williams (dylan.williams@ucl.ac.uk).

Abstract

The impact of long COVID is increasingly recognised, but risk factors are poorly characterised. We analysed questionnaire data on symptom duration from 10 longitudinal study (LS) samples and electronic healthcare records (EHR) to investigate sociodemographic and health risk factors associated with long COVID, as part of the UK National Core Study for Longitudinal Health and Wellbeing.

Methods

Analysis was conducted on 6,899 adults self-reporting COVID-19 from 45,096 participants of the UK LS, and on 3,327 cases assigned a long COVID code in primary care EHR out of 1,199,812 adults diagnosed with acute COVID-19. In LS, we derived two outcomes: symptoms lasting 4+ weeks and symptoms lasting 12+ weeks. Associations of potential risk factors (age, sex, ethnicity, socioeconomic factors, smoking, general and mental health, overweight/obesity, diabetes, hypertension, hypercholesterolaemia, and asthma) with these two outcomes were assessed, using logistic regression, with meta-analyses of findings presented alongside equivalent results from EHR analyses.

Results

Functionally limiting long COVID for 12+ weeks affected between 1.2% (age 20), and 4.8% (age 63) of people reporting COVID-19 in LS. The proportion reporting symptoms overall for 12+ weeks ranged from 7.8 (mean age 28) to 17% (mean age 58) and for 4+ weeks 4.2% (age 20) to 33.1% (age 56). Age was associated with a linear increase in long COVID between age 20-70. Being female (LS: OR=1.49; 95%CI:1.24-1.79; EHR: OR=1.51 [1.41-1.61]), poor pre-pandemic mental health (LS: OR=1.46 [1.17-1.83]; EHR: OR=1.57 [1.47-1.68]) and poor general health (LS: OR=1.62 [1.25-2.09]; EHR: OR=1.26; [1.18-1.35]) were associated with higher risk of long COVID. Individuals with asthma also had higher risk (LS: OR=1.32 [1.07-1.62]; EHR: OR=1.56 [1.46-1.67]), as did those categorised as overweight or obese (LS: OR=1.25 [1.01-1.55]; EHR: OR=1.31 [1.21-1.42]) though associations for symptoms lasting 12+ weeks were less pronounced. Non-white ethnic minority groups had lower 4+ week symptom risk (LS: OR=0.32 [0.22-0.47]), a finding consistent in EHR. Associations were not observed for other risk factors. Few participants in the studies had been admitted to hospital (0.8-5.2%).

Conclusions

Long COVID is clearly distributed differentially according to several sociodemographic and pre-existing health factors. Establishing which of these risk factors are causal and predisposing is necessary to further inform strategies for preventing and treating long COVID.

Introduction

SARS-CoV-2 infection can lead to sustained or recurrent multi-organ symptoms in some individuals.[1–3] Extended COVID-19 symptomatology over weeks to months has been defined by individuals as 'long COVID'.[4] More formally, the UK's National Institute for Health Care and Excellence defined acute COVID-19 (AC; lasting <4 weeks), ongoing symptomatic COVID-19 (OSC; 4-12 weeks), and post-COVID-19 syndrome (PCS; >12 weeks), with the latter two categories combined as 'long COVID'.[1] Estimates of long COVID prevalence range from 13.3% in highly selected, community-based survey respondents with test-confirmed COVID-19, to at least 71% among those hospitalised by the infection.[5–7] Given the scale of the pandemic, even a low proportion of individuals with long COVID will generate a major burden of lingering illness.[8]

In order to target appropriate support and focus research on possible causal mechanisms, we first need to understand risk factors for the disease. Current understanding of frequency of, and risk factors for, long COVID remains poor, impeding mechanistic investigation for intervention development and constraining service planning. Obtaining accurate estimates of association and risk requires large generalisable samples with comprehensive measures of pre-morbid health. UK national primary care records, which cover >95% of the population, afford one such data source, but are limited to those who consult with symptoms and depend on diagnosis and recording of long COVID. Furthermore, risk factor and co-morbidity data are limited to those who consult and are tested. Population-based longitudinal studies (LS), established decades before the pandemic, overcome these limitations as they collect data from all participants agnostically, regardless of healthcare attendance, with detailed and, where possible, objective measures of pre-pandemic health. The limitation of LS is that they are relatively small and separately may yield imprecise estimates. Combined analysis of primary care records and LS provides a powerful tool to compensate for their different limitations and biases.

This work aimed to satisfy the clinical and policy need to better understand factors reliably associated with OSC and PCS (long COVID). To do this, we identified individuals with these specifically pre-defined COVID-19 outcomes in: 1) a consortium of population-based LS which captured coordinated repeat questionnaire data on COVID-19 using harmonised measures from the Wellcome Trust's Covid-19 Questionnaire, and 2) the OpenSAFELY dataset of primary care records (https://www.opensafely.org/). Within these data sets, we examined the frequency of long COVID among individuals with suspected and test-confirmed COVID-19 and examined associations of sociodemographic and pre-pandemic health risk factors.

Methods

Design

The UK National Core Studies – Longitudinal Health and Wellbeing programme draws together data from multiple UK population-based LS and electronic health records (EHR) to answer questions relevant to the pandemic. We coordinated analyses within each LS, then pooled results statistically to provide more robust estimates and to identify explanations for between-LS heterogeneity. Parallel coordinated investigation in EHR enabled comparison of population-based findings with those in individuals who sought healthcare.

Sample: Longitudinal Studies

Data were drawn from 10 UK LS that had conducted surveys before and during the COVID-19 pandemic (ten samples were yielded in total as one parent-offspring cohort was split into two samples by generation). These included five age-homogenous samples: the Millennium Cohort Study (MCS);[9] the Avon Longitudinal Study of Parents and Children (ALSPAC (generation 1, "G1"));[10] Next Steps (NS);[11] the 1970 British Cohort Study (BCS);[12], and the National Child Development Study (NCDS).[13,14] Five further age-heterogeneous samples (each covering a range of age groups) were included: the Born in Bradford study (BIB);[15,16] Understanding Society (USOC);[17] Generation Scotland: the Scottish Family Health Study (GS);[18] the parents of the ALSPAC-G1 cohort, whom we refer to as ALSPAC-G0;[19] and the UK Adult Twin Registry (TwinsUK);[20,21]. Details of design, sample frames, current age range, timing of the most recent pre-pandemic and COVID-19 surveys, and analytical sample sizes are shown in Supplementary Table 1. Studies were selected to allow derivation of harmonised measures of long COVID and for prospectively collected measures of health pre-pandemic. Minimum inclusion criteria included self-reported COVID-19, self-reported duration of COVID-19 symptoms, and age, sex, and ethnicity.

Measures: LS

COVID-19 case definition (self-report)

Cases of COVID-19 were defined as those who self-reported COVID-19. Information to substantiate case definition included testing and health care professional confirmation (see Supplementary File 1 for full details of the questions and coding used within each study).

Long COVID definitions

Self-reported symptom length

Long COVID was defined using guidelines developed jointly by NICE, the Scottish Intercollegiate Guidelines Network (SIGN) and the Royal College of General Practitioners (RCGP).[1] Most of the LS questionnaires specifically reflected the categories used by these guidelines, asking respondents to self-report duration of symptoms with categories that could be designated as either AC, OSC or PCS. Based on these categories, we defined two primary outcomes: i) durations lasting 4+ weeks (combining OSC and PCS), with individuals reporting symptoms lasting 0-4 weeks as reference, and ii) 12+ weeks of symptoms (PCS specifically), with reference being individuals with symptoms lasting 0-12 weeks. Full details of the questions and coding are available in Supplementary file 1. In addition, two studies derived an alternative estimate of long COVID based on whether symptoms were present for more than 4 or 12 weeks in total over at least six months (BiB, TwinsUK). BiB study members who self-reported COVID-19 were asked to report whether any particular symptoms (27 in total) were present during March-September 2020. Three of these symptoms (runny nose, sneezing and blocked nose) were considered non-specific for COVID-19 and were removed. Data were used to derive symptom length categories above by summing included symptoms present for 0-4 weeks; 4-12 weeks or 12+ weeks. In TwinsUK, all study members were asked to report whether they had experienced particular symptoms (33 in total) between February and November 2020. Five symptoms (runny nose, sneezing, blocked nose, shaking or difficulty while walking, and phlegm production/chesty cough) were considered non-specific for COVID-19 and were removed. Similar to BIB, data were used to derive the symptom length categories above through summing whether any of the included symptoms were present for 0-4 weeks; 4-12 weeks or 12+ weeks, at any point in time over the specified period. This was performed for people who had had COVID-19 and those who had not (confirmed by negative antibody testing).

Exposures

Pre-pandemic risk factors were restricted to measures more than 6 months but less than 5 years prior to 23rd March 2020 wherever possible. Full details of the questions and coding are available in Supplementary file 1**.**

Sociodemographic factors.

These included: age, sex (female/male), ethnicity (white, non-white ethnic minority; in studies where possible), and socioeconomic position measured by highest education levels (degree, no degree), Index of Multiple Deprivation (IMD, a widely used geographical based measure of relative deprivation based on factors such as income, employment and education), and occupational class of own current/recent job (or parental occupational class for younger cohorts; four categories: managerial/professional; intermediate; routine; or not working/not available).

Mental health.

Mental health was captured in the most recent pre-pandemic survey using validated continuous scales of psychological distress that assessed symptoms of common mental health difficulties such as anxiety and depression (e.g., Hospital Anxiety and Depression scale, TwinsUK; Short Mood and Feelings Questionnaire, ALSPAC-G1; Edinburgh Postnatal Depression Scale, ALSPAC-G0; General Health Questionnaire-12, USOC). For analyses, each scale was transformed into standard deviation units (z-scores) within each study, and a dichotomous variable was derived using established cut-offs for each measure (see Supplementary file 1 for details).

Self-rated health.

All study participants had been asked about their general health prior to the pandemic using 5 categories (1= Excellent; 2 = Very good; 3 = Good; 4 = Fair; 5 = Poor). A binary variable was derived for self-reported health, grouping those with excellent-good health (categories 1-3) and those with fair-poor health (categories 4-5).

Health conditions.

Body mass index (BMI = weight [kg]/(height [m]^2)) of study participants was obtained prior to the pandemic. For analyses a binary weight variable were categorised as those who had a BMI between 0-24.9 (underweight/normal weight) and those who had a BMI of 25 or more (overweight/obese). Pre-pandemic asthma, diabetes, hypertension, and high cholesterol status was captured through self-report.

Statistical analysis

Main analyses were conducted in studies with a direct self-reported measure of COVID-19 symptom length. Using this measure, two separate binary variables were created. The first grouped those who had symptoms from 0-4 weeks (reference group) and those who had symptoms for 4+ weeks (long COVID). The second grouped those who had had symptoms from 0-12 weeks (reference group) and those who had symptoms for 12+ weeks (post-COVID-19 syndrome). The association between each sociodemographic or pre-pandemic health risk factor and each long COVID outcome was assessed in separate multivariable logistic regression models within each study. We adjusted for a minimal set of confounders across all studies, where relevant: age (as a continuous variable), sex, and ethnicity. Odds ratios (ORs) and 95% confidence intervals (CIs) were the main measure of association.

We modelled the relationship of age with long COVID risk in two ways, given that there were diverse age structures between studies. First, in age-heterogeneous samples, we analysed long COVID within age categories relative to pre-defined baseline groups, given an *a priori* rationale that association between age and long COVID may not be linear. Categories within each study are shown in

Supplementary figures 1 and 2. Second, in a subset of LS birth cohorts with participants of near-identical ages and who were issued fully harmonised long COVID questionnaires (MCS, NS, BCS70 and NCDS), we analysed the trend in absolute risk of long COVID with increasing age between studies using meta-regression.

Data synthesis

To synthesise effect sizes across studies, fixed-effect meta-analyses with restricted maximum likelihood were carried out and repeated with random-effects modelling for comparison. We report heterogeneity using the I^2 statistic to examine the percentage of variability in effect estimates accounted for by heterogeneity rather than sampling error (0% indicates no variation between estimates across studies; values closer to 100% indicate greater heterogeneity).

Attrition and design weights

Selective attrition in LS (compounded by conventional problems of self-selection) can affect the representativeness of retained samples and introduce bias in estimates of association. To address this, where possible, most studies were weighted to be representative of their target population. This attempted to account for survey design and differential non-response to the COVID-19 surveys. Weights were not available for GS, BiB or TwinsUK.

Sensitivity analysis

To mitigate index response bias,[22] inverse probability weights (IPW) were derived for COVID-19 status. These were derived in each LS separately but following a common approach used previously.[14] Self-reported COVID-19 status was regressed on each exposure to assess whether COVID-19 was associated with each socio-demographic or pre-pandemic health risk factor. To determine what variables to include across LS, observed associations were meta-analysed to identify consistent predictors of COVID-19 self-report status (see Supplementary Information 2 for list of covariates used to derive IPWs). To avoid missingness on IPWs, covariates included in each model were imputed using multiple imputation by chained equations (MICE) and IPWs were derived across multiple imputed data sets. Derived weights were then applied in all analysis models as a sensitivity check.

For studies in which we were able to verify SARS-CoV-2 infection through collected serology data in summer/autumn 2020 (TwinsUK and ALSPAC-G0 and -G1), analyses were replicated on a sub-sample of those who had positive polymerase chain reaction (PCR) obtained through linkage to testing data and/or lateral flow antibody testing (ALSPAC) and enzyme-linked immunosorbent assay (ELISA) (TwinsUK)[23] confirming exposure to COVID-19. All statistical analyses on the LS were performed in Stata version 16 or R (release 3.6.0 or later).

Data sources: EHR

Working on behalf of NHS England, we conducted a population-based cohort study to measure long COVID recording in electronic health record (EHR) data from primary care practices using TPP SystmOne software, linked to Secondary Uses Service (SUS) data (containing hospital records) through OpenSAFELY. This is a data analysis platform developed during the COVID-19 pandemic, on behalf of NHS England, which includes all practices using TPP SystmOne software linked to Secondary Uses Service (SUS) data (containing hospital records) to allow near real-time analysis of pseudonymised primary care records at scale, operating within the EHR vendor's highly secure data environment. Details on Information Governance for the OpenSAFELY platform can be found in the Supplementary Information 1.

Sample: EHR

From a population of all people alive and registered with a general practice on 1 December 2020, we selected all patients who had evidence of a COVID related code, either: testing positive for SARS-CoV-2, being hospitalised with an associated COVID diagnostic code, or having a recorded diagnostic code for COVID in primary care.

Outcome: EHR

The outcome was any record of long COVID in the primary care record, as a binary variable. This was defined using a list of 15 UK SNOMED codes, which are categorised as diagnostic (2 codes), referral (3) and assessment (10) codes. SNOMED is an international structured clinical coding system for use in electronic health records. The outcome was measured between the study start date (2020-02-01) and the end date (2021-05-09).

Measures: EHR

Sociodemographic variables

Demographic variables included age (in categories), sex, geographic region, IMD (divided into quintiles), and ethnicity. Details on coding have been reported elsewhere (see: Mathur et al. 2021).[24]

BMI.

People were categorised as not obese or obese using their most recent BMI measurement, with those in the obese category further categorised into Obese I (BMI 30-34.9), Obese II (BMI 35-39.9), or Obese III (BMI 40+). Those with a missing BMI were assumed to be not obese.

Health conditions.

A previous code six months to five years before March 2020, for one or more of: diabetes; cancer; haematological cancer; asthma; chronic respiratory disease; chronic cardiac disease; chronic liver disease; stroke or dementia; other neurological condition; organ transplant; dysplasia; rheumatoid arthritis, systemic lupus erythematosus or psoriasis; or other immunosuppressive conditions. Those with no relevant code for comorbidities were assumed not to have that condition. Number of comorbidities was categorised into "0", "1", and "2 or more".

Mental health.

Evidence of a pre-existing mental health condition was defined using a prior code for one of: psychosis; schizophrenia; bipolar disorder; or depression.

Statistical methods: EHR

The number of people with or without a long COVID code was recorded amongst the selected sample and stratified by each of the measures. The proportion of people with long COVID codes was calculated overall and within each code category. The percentage of long COVID events across the different measure categories was also reported.

We conducted multivariable logistic regression to assess whether GP-recorded long COVID was associated with each sociodemographic or pre-pandemic health characteristic. We adjusted for the same set of confounders as used in the LS analyses: age (as categorical variable), sex, ethnicity. Odd ratios (ORs) and 95% confidence intervals (CIs) were again the main measure of association.

In further analyses of age as a risk factor for long COVID in the EHR data, we assigned individuals within 10-year categories an age at the midpoint of each group, then assessed the trend in long COVID frequency with age using linear and non-linear meta-regression.

All code for the OpenSAFELY platform for data management, analysis and secure code execution is shared for review and re-use under open licenses at https://github.com/opensafely. Codelists describing the definition of all the above conditions can be found at: https://github.com/opensafely/long-covid-historical-health/tree/main/codelists All code for data management and analysis for this paper is shared for scientific review and re-use under open licenses on GitHub https://github.com/opensafely/long-covid-historical-health

Results

Longitudinal studies

Of 45,096 individuals surveyed in LS, 6866 (15.2%) self-reported suspected or confirmed COVID-19. Within cases, the percentage of females ranged from 55% (NCDS) to 96% (BiB) and the mean age across studies ranged from 19.9 years (MCS) to 63.0 years (NCDS) (**Table 1)**. Ethnicity differed across LS, with members identifying as 'White' ranging from 43.8% in BiB to 98.4% for ALSPAC G0. The percentage of those with a degree ranged from 7.5% (BiB) to 49.5% (TwinsUK). Within each LS, most participants managed their illness at home and were not admitted to hospital (range 0.8%-5.2%, see Table 1). Descriptives for the cases and base sample populations are provided in Supplementary Table 1.

In studies ascertaining long COVID of any functional severity, between 7.8% (ALSPAC G1) and 17.0% (ALSPAC G0) of self-reported COVID cases reported symptoms they attributed to COVID-19 for 12+ weeks (PCS). Between 14.5% (ALSPAC G1) and 18.7% (TwinsUK) reported symptoms for 4-12 weeks (OSC) (Table 2). Figures varied considerably within LS comparing self-reported confirmed and suspected cases (Supplementary Table 2). However, in ALSPAC and TwinsUK, both with SARS-CoV-2 antibody testing, the frequency of PCS and OSC were broadly similar (TwinsUK PCS 20%, OSC 20%; ALSPAC G0 PCS 14%, OSC 8.8%; ALSPAC G1 PCS 11%, OSC 11% respectively, Supplementary Table 3). In studies ascertaining long COVID with symptoms limiting day-to-day function, frequencies were lower, ranging from 1.2-4.8% for PCS and 3.0-13.7% for OSC (Table 2). BiB used an individual symptoms approach (recorded retrospectively over several months) to ascertain long COVID and found 40.7% of study members reported symptoms for 12+ weeks and 22.7% reported symptoms for 4-12 weeks (See Table 2 and Supplementary Table 6). This analysis was repeated in TwinsUK and results were similar for COVID-19 cases (12+ weeks, 45.6% and 4-12 weeks, 25.8%), but also high in non-COVID-19 cases ascertained at the same time (12+ weeks, 28.8%, 4-12 weeks, 21.8%, and 0-4 weeks, 17.8% Supplementary Table 6). Therefore, these results were not taken forward to risk factor analysis due to uncertainty as to whether the majority of these symptoms were attributable to COVID-19 itself.

In the age-heterogeneous LS, increasing trends in risk of symptoms lasting both 4+ weeks and 12+ weeks with higher age were observed across participants ranging from young adulthood to approximately 70 years (Supplemental Figures 1 and 2). In meta-regression analyses to assess absolute differences in long COVID frequency with age, a clear linear trend in reporting of symptoms for 4+ weeks was present across the four national cohorts (MCS, NS, BCS70 and NCDS), corresponding to a 3.02% (95% CI: 1.86, 4.17) higher proportion of individuals with COVID-19 reporting OSC or PCS per decade between 20 to 63 years (Figure 1, left panel). A more modest linear trend in the reporting of symptoms for 12+ weeks was observed, and with less precision due to a lower number of cases reporting PCS alone (0.68% per decade; 95% CI: -0.15, 1.51).

Pooled associations between other sociodemographic and health traits and each binary long COVID outcome (4+ vs 0-4 weeks (OSC and PCS combined) and 12+ vs 0-12 weeks (PCS specifically)) are presented as part of Figure 2, and in full detail in Supplementary Figures 3 to 6. This synthesised analysis included the 10 LS samples with a total of 6754 participants.

Females had higher risk of both long COVID outcomes (4+ weeks: OR=1.49; 95%CI: 1.24-1.79; 12+ weeks: OR=1.60; 95%CI: 1.23-2.07). No clear evidence was found for individuals of non-white ethnicity (compared to individuals of white ethnicity) having differential risk of OSC and PCS combined (OR for symptoms lasting 4+ weeks =0.80; 95%CI: 0.54-1.19). Non-white ethnicity was associated with lower risk of PCS specifically (OR=0.32; 95%CI: 0.22-0.47) after meta-analysis, but these study-level findings displayed a high degree of heterogeneity (I^2=75%, *P*<0.001; Supplementary figure 5). Across LS, no strong evidence was found for associations of IMD with either outcome. Having not attained a degree from higher education was associated with lower risk of PCS specifically (OR: 0.73; 95% CI: 0.57-0.94), but not with OSC and PCS in combination (OR: 0.95: 95% CI: 0.80-1.14).

When synthesising associations for health characteristics across LS, those with poor or fair pre-pandemic self-reported general health were found to have greater odds of having symptoms for both long COVID outcomes (4+ weeks: OR=1.62; 95%CI: 1.25-2.09; 12+ weeks: OR=1.66; 95%CI: 1.14-2.40). Greater pre-pandemic psychological distress was also associated with higher risk of both long COVID outcomes (4+ weeks: OR=1.45; 95%CI: 1.16-1.82; PCS: OR=1.58; 95%CI: 1.15-2.17). No strong evidence was observed for a linear association of BMI with either outcome. In models to examine the potential importance of a BMI threshold in relation to long COVID, overweight/obesity was associated with increased odds of symptoms lasting for 4+ weeks (OR= 1.24; 95%CI: 1.01-1.53) threshold but not with PCS specifically (OR 0.95, 95% CI: 0.70-1.28). Associations were not found for diabetes, hypertension, or high cholesterol with either outcome, although modest point estimates were on the side of higher long COVID risk in several instances (**Supplementary figures 4 and 6**. Asthma was the only specific medical condition associated with increased odds of having symptoms for 4+ weeks (OR=1.31; 95%CI: 1.06-1.62), although the association with PCS specifically was closer to the null (OR=1.13;95%CI: 0.80-1.58).

Sensitivity analyses

When including IPWs for risk of COVID-19 status, all identified associations persisted and, in some instances, associations increased slightly in magnitude **(Supplementary figures 7 to 10)**. Notably hypercholesterolaemia was associated with both long COVID outcomes in the LS meta-analyses weighted for probability of reporting COVID-19.

Electronic Health Records

Within 1,199,812 individuals with any acute COVID-19 code, 3327 individuals also had a recorded long COVID code, constituting 0.27% of COVID-19 cases.

An inverted U-shaped association of recording of long COVID with age was observed (Supplemental Figure 1), where long COVID reporting was highest among those aged 45-54 and 55-69 years, whereas individuals aged 80 or older were at no higher risk of having a long COVID code than the reference group aged 18-24 years. There was a linear increase of absolute risk of long COVID of 0.12% per decade (95% CI: 0.08-0.17) between 18 and 70 years, aligning with LS results (Figure 1, right panel), although a quadratic trend for long COVID reporting was a closer fit for this full range of age data in OpenSAFELY.

In keeping with the LS results, females had higher risk of long COVID than males (OR=1.51; 95%CI:1.41-1.61), while odds were lower in individuals of South Asian (compared to (OR=0.75; 95%CI:0.67-0.84) or black ethnicity, relative to white ethnicity (OR=0.66; 95%CI:0.52-0.83) (Table 3 and Figure 2). Individuals living in areas with the least deprivation had higher odds of having a long COVID code compared to those in the most deprived IMD quintile (Figure 2).

In EHRs, increased odds of having a long COVID code was seen in individuals with pre-existing comorbidities (OR=1.26; 95%CI:1.18-1.35) and psychiatric conditions (OR=1.57; 95%CI:1.47-1.68). Again, as with the population-based studies an increased risk was observed in individuals with a pre-pandemic diagnosis of asthma (OR=1.56; 95%CI:1.46-1.67) and overweight and obesity (OR=1.31, 95%CI:1.21-1.42). No increase in risk was observed for diabetes.

Discussion

This research aimed to provide information useful to both clinical practice and policy given the lack of characterisation of factors reliably associated with OSC and PCS (together termed long COVID). Using data from a consortium of population-based LS which captured coordinated repeated questionnaire data on COVID-19 and the OpenSAFELY resource, we examined the frequency of long COVID and associations with sociodemographic and pre-pandemic health risk factors.

Main findings

The frequency of those with apparent OSC specifically ranged from 3% to 18% and the frequency of those with PCS specifically ranged from 1% to 17% across studies in young adult LS and late midlife LS respectively. Using a stricter definition of symptoms affecting day-to-day function that was recorded by a subset of our LS questionnaires, proportions with both conditions were lower (OSC: 3.0 to 13.7; PCS: 1.2% to 4.8% in young adults and those in late midlife respectively). In individuals both presenting to and diagnosed with COVID-19 by primary care practitioners, the proportion recorded with long COVID of any duration was substantially lower at 0.27%.

In both LS and EHR, long COVID reporting by any definition increased with age. Unlike risk of severe COVID-19, this appeared to be an apparently linear (and not exponential) relationship across most adult age groups. Women were approximately 50% more likely to report long COVID than men, while those of non-white ethnicity were approximately a third to a quarter less likely than those of white ethnicity to report long COVID in EHR, and PCS specifically in LS. Greater socioeconomic advantage (measured from area of residence) was associated with a greater risk of long COVID in primary care data, but not in LS.

In both LS and EHR, pre-existing adverse mental health was associated with an approximate 50% increase in the odds of reporting long COVID, while estimates of the association with poorer general health ranged between 1.26 for EHR to 1.62 for LS. Asthma was the only specific prior health condition associated with greater odds of persistent symptoms; in LS by a third, and in primary care by a half.

Reports on the proportions of infected individuals going on to experience long COVID have varied. Current estimates from Office of National Statistics (ONS) estimate that by May 2021, among 1.0 million individuals living in the UK, 1.6% self-report long COVID (defined by symptoms persisting for more than four weeks after the first suspected COVID-19infection that were not explained by something else).[25] In this report ONS ascertained long COVID estimates using a self-reported question similar to those used in our LS.

Many currently available studies reported on selected, hospitalised or outpatient populations find higher proportions with long COVID. We found only two population-based samples in the literature to date, which assessed the presence of ≥1 symptom at 60 days (n=594)[26] or 125 days (n=180)[27] respectively, and showed high reporting at these time points (53.1% and 35.0%) when counting all symptoms (some of which may be attributable to other conditions). Previous ONS data using symptom counts also found higher proportions of persistent symptomatology (OSC and PCS combined: 21.1%, PCS specifically: 9.9 %).[28] These studies did not ascertain symptoms in individuals without history of COVID-19; and there are multiple long COVID symptoms which overlap with other conditions. Defining long COVID in the same way in two of our studies produced similarly high proportions (41.1-45.6%). However, critically, proportions in individuals with no previous self-report diagnosis of COVID were also high (12+ weeks, 28.8%, 4-12 weeks, 21.8% and 0-4 weeks, 17.8%) during the same time window. While symptom reporting in COVID-19 could reflect other sequelae of COVID-19, such as new alternative diagnoses triggered by COVID-19, impaired recall, or misattribution, the high frequency in symptom reporting in the unaffected population suggests many symptoms may not relate to COVID-19 itself. Therefore, we focused on estimates of duration of symptoms attributed to COVID-19 by the individuals themselves. In LS, we show that the proportion of people self-reporting a COVID-19 illness who experienced prolonged symptoms differed between studies depending on the age of the study participants and whether the definition specified symptoms impairing day-to-day activity.

Despite these differences and the markedly lower risk of long COVID diagnosis in primary care versus LS, several risk factor associations were consistent between various LS and in EHR. Findings that long COVID was more common with each decade of age from age 20 to age 70, and was 50% higher in women than men, are consistent with reports from most[4,29–33] but not all previous studies.(34,35) There was an approximate linear increase in risk with age between 18 and 70 years. Over the age of 70, we observed a sharp decline in risk in most LS and EHR. This decline in risk for older adults which has been observed in other studies,[4,29,32] may be explained by selective competing risk of mortality, non-response bias, individuals misattributing long COVID to other illnesses, or a combination of these factors.

We observed a counterintuitive reduction in odds for long COVID for demographic factors which are commonly associated with increased morbidity, such as lower education (associated with lower risk of PCS only). This contrasts with a population-based study in the US[26] which found no strong evidence for associations with ethnicity or socioeconomic status and a Swedish study which found no socioeconomic gradient in long COVID.[34] While we found no strong evidence for a relationship between area-level socioeconomic status in LS, in primary care EHR there was an apparent gradient of higher risk in individuals from the least deprived areas. This likely reflects unmet need in those who live in socioeconomically deprived areas, given that both pre-existing adverse mental and physical health is associated with greater risks of long COVID, and that these conditions are likely more prevalent in those who are less advantaged. We found an apparent reduction in odds in minority ethnic groups for PCS specifically in LS and long COVID code reporting in EHR. Further research is needed to understand the reasons for this.

Both LS and EHR have rich pre-pandemic data on health and disease on their participants, which most published studies of long COVID lack.[25,35] Therefore, we were able to disaggregate mental and physical health characteristics caused by the pandemic, from those that were pre-existing. Our finding of a greater risk of long COVID related to adverse prior mental health, has been reported elsewhere,[26] but pre-pandemic general health has not previously been highlighted.[29,31,36] These findings were robust in all analyses including inverse probability weighting for risk of COVID-19. The finding of an excess risk of long COVID in association with asthma across cohorts and primary care records resolves previous conflicting and limited findings,[26,29,35] and provides considerable support for focusing on asthma as a high-risk condition, for example by investigating into whether immune processes seen in asthma play a role in the development of long COVID. In our analysis weighted for risk of COVID-19 onset, high cholesterol measures were associated with a greater risk of long COVID which has not been reported previously, although only two LS contributed data for meta-analyses of this factor and further studies will be required to confirm or refute this finding. We found no association between diabetes or hypertension and long COVID.[26,31,35,37] Findings for overweight/obesity were suggestive of an increased risk again resolving previous uncertainty.[29,35,36]

The markedly lower reporting of long COVID in primary care compared to LS suggests only a minority of people with long COVID seek care and subsequently receive a code. Diagnostic codes for long COVID have only recently been instituted and uptake by primary care practitioners has not been uniform.[38] Additionally, the analyses here are based on practices that use TPP SystmOne software, noting that these practices had a 2- to 3-fold lower rate of long COVID recording than those that use EMIS software.[38]

Strengths and limitations

This analysis brings together data from 10 longitudinal study samples and EHR, with rich information on pre-pandemic risk factors and COVID-19 symptom length. Although several recent surveys are available, the lack of pre-pandemic measures makes it difficult to assess directional effects of risk factors on outcomes. This study is strengthened by the coordinated investigation in multiple LS that are each susceptible to different sources of bias, with differing study designs, target populations, and selection and attrition processes. Moreover, the use of multiple studies increased statistical power to look at subpopulations, such as ethnic minority groups, and allowed for greater examination of the influence of age on long COVID. Our novel approach to harnessing multiple datasets allowed research questions to be addressed which would not otherwise be possible. Differences between studies in a range of factors - including measurement of risk factors, timing of surveys, design, response rates, and differential selection into the COVID-19 sweeps - are potentially responsible for heterogeneity in estimates. However, despite this heterogeneity, the key findings were consistent across most datasets. Unmeasured/residual confounding bias cannot be ruled out in either LS or EHR; and our analysis was not able to assess causation. We attempted to assess any index event bias using a systematic, structured approach across LS, which produced consistent results, but there remains the possibility that we have not fully accounted for this, due to the presence of unobserved factors and imperfect measurement of observed factors. Further, analysis of case series alone may yield bias as a result of generating an artificial sampling frame within which observed associations do not reflect whole population truths.[22] This may equally be the case for other published manuscripts that confine their samples to hospitalised/Emergency Department patients.[31,36,39] Our samples were population-based, and only a small number of individuals were admitted to hospital. We did not adjust for severity of initial disease which others have reported as relevant.[4,26] However, the persistence of associations across studies of scale and with heterogeneous characteristics lends confidence in our findings. Lastly, it should be noted that associations presented here represent those specific to case status as defined in our collections.

Implications

It has been possible to identify risk factors at the level of the population. Although causal inferences cannot be made at this stage, this provides evidence to support investigation, in particular of the role of sex differences, biological ageing, and immunity in the development of long COVID. Targeting services to those most in need may be warranted. Our data suggest that improved diagnosis within primary care is needed, both to facilitate research but also to allow rolling out of future interventions when effective support becomes available. Further research on prevention and treatment of long COVID is urgent and critical given the scale of the pandemic and the functional consequences of the condition. Individuals in older working age may particularly require support and given high levels of comorbidity in this group, will require holistic approaches that incorporate potential multimorbidity's. Trials should therefore ensure inclusivity of older people, and people with prior mental and physical health diagnoses. In this work we have demonstrated the benefits of cross-cohort collaborations and harmonised analyses which has accelerated the return of robust reproducible findings to the scientific community and the public. Future efforts to link EHRs to LS could yield even greater insights.

References

1. National Institute for Health and Care Excellence, Practitioners RC of G, Scotland HI. COVID-19 rapid guideline : managing the long-term effects of COVID-19. *NICE Guidel*. 2020;(18 December 2020):1-35.

2. Ayoubkhani D, Khunti K, Nafilyan V, et al. Epidemiology of post-COVID syndrome following hospitalisation with coronavirus: a retrospective cohort study. *medRxiv*. Published online 2021:2021.01.15.21249885. https://doi.org/10.1101/2021.01.15.21249885

3. Callard F, Perego E. How and why patients made Long Covid. *Soc Sci Med*. 2021;268(October 2020):113426. doi:10.1016/j.socscimed.2020.113426

4. Sudre CH, Murray B, Varsavsky T, et al. Attributes and predictors of long COVID. *Nat Med*. 2021;27(April). doi:10.1038/s41591-021-01292-y

5. Savarraj JPJ, Burkett AB, Hinds SN, et al. Three-month outcomes in hospitalized COVID-19 patients. *medRxiv*. Published online January 1, 2020:2020.10.16.20211029. doi:10.1101/2020.10.16.20211029

6. Carfì A, Bernabei R, Landi F, Group for the GAC-19 P-ACS. Persistent Symptoms in Patients After Acute COVID-19. *JAMA*. 2020;324(6):603-605. doi:10.1001/jama.2020.12603

7. Ward H, Cooke G, Whitaker M, et al. REACT-2 Round 5: increasing prevalence of SARS-CoV-2 antibodies demonstrate impact of the second wave and of vaccine roll-out in England. *medRxiv*. Published online 2021. https://www.medrxiv.org/content/10.1101/2021.02.26.21252512v1

8. Maxwell E. Living with COVID-19. A dynamic review of the evidence around ongoing covid-19 symptoms (often called long covid). *Natl Inst Heal Res*. 2020;31(3):197-200. doi:10.22365/jpsych.2020.313.197

9. Joshi HE, Fitzsimons E. The UK Millennium Cohort Study: the making of a multi- purpose resource for social science and policy in the UK. *Longit Life Course Stud*. 2016;7(4):409-430. doi:10.14301/llcs.v7i4.416

10. Boyd A, Golding J, Macleod J, et al. Cohort profile: The ’Children of the 90s’-The index offspring of the avon longitudinal study of parents and children. *Int J Epidemiol*. 2013;42(1):111-127. doi:10.1093/ije/dys064

11. Calderwood L, Sanchez C. Next Steps (formerly known as the Longitudinal Study of Young People in England). Published online 2016:2-4.

12. Elliott J, Shepherd P. Cohort profile: 1970 British Birth Cohort (BCS70). *Int J Epidemiol*. 2006;35(4):836-843. doi:10.1093/ije/dyl174

13. Power C, Elliott J. Cohort profile: 1958 British birth cohort (National Child Development Study). *Int J Epidemiol*. 2006;35(1):34-41. doi:10.1093/ije/dyi183

14. Brown M, Goodman A, Peters A, et al. COVID-19 Survey in Five National Longitudinal Studies: Wave 1, 2 and 3. User Guide (Version 3). *UCL Cent Longitud Stud MRC Unit*

Lifelong Heal Ageing London, UK. 2020;(June):1-62. https://cls.ucl.ac.uk/wp-content/uploads/2021/01/UCL-Cohorts-COVID-19-Survey-user-guide.pdf

15. Wright J, Small N, Raynor P, et al. Cohort profile: The born in bradford multi-ethnic family cohort study. *Int J Epidemiol*. 2013;42(4):978-991. doi:10.1093/ije/dys112
16. Dickerson J, Bird PK, McEachan RRC, et al. Born in Bradford's Better Start: An experimental birth cohort study to evaluate the impact of early life interventions. *BMC Public Health*. 2016;16(1):1-14. doi:10.1186/s12889-016-3318-0
17. University of Essex, Institute for Social and Economic Research, NatCen Social Research KP. Understanding Society: Waves 1-9, 2009-2019 and Harmonised BHPS: Waves 1-18, 1991-2009. [data collection].
18. Smith BH, Campbell A, Linksted P, et al. Cohort profile: Generation scotland: Scottish family health study (GS: SFHS). The study, its participants and their potential for genetic research on health and illness. *Int J Epidemiol*. 2013;42(3):689-700. doi:10.1093/ije/dys084
19. Fraser A, Macdonald-wallis C, Tilling K, et al. Cohort Profile : The Avon Longitudinal Study of Parents and Children : ALSPAC mothers cohort. 2013;(April 2012):97-110. doi:10.1093/ije/dys066
20. Verdi S, Abbasian G, Bowyer RCE, et al. TwinsUK: The UK Adult Twin Registry Update. *Twin Res Hum Genet*. 2019;(May 2007):1-7. doi:10.1017/thg.2019.65
21. Suthahar A, Sharma P, Hart D, et al. TwinsUK COVID-19 personal experience questionnaire (CoPE): wave 1 data capture April-May 2020 [version 1 ; peer review : awaiting peer review]. 2021;(May 2020):1-10.
22. Griffith GJ, Morris TT, Tudball MJ, et al. Collider bias undermines our understanding of COVID-19 disease risk and severity. *Nat Commun*. 2020;11(1):1-12. doi:10.1038/s41467-020-19478-2
23. Pickering S, Betancor G, Galão RP, et al. Comparative assessment of multiple COVID-19 serological technologies supports continued evaluation of point-of-care lateral flow assays in hospital and community healthcare settings. *PLoS Pathog*. 2020;16(9):e1008817. doi:10.1371/journal.ppat.1008817
24. Mathur R, Rentsch CT, Morton CE, et al. Ethnic differences in SARS-CoV-2 infection and COVID-19-related hospitalisation, intensive care unit admission, and death in 17 million adults in England: an observational cohort study using the OpenSAFELY platform. *Lancet*. 2021;397(10286):1711-1724. doi:10.1016/S0140-6736(21)00634-6
25. Ayoubkhani D. Prevalence of ongoing symptoms following coronavirus (COVID-19) infection in the UK : 1 April 2021. *Off Natl Stat*. 2021;(April):1-16.
26. Hirschtick JL, Titusa AR, Slocum E, et al. Population-based estimates of post-acute sequelae of SARS-CoV-2 infection (PASC) prevalence and characteristics Jana. *Clin Infect Dis*. Published online 2021.

27. Petersen MS, Kristiansen MF, Hanusson KD, et al. Long COVID in the Faroe Islands - a longitudinal study among non-hospitalized patients. *Clin Infect Dis*. Published online 2020:1-18. doi:10.1093/cid/ciaa1792

28. The prevalence of long COVID symptoms and COVID-19 complications. Office for National Statistics.

29. Augustin M, Schommers P, Stecher M, et al. Post-COVID syndrome in non-hospitalised patients with COVID-19: a longitudinal prospective cohort study. *Lancet Reg Heal Eur*. 2021;6:100122. doi:10.1016/j.lanepe.2021.100122

30. Makaronidis J, Firman C, Magee CG, et al. Distorted chemosensory perception and female sex associate with persistent smell and/or taste loss in people with SARS-CoV-2 antibodies: a community based cohort study investigating clinical course and resolution of acute smell and/or taste loss in people. *BMC Infect Dis*. 2021;21(1):1-11. doi:10.1186/s12879-021-05927-w

31. Peghin M, Palese A, Venturini M, et al. Post-COVID-19 symptoms 6 months after acute infection among hospitalized and non-hospitalized patients. *Clin Microbiol Infect*. Published online 2021. doi:10.1016/j.cmi.2021.05.033

32. Huang Y, Pinto MD, Borelli JL, et al. COVID Symptoms, Symptom Clusters, and Predictors for Becoming a Long-Hauler: Looking for Clarity in the Haze of the Pandemic. *medRxiv*. Published online January 1, 2021:2021.03.03.21252086. doi:10.1101/2021.03.03.21252086

33. Mahmud R, Rahman MM, Rassel MA, et al. Post-COVID-19 syndrome among symptomatic COVID-19 patients: A prospective cohort study in a tertiary care center of Bangladesh. *PLoS One*. 2021;16(4 April):1-13. doi:10.1371/journal.pone.0249644

34. Westerlind E, Palstam A, Sunnerhagen KS, Persson HC. Patterns and predictors of sick leave after Covid-19 and long Covid in a national Swedish cohort. *BMC Public Health*. 2021;21(1023):1-9.

35. Sudre CH, Murray B, Varsavsky T, et al. Attributes and predictors of long COVID. *Nat Med*. 2021;27(4):626-631. doi:10.1038/s41591-021-01292-y

36. Moreno-pérez O, Merino E, Leon-ramirez J, et al. Post-acute COVID-19 syndrome. Incidence and risk factors: A Mediterranean cohort study. *J Infect J*. 2021;82(January):373-378.

37. Chioh FWJ, Fong SW, Young BE, et al. Convalescent covid-19 patients are susceptible to endothelial dysfunction due to persistent immune activation. *Elife*. 2021;10:1-23. doi:10.7554/eLife.64909

38. 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. *medRxiv*. Published online 2021:2021.05.06.21256755. http://medrxiv.org/content/early/2021/05/13/2021.05.06.21256755.1.abstract

39. Osikomaiya B, Erinoso O, Wright KO, et al. 'Long COVID': persistent COVID-19 symptoms

in survivors managed in Lagos State, Nigeria. *BMC Infect Dis*. 2021;21(1):1-7. doi:10.1186/s12879-020-05716-x

People acknowledgments:

OpenSAFELY Collaborative Group: Alex J Walker, Brian MacKenna, Peter Inglesby, Christopher T Rentsch, Helen J Curtis, Caroline E Morton, Jessica Morley, Amir Mehrkar, Seb Bacon, George Hickman, Chris Bates, Richard Croker, David Evans, Tom Ward, Jonathan Cockburn, Simon Davy, Krishnan Bhaskaran, Anna Schultze, Elizabeth J Williamson, William J Hulme, Helen I McDonald, Laurie Tomlinson, Rohini Mathur, Rosalind M Eggo, Kevin Wing, Angel YS Wong, Harriet Forbes, John Tazare, John Parry, Frank Hester, Sam Harper, Ian J Douglas, Stephen JW Evans, Liam Smeeth, Ben Goldacre.

Studies

Generation Scotland: Drew Altschul, Chloe Fawns-Ritchie, Archie Campbell, Robin Flaig.

ALSPAC: Daniel J Smith.

Understanding Society: Michaela Benzeval.

TwinsUK: Deborah Hart, María Paz García, Rachel Horsfall

Centre for Longitudinal Studies: Matt Brown, Lisa Calderwood, Emla Fitzsimons, Alissa Goodman, Aida Sanchez

Born in Bradford: John Wright, Dan Mason

Funding acknowledgements:

This work was supported by the National Core Studies, an initiative funded by UKRI, NIHR and the Health and Safety Executive. The COVID-19 Longitudinal Health and Wellbeing National Core Study was funded by the Medical Research Council (MC_PC_20030).

The contributing studies have been made possible because of the tireless dedication, commitment and enthusiasm of the many people who have taken part. We would like to thank the participants and the numerous team members involved in the studies including interviewers, technicians, researchers, administrators, managers, health professionals and volunteers. We are additionally grateful to our funders for their financial input and support in making this research happen.

Studies acknowledgements:

Data gathered from questionnaire(s) was provided by Wellcome Longitudinal Population Study (LPS) COVID-19 Steering Group and Secretariat (221574/Z/20/Z).

Understanding Society is an initiative funded by the Economic and Social Research Council and various Government Departments, with scientific leadership by the Institute for Social and Economic Research, University of Essex, and survey delivery by NatCen Social Research and Kantar Public. The Understanding Society COVID-19 study is funded by the Economic and Social Research Council (ES/K005146/1) and the Health Foundation (2076161). The research data are distributed by the UK Data Service.

The Millennium Cohort Study, Next Steps, British Cohort Study 1970 and National Child Development Study 1958 are supported by the Centre for Longitudinal Studies, Resource Centre 2015-20 grant (ES/M001660/1) and a host of other co-funders. The COVID-19 data collections in these five cohorts were funded by the UKRI grant Understanding the economic, social and health impacts of COVID-19 using lifetime data: evidence from 5 nationally representative UK cohorts (ES/V012789/1).

The UK Medical Research Council and Wellcome (Grant Ref: 217065/Z/19/Z) and the University of Bristol provide core support for ALSPAC. A comprehensive list of grants funding is available on the ALSPAC website (http://www.bristol.ac.uk/alspac/external/documents/grant-acknowledgements.pdf). We are extremely grateful to all the families who took part in this study, the midwives for their help in recruiting them, and the whole ALSPAC team, which includes interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists and nurses. Please note that the study website contains details of all the data that is available through a fully searchable data dictionary and variable search tool" and reference the following webpage: http://www.bristol.ac.uk/alspac/researchers/our-data/. Ethical approval for the study was obtained from the ALSPAC Ethics and Law Committee and the Local Research Ethics Committees. Part of this data was collected using REDCap, see the REDCap website for details https://projectredcap.org/resources/citations/

TwinsUK receives funding from the Wellcome Trust (WT212904/Z/18/Z), the National Institute for Health Research (NIHR) Biomedical Research Centre based at Guy's and St Thomas' NHS Foundation Trust and King's College London.The TwinsUK COVID-19 personal experience study was funded by the King's Together Rapid COVID-19 Call award, under the projects original title ‘Keeping together through coronavirus: The physical and mental health implications of self-isolation due to the Covid-19 TwinsUK is also supported by the Chronic Disease Research Foundation and Zoe Global Ltd. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Generation Scotland received core support from the Chief Scientist Office of the Scottish Government Health Directorates [CZD/16/6] and the Scottish Funding Council [HR03006]. Genotyping of the GS:SFHS samples was carried out by the Genetics Core Laboratory at the Wellcome Trust Clinical Research Facility, Edinburgh, Scotland and was funded by the Medical Research Council UK and the Wellcome Trust (Wellcome Trust Strategic Award “STratifying Resilience and Depression Longitudinally” (STRADL) Reference 104036/Z/14/Z). Generation Scotland is funded by the Wellcome Trust (216767/Z/19/Z).

Born in Bradford (BiB) receives core infrastructure funding from the Wellcome Trust (WT101597MA), and a joint grant from the UK Medical Research Council (MRC) and UK Economic and Social Science Research Council (ESRC) (MR/N024397/1) and one from the British Heart Foundation (BHF) (CS/16/4/32482). The National Institute for Health Research Yorkshire and Humber ARC, and Clinical Research Network both provide support for BiB research. Born in Bradford is only possible because of the enthusiasm and commitment of the children and parents in BiB. We are grateful to all the participants, health professionals, schools and researchers who have made Born in Bradford happen.

OpenSAFELY is jointly funded by UKRI, NIHR and Asthma UK-BLF [COV0076; MR/V015737/] and the Longitudinal Health and Wellbeing strand of the National Core Studies programme. EMIS and TPP provided technical expertise and infrastructure within their data environments pro bono in the context of a national emergency. The OpenSAFELY software platform is supported by a Wellcome Discretionary Award. BG's work on clinical informatics is supported by the NIHR Oxford Biomedical Research Centre and the NIHR Applied Research Collaboration Oxford and Thames Valley. Funders had no role in the study design, collection, analysis, and interpretation of data; in the writing of the report; and in the decision to submit the article for publication. The views expressed are those of the authors and not necessarily those of the NIHR, NHS England, Public Health England or the Department of Health and Social Care.

People funding:

NJT is a Wellcome Trust Investigator (202802/Z/16/Z), is the PI of the Avon Longitudinal Study of Parents and Children (MRC & WT 217065/Z/19/Z), is supported by the University of Bristol NIHR Biomedical Research Centre, the MRC Integrative Epidemiology Unit (MC_UU_00011/1) and works within the CRUK Integrative Cancer Epidemiology Programme (C18281/A29019). 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). ASFK acknowledges funding from the ESRC (ES/V011650/1). EJT acknowledges funding from the Wellcome Trust (WT212904/Z/18/Z). RM acknowledges support from the Elizabeth Blackwell Institute for Health Research, University of Bristol, and the Wellcome Trust Institutional Strategic Support Fund (204813/Z/16/Z). GBP acknowledges funding from the Economic and Social Research Council (ES/V012789/1). CLN acknowledges funding from the Medical Research Council (MR/R024774/1). KT works in a Unit that is supported by the University of Bristol and UK Medical Research Council (MC_UU_00011/3). DMW is supported by funding from UK Medical Research Council (MC_PC_20030). NC is supported by funding from the UK Medical Research Council (MC_UU_00019/2)

Declaration of interests:

No conflicts of interest were declared by EJT, DMW, AJW, REM, CLN, TCY, CFH, ASFK, RJS, GDG, RCEB, KN. BH, MJG, BD, KJD, ELD, FMKW, AS, DJP, RRCM, LT, BG, PP, GBP, KT, CTR, NJT, NC, CJS. SVK is a member of the Scientific Advisory Group on Emergencies subgroup on ethnicity and COVID-19 and is co-chair of the Scottish Government's Ethnicity Reference Group on COVID-19. NC serves on a data safety monitoring board for trials sponsored by AstraZeneca. CJS is an academic lead on KCL Zoe Global Ltd. COVID symptoms study.

List of Tables and Figures

Table 1: Characteristics of the analytic samples from the longitudinal studies (self-reported COVID-19 cases with data on duration of symptoms)

	MCS	ALSPAC G1	NS	BiB	USoc	BCS70	TwinsUK	GS	ALSPAC G0	NCDS
Sample size	1055	668	848	110	1033	889	806	343	446	709
Age, mean years (SD)	19.9 (0.3)	28.4 (0.5)	31.0 (0.3)	40.7 (5.9)	48.5 (14.8)	51 *	52.7 (15.8)	56.0 (10.6)	58.3 (4.4)	63 *
Female sex, N (%)	652 (61.8)	426 (63.8)	539 (64.6)	106 (96.4)	675 (65.3)	507 (57.0)	709 (88)	219 (63.9)	303 (67.9)	389 (54.9)
Ethnicity, N (%)										
White	862 (81.7)	638 (95.5)	574 (67.7)	49 (44.5)	879 (85.1)	747 (84.0)	776 (96.3)	330 (96.2)	439 (98.4)	652 (92.0)
Non-white ethnic minority	192 (18.2)	30 (4.5)	254 (30.0)	56 (50.9)	136 (13.2)	27 (3.0)	30 (3.7)	5 (1.5)	6 (1.3)	19 (2.7)
Missing	1 (0.1)	0	20 (2.4)	5 (4.6)	18 (1.7)	115 (12.9)	1 (0.1)	8 (2.3)	1 (0.2)	35 (5.4)
Education, N (%)										
Degree	494 (46.8)	338 (50.6)	396 (49.7)	11 (10)	500 (48.4)	377 (42.4)	402 (49.9)	168 (49.0)	106 (23.8)	284 (40.1)
No degree	502 (47.6)	149 (22.3)	358 (42.2)	82 (74.5)	429 (41.5)	444 (49.9)	224 (27.8)	168 (49.0)	307 (68.8)	415 (58.5)
Missing	59 (5.6)	181 (27.1)	94 (11.1)	17 (15.5)	104 (10.1)	68 (7.7)	180 (22.3)	7 (2.0)	33 (7.4)	10 (1.4)
Social class , N (%)										
Managerial, Admin, Professional	..	120 (18.0)	..	26 (23.6)	402 (38.9)	..	..	53.1 (182)	57 (12.8)	..
Intermediate	..	280 (41.9)	..	36 (32.7)	171 (16.6)	..	..	17.8 (61)	130 (29.1)	..
Manual/Routine	..	171 (25.6)	..	21 (19.1)	220 (21.3)	..	..	11.4 (39)	190 (42.6)	..
Not in employment	..	2 (0.3)	..	..	212 (20.5)	..	..	..	5 (1.1)	..
Missing	..	95 (14.2)	..	27 (24.5)	28 (2.7)	..	..	61 (17.8)	64 (14.3)	..
Country, N (%)										

England	746 (70.7)	668 (100)	828 (97.6)	110 (100)	866 (83.8)	770 (86.6)	747 (92.7)	4 (1.2)	446 (100)	613 (86.5)
Scotland	93 (8.8)	..	5 (0.6)	..	62 (6.0)	57 (6.4)	26 (3.2)	339 (98.8)	..	45 (6.4)
Wales	136 (12.9)	..	9 (1.1)	..	69 (6.7)	44 (5.0)	24 (3.0)	..	..	38 (5.4)
Northern Ireland	75 (7.1)	..	1 (0.1)	..	36 (3.5)	0	1 (0.1)	..	..	2 (0.3)
Missing / Other	5 (0.5)	..	5 (0.6)	..	0 (0)	18 (2.0)	8 (1)	..	..	11 (1.6)
Hospitalised with COVID-19, N (%)	8 (0.8)	..	23 (2.7)	..	21 (2.0)	40 (4.5)	27 (3.3)	..	..	37 (5.2)

Study acronyms: ALSPAC – Avon Longitudinal Study of Parents and Children (Generations 0 and 1); BCS70 – 1970 British Cohort Study; BiB – Born in Bradford; GS – Generation Scotland; MCS – Millennium Cohort Study; NCDS – 1958 National Child Development Study; NS – Next Steps; USoc – Understanding Society. Studies are ordered left to right from youngest to oldest mean age.

Table 2: Symptoms duration among self-reported COVID-19 cases in the longitudinal studies

Study	COVID-19 cases with symptom duration data	Mean age	Duration of symptoms, N (%)		
			Acute (0-4 weeks)	Ongoing symptomatic COVID-19 (4-12 weeks)	Post COVID-19 syndrome (12+ weeks)
Studies ascertaining long COVID of any severity					
ALSPAC G1	668	28.4	519 (77.7)	97 (14.5)	52 (7.8)
USoc	1033	48.5	742 (71.8)	182 (17.6)	109 (10.6)
TwinsUK	806	52.7	579 (71.8)	146 (18.1)	81 (10)
GS	335	56.0	224 (66.9)	54 (16.1)	57 (17.0)
ALSPAC G0	446	58.3	302 (67.7)	68 (15.2)	76 (17.0)
*Studies ascertaining severe long COVID only **					
MCS	1055	19.9	1010 (95.7)	32 (3.0)	13 (1.2)
Next Steps	848	31.0	773 (91.2)	51 (6.0)	24 (2.8)
BCS70	889	51.0	757 (85.2)	84 (9.5)	48 (5.4)
NCDS	709	63.0	578 (81.5)	97 (13.7)	34 (4.8)

*Studies ascertaining long COVID by monthly symptom reporting ***					
BiB	110	41.1	40 (36.4)	26 (22.7)	46 (40.9)
TwinsUK	953	54	272 (28.5)	246 (25.8)	435 (45.6)

Study acronyms: ALSPAC – Avon Longitudinal Study of Parents and Children (Generations 0 and 1); BCS70 – 1970 British Cohort Study; BiB – Born in Bradford; GS – Generation Scotland; MCS – Millennium Cohort Study; NCDS – 1958 National Child Development Study; NS – Next Steps; USoc – Understanding Society.

Footnotes: Studies are ordered from youngest to oldest mean age within categories of method of long COVID ascertainment

* Questionnaires in these four cohorts asked respondents to report duration for which COVID-19 symptoms impeded normal function, rather than simply the duration of any symptoms (however mild) as in other studies. Hence proportions reporting long COVID in them are expected to be lower when compared to other cohorts with similar characteristics

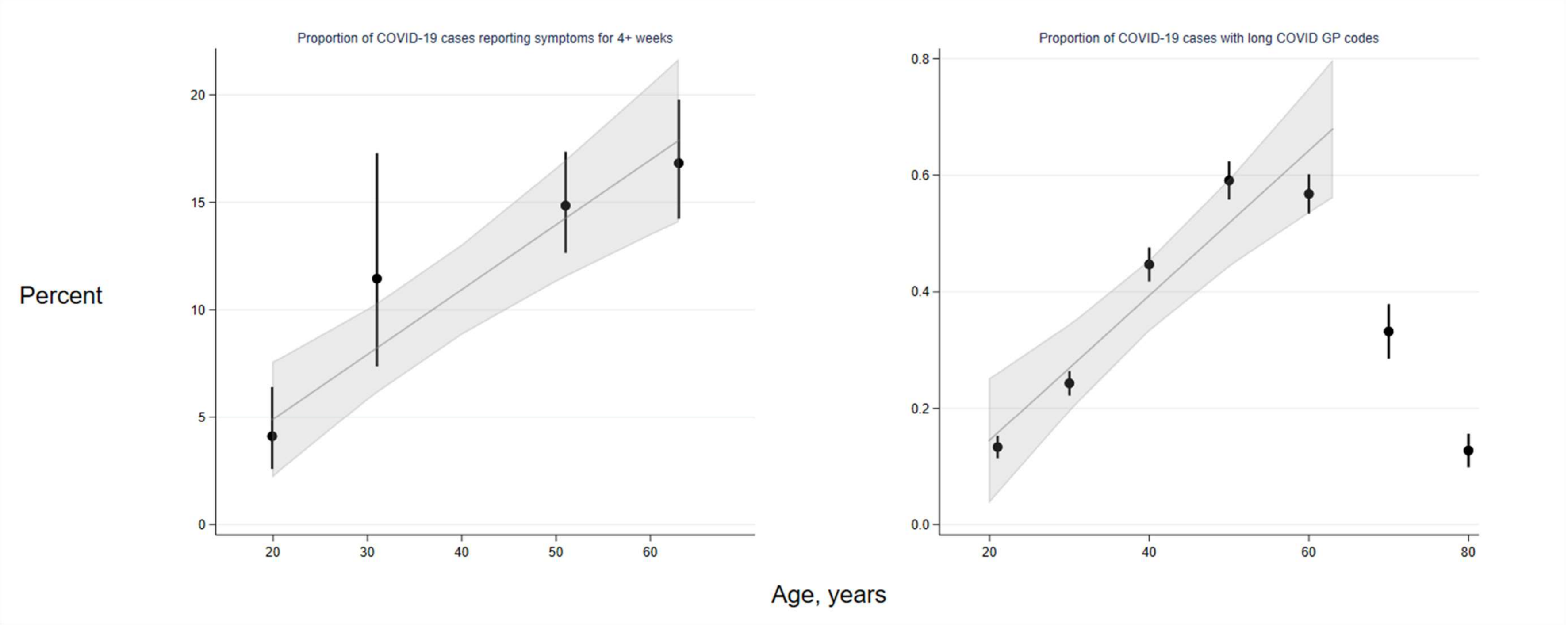
** Based on symptom-counting approach over months, rather than self-reported duration of symptoms as in all other cohorts, which yields higher proportions of individuals being designated long COVID categories

Table 3: Characteristics of individuals reported to have had COVID-19 and long COVID by general practitioners in OpenSAFELY

	Acute COVID-19	Long COVID	Long COVID rate per 100,000 cases	Proportion of long COVID cases in category (%)
Sample size	1,064,491	4,189	392	
Age, years				
18-24	137,997	184	133.2	4.4
25-34	211,479	515	242.9	12.3
35-44	199,750	897	447.1	21.4
45-54	208,351	1,238	590.7	29.6
55-69	190,616	1,088	567.5	26
70-79	57,886	193	332.3	4.6
80+	58,412	74	126.5	1.8
Sex				
Female	582,220	2,678	457.9	63.9
Male	482,271	1,511	312.3	36.1
Ethnicity				
White	635,414	2,647	414.9	63.2
Mixed	12,498	49	390.5	1.2
South Asian	111,026	340	305.3	8.1
Black	25,886	73	281.2	1.7
Other	16,521	53	319.8	1.3
IMD quantile				
0	22,104	75	338.2	1.8
1	255,431	787	307.2	18.8
2	226,760	850	373.4	20.3
3	208,684	932	444.6	22.2
4	188,224	814	430.6	19.4
5	163,288	731	445.7	17.5
BMI category				
Not obese	800,439	2,694	335.4	64.3
Obese I (30-34.9)	151,782	787	515.8	18.8
Obese II (35-39.9)	67,470	411	605.5	9.8
Obese III (40+)	44,800	297	658.6	7.1

Comorbidities				
0	661,200	2,336	352.1	55.8
1	291,106	1,335	456.5	31.9
2 or more	112,185	518	459.6	12.4
Mental health disorder(s)				
0	835,361	2,772	330.7	66.2
1 or more	229,130	1,417	614.6	33.8

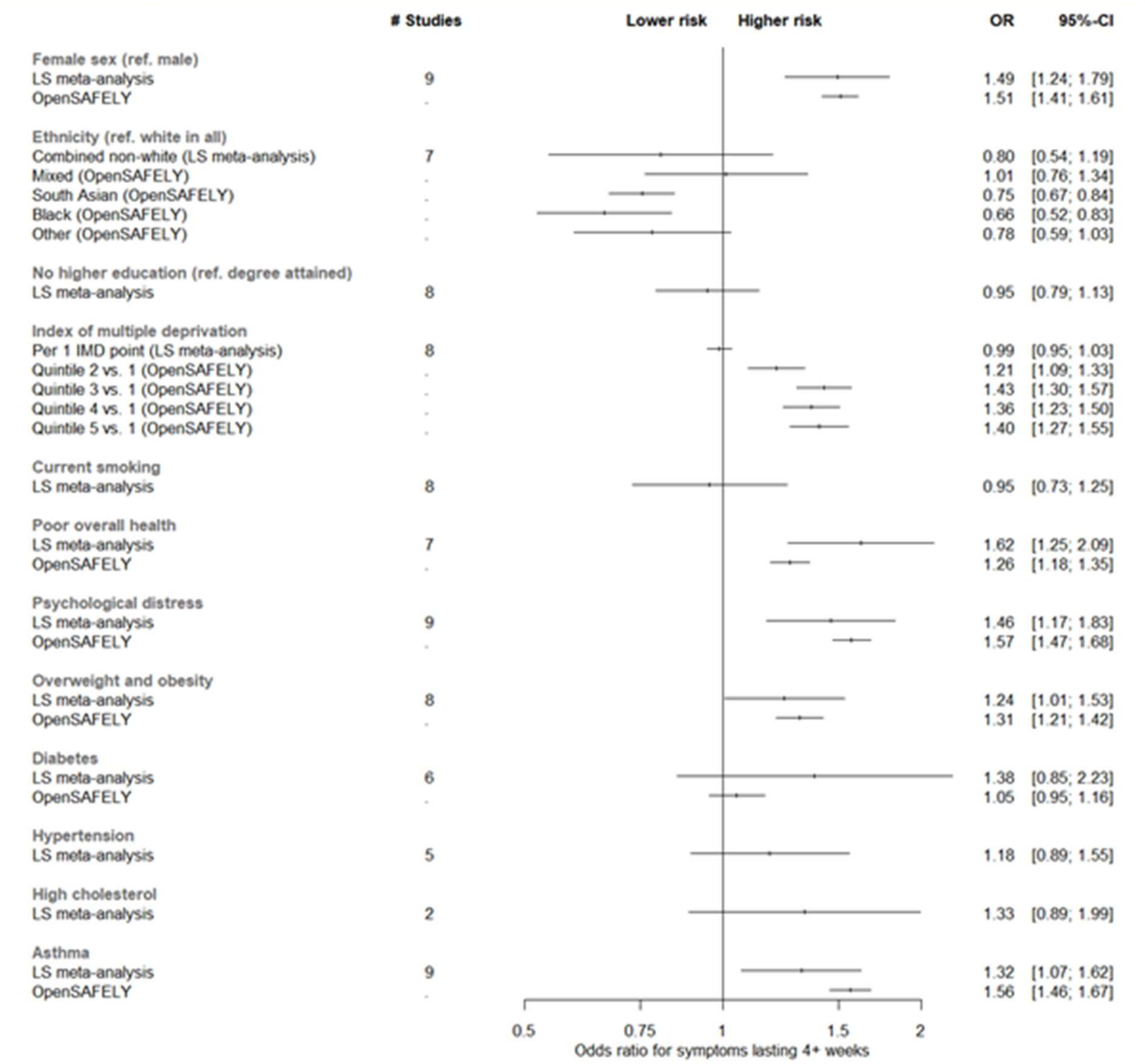
Figure 1: trends in long COVID frequency among COVID-19 cases by age, in four age-homogeneous longitudinal studies (left) and EHRs (right)



Legend:

Left -- in four longitudinal studies where participants are of near-identical ages (the cohorts MCS, NS, BCS70 and NCDS), proportions reporting symptom length of four or more weeks in COVID-19 cases were ascertained from questionnaire responses. Right -- in OpenSAFELY, proportions represent individuals within 10-year age categories (with estimates grouped at the mid-point of each category) who have long COVID codes in GP records, hence the proportions are substantially lower than in the corresponding cohort data. Trend lines and 95% confidence interval shading represent absolute differences in long COVID frequencies with increasing age, estimated by linear meta-regression of data from the four cohorts and from 18 to 70 year olds in OpenSAFELY (data from older individuals were not modelled; refer to results text for further explanation).

Figure 2: Risk factors associated with long COVID from meta-analyses of longitudinal study findings alongside corresponding analyses from EHRs



Legend:

All associations were adjusted for age and sex, except where redundant. In all instances where it was possible to derive results from both meta-analyses of longitudinal studies and analysis of EHRs, the corresponding results are plotted side-by-side for comparison. The outcome used for longitudinal study fixed-effect meta-analysis estimates presented here was symptoms lasting for 4+ weeks, and the outcome in EHRs was any reporting of a long COVID read code in GP records (regardless of duration of symptoms). Full study-level results, heterogeneity statistics and random-effect estimates for the longitudinal study meta-analyses are presented in supplemental figures 3 and 4. The equivalent meta-analyses of longitudinal study data where symptom duration of 12+ weeks was instead used as the outcome are depicted in supplemental figures 5 and 6. ‘Poor overall health’ represents the self-rated health exposure in the LS meta-analysis, and comorbidities in OpenSAFELY. The outcome ‘Overweight and obesity’ represents combined BMI categories over 25 in the LS, and solely individuals with BMI 30-34.9 in OpenSAFELY.